PSYCHOMETRIC PERSPECTIVES ON THE ASSESSMENT OF PRE-SERVICE TEACHERS IN INSTRUMENTAL MUSIC EDUCATION

by

DOROTHY JEAN MUSSELWHITE

(Under the Direction of Brian C. Wesolowski)

ABSTRACT

Teacher accountability and student growth have become the forefront of public education. Therefore, the need for valid, reliable, and fair assessments must be developed in order to accompany the change in teacher accountability. The purpose of this dissertation is to develop assessments using Rasch measurement theory. Specifically, these assessments were developed in the context of assessing pre-service music education students in lesson planning and teaching ability. It is the intent of these studies to guide pre-service teacher preparation in aligning pre-service expectations with in-service teaching expectations.

INDEX WORDS: Assessment, Pre-service Teachers, Psychometrics, Music Education, Scale Development

PSYCHOMETRIC PERSPECTIVES ON THE ASSESSMENT OF PRE-SERVICE TEACHERS IN INSTRUMENTAL MUSIC EDUCATION

by

DOROTHY JEAN MUSSELWHITE

B.Mus., University of Georgia, 2008

M.M.Ed., University of Georgia, 2015

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

2018

PSYCHOMETRIC PERSPECTIVES ON THE ASSESSMENT OF PRE-SERVICE TEACHERS IN INSTRUMENTAL MUSIC EDUCATION

by

DOROTHY JEAN MUSSELWHITE

Major Professors:	Brian C. Wesolowski George Engelhard, Jr.
Committee:	Alison P. Farley Peter J. Jutras

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2018

DEDICATION

This work is dedicated to my fiancee Eric for his unconditional love and constant encouragement. It is with gratitude that I also dedicate this dissertation to my inspiring parents Donna Musselwhite, and Harry and Laura Musselwhite. Lastly, I dedicate this work to Gloria and Clayton Byrd for their unwavering support.

ACKNOWLEDGEMENTS

This work would not be possible without the selfless assistance of my mentor and advisor, Dr. Brian Wesolowski. I am grateful that he never gave up on me, and challenged me to explore an unfamiliar realm of music education research. He has taught me that hard work is the pinnacle of success, and that you should never give up if you truly believe in something.

I am thankful that I have been blessed with professors who are willing to go beyond the requirements of their job to teach and to mentor. I would like to acknowledge the professional guidance given by Dr. George Engelhard, Jr. Because of his passion for teaching, he has opened a new world of measurement possibilities, and I know my life and my research will not be the same. I am also thankful for the support from Dr. Alison Farley and Dr. Peter Jutras, who have bestowed their time to my studies and research.

Lastly, my family has been more supportive and encouraging than I could ever have imagined. I would like to thank my parents and my grandparents, who have supported me in my pursuit of this degree. Moreover, I wish to thank my love, Eric Thompson, for his encouragement, compassion, and unwavering love. Without these people believing in me, I would never have been able to follow my passion.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Over the past several decades, changes in federal and state regulations have led to a significant shift in the requirements and expectations of educators (107th Congress, 2001; ESSA, 2015). The *No Child Left Behind Act* (107th Congress, 2001) required that all teachers be "highly qualified," having a bachelor's degree and certification in the specific subject matter (p. 360). Each state, however, was allowed the autonomity to define "highly qualified," leading to more in-depth teacher evaluation programs. In 2015, the *Every Student Succeeds Act* allowed states to be even more independent in implementing teacher evaluation in instances where the U.S. Department of Education no longer had direct control into this aspect of public education (ESSA, 2015).

The changing federal and state regulations for in-service teachers directly impact the training of pre-service teachers. If expectations and standards are being raised at the state and national level, the same expectations and standards must be raised at the preparatory level in order to prepare pre-service teachers for entering the teaching profession. Campbell and Thompson (2007) explain that aligning the curriculum to closely resemble real-world teaching will aid in retention of teachers once they reach the real classroom. Therefore, the instruction and assessment of pre-service teachers must become more detailed, comprehensive, and authentic. It is important that the assessment of pre-service teachers addresses the needs of the pre-service teachers themselves, the students within the classroom, and the standards set forth at district, state, and national levels.

Purpose

The standards set forth by the National Association of Schools of Music state that students obtaining teacher certification are expected to gain a common body of knowledge and set of skills (NASM, 2016). However, these standards do not provide a common set of assessments. The absence of common assessments means that teachers from various universities may look drastically different upon exiting preparatory programs.

The presence of assessments alone does not indicate that proper measures are in place. Therefore, it is imperative that all assessments are developed using appropriate methods of measurement. Proper steps must be taken to ensure measures are valid, reliable, and fair (AERA, APA, NCME, 2014). The purpose of the subsequent chapters within this dissertation is to develop measures for the assessment of pre-service teachers, using Rasch Measurement Theory. The dissertation focuses on two specific aspects of pre-service teaching: planning and instruction.

Background

In developing measures of assessment for pre-service music teachers, it is necessary to investigate the assessment protocols of pre-service general education teachers. This section will review assessments and research in the assessment of pre-service teachers, both in and out of the music curriculum.

Assessment of Pre-service Teachers

In the 1970s, the Coleman Report brought about a focus on teacher effectiveness, and "process-product" research (Coleman, 1966; Grant & Drafall, 1991). This led to a search for the fundamental qualities that would define effectiveness. Within music education, this effectiveness

has long been indicated by ratings at competitions or performance evaluations, student perceptions, or the size of a program's enrollment (Grant & Drafall, 1991).

The 1980s brought about qualitative investigations into the preparation and success of teachers. The literature emphasis on *learning-to-teach* refocused research on the mental processes involved in teaching effectiveness (Kagan, 1992). Kagan used aspects of the *learning-to-teach* literature to fuel her research on the professional growth of teachers. In regards to pre-service teachers, Kagan (1992) emphasized the need for teacher preparatory programs to provide pre-service teachers with more procedural knowledge, a better sense of self-reflection, and a focus on understanding the students in the classroom.

Porter and Brophy (1988) used research on teacher effectiveness to determine traits and expectations among the most effective teachers. These traits include being knowledgeable in specific content area and teaching strategies, being knowledgeable about students and the instructional needs of those students, clearly expressing instructional goals and expectations, teaching beyond the book, using differentiated instruction, teaching for cross-curricular connections, and using reflective teaching practices.

The competency-based movement in education in the 1960s led to a shift in focus for teacher educators (Nodine, 2016). While numerical grades had long been an indicator of success, this movement emphasized the need for more detailed feedback and more in-depth reporting of pre-service teachers' knowledge, skills, and dispositions (McDonald, 1978). In his 1978 article, McDonald indicates that although this movement brought about needed change in the evaluation of pre-service teachers, little progress in the area of evaluation was being made. The needs of evaluation and assessment require substantial time to garner effective results. Specifically, the

nationally-approved music curriculum does not leave adequate time resources to allow pre-service teachers to become highly effective teachers before they graduate.

The National Commission on Teaching and America's Future (1996) communicated the need to develop more specific and demanding teaching standards, and to improve education to focus on student learning. Later, the National Council for Accreditation of Teacher Education (NCATE, 2000) continued this movement to create accreditation standards for pre-service teaching candidates.

Denner, Salzman, and Bangert (2001) developed an assessment of pre-service teachers using the Teacher Work Sample Methodology (TWSM) (Schalock, 1998; Schalock, Cowart, & Staebler, 1993; Schalock, Schalock, & Girod, 1997). In the adapted TWSM, pre-service teachers compiled documentation from actual lessons taught, including lesson plans, analysis of student learning, and reflection. These candidates were required to develop specific learning goals and the sequence of instruction needed to reach such goals. This study determined that teacher work samples could be linked to the assessment of pre-service teachers using indicators of student learning.

Observation instruments are often employed to determine growth and performance among pre-service teachers. Sportsman (1986) used a catalog of effective behaviors including developing an anticipatory set, stating objectives, adjusting teaching based on input, modeling ideal behavior, checking for comprehension, allowing students both guided and independent practice, and closing the lesson. In another observation instrument, White, Wyne, Stuck, and Coop (1987) classified performance objectives to determine pre-service teachers' management of instructional time, management of student behavior, presentation of instructional material, monitoring of instruction, and ability to give feedback.

The concerns of pre-service teachers themselves should be considered when designing curriculum and assessment of teaching practices. Reeves and Kazelskis (1985) found that both in-service and pre-service teachers were more concerned with the impact they had on their students than with the overwhelming nature of the job or self-focused thinking. Overall, their highest concerns involved meeting the needs of each student, challenging unmotivated students, and guiding students toward both intellectual and emotional growth.

Action research is an important component of the pre-service teaching curriculum, as it specifically pertains to self-monitoring and improving teaching practices (Gore & Zeichner, 1991). Gore & Zeichner (1991) integrated action research with the premise of reflective teaching. This practice is meant to promote student learning through more in-depth content knowledge, use research to apply a variety of meaningful teaching strategies, and focus the style and content of teaching to appeal to the students' interests and development.

A research-based teaching model is necessary to keep up with the demands of a national- and state-mandated curriculum (Grant & Drafall, 1991). Educational researchers approach this need by determining a common concern. Then preliminary testing begins to determine if there is the potential for addressing the concern in multiple classroom environments. After revisions, testing is completed in a larger scope, with results or tools implemented into classrooms. This process is more difficult for a classroom teacher, as the time constraints of a typical teacher allow for little outside research. Therefore, it is imperative that music educational researchers assist classroom teachers in researching solutions to problems.

Assessment of Pre-service Music Teachers

Early assessments of pre-service music teachers focused on a variety of teaching needs, not limited to teaching effectiveness (Grant & Drafall, 1991). Doane (1981) developed an

assessment of pre-service teachers' characteristics, all assumed to affect student learning. Within his research, he investigated the traits of music teachers and specific musical behaviors that indicate good teaching. As pre-service teachers would have had less experience with the latter, Doane focused on specific teacher traits. Within a number of teaching episodes, the least viewed traits included "uses a variety of questioning techniques," "encourages students to analyze or evaluate what they hear or perform," and "clearly explains or demonstrates ways of achieving a musical performance."

Taylor (1980) surveyed both elementary and secondary choral teachers to investigate certain competencies in music and in teaching behavior. Overall she found that communication and human relations were the most important competencies. Additionally, Taebel (1980) explored similar competencies, finding that error detection, conducting, and vocal modeling were the most important musical competency factors. Self-evaluation, classroom environment, and professionalism were among the most important teaching competency factors.

Although the NASM (2016) has created a list of expectations, Teachout (1997) investigated which skills and behaviors were most important in the first three years of teaching music. Based on the opinions of both in-service and pre-service teachers, the traits considered to be most important included maturity, leadership skills, ability to motivate students, organization, confidence, and involvement of students in the learning process. Because these are seen as integral by both levels of teachers, these traits should be considered when evaluating pre-service teachers.

Teacher effectiveness literature and research therein have garnered a wide range of recommendations in regards to pre-service music teacher assessment. Farmilo (1981) determined that creativity did not significantly impact teaching effectiveness or teaching style in elementary

teachers. Erbes (1983) recommended that the incorporation of student ideas, the appropriate use of approval, and the promotion of student interaction will build a strong classroom climate. Brand (1985b) described effective music teachers as those who paced lessons well, demonstrated a high level of musicianship, and proficiently connected music lessons to the interests of the students.

The literature regarding the assessment of pre-service teachers tends to be focused on desirable teacher traits. Less common, and perhaps equally as important, is the assessment of pre-service teachers' delivery of pedagogical content. Several more recent studies have investigated this aspect of pre-service teaching (Duke, 2009; Millican, 2012; Millican, 2016; Raiber & Teachout, 2014). Specifically, Millican (2016) investigated the identification, causes, and solutions of performance issues at the pre-service level. The performance problems most frequently identified included tone quality, movement of air, posture, stopping and starting, and tempo. Although problems themselves were identified, pre-service teachers had trouble identifying the causes of these problems. In addition, pre-service teachers would prescribe solutions to problems that did not relate to the performance issue.

There have been numerous studies that examine the behaviors of effective teachers using various evaluation forms (Brandt, 1986; Grant & Drafall, 1991; Madsen, 1988; Price, 1983; Yarbrough, 1975; Yarbrough, Price, & Bowers, 1991). Results and significance within these studies were all determined using classical test theory (CTT). While CTT can provide valuable information about a set of data, the information given is sample-dependent and test-dependent. Madsen, Standley, Byo, & Cassidy (1992) assessed effective teaching through videos using a 10-point scale. The concern with this study was that the researchers never defined "effective

teaching," as the literature did not precisely state the necessary behaviors to merit an effective teacher.

Methodology

It has been common, and long accepted, for much of research in music assessment to use methods of classical test theory. In the development of scales, factor analysis is a commonly used method for scale development (Aubrecht, Hanna, & Hoyt, 1986; Bergee, 1992; Brand, 1985a; Cocetti, 1985; Gorsuch, 1983; Hosler & Schmidt, 1985; Miksza, 2012; Nichols, 1991; Smith, 2009; Smith & Barnes, 2007; ten Holt, et al., 2010; Zdzinski & Barnes, 2002). Specifically, a facet-factorial analysis is often employed (Butt & Fiske, 1968; Greene, 2012; Russell, 2010; Wesolowski, 2016; Wesolowski, 2017). Factor analysis is a statistical method that focuses on data-reduction by using raw scores and covariance matrices as a means to identify commonality and divergence between items. A common misconception is the use of factor analysis as a measure construction method, when the use of raw scores are not indicative of measurement as they are neither linear, additive, or unidimensional (Wright & Stone, 1999). The misconception of falsely using raw scores as linear measures inhibits a scale or test from being generalizable, by locking the construction process into sample dependency, and confounds raters to performances if the assumption of independence is adequately followed (Wright, 1991).

In the development of scales, Item Response Theory (IRT) can provide a model for invariance where persons and items are organized on a latent variable. Invariance allows for measures to be compared, and for the assumption of uniformity along the latent variable in a specified frame of reference (Christensen, et. al, 2012). Under the umbrella of IRT is the Rasch Measurement Model, consisting of five unique requirements of invariance. These requirements provide a stricter model, more appropriate for the development of measures in music

performance (Engelhard & Perkins, 2011). The five requirements for invariant measurement include: (a) the calibration of the items must be independent of the particular persons used for calibration; (b) any person must have a better chance of success on an easy item than on a more difficult item; (c) the measurement of persons must be independent of the particular items that happen to be used for the measuring; (d) a more able person must always have a better chance of success on any item than a less able person; and (e) items must be measuring a single underlying latent variable. Invariant measurement is achieved when adequate fit of the data to the model is observed (Engelhard and Perkins, 2011). The methodologies used in each of the articles of this paper will employ the Rasch Measurement Model.

Organization of the Dissertation

The following three chapters are to be considered as individual research manuscripts. Each manuscript is followed by the specific resources, tables, and figures applicable to that individual study. The first manuscript details the theory behind and the steps necessary in building a measure of performance assessment. In the development of a measure to assess music performance, the traditional method of data analysis is factor analysis (Miksza, 2012; Russell, 2010; Nichols, 1991; Brand, 1985a, for example). However, Rasch Measurement Theory is a branch of item response theory that is underscored by properties of invariance using a fixed model across independent items, persons, and raters (Engelhard, 2013). The properties of invariance within Rasch Measurement Theory make Rasch the preferred method for the development of measures in the context of performance assessment. The purpose of this first manuscript is to provide a clearly defined, thirteen-step methodology for developing and validating music performance measures using Rasch Measurement Theory.

The second manuscript investigates the lesson planning practices of pre-service teachers through the lens of music education professors and public school administrators. The purpose of this study is to evaluate the psychometric quality (i.e. validity, reliability, and fairness) of a rating scale to assess pre-service teachers' lesson plan development in the context of a secondary-level music performance classroom. The research questions that guided this study included: (a) what items demonstrate acceptable model fit for the construct of lesson plan development in the context of a secondary-level music performance classroom? (b) how does the structure of the rating scale vary across items? and (c) does differential severity emerge for academic administrators or music education content specialists across items? Using multiple teacher effectiveness frameworks, plans in this study were evaluated using a four-point Likert-type rating scale (e.g., *strongly agree, agree, disagree, strongly disagree*) consisting of five domains: (a) instructional planning; (b) instructional delivery; (c) differentiated instruction; (d) assessment uses; and (e) assessment strategies. Secondary-level school administrators (n = 8) and music education content specialists (n = 8) rated 32 lesson plans. The Multifaceted Rasch Measurement Partial Credit Model was used in this study. Results suggest higher rater severity among administrators than music specialists. Implications for student teacher preparation, teacher effectiveness, and the validity of measures are discussed.

The purpose of the third manuscript is to develop of a measure to assess pre-service teachers during their student teaching experiences. Many studies regarding pre-service teachers in the student teaching setting involve perceptions and the difficulties of changing roles from student to teacher (Kelly, 2015; Paul, 1998; Teachout, 1997). In addition, other studies focus only on the rehearsal strategies used by pre-service music teachers (Bergee, 1992; Witt, 1986). This study will incorporate the expectations of teacher evaluation systems, as well as music-

specific pedagogical teaching expectations, to develop a valid, reliable, and fair measure of assessment for pre-service music teachers.

The dissertation will conclude with a discussion of implications for the field of music education and further steps to be taken. It will discuss the implementation of the development of measures within the collegiate music education classroom. Specifically, the conclusion will explore the expansion of the second manuscript into a rubric. This study will be developed further to aid students in their understanding of building individual lesson plan components. The conclusion will also focus on the importance of using Rasch Measurement Theory as future performance assessments are developed.

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Aubrecht, J. D., Hanna, G. S., & Hoyt, D. P. (1986). A comparison of high school student ratings of teaching effectiveness with teacher self-ratings: Factor analytic and multitrait-multimethod analyses. *Educational and Psychological Measurement, 46,* 223-231.

Bergee, M. J. (1992). A scale assessing music student teachers' rehearsal effectiveness. *Journal of Research in Music Education*, *40*(1), 5–13.

Brand, M. (1985a). Development and validation of the home musical environment scale for use at the early elementary level. *Psychology of Music*, *13*(1), 40-48.

Brand, M. (1985b). Research in music teacher effectiveness. *Update: The Applications of Research in Music Education, 3*(2), 13-16.

Brandt. R. S. (1986). On the expert teacher: A conversation with David Berliner. *Educational Leadership, 44*(2), 4-9.

Butt, D. S., & Fiske, D. W. (1968). Comparison of strategies in developing scales for dominance. *Psychological Bulletin, 52,* 281-302.

Campbell, M. R., & Thompson, L. K. (2007). Perceived concerns of preservice music education teachers: A cross-sectional study. *Journal of Research in Music Education, 55*, 162–176. doi:10.1177/002242940705500206

Christensen, K. B., Engelhard, Jr., G., & Salzberger, T. (2012). Ask the experts: Rasch vs. factor analysis. *Rasch Measurement Transactions*, *26*(3).

Cocetti, R. A. (1985, April). *Communication stye or leadership: The validation and interpretation of an instrument.* Paper presented at the meeting of the Central States Speech Association, Indianapolis, IN.

Coleman, J. S. (1966). *Equality of educational opportunity*. *National Center for Educational Statistics*. Retrieved from http://www.eric.ed.gov/PDFS/ED012275.pdf

Denner, P. R., Salzman, S. A., & Bangert, A. W. (2001). Linking teacher assessment to student performance: A benchmarking, generalizability, and validity study of the use of teacher work samples. *Journal of Personnel Evaluation in Education*, *15*(4), 287–307. http://doi.org/10.1023/A:1015405715614

Doane, C. P. (1981). The development and evaluation of a test to assess selected characteristics of prospective music educators. (Doctoral dissertation, Ohio State University). *Dissertation Abstracts International, 42*(10A), 4346.

Duke, R. A. (2009). *Intelligent music teaching: Essays on the core principles of effective instruction*. Austin, TX: Learning and Behavior Resources.

Engelhard Jr., G., & Perkins, A. F. (2011). Person response functions and the definition of units in the social sciences. *Measurement: Interdisciplinary research and perspectives*, *9*(1), 40–45.

Erbes, R. (1983). Teaching effectiveness: Developing a climate for music learning. *Update: The Applications of Research in Music Education, 1*(4), 7-9.

Every Student Succeeds Act, S. 1177, 114th Cong. (2015).

Farmilo, N. R. (1981). The creativity, teaching style, and personality characteristics of the effective elementary music teacher. (Doctoral dissertation, Wayne State University). *Dissertation Abstracts International, 42*(02A), 591.

Gore, J. M., & Zeichner, K. M. (1991). Action research and reflective teaching in preservice teacher education: A case study from the United States. *Teaching and Teacher Education*, *7*(2), 119–136. http://doi.org/10.1016/0742-051X(91)90022-H

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grant, J. W., & Drafall, L. E. (1991). Teacher effectiveness research: A review and comparison. *Bulletin of the Council for Research in Music Education*, *108*, 31–48.

Greene, T. (2012). An application of the facet-factorial approach to scale construction in development of a rating scale for high school marching band performance. (Doctoral dissertation, Indiana University). *ProQuestLLC, ED550211.*

Hosler, A. M., & Schmid, J. (1985). Relating factor traits of elementary, secondary, and college teachers. *Journal of Experimental Education, 53,* 211-215.

Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research*, *62*(2), 129–169.

Kelly, S. N. (2015). The influence of student teaching experiences on preservice music teachers' commitments to teaching. *Journal of Music Teacher Education*, *24*(2), 10–22. http://doi.org/10.1177/1057083713506120

Madsen, C. K. (1988). Intensity as an attribute of effective teaching. In A. Kemp (Ed.), *Research in music education: A Festschrift for Arnold Bentley* (pp. 45-52). Wiltshire, Great Britain: International Society for Music Education.

Madsen, C. K., Standley, J. M., Byo, J. L., & Cassidy, J. W. (1992). Assessment of effective teaching by instrumental music student teachers and experts. *Update: Applications of Research in Music Education*, *10*(2), 20–24.

McDonald, F. J. (1978). Evaluating preservice teachers' competence. *Journal of Teacher Education*, *29*(2), 9–13. http://doi.org/10.1177/002248717802900204

Miksza, P. (2012). The development of a measure of self-regulated practice behavior for beginning and intermediate instrumental music students. *Journal of Research in Music Education, 59*(4), 321-338. Retrieved from http://www.jstor.org.proxy-remote.galib.uga.edu/stable/41348841.

Millican, J. S. (2012). *Starting out right: Beginning band pedagogy.* Lanham, MD: Scarecrow Press.

Millican, J. S. (2016). Describing preservice instrumental music educators' pedagogical content knowledge. *Update: Applications of Research in Music Education*, *34*(2), 61–68. http://doi.org/10.1177/8755123314552664

National Association of Schools of Music (2016). *National Association of Schools of Music Handbook*.

National Commission on Teaching and America's Future. (1996). *What Matters Most: Teaching and America's Future.* New York: Teachers College.

National Council for Accreditation of Teacher Education. (2000). *NCATE 2000 Unit Standards.* Washington, DC: Author.

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

Nodine, T. R. (2016). How did we get here? A brief history of competency-based higher education in the United States. *The Journal of Competency-Based Education, 1*(1), 5-11. DOI: 10.1002/cbe2.1004

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

Porter, A. C., & Brophy, J. (1988). Synthesis of research on good teaching: Insights from the work of the Institute for Research on Teaching. *Educational Leadership*, *45*(8), 74-85.

Price, H. E. (1983). The effect of conductor academic task presentation. conductor reinforcement, and ensemble practice on performers: Musical achievement, attentiveness, and attitude. *Journal of Research in Music Education, 31*, 245-257.

Public Law 107–110, No Child Left Behind of 2001, H. Res. 1, 107th Cong., 115 Stat. 1425 (2002) (enacted).

Raiber, M., & Teachout, D. (2014). *From music student to teacher*. New York, NY: Routledge.

Reeves, C. K., & Kazelskis, R. (1985). Concerns of preservice and inservice teachers. *The Journal of Educational Research*, *78*(5), 267–271. http://doi.org/10.1080/00220671.1985.10885614

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

Schalock, M. (1998). Accountability, student learning and the preparation and licensure of teachers: Oregon's Teacher Work Sample Methodology. *Journal of Personnel Evaluation in Education, 12,* 269-285.

Schalock, M., Cowart, B., & Staebler, B. (1993). Teacher productivity revisited: Definition, theory, measurement, and application. *Journal of Personnel Evaluation in Education, 7,* 179-196.

Schalock, H. D., Schalock, M., & Girod, G. (1997). Teacher Work Sample Methodology as used at Western Oregon State College. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* (pp. 15-45). Thousand Oaks, CA: Corwin, Press.

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

Smith, D. T. (2009). Development and validation of a rating scale for wind jazz improvisation performance. *Journal of Research in Music Education*, (3), 217.

Sportsman, M. A. (1986). Evaluating teacher effectiveness fairly. *Curriculum Review, 25*(4), 8-10.

White, K., Wyne, M. D., Stuck, G. B., & Coop, R. H. (1987). Assessing teacher performance using an observation instrument based on research findings. *NASSP Bulletin, 71,* 89-95.

Witt, A. C. (1986). Use of class time and student attentiveness in secondary instrumental music rehearsals. *Journal of Research in Music Education*, *34*(1), 34–42. http://doi.org/10.2307/3344796

Taebel, D. K. (1980). Public school music teachers' perceptions of the effect of certain competencies on pupil learning. *Journal of Research in Music Education, 28,* 185-197.

Taylor, B. P. (1980). The relative importance of various competencies needed by choral-general music teachers in elementary and secondary schools as rated by college supervisors, music supervisors, and choral-general music teachers. (Doctoral dissertation, Indiana University). *Dissertation Abstracts International, 41*(07A), 2990.

Teachout, D. J. (1997). Preservice and experienced teachers' opinions of skills and behaviors important to successful music teaching. *Journal of Research in Music Education*, *45*(1), 41–50. http://doi.org/10.1080/13540600601152546

ten Holt, J. C., van Duijn, M. J., & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling*, *52*(3), 272-297.

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.

Wesolowski, B. C. (2017). A facet-factorial approach towards the development and validation of a jazz rhythm section performance rating scale. *International Journal of Music Education, 25*(1), 17-30.

Wright, B. D. (1991). Factor analysis versus Rasch analysis of items. *Rasch Measurement Transactions, 5*(1), 134-135. Retrieved from https://www.rasch.org/rmt/rmt51i.htm.

Wright, B. D. & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, Delaware.

Yarbrough, C. (1975). The effect of magnitude of conductor behavior on performance, attentiveness, and attitude of students in selected mixed choruses. *Journal of Research in Music Education, 23*, 134-146.

Yarbrough, C., Price, H. E., & Bowers J. K. (1991). The effect of knowledge of research on rehearsal skills and teaching values of experienced teachers. *Update: Applications of Research in Music Behavior, 9*(2), 17-20.

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, (3), 245.

.

CHAPTER 2

OLD WINE IN NEW BOTTLES: A METHODOLOGY FOR DEVELOPING AND VALIDATING PERFORMANCE MEASURES USING MODERN MEASUREMENT THEORY[1]

[1] Musselwhite, D. J., & Wesolowski, B. C. (in press, 2018). Old wine in new bottles: a methodology for developing and validating performance measures using modern measurement theory. In T. Brophy, M. Fautly (Eds.), *Context Matters: Proceedings from the 6th International Symposium on Assessment In Music Education (ISAME)*, Birmingham, UK.

Abstract

In the development of a measure to assess music performance, the traditional method of data analysis has been factor analysis. However, Rasch Measurement Theory is a branch of item response theory that is underscored by properties of invariance using a fixed model across independent items, persons, and raters. It is because of the properties of invariance that Rasch Measurement Theory is the preferred method for the development of measures in the context of performance assessment. The purpose of this paper is to provide a clearly defined, thirteen-step methodology for developing and validating music performance measures using Rasch Measurement Theory.

Keywords: Invariance, Measurement, Rasch, Rating Scale, Validity

OLD WINE IN NEW BOTTLES: A METHODOLOGY FOR DEVELOPING AND VALIDATING PERFORMANCE MEASURES USING MODERN MEASUREMENT THEORY

In the development of a measure to assess music performance, the traditional method of data analysis is factor analysis (Miksza, 2012; Russell, 2010; Ten Holt, et al., 2010; Smith, 2009; Smith & Barnes, 2007; Zdzinski & Barnes, 2002; Nichols, 1991; Brand, 1985, for example). Factor analysis is a statistical method rooted in the Classical Test Theory tradition with the purpose of describing the variability among correlated variables, using raw scores and covariance matrices. Raw scores are not indicative of measurement because they are not linear, additive, or unidimensional (Wright & Stone, 1999). The use of factor analysis is an acceptable data analysis method for very specific purposes (for example, when interest is in reducing data while also defining a latent variable).

Rasch Measurement Theory (RMT) is a method of analysis that offers valid measures that, when developed, are independent from the sample used (Granger, 2008). RMT is a branch of item response theory that is underscored by properties of invariance using a fixed model across independent items, persons, and raters. Due to the requirements of invariance, students' level of achievement, items' level of difficulty, and rater severity in the context of performance evaluation will not affect the overall model. It is because of the properties of invariance that RMT is the preferred method for the development of measures in the context of performance assessment. In particular, there are five requirements for invariant measurement (Engelhard & Perkins, 2011): (a) the calibration of the items must be independent of the particular persons used for calibration; (b) any person must have a better chance of success on an easy item than on a more difficult item; (c) the measurement of persons must be independent of the particular items that happen to be used for the measuring; (d) a more able person must always have a better

chance of success on any item than a less able person; and (e) items must be measuring a single underlying latent variable. Specifically, this paper addresses the development of performance assessments where raters are used to gather data. Invariance is a property that is defined by empirical data, specifically model-data fit. With the inclusion of raters, rater-invariant measurement must also be determined. This property implies that persons and raters are independent (Wind & Engelhard, 2013).

The process of constructing a measure for music performance should be guided by two underlying questions:

1. How can raw score data be collected from raters in a valid and meaningful way?
2. How can test construction and development be handled in order to make inferences that are valid, reliable, and fair?

The thirteen-step methodology described in this paper provides a framework for developing measures in the context of music performance. Due to the limitations of the length of this paper, this methodology should be considered a basic framework. Throughout the paper, aspects of decision making will be addressed in relation to the process of test construction and development. The purpose of this paper is to provide a clearly defined methodology for developing and validating music performance measures using Rasch Measurement Theory.

Step 1: Observational Design. The observational design refers to the content and design of the items. The researcher must envision the construct he/she wants to build, then think about the items that would best describe that construct. After consulting subject matter experts and various pedagogical and methodological resources, item construction can begin, and items can be

qualitatively grouped into a priori domains. These domains and related items become the framework for the measure.

Step 2: Decide Between Using a Rating Scale or Rubric. There are two types of preferred response formats: rating scales or rubrics. The researcher may choose to use ether a rating scale response format or a rubric-based response format based upon the needs and requirements of the assessment context (time, detail, test requirements, requirements of stakeholders, for example). By choosing a rubric-based response from the start, more work is required up front. In a rubric, categories of performance are listed (i.e., tone, articulation, posture) with accompanying levels of performance. Also important is the terminology used across the categories within the rubric. The language used in each category must be consistent. For example, the type of language to address tone could be level of desirability (e.g., very undesirable, undesirable, desirable, and very desirable). The type of language to address appropriate use of articulation could be level of acceptability (e.g., totally unacceptable, slightly unacceptable, slightly acceptable, and perfectly acceptable) (see Vagias, 2006). It is preferred for there to be between three and five levels at most (Dumas, 1999; Wright, 1977).

A rating scale is different in that a statement is given to the rater, then the rater must decide the level of agreement based on the performance (i.e., strongly agree, agree, disagree, and strongly disagree). The previously discussed domains and subsequent items can be paired with a Likert-type scale in order to meaningfully design a rating scale structure. While a five-category Likert scale is most common, it does not provide meaningful feedback. By removing a middle category (i.e. undecided, neutral), the rater is forced to make a choice, and the researcher is provided with a more accurate picture of the performance (Cox, 1980).

Step 3: Design a Judging Plan. Raters may be organized in a variety of ways. The type of linking design chosen will have an effect on the amount of information and related standard error of the assessment context (Wind, Engelhard, & Wesolowski, 2016). Rater variability is a necessary component in the development of a measure, as multiple perspectives serve to improve the validity of the measurement instrument (Wilson, 2005). In a complete linking design, every judge or rater will evaluate every performance. While this is the most reliable of the linking designs, a complete system has drawbacks. There may be an increased cost due to the workload of every rater having to evaluate every performance. Potentially, raters may drop out due to time and energy constraints. In addition, the time requirement could impact consistency among raters. In an incomplete design, there are more raters and more performances to evaluate, therefore more information will be provided by the design. Here, all raters are involved, but they will not evaluate every performance. There are multiple incomplete designs. For example, Rater 1 will evaluate performances 1, 2, 3, and 4. Rater 2 will evaluate performances 3, 4, 5, and 6. Every performance will be judged by at least two raters, but raters are not having to spend a large amount of time rating.

Step 4: Collect Rater Data. Using the predetermined items and specified rating design, the researcher must develop a pilot measure to conduct with a sample group. This is the first time the raters are interacting with the measure. The raters must be instructed as to word choice, meaning, and the overall operational procedure of the performance assessment. Data must then be collected in a systematic way.

Step 5: Analyze the Data. Two models may be considered based on the qualitative decision-making of the researcher: The Rating-Scale Model (RSM) (Wright & Masters, 1982) or the

Partial-Credit Model (PCM) (Masters, 1982). Linacre (2000) outlines the decision making process between the Rating-Scale Model and the Partial-Credit (Linacre, 2000):

1. Design of the items: If the items are clearly intended to use the same rating scale throughout (e.g., a Likert-type scale), then the RSM should be used. If each item is intended to have a different rating scale, then the PCM should be used.
2. Communication: Each item should match the response-options. A question/item that merits a yes/no response should not be followed by four Likert-scale responses.
3. Size of the dataset: There should be at least 10 observations in each category. This will prevent accidents in the data. However, if the sample size does not allow for 10 observations per category, the RSM should be considered over the PCM.
4. Construct and Predictive Validity: If there is a meaningful difference between the item abilities and between the person abilities, the PCM should be used.
5. Fit Considerations: Underfit is a greater threat to validity than overfit. It is imperative to examine parameter-level fit statistics in addition to the fit statistics for each element. If the fit is poor, then better data is needed for the intended purposes. This is not an indication of the need for a better model (see Step 6).
6. Category Thresholds: In the PCM, category thresholds (i.e., step difficulties between rating scale categories) are unknown before data collection. In the RSM, the thresholds are set in advance.
7. Unobserved categories: In the PCM, unused categories will distort the structure of the rating scale. When there is an unobserved category in the RSM, its function is inferred from other items that employ the same category.

8. Statistical information: Both the PCM and RSM provide the same statistical information, therefore there is no benefit of choosing one over the other in this regard.
9. Optimization: Optimization refers to a process where careful examination of the items will lead to more effective use of the rating scale structure (see Step 10). Specifically, categories may need to be collapsed in order to achieve a lower standard error. In order to optimize the rating scale structure, the PCM should be used (Linacre, 2000).

Step 6: Evaluate Parameter-Level Fit Statistics. Parameter-level fit statistics will help determine overall how the components are working. A parameter refers to a measurable factor that is essential to understanding a set of data. Parameters may include items, persons, and raters. Specifically, in the context of music performance assessment, parameter-level statistics look at the student performances, items, and raters within the music performance assessment to see how these components are performing in the model. The range of reasonable mean-square fit values can change depending on the context of assessment (Wright & Linacre, 1994). As an example, there are five contexts: (a) high stakes, (b) run of the mill, (c) survey, (d) clinical observation, and (e) judged test, where agreement is encouraged. The choice of fit statistic thresholds is a qualitative decision.

Fit statistics (e.g., infit and outfit) describe the degree to which invariant measurement is achieved. Infit Mean Squares refers to data fit that is sensitive to inliers (Linacre, 2002). This statistic focuses on the configuration of responses to items aimed on the person. Outfit Mean Squares refers to data fit being outlier-sensitive. This statistic looks at any data that may lie far from the person and looks at what may affect the patterning of responses. Mean squares show how much randomness occurs in the specified data set. The expected mean square error statistics should be close to 1.00 with very little variation within the linear scale: usually a standard

deviation of 0.20 at the most. Infit problems can be seen as a bigger threat to measurement, and therefore should be evaluated first (Linacre, 2002). For example, if the infit and outfit statistics fall within the range of 0.80-1.20, it can be concluded that the data demonstrates acceptable levels of invariance for that context. If the infit and outfit statistics fall outside the range of 0.80-1.20, it can be concluded that the data demonstrates unacceptable levels of invariance for that context should be qualitatively evaluated as to how the parameter can be improved (Wright & Linacre, 1994).

Step 7: Evaluate Fit Statistics for Elements. After fit statistics for the entire measure have been obtained, each facet can be examined to determine individual fit statistics for each element. An element is the individual component of the parameter. For example, in the item parameter, a specific item would be an element. The same thresholds from the parameter level will hold true within each element. The focus here, however, is on individual items, individual performances, and/or individual raters. The evaluation of fit statistics at the element level provides important diagnostic and qualitative information on how individual performers, items, and raters performed within the model.

Step 8: Manage Misfit. Misfit, quantitatively, means the item, student, or rater lies outside of the specified threshold described in Step 6 and Step 7. In the analysis of items, misfit should not be viewed as a "bad item." Rather, misfit should be valued as an opportunity to learn and investigate. The same is true for misfitting raters and misfitting performances. Misfit should fuel the rewriting of items and draw attention to content and construct validity concerns.

Step 9: Refine the Measure. Misfitting items should either be removed from the measure or rewritten based upon qualitative decision making. Once the items have been removed or

rewritten, the items should go through a follow-up pilot test. In the follow-up study, the same considerations for fit should be applied.

Step 10: Evaluate and Optimize the Rating Scale Structure. The structure of the rating scale can be evaluated when specifically using the Partial Credit Model (PCM). Linacre (2002) provides a set of nine guidelines for optimizing this rating scale structure.

1. All items within the rating scale should align with one latent variable.
2. There should be at least 10 observations per rating scale category.
3. There should be a uniform distribution of observations across categories.
4. Average measures advance monotonically with each category.
5. Outfit Mean Squares are less than 2.00.
6. Step calibrations should advance (showing that category usage is regular).
7. The ratings imply measures, and the measures imply ratings.
8. Step difficulties advance by at least 1.4 logits.
9. Step difficulties advance by at most 5.0.

In this step, each item is examined individually to find out how the categories were used. Each item must meet all of the guidelines listed above in order to justify the use of each individual category within the context of the rating scale structure.

Step 11: Refine the Learning Outcomes. If a rating scale has been used, and the developer wants to transition into a rubric, rewriting of the items is necessary. Item stems should be rephrased without directionality as to resemble outcome criteria. For example, an item worded as, “Student performs excerpt with desired tone,” could be rephrased as simply “Tone.” This step aids in the process of transitioning from rating scale to rubric.

Step 12: Write Qualitative Descriptors. Each of the four levels of performance should now be written to describe a specific outcome related to the item stem. Using the aforementioned item stems, all scale categories must be represented with similar tone and language (see Vagias, 2006).

Step 13: Begin the Revalidation Process. The rubric should be revalidated using all previous steps. Once the rating scale items have transformed into a rubric, the rubric must once again be tested for reliability and validity in the same manner described above. Assessment contexts can have an effect on performance of a measurement instrument, therefore it is important to consistently be evaluating items, performances, and raters in the context of a performance assessment. A rubric is a living breathing organism that can change based upon objects of measurement, raters, context, standards change.

In a performance based assessment in psychological sciences (i.e. music), constructs must be defined and inferred through secondary behaviors (i.e., tone, articulation, posture). Music performance can be adequately assessed through the inferences from these secondary behaviors. In order to make inferences that are valid, reliable and fair, researchers and educators should be using tools that have been well-developed and maintained. The use of a measurement instrument in any context should be closely monitored and evaluated for its properties of invariance.

References

Brand, M. (1985). Development and validation of the home musical environment scale for use at the early elementary level. *Psychology of Music*, *13*(1), 40-48.

Cox, III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*, 407-422.

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

Engelhard Jr., G., & Perkins, A. F. (2011). Person response functions and the definition of units in the social sciences. *Measurement: Interdisciplinary research and perspectives*, *9*(1), 40–45.

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

Granger, C. V. (2008). Rasch analysis is important to understand and use for measurement. *Rasch Measurement Transactions, 21*(3), 1122-1123.

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

Linacre, J.M. (2000). Comparing "partial credit models" (PCM) and "rating scale models" (RSM). *Rasch Measurement Transactions, 14*(3), 768.

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

Miksza, P. (2012). The development of a measure of self-regulated practice behavior for beginning and intermediate instrumental music students. *Journal of Research in Music Education, 59*(4), 321-338. Retrieved from http://www.jstor.org.proxy-remote.galib.uga.edu/stable/41348841.

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

Russell, B. (2010). The development of a guitar performance rating scale using a facet-factorial approach. *Bulletin of the Council for Research in Music Education,* (184), 21-34. Retrieved from http://www.jstor.org.proxy-remote.galib.uga.edu/stable/27861480.

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

Smith, D. T. (2009). Development and validation of a rating scale for wind jazz improvisation performance. *Journal of Research in Music Education*, (3), 217.

Ten Holt, J. C., van Duijn, M. J., & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling*, *52*(3), 272-297.

Vagias, W. M. (2006). *Likert-type scale response anchors.* Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson Management.

Wind, S. A. & Engelhard, Jr., G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18*, 278-299.

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

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

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

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

Wright, B. D. & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, Delaware.

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, (3), 245.

CHAPTER 3

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[2]

[2] Musselwhite, D. J., & Wesolowski, B. C. (in press, 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.*
Reprinted here with permission from the publisher.

Abstract

The purpose of this study is to evaluate the psychometric quality (i.e. validity and reliability) of a rating scale to assess pre-service teachers' lesson plan development in the context of a secondary-level music performance classroom. The research questions that guided this study include: (a) what items demonstrate acceptable model fit for the construct of lesson plan development in the context of a secondary-level music performance classroom? (b) how does the structure of the rating scale vary across items? and (c) does differential severity emerge for academic administrators or music education content specialists across items? Using multiple teacher effectiveness frameworks, plans in this study were evaluated using a four-point Likert-type rating scale (e.g., strongly agree, agree, disagree, strongly disagree) consisting of five domains: (a) instructional planning; (b) instructional delivery; (c) differentiated instruction; (d) assessment uses; and (e) assessment strategies. Secondary-level school administrators (n = 8) and music education content specialists (n = 8) rated 32 lesson plans. The Multifaceted Rasch Measurement Partial Credit Model was used in this study. Results suggest higher rater severity among administrators than music specialists. Implications for student teacher preparation, teacher effectiveness, and the validity of measures are discussed.

Keywords: Lesson Plan, Rating Scale, Rasch model, Reliability, Validity

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

Lesson plan development is an integral component of the teaching process (Butt, 2006; Coppola, et al., 2004). In this study, lesson plan development involves defining the learning outcomes and the methodological process that is to be taken by the students and teacher in order to reach such outcomes. The practice of pre-planning objectives, assessments, appropriate materials, teaching sequences, and student pacing is important for the establishment of a learning environment conducive toward optimizing student success (Frey, Fisher, & Moore, 2005; Brittin, 2005). High-quality lesson-planning skills are associated with more successful teaching practices and higher teaching competencies (Brittin, 2005; Butt, 2006; Lane & Talbert, 2013; Miksza & Berg, 2013; Schmidt, 2005; Scott, 2012). Furthermore, a teacher's prioritized attention to planning is vital in order to reach the needs of diverse students (Houston & Beech, 2002). Specifically, learning to use time effectively to plan is a skill with which pre-service teachers seem to struggle (Houston & Beech, 2002).

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

opportunities-to-learn in the arts, but may also affect consistency of teacher evaluations due to the lack of cross-district coherence in music curricula.

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

A central content standard of national undergraduate curricula, and more specifically undergraduate music education curricula, is for pre-service teachers to demonstrate competency designing effective musical instruction through the development of lesson plans (National Association of Schools of Music, 2016; Council for the Accreditation of Educator Preparation, 2016). In the context of secondary-level music performance methods classes, lesson planning is often geared toward a mock student audience because many students are planning lessons for their peers (Paul, 1998). The resulting lesson plans at the pre-service level regularly reflect inconsistencies in sequencing, assessment, and the methodologies of sequence-based rehearsal strategies or conceptual lessons (Lane, 2006; Schleuter, 1991; Schmidt, 2005). In addition, pre-service teachers tend to be vague in their procedural descriptions and are not specific in their learning goals for students (Brittin, 2005; Lane, 2006; Schmidt, 2005). Therefore, the evaluation

and measurement of students' lesson plan writing must be integrated into the undergraduate curriculum and guided by valid and reliable measurement instruments.

Achievement in lesson plan development is not directly observable and is therefore considered to be a latent (i.e., unobservable) construct (Baghaei, 2008). Therefore, secondary observable behaviors are needed to operationally define the construct intended to be measured, that in the context of this study, come in the form of criteria, or judgmental cues, within the measurement instrument (Wesolowski, Wind, & Engelhard, 2016). The criteria set forth within a measurement instrument operationally define the latent construct and help support construct validity arguments. In order to properly evaluate and measure pre-service teachers' "performances" of lesson plan development and related levels of "achievement," a validated measure is needed to outline the secondary, observable behaviors that define the construct of "lesson plan development." The purpose of this study is to evaluate the psychometric quality (i.e., validity and reliability) of a rating scale to assess pre-service teachers' lesson plan development in the context of a secondary-level music performance classroom. The research questions that guided this study include:

1. What items demonstrate acceptable model fit for the construct of lesson plan development in the context of a secondary-level music performance classroom?
2. How does the structure of the rating scale vary across items?
3. Does differential severity emerge for academic administrators (e.g., principals and assistant principals) or music education content specialists (e.g., university music education faculty) across items?

Background

Teacher Accountability

The National Education Association (NEA) indicates that the implementation of high quality teacher evaluation systems leads to better teaching practices, thereby advancing student learning (NEA, 2011). The NEA further recommends that "highly trained evaluators" should conduct the evaluation of teachers (NEA, para. 3). These evaluators should use clear, rigorous standards that explicitly specify the depth of knowledge, skills, abilities, and responsibilities of teachers (NEA, 2011). Models for teacher evaluation can come from national models such as the NEA Principles of Professional Practice, or from state-adopted, research-based models such as the Danielson (2013), Marzano (2013), Stronge (2013), and Mid-continent Research for Evaluation and Learning (McREL) (The Center for Educator Effectiveness, 2013) frameworks. It is important to note that these frameworks are intended for the in-service teacher and will differ from the expectations within a pre-service teaching curriculum. However, teaching frameworks are important for the development of pre-service teachers and for expectations for achievement in the field. Specifically, for the pre-service or early-career in-service teacher, teacher effectiveness frameworks provide structure in a complex field (Danielson, 2007).

With the implementation of new teacher certification processes such as edTPA (2015) and with heavy reliance on teacher evaluation frameworks, both pre-service and early-career teachers must quickly synthesize and demonstrate marked achievement of the various framework expectations of lesson plan development. Although the edTPA does not specifically employ one of the aforementioned frameworks, pre-service teachers may benefit from an introduction to these systems. Regarding lesson plan development in particular, the overarching goal of these frameworks is to increase student achievement through the clear documentation of teaching

practices and through gathering evidence of student learning. Teaching, in this context, is an intricate task that links a teacher's knowledge, skills, and character to meet the educational needs of the students (The Center for Educator Effectiveness, 2013).

Lesson Planning Dimensions of Teacher Evaluation Frameworks

Lesson planning is often emphasized as a pivotal aspect of the teaching process (Akyuz, Dixon, & Stephan, 2012). Specifically, lesson planning allows for the thoughtfulness of detailed methodologies, where the teacher can continually adjust and improve instruction (Kilpatrick, Swafford, & Findell, 2001). Teacher effectiveness frameworks aim to diagnose strengths and weaknesses not only in lesson planning, but also in the effectiveness of teaching practices.

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

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

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

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

Although these frameworks have a practical application toward the traditional classroom, concerns of validity have been raised in the context of music teaching (Wesolowski, 2014; Wesolowski, 2015). A disparity can potentially occur in the observation process when administrators evaluate teachers of the arts. Unless an administrator has had prior training or experience in the performing arts, a performing arts teacher may not be evaluated fairly. Music teachers may be assessed with the expectation that their classroom should mirror that of a

traditional, academic teacher (i.e., mathematics, science, history), for example, with more transparent differentiation. Therefore, one important research question of this study is to investigate the difference of ratings between academic administrators (i.e., principals and assistant principals) and music education content specialists (i.e., university music education faculty).

Psychometric Considerations

Item Response Theory is a branch of test theory where the specific qualities of an individual or group, and the qualities of specific items, will have an impact on an individual's or a group's response to an item (Furr & Bacharach, 2007). The Rasch Measurement Model is a specific version of the one-parameter-logistic 1-PL model under the umbrella of Item Response Theory (IRT). The Rasch Measurement Model was used in this study in order to construct a linear measure from raw scores. The benefit of Rasch Measurement is that when the data adequately fit the model, invariant measurement is achieved. Engelhard and Perkins (2011) define invariant measurement through five requirements: (a) the calibration of the items must be independent of the particular persons used for calibration (i.e., person-invariant calibration of test items); (b) any person must have a better chance of success on an easy item than on a more difficult item (i.e., non-crossing item response functions); (c) the measurement of persons must be independent of the particular items that happen to be used for the measuring (i.e., item-invariant measurement of persons); (d) a more able person must always have a better chance of success on any item than a less able person (i.e., non-crossing person response functions); and (e) items must be measuring a single underlying latent variable: (i.e., unidimensionality as evidenced through a variable map). These requirements are defined in the context of cognitive-based exams, where the test-taker (i.e., person) directly interacts with the items on an exam. In

the context of this study, items refer to the rubric criteria, and "persons" refer to the lesson plans. Model-data fit is achieved when all of these requirements are met. Evidence of model-data fit is necessary for providing: (a) an interpretation of construct and content reliability of the measurement instrument (research question 1); (b) a definition of the locations of the thresholds for each rating scale category across each individual item (research question 2); and (c) evidence of systematic differential severity between rater-type (e.g., academic administrators and music education content specialists) across items (research question 3).

The Rasch-based statistics explored in this study were calculated using *FACETS* (Linacre, 2014). Specifically, this study employs the Multifaceted Rasch Partial Credit Model (MFR-PC). This model requires that all achievement levels available to raters on a measurement instrument be identified and ordered prior to the distribution of the items (Masters, 1982). These levels of achievement only indicate an ordering, and do not imply any categorical weighting. The PC version of the MFR model treats each rating scale category for each item independently, providing a more precise outcome estimate than the MFR model alone. The Partial Credit Model is as specified as follows:

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

where

$\ln\left[\frac{P_{nijmk}}{P_{nijmk-1}}\right]$= the natural log of the probability that Performance *n* rated by Rater *I* on Item *j* in level *m* receives a rating in category *k* rather than category *k*-1;

θ_n = achievement level of lesson plan *n*;

λ_i = severity of rater *I*;

δ_j = difficulty of item *j*;

γ_m = rater type m (e.g., academic administrator or music education content specialist);

τ_{ik} = the location on the logit scale where rating scale categories k and $k - 1$ are equally probable for Rater i.

$\lambda_i\gamma_m$ = interaction term between rater severity and rater type.

In this study, each of the rubric criteria contain four response levels within the rating scale structure: Strongly Disagree, Disagree, Agree, and Strongly Agree.

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

Method

Initial item pool generation, Raters, and Judging Plan

Items for evaluating lesson plans were gathered from performance standards from each of the teacher evaluation frameworks (reviewed earlier, Danielson, 2013; Marzano, 2013; Stronge, 2013; The Center for Educator Effectiveness, 2013; Woods, 2015). Four areas were found relevant to be assessed using only a pre-service teacher's lesson plan of the various performance standards in each of the frameworks: (a) instructional planning; (b) instructional strategies; (c) differentiated instruction; and (d) assessment strategies. These indicators, combined with performance indicators from other frameworks, became the structure for the preliminary lesson plan rating scale (see Figure 3.1). Relevant items from each of the frameworks were removed and transformed into statements applicable for the assessment of a music-specific lesson plan. In order to inspect face validity of the criteria, the authors and one outside university music education professor screened the item pool for clarity, writing style, and redundancy. Any items that appeared unclear or redundant were removed from the overall item pool. The remaining items ($N = 34$) were listed in a randomized order.

Anonymous lesson plans were collected from undergraduate music education majors at a large southern university, and students were given informed consent (see Figures 3.2, 3.3, 3.4). These students ranged from second to fifth year undergraduate students. All identifying information was removed from each lesson plan to maintain student anonymity. Lesson plans were written for both middle school and high school level, including band, orchestra, and choral content matter. A total of 32 lesson plans were used in the study, meeting the minimum sample requirement to produce statistically stable measures with a 95% confidence interval (Linacre, 1994).

The lesson plans were sent to 16 volunteer raters: (a) university music education faculty (n = 8), and (b) academic administrators (principals, n = 1; assistant principals, n = 7). Raters were solicited based upon reputation, record of success within their field, and availability. Accompanying each lesson plan was the initial rating scale (Figure 3.1). Each rater independently evaluated each of four lesson plans using the 34 rating scale items on the included rating scale. The rating scale structure for each item on the rating scale was based upon a four-point Likert-type scale. The response alternatives included "Strongly Agree," "Agree," "Disagree," and "Strongly Disagree." A four-point rating scale structure was chosen specifically due to its absence of a neutral category, thereby requiring a forced-choice, resulting in a better estimate of raters' attitudes (Dumas, 1999; Wright, 1977).

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

Rater Judging Plan

The judging plan was a balanced incomplete assessment network (Engelhard 1997). This judging plan ensures reliability and validity both within and between facets, as recommended by the judging plan recommended by Linacre and Wright (2004) and Wright and Stone (1979). Each rater evaluated four lesson plans. For example, Rater 1 evaluated Lesson Plans 1, 2, 3, and 4. In this particular judging plan, overlap was needed to ensure there is no bias in the rating. So,

Rater 2 evaluated Lesson Plans 3, 4, 5, and 6. This pattern would continue until Rater 16 evaluated Lesson Plans 31, 32, 1, and 2, when the circuit was complete. Therefore, every lesson plan was evaluated twice by each type of rater, and no single rating weighed more heavily than another. These lesson plans have been linked sufficiently based upon a sound data collection design (Engelhard, 1997; Kirk, 1995; Wind, Engelhard, & Wesolowski, 2016). This form of an incomplete assessment network was verified to demonstrate the best model data fit among multiple incomplete assessment network structures (Wesolowski, Wind, & Engelhard, 2016).

Wright Map

The Rasch model indicates its unidimensionality by displaying all facets on a linear scale. This display is the Wright Map, which depicts the operational definition of the latent construct. This Wright Map displays lesson plan difficulty, rater severity, item difficulty, and rater type on one scale (see Figure 3.5). The first column of the Wright Map is the logit-scale measure, which is the underlying scale for all facets. This scale is composed of equally spaced units representing the unidimensional latent construct. The second column indicates the distribution of lesson plans using asterisks, from high achieving to low achieving. The third column is the location of raters, from most severe to most lenient. The fourth column is location of rater type, from most severe to most lenient. The fifth column is the location of items from the rating scale, from most difficult to easiest.

Results

In this study, the MFR-PC model was used to evaluate the validity and reliability of a rating scale to assess pre-service teachers' lesson plan development in the context of a secondary-level music performance classroom. The descriptions provided in this section are focused on separation, as evidenced through chi-square statistics and their related reliability of

separation statistics, model-data fit, and logit-scale locations, shown on the Wright map and through the calibration of elements (i.e., each lesson plan, each rater, each item, for example) within each facet.

Summary Statistics

Table 3.1 provides the summary statistics for the MFR-PC model using *FACETS* (Linacre, 2014) for lesson plans (θ), raters (λ), items (δ), and rater type (γ). The analysis indicated overall significant differences for lesson plans ($\chi^2 = 499.2$, $p < .01$), raters ($\chi^2 = 621.8$, $p < .01$), items ($\chi^2 = 296.2$, $p < .01$), and rater type ($\chi^2 = 118.5$, $p < .01$). Reliability of separation is also reported for each facet. Specifically, reliability of separation refers to the reproducibility of the relative measure location (Linacre, 2017). This characteristic is interpreted similarly to Cronbach's alpha in its estimation of the spread of elements within a facet. Overall, high reliabilities of separation between lesson plans ($REL_{\text{LessonPlans}} = .94$), raters ($REL_{\text{Raters}} = .97$), items ($REL_{\text{Items}} = .89$), and rater type ($REL_{\text{RaterType}} = .98$) indicate that the Lesson Plan Evaluation Rating Scale was able to reliably separate each facet from the underlying latent trait of lesson plan achievement. More specifically, the lesson plans were able to be reliably separated at varying achievement levels across the unidimensional continuum. Both raters (.97) and items (.89) were able to separate lesson plans based on variability in achievement with reasonable reliability. Regardless of rater type, varying achievement levels of lesson plans were able to be distinguished.

Model-data fit. Fit statistics indicate the degree to which invariant measurement is achieved. Specifically, infit and outfit statistics are used to determine how invariant the data is. Infit Mean Squares refers to the fit of the data that is sensitive to inliers, focusing on individual person responses (Linacre, 2002). Outfit Mean Squares look at the fit of the data in response to

outliers, focusing on potential effects on response patterns. Overall, mean squares seek to determine randomness that occurs in the data set. Table 3.1 and Table 3.2 indicate that the Mean Infit and Outfit MSE are centered near 1.00. In the strictest view, infit and outfit statistics should fall within the range of 0.8 – 1.2, implying acceptability of invariance among the data. If a statistic falls outside of this invariant range, the statistic and its related element should be carefully evaluated (Wright & Linacre, 1994). An indication of good model-data fit is evidenced through fit statistics falling within Wright and Linacre's specified range. As a result, evidence of good model-data fit indicates a degree of reasonable invariant measurement that produces interpretable estimates of measurement. When invariant measurement is achieved, along with high reliability of separation, we can infer the trustworthiness of the score interpretation (Baghaei, 2008). More broadly, the presence of invariant measurement, and therefore the trustworthiness of score interpretation, yields a strong argument for construct validity, as depicted in the Wright Map (see Figure 3.5).

Table 3.3 indicates the function of the rating scale categories for each item. In other words, this table shows how the categories from each item were used by the raters. Items are listed in numerical order, as they appeared to each rater. Columns 2 – 5 indicate the raw score of instances when each category was used on a particular item. The percentage is shown in parentheses. Columns 6 – 9 indicate the average observed measure. This number indicates where the item falls on the logit scale. Columns 10 – 13 detail each item's Outfit MSE per category. Again, this statistic should fall in the range of 0.8 – 1.2, so misfit items may be detected here. Misfit items include, but are not limited to: Item 3 in category 1, Item 5 in category 1, and Item 8 in category 4. Each category that is found to be misfit is first evaluated and potentially eliminated from the rating scale category structure. Critical judgments must be made in this

process, as the stepwise ordering between response categories must remain intact. For example, if an item demonstrates evidence of misfit in category 2, it would warrant the consideration of collapsing (i.e., combining) into an adjacent category in order to maintain a stepwise ordering within the category structure.

There are multiple approaches to the collapsing of categories. The purpose of collapsing categories is to properly organize disordered thresholds. Linacre (2004) suggests that every response category should have at least 10 observations, and that observations should be distributed somewhat evenly among the categories. Bond and Fox (2015) suggest only collapsing categories when it makes substantive sense. Items leaning in only one direction (items 12, 13, 23) should also be carefully evaluated. Although the rating scale allows the investigator to determine the level of agreement or disagreement, it does not allow the dichotomous separation of ability. It is under the suggestions of Bond and Fox that categories were collapsed in this study, resulting in the Revised Lesson Plan Rating Scale (Figure 3.6).

Table 3.4 shows a summary for the calibration of all raters. Rater severity ranged from 1.98 (rater 13, most severe) to -1.78 (rater 1, most lenient). Raters 3, 5, 7, 8, 9, 10, and 11 all have Infit MSE values less than 0.8, indicating muted response patterns. Rater 16, however, resulted in an Infit MSE of 2.13, which indicates an irregular, or unexpected, response pattern.

Table 3.5 displays a summary of the statistics for rater type. Administrators were placed on the logit scale at 0.41, while the music content experts were placed at -0.42. These measures indicate that overall, administrators demonstrate higher severity in scoring than music content experts. Infit and Outfit MSE values fall within the required range (0.8-1.2), implying acceptability of fit in regards to the rater type.

Table 3.6 is a display of the differential rater functioning (DRF) statistics. DRF is exhibited when raters show systematic levels of severity or leniency among different subgroups (Engelhard, 2008). DRF is indicated by a *Z*-score higher than 2.00, or below -2.00. This table shows specific items in which raters exhibited highly unexpected (overly lenient or severe) behavior. There were a total of 68 interaction terms, 5 of which are indicated by a Z-score +/- 2.00. Specifically, item 8 shows opposite behavior depending on the rater type. Administrators were far more lenient on item 8, while music specialists were highly irregular in their ratings.

Overview of Results

The first research question investigated which items demonstrate acceptable model fit for the construct of lesson plan development in the setting of a secondary-level music performance classroom. Overall, the majority of items demonstrated good model fit (see Table 3.2 and Figure 3.1). However, a total of seven items did not adequately fit the model. First, item 8 read "activities permit student choice." Administrators and music specialists did not treat this item similarly, perhaps the reason why item did adequately fit the model. Item 9 addressed the teacher's statement of connection to other disciplines. This item may not be explicitly stated within a lesson plan, and may come more organically if a teacher were to be observed. Items 10 and 20 both addressed differentiation, which may be more difficult for pre-service teachers to explicitly state in a lesson plan. In addition, differentiation within a music classroom may be overlooked by an administrator as a performance-based classroom looks different from a content-driven classroom. Item 18 addressed authentic learning through real-life examples. This item may not fit due to an unclear definition of authentic learning, or a lack of transparency in providing students with a connection to the outside world. Assessment was addressed in items 28

and 33. As pre-service teachers have reported feeling ill-prepared in this area, it can be concluded that clear plans for assessment are not seen within the lesson plan.

The second research question investigated how the rating scale changed structure as the raters showed inconsistent usage of particular categories. For the majority of the items in the Revised Lesson Plan Rating Scale (see Figure 3.6), only three out of four categories were used. In general, the categories most eliminated were the extremes of the categories, either "Strongly Disagree" or "Strongly Agree." Only four cases existed where all four rating categories were used consistently (items 11, 15, 19, and 26). In addition, four cases existed where two categories were eliminated (items 12, 13, 23, and 34). Some items showed a general positive leaning for all raters (items 12, 13, 23). The revised rating scale allowed for a more accurate evaluation of pre-service teachers' lesson plans, as only applicable items and rating scale categories remained.

The final research question investigated the presence of differential rater severity between administrators and music specialists across items. This differential severity was present among three items only. First, item 8 addressed student choice. Administrators were likely expecting a clear plan for students to make clear choices, while music specialists may not have expected student choice to be included in the performance of music. Next, item 19 addressed the reference to curricular frameworks and accurate sequencing. Administrators demonstrated severity on the rating of this item, likely because lesson plans did not explicitly state a sense of sequencing. Music specialists, who understood which skills are required to move from task to task, could see the sequencing without a clear statement from the pre-service teacher. Last, item 23 referred to planning at an appropriate content level.

Discussion

Pre-service music educators have reported feeling inadequately prepared for teaching in the performing arts classroom, and especially have a perceived lack of understanding of teacher evaluation (Duncan, 2011). In addition, pre-service teachers have reported the need for greater attention during preparation programs in the areas of music curriculum, lesson planning, and student assessment (Berg & Miksza, 2010; Conway, 2002a; Snyder, 1998). To develop an instrument that could evaluate students' lesson plans, a multi-step process for scale development was employed. After creating the observational design, a Likert-type scale was used for raters to evaluate each item. The judging plan was formed and rater data collected. When analyzing data, misfit is extremely sensitive and must be handled with careful scrutiny. Once the misfit was managed, the original measure had to be refined (Figure 3.6). Finally, the rating scale structure was evaluated and optimized.

As discussed earlier, good model-data fit and high reliability of separation indicate a strong argument for construct validity. Therefore, any changes to the rating scale itself include the eliminating of misfit items and changes to the rating scale category structure. We acknowledge the divergence of response based upon the type of rater (e.g., academic administrators vs. music education content specialists). Ideally, both administrators and music specialists would undergo some sort of rater training protocol in order to align their ratings to fair and equitable rating practices. Music specialists would likely have a more accurate understanding of what type of planning is most appropriate for each level of teaching. Particular musical training for administrators would likely aid in this assessment. However, these types of training most likely are not feasible due to challenges related to time, money, standards, etc.

Although it is not ideal, we can control for rater differences in the measurement model itself to the best of our ability.

Face validity is a qualitative, contextual way to approach the validity of a rubric. Empirical data, however, is the best way to examine construct validity. With face validity, items cannot be added or dropped based on perceptions. In order to substantiate such processes, the rubric must be revalidated with empirical evidence of how well the overall construct functions. The revised instrument (see Figure 3.6) in this study does not maintain consistent categories among items. The decisions leading to the inclusion of specific categories should be empirically based. The instrument gives us the strongest interpretation of the function of items and rating scale categories that ultimately defines the construct. Any changes made to the instrument due to face validity would be speculation at best and were not considered as part of this study. A future revalidation study to include considerations of the perceptions and use of the final instrument resulting in this study is therefore suggested.

As pre-service teachers transition from the college setting to the classroom setting, a significant shift occurs as evaluation moves from the hands of music education specialists to school administrators. These administrators often have teaching backgrounds in non-performing related subjects, such as language arts or social studies. More specifically, administrators in this study came from backgrounds in career and technical education, counseling, language arts, mathematics, foreign language, science, and social studies. This study suggests that academic administrators are, overall, more severe evaluators than music education content specialists. This gap in severity could stem from the lack of content-specific knowledge by the administrator. The gap in music teaching expectations may also stem from the location of music teacher preparation programs. In most colleges and universities, preparation for music teachers is housed in the

school of music. Preparation for teachers of other subjects, such as science, mathematics, and social studies, is housed in the college of education. These pre-service teachers are trained with a common set of standards and expectations, whereas pre-service music teachers may not be trained with similar standards. However, these are speculative considerations and warrant further phenomenographic investigations.

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

In-service teacher evaluation procedures focus on the improvement of teaching behaviors and overall student learning. However, these procedures may have other consequences. Nelson (2012) discusses teachers' concerns about the additional time and work needed to prepare for the evaluation process. Stresses related to evaluation may impact teacher retention. In addition, some states use teacher evaluation as a means to determine teacher salary.

The adoption of the Common Core curriculum has directly impacted teacher accountability. Classroom teachers are even more responsible to document student learning specifically in the form of individual growth. Potentially, administrators may view this integration of common core into the arts curriculum as a need for project-based learning (Taylor, 2014). Teachers are being asked to go beyond the daily lesson and rehearsal. Pre-service teachers

should approach this not as a complication, but as a way to provide engaging activities for every type of learner in the classroom.

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

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

The Revised Lesson Plan Rating Scale (Figure 3.6) provides an opportunity to help pre-service teachers become more familiar with classroom expectations. By incorporating this Rating Scale into the curriculum, pre-service teachers will have the potential to provide more comprehensive lesson plans to evaluate and promote student learning within the music classroom. This Rating Scale should go through the revalidation process to ensure accurate outcomes, and then should be transferred to a rubric format for implementation in the classroom. Therefore, constant use and monitoring will only aid in its ability to provide accurate feedback to the user. This Lesson Plan Rating Scale serves as a way to communicate between a teacher's expectations and a student's performance, and should be used to further the discussion of quality teaching.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lehman, P. (2014). How are we doing? In T. S. Brophy, M.-L. Lai, & H.-F. Chen (Eds.), *Music Assessment and Global Diversity: Practice, Measurement and Policy,* (pp.3-17). Chicago, IL: GIA Publications, Incorporated.

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

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

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

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

Linacre, J. M. (2017). *Reliability and separation of measures.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling.* New York: Routledge, Taylor & Francis Group.

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

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

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

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

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

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

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

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

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

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

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

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016). Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted Rasch partial credit model. *Music Perception, 33*(5), 662-678.

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

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

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

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

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

Tables

Table 3.1. *Summary Statistics from MFR-PC Model*

		Facets			
		Lesson Plan (θ)	Rater (λ)	Item (δ)	Rater Type (γ)
Measure (Logits)					
	Mean	0.00	0.00	0.00	0.00
	SD	0.88	0.95	0.69	0.41
	N	32	16	34	2
Infit *MSE*					
	Mean	1.00	0.99	1.01	1.00
	SD	0.39	0.44	0.21	0.15
Std. Infit *MSE*					
	Mean	-0.20	-0.30	0.00	-0.30
	SD	2.20	3.40	1.20	3.50
Outfit *MSE*					
	Mean	1.01	1.01	1.01	1.01
	SD	0.40	0.45	0.22	0.14
Std. Outfit *MSE*					
	Mean	-0.10	-0.20	0.00	0.00
	SD	2.20	3.30	1.20	3.00
Separation Statistics					
Reliability of Separation		0.94	0.97	0.89	0.98
Chi-Square		499.2*	621.8*	296.2*	118.5*
Degrees of Freedom		31	15	33	1

Note. * $p < 0.01$. *Infit and Outfit Mean Square Error (MSE) should exist within the range of 0.8-1.2. Std. refers to a weighted mean square residual, where the mean is set to zero, and the variance is set to 1 (Wright & Masters, 1982). SD = Standard deviation. Reliability of separation refers to the spread of elements within the facet.*

Table 3.2. *Calibration of Items*

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

Note. Items are ordered according to Measure, from highest value to lowest value.

Table 3.3. *Rating Scale Category Functioning: Category Usage, Average Observed and Expected Measures, and Outfit MSE*

Item	Category Usage (%)				Average Observed Measure (Average Expected Measure)				Outfit *MSE*			
	1	2	3	4	1	2	3	4	1	2	3	4
1	0 (0)	15 (23)	35 (55)	14 (22)	-	-0.75 (-1.24)	-0.31 (0.01)	1.24 (0.98)	-	1.30	1.50	0.80
2	0 (0)	10 (16)	40 (63)	14 (22)	-	-1.01 (-1.15)	0.16 (0.22)	1.35 (1.25)	-	1.00	0.90	1.00
3	1 (2)	6 (9)	42 (66)	15 (23)	0.96 (-1.03)	-0.05* (0.25)	1.53 (1.63)	2.92 (2.68)	3.70	0.80	1.00	0.80
4	0 (0)	14 (22)	42 (66)	8 (13)	-	-1.05 (-1.57)	-0.44 (-0.23)	0.98 (0.78)	-	1.30	1.60	0.90
5	3 (5)	7 (11)	44 (69)	10 (16)	-0.97 (-1.45)	-0.41 (-0.16)	1.02 (1.05)	2.19 (2.06)	1.20	0.90	1.10	1.00
6	4 (6)	14 (22)	36 (56)	10 (16)	-1.80 (-1.49)	-0.07 (-0.16)	0.87 (0.88)	1.85 (1.81)	0.50	1.10	1.00	1.00
7	4 (6)	10 (16)	41 (64)	9 (14)	-1.60 (-1.53)	0.08 (-0.23)	0.76 (0.87)	1.98 (1.83)	0.60	1.40	1.00	0.90
8	11 (17)	30 (47)	16 (25)	7 (11)	-2.12 (-1.75)	-0.13 (-0.53)	0.13 (0.34)	0.41 (1.10)	0.70	1.30	1.20	2.50
9	13 (20)	32 (50)	17 (27)	2 (3)	-2.39 (-2.27)	-1.12 (-1.03)	0.08 (-0.10)	1.33 (0.64)	0.80	1.10	0.70	0.60
10	9 (14)	30 (47)	24 (38)	1 (2)	-2.91 (-2.48)	-1.14 (-1.17)	-0.09 (-0.18)	1.61 (0.62)	0.70	0.60	0.90	0.70
11	7 (11)	15 (23)	35 (55)	7 (11)	-1.96 (-1.68)	-0.45 (-0.43)	0.60 (0.52)	1.33 (1.42)	0.70	0.80	0.70	1.00
12	0 (0)	3 (5)	45 (70)	16 (25)	-	-1.42 (-0.70)	0.93 (0.92)	2.20 (2.07)	-	0.70	1.00	0.80
13	0 (0)	5 (8)	45 (70)	14 (22)	-	-0.39 (-0.96)	0.59 (0.59)	1.51 (1.70)	-	1.40	1.30	1.10
14	1 (2)	12 (19)	46 (72)	5 (8)	0.38 (-1.50)	0.13* (-0.07)	1.09 (1.26)	2.96 (2.29)	2.60	1.20	1.40	0.70
15	10 (16)	30 (47)	19 (30)	5 (8)	-1.95 (-1.89)	-0.48 (-0.63)	0.03 (0.28)	1.22 (1.06)	0.90	1.40	1.40	1.00
16	13 (20)	25 (39)	23 (36)	3 (5)	-2.11 (-2.07)	-0.57 (-0.89)	-0.40 (0.00)	1.35 (0.79)	0.90	1.40	1.80	0.70
17	10 (16)	25 (39)	26 (41)	3 (5)	-2.14 (-2.06)	-0.74 (-0.82)	0.08 (0.11)	0.82 (0.93)	0.90	0.80	1.40	1.00
18	6 (9)	18 (28)	35 (55)	5 (8)	-0.79 (-1.82)	-0.32 (-0.51)	0.25 (0.49)	1.18 (1.39)	1.70	1.50	2.10	1.10
19	8 (13)	18 (28)	29 (45)	9 (14)	-1.79 (-1.59)	-0.29 (-0.36)	0.43 (0.53)	1.73 (1.38)	0.70	1.10	1.00	0.70
20	9 (14)	33 (52)	18 (28)	4 (6)	-2.54 (-1.99)	-0.74 (-0.68)	0.64 (0.26)	1.13 (1.04)	0.60	0.50	0.50	0.90
21	9 (14)	31 (48)	22 (34)	2 (3)	-2.53 (-2.24)	-0.92 (-0.93)	0.11 (0.04)	1.35 (0.84)	0.80	0.80	0.90	0.70
22	7 (11)	28 (44)	21 (33)	8 (13)	-1.95 (-1.66)	-0.09 (-0.35)	0.59 (0.56)	0.66 (1.38)	0.80	0.80	1.00	2.7
23	0 (0)	4 (6)	52 (81)	8 (13)	-	-1.22 (-1.22)	0.53 (0.43)	0.97 (1.59)	-	1.20	1.00	1.30
24	2 (3)	15 (23)	40 (63)	7 (11)	-0.79 (-1.52)	-0.06 (-0.11)	1.04 (1.07)	1.91 (2.04)	1.30	1.00	1.40	1.10

25	12 (19)	32 (50)	17 (27)	3 (5)	-2.06 (-2.11)	-1.00 (-0.87)	0.24 (0.06)	0.94 (0.81)	1.00	1.20	0.70	0.90
26	4 (6)	14 (22)	42 (66)	4 (6)	-1.54 (-1.86)	-0.62 (-0.51)	0.56 (0.61)	2.16 (1.57)	1.20	0.90	1.30	0.80
27	10 (16)	34 (53)	17 (27)	3 (5)	-2.62 (-2.11)	-0.66 (-0.82)	0.09 (0.14)	1.12 (0.90)	0.60	0.70	1.10	0.80
28	10 (16)	28 (44)	24 (38)	2 (3)	-2.82 (-2.23)	-0.86 (-0.96)	0.07 (-0.01)	1.57 (0.80)	0.50	0.60	0.80	0.70
29	9 (14)	32 (50)	21 (33)	2 (3)	-2.23 (-2.24)	-1.06 (-0.93)	0.18 (0.05)	1.45 (0.84)	0.90	1.00	0.80	0.70
30	2 (3)	8 (13)	48 (75)	6 (9)	-1.64 (-1.59)	0.02 (-0.25)	1.01 (1.07)	2.18 (2.11)	0.60	1.30	1.20	1.00
31	0 (0)	12 (19)	45 (70)	7 (11)	-	-1.74 (-1.61)	-0.15 (-0.21)	0.64 (0.83)	-	.80	1.10	1.10
32	5 (8)	35 (55)	21 (33)	3 (5)	-2.14 (-2.06)	-0.52 (-0.61)	0.23 (0.42)	1.76 (1.24)	1.00	0.70	1.50	0.60
33	9 (14)	30 (47)	21 (33)	4 (6)	-2.44 (-1.97)	-0.69 (-0.68)	0.39 (0.26)	1.49 (1.05)	0.70	0.50	0.80	0.60
34	4 (6)	26 (41)	33 (52)	1 (2)	-2.20 (-2.38)	-0.87 (-0.94)	0.02 (0.15)	2.13 (1.05)	1.00	0.90	1.50	0.80

Note. Category 1 = “strongly disagree;” Category 2 = “disagree;” Category 3 = “agree;” Category 4 = “strongly agree;” *Violation of monotonicity

Table 3.4. *Calibration of Raters*

Rater Number	Observed Avg. Rating	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
13	1.86	1.98	0.15	1.50	4.0	1.46	3.6
16	2.03	1.34	0.15	2.13	7.9	2.17	7.9
2	2.32	0.80	0.16	1.07	0.5	1.11	0.7
3	2.44	0.78	0.15	0.67	-2.9	0.80	-1.5
8	2.48	0.61	0.15	0.49	-4.6	0.48	-4.4
6	2.53	0.49	0.15	0.80	-1.6	0.90	-0.7
7	2.71	0.15	0.15	0.5	-4.5	0.43	-5.0
5	2.78	0.04	0.15	0.65	-3.0	0.62	-3.2
15	2.54	-0.10	0.15	1.32	2.2	1.30	1.9
10	2.82	-0.28	0.15	0.76	-1.9	0.79	-1.5
4	2.51	-0.53	0.15	1.67	4.3	1.71	4.1
11	2.61	-0.70	0.15	0.73	-2.2	0.73	-2.0
14	3.00	-0.75	0.15	1.05	0.4	1.02	0.1
9	3.07	-0.84	0.15	0.63	-3.5	.66	-3.1
12	3.07	-1.21	0.15	0.86	-1.1	0.81	-1.6
1	3.02	-1.78	0.15	1.07	0.6	1.13	1.1
Mean	2.61	0.00	0.15	0.99	-0.3	1.01	-0.2
SD	0.35	0.98	0.00	0.46	3.5	0.47	3.4

Note. The raters are arranged in Measure (severity) order, from severe to lenient.

Table 3.5. *Calibrations of the Rater Type Facet*

Rater Type	Observed Avg. Rating	Measure	*SE*	Infit *MSE*	Std. Infit *MSE*	Outfit *MSE*	Std. Outfit *MSE*
Administrator	2.82	0.41	0.05	0.84	-3.7	0.87	-2.9
Music Content Expert	2.41	-0.42	0.05	1.15	3.2	1.15	2.9
Mean	2.61	0.00	0.05	1.00	-0.3	1.01	0.00
SD	0.29	0.59	0.00	0.22	4.9	0.20	4.2

Note. The rater type is arranged by Measure (severity), from severe to lenient.

Table 3.6. *Summary of differential rater functioning statistics (rater interactions) for selected raters exhibiting* **| Z |** > = 2.0.

Item Number	Rater Type	Infit MSQ	Outfit MSQ	Total observ-ed	Total expect-ed	Std. Mean Res. (obs-exp)	Bias logit	SE	Z
23	Music Content Specialist	0.90	1.10	100	95.15	0.15	1.11	0.45	2.47
19	Admini-strator	0.80	0.90	100	92.42	0.24	0.62	0.29	2.10
8	Music Content Specialist	1.60	1.70	72	63.89	0.25	0.58	0.26	2.24
8	Admini-strator	0.60	0.60	75	83.12	-0.25	-0.53	0.26	-2.06
23	Admini-strator	1.20	1.00	96	100.83	-0.15	-1.03	0.48	-2.12

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

Figures

1. Develops plans that are clear.	SD	D	A	SA
2. Develops plans that are logical.	SD	D	A	SA
3. Develops plans that are sequential.	SD	D	A	SA
4. Plans instruction effectively for pacing.	SD	D	A	SA
5. Transitions are logical and sequential.	SD	D	A	SA
6. Aligns and connects lesson objectives to standards.	SD	D	A	SA
7. Develops appropriate daily plans.	SD	D	A	SA
8. Activities permit student choice.	SD	D	A	SA
9. Learning experiences connect to other disciplines.	SD	D	A	SA
10. Provides a variety of appropriately challenging resources that are differentiated for students in the class.	SD	D	A	SA
11. Organizes the lesson to progress toward a deep understanding of content (i.e. content mastery).	SD	D	A	SA
12. Engages students in active learning.	SD	D	A	SA
13. Builds upon students' existing knowledge and skills.	SD	D	A	SA
14. Reinforces learning goals throughout the lesson.	SD	D	A	SA
15. Effectively uses appropriate instructional technology to enhance student learning.	SD	D	A	SA
16. Develops higher order thinking through questioning.	SD	D	A	SA
17. Encourages critical thinking through problem solving activities.	SD	D	A	SA
18. Engages students in authentic learning by providing real-life examples.	SD	D	A	SA
19. Teacher's plans reference curricular frameworks or blueprints to ensure accurate sequencing.	SD	D	A	SA
20. Differentiates the instructional process to meet individual developmental needs.	SD	D	A	SA
21. Provides remediation and enrichment to further student understanding of material.	SD	D	A	SA
22. Uses flexible grouping strategies to encourage appropriate peer interaction and to accommodate learning needs/goals.	SD	D	A	SA
23. Plans at a content level that is appropriate for students.	SD	D	A	SA
24. Demonstrates high expectations for all students in content mastery.	SD	D	A	SA
25. Plans follow-up activities designed to meet varied abilities of students.	SD	D	A	SA
26. Aligns student assessment with established objective.	SD	D	A	SA
27. Involves students in setting learning goals and monitoring their own progress.	SD	D	A	SA
28. Varies and modifies assessments to determine individual student needs and progress.	SD	D	A	SA
29. Uses formal assessments for diagnostic, formative, and summative purposes.	SD	D	A	SA

30. Uses informal assessments for diagnostic, formative, and summative purposes.	SD	D	A	SA
31. Uses assessment techniques that are appropriate for the developmental level of students.	SD	D	A	SA
32. Uses diagnostic assessment data to develop learning goals for students.	SD	D	A	SA
33. Uses diagnostic assessment data to differentiate instruction.	SD	D	A	SA
34. Uses diagnostic assessment data to document learning.	SD	D	A	SA

Figure 3.1. Initial 34-item Lesson Plan Evaluation Rating Scale

Subject: Beginning Orchestra Grade Level: 6th grade School: Middle

National Standards & GPS	Conceptual Objective:	Procedures:	Materials:	Assessment:
MMSBO.2 – Performing on instruments, alone and with others, a varied repertoire of music	Students will be able to a. Demonstrate right hand position (bow hold), posture, instrument position and bow placement. b. Produce a characteristic sound using legato,	1. Review bow holds: Go over the step by step process of holding a bow for each instrument (thumb, pinky, etc.) Student will participate both orally and physically. 2. Students will be asked to find a partner and to demonstrate a proper bow hold to there stand partner. Partner will correct mistakes and vice versa. Again repeating the step-by-step process. Teacher will observe to correct general mistakes. 3. Once students have correct bow hold adjustments, they will shadow bow (on shoulders for violins and violas, arm for cellos and basses). Ask students to self assess and partner asses that their bow is moving from frog to tip and that a) movement in wrist b) proper extension of elbow c) bow hold remains unchanged (all in half notes) 4. Increase speed of shadow bowing from half notes, to quarters still using full bow. Reassess both from peers and teacher. 5. Apply to open strings D and A again in half notes and then quarters keeping in mind wrist, elbow, and hand using full bows in a legato motion. Teacher will look for proper posture, right hand position, and also smooth bow changes. (wrist movement) 6. Then apply to # 12 in EE Book 1	Instruments Bow Partners Essential Elements	Through, Modeling, Peer Review, and Teacher Observation, Students were able to have several steps in assessment and evaluation to reinforce proper bow hold and bow movement to create a characteristic legato tone.

Figure 3.2. High Achieving Lesson Plan

SCHOOL OF MUSIC MUSIC EDUCATION LESSON PLAN				Teacher: Grade/Level: 8th Grade National Standard: Performing on instruments, alone and with others, a varied repertoire of music.
CENTRAL FOCUS: EXPRESSION: ACCENT GROUPINGS, RHYTHM OF THE MELODY, AND THE REGULAR OR IRREGULAR MOVEMENT OF RHYTHMIC PATTERNS CONTRIBUTE TO EXPRESSIVENESS. LEVEL: 8TH GRADE CONCERT BAND PRIOR KNOWLEDGE: - PRIOR EXPERIENCE WITH ACCENT MARKINGS IN BAND CLASS, AS WELL AS IN LESSON PREVIOUS TO MINE (STACCATO LESSON).	L. OBJECTIVES	ASSESSMENT	I/F	PROCEDURES/LEARNING TASKS
	- students will define tenuto markings with 85% accuracy. - Students will demonstrate legato playing on instruments with 80% accuracy. - Students will differentiate between legato and staccato with 90% accuracy.	Teacher will evaluate student performance of tenuto markings of legato playing on instruments in rhythm exercises.		- A.O.: Talk about the short sounds of candy (like M&Ms) hitting the ground versus the long sound of a stream of chocolate milk or water hitting the inside of a cup. Also, bring up the terminology and say it in a way that reinforces its meaning (say sta-cca-to very light and separated, le-ga-to very long and connected). - Reinforce the definition of staccato (taught previously by Matt), introduce legato and tenuto. - Write exercises utilizing tenuto and staccato markings on the board, and have students tizzle breathe them, demonstrating the note lengths dramatically. - Use these same exercises and have the students say “loooong” and “short” on them, demonstrating first. - Have the students play the exercises on concert Bb.
POSSIBLE PROBLEMS AND PLANNED SUPPORTS				EVALUATION CRITERIA
- Possible overload of terminology, so make sure to say the words in a way that reinforces their meaning (sta-cca-to very light and separated, le-ga-to very long and connected). - Tonguing too heavily on the staccato notes, creating a “thwack” sound; emphasize lightness of short staccato sound. - Students may just hold out a singular note instead of re-articulating a legato passage; make sure to emphasize the length of the note, while reminding them to tongue lightly in-between two notes to just barely separate them.				Various exercises written on the board in 4/4 time, in quarter notes and/or eighth notes, alternating tenuto and staccato markings. MATERIALS/EQUIPMENT Whiteboard and marker

Figure 3.3. Middle Achieving Lesson Plan

Subject: Band Grade Level: Wind Symphony School: High School

National Standards & GPS	Conceptual Objective:	Procedures:	Materials:	Assessment:
MHSIB.2 - Performing on instruments, alone and with others, a varied repertoire of music. d. Use the following ensemble skills as a means of interpreting the performance of musical literature: dynamic expression, style, blend and balance, steady tempo, rhythmic accuracy, intonation, and rehearsal etiquette	The beat can be divided or augmented evenly. Students will demonstrate knowledge of beat divisions through playing through passing eighth notes.	1. A.O: Greeting 2. Play through Concert Bb Major Scale 3. Play through concert Bb Major Scale on legato eight notes. Four eighth notes per pitch 4. Play from F to H in Carols Three as written 5. Ask them to use beat divisions (play legato eighth notes for every note) from F to G. Make sure they are moving together. 6. Do the same with G to H. Then have them “bop” the section to further insure their understanding of beat divisions. 7. Play as written.	Carols Three by Luigi Zaninelli	Aural assessment of even eighth notes

Figure 3.4. Low Achieving Lesson Plan

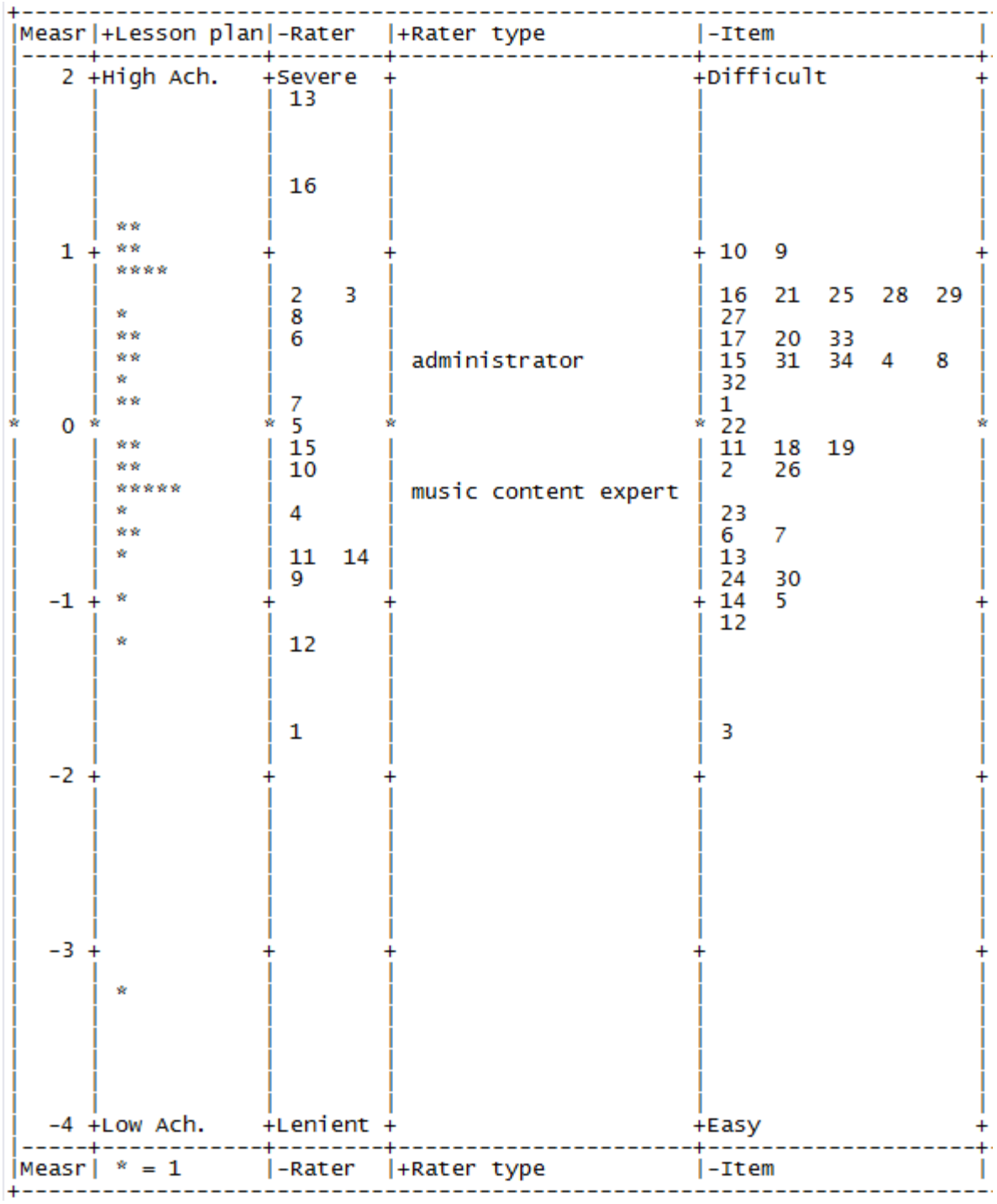

Figure 3.5. Wright Map

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

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

Figure 3.6. Revised 26-item Lesson Plan Evaluation Rating Scale

CHAPTER 4

DEVELOPMENT OF A RATING SCALE TO ASSESS PRE-SERVICE INSTRUMENTAL MUSIC TEACHERS' CLASSROOM TEACHING PERFORMANCE IN THE STUDENT TEACHING SETTING[1]

[1] Musselwhite, D. J. To be submitted.

Abstract

The purpose of this study is first to determine which content- and performance-based behaviors and skills are relevant in the assessment of pre-service instrumental music teachers. Additionally, this exploratory study served to build a measure to assess those pre-service teachers' performance in the student teaching setting. The research questions addressed the psychometric quality of the scale, item fit, and the rating scale category structure. Results regarding the psychometric quality of the scale indicate sufficient reliability of teaching episodes, but a larger sample size is needed to determine the reliability of items and raters. Results also indicated overfit among three items related to the use of technology, the use of a warm-up, and the teacher's response to classroom disruptions. The rating scale structure violated assumptions of monotonicity and guidelines regarding category usage and Outfit MSE. The implications for pre-service teacher preparation and assessment will be discussed.

Keywords: Assessment, Pre-service teaching, Rasch Measurement, Rating Scale, Reliability

DEVELOPMENT OF A RATING SCALE TO ASSESS PRE-SERVICE TEACHERS' CLASSROOM TEACHING PERFORMANCE IN THE STUDENT TEACHING SETTING: A PILOT STUDY

Student teaching is considered to be one of the most valuable experiences in teacher preparation programs (Brand, 1982; Conway, 2002; Conway 2012; LaParo et al., 2014). The process of learning to teach involves the acquisition of both content-based knowledge and performance-based knowledge (Munby, Russell, & Martin, 2001). One important concern of supervising teachers and cooperating professors is how to assess pre-service instrumental music teachers' performance in student teaching (i.e., pre-service teaching) settings. The purpose of assessment in this context is to provide pre-service teachers with feedback about the intersection between teaching and learning (Roegman, et al., 2016). Teaching performance assessments allow for inferences to be made regarding student achievement and teacher ability, and these assessments must be continuously monitored for reliability and validity concerns (Pecheone & Chung, 2006).

Content Validity of Peformance Assessments

One particular focus of research involving observations of effective teacher behaviors addresses verbal versus non-verbal behavior (Abeles, 1975; Gipson, 1978, Kostka, 1984; Hepler, 1986; O'Neill, 1993; Wang, 2001). As an example, Abeles (1975) developed a teacher evaluation instrument with 30 items, divided into five domains: (a) rapport, (b) instructional systemization, (c) instructional skill, (d) musical knowledge, and (e) general instructional competence. Interestingly, four of those five domains focused on verbal behavior. Kostka (1984) found the most frequent lesson behaviors involved teacher talk and student performance. O'Neill (1993) focused on three domains of teacher behavior: (a) non-verbal, (b) musical, and (c) verbal behavior. Wang emphasizes the importance of observing both verbal and non-verbal in music

teaching. However, the introduction of musical behavior by O'Neill (1993) into the needed teaching performance skills emphasizes its existence beyond the verbal/non-verbal communication spectrum. The domains used to group items in the previous studies served as a base for the construction of this pre-service teaching rating scale.

While the previous studies focused on communicative behaviors, Sparks (1984) specifically explored teaching strategies. Sparks (1984) developed a teaching evaluation system that included five domains: (a) diagnosis, (b) prescriptions, (c) presentation, (d) monitoring, and (e) feedback. Lessons of effective teachers were quickly paced, driven by clear directions (Sparks, 1984). Other studies in teaching effectiveness suggest the importance of context. Research advises that effective teaching strategies must be used in moderation, as overuse may have negative effects on learning (Coker, Medley, & Soar, 1980; Soar & Soar, 1983). Additionally, effective teaching strategies may be reliant upon the objectives and the lesson content (Rosenshine, 1987; Taebel, 1990).

The development of a measure to assess pre-service teacher performance in music education poses special considerations beyond that of other classroom teachers. One important consideration is that the measure must incorporate both general teaching strategies and music rehearsal strategies. For example, where general teaching strategies involving teacher communication may be focused broadly on verbal and non-verbal behavior, music teaching strategies must include communication specifically through conducting gestures. Another consideration is that music content must play a meaningful role in the development of the measure. Teacher effectiveness research suggests that music teachers who demonstrate thorough knowledge of the content have more engaged students and are more effective in their instruction (Madsen, 2003; Young & Shaw, 1999). The specific strategies included in a teaching

performance assessment must be vetted by current music education professionals in order to determine the relevant skills all pre-service teachers should possess. These strategies come in the form of items, and must go through a piloting process, which acts as a pre-testing of the items being used in the measurement instrument (Teijlingen & Hundley, 2001).

Performance Assessments as Communication

Another consideration in the development of a measure of pre-service teaching is communication, as the measure will be used by supervising and cooperating teachers. Pre-service teacher feedback may come in the form of written observations (e.g., prose) or in the form of a pre-developed rating scale or rubric. Vavrus (1999) developed a performance-based assessment instrument specifically for the student teaching program at a college in the northwestern United States. He observed a discrepancy between the knowledge of the cooperating teachers and the expectations of the state (e.g., teaching standards). Upon development of the instrument, results suggested that cooperating teachers were hesitant to implement the instrument in their normal observations. Through observing pre-service teachers, cooperating and supervising teachers found differences between actual lessons and pre-service teacher perceptions of those lessons, especially as related to student-centered learning and multicultural education (Vavrus, 1999).

The pre-service teacher is assessed from two viewpoints: the cooperating teacher and the supervising teacher. One of the key benefits of having multiple observers or evaluators is having multiple perspectives (Roegman, et al., 2016). The divergence of multiple responses is welcomed. The cooperating teacher provides a perspective from a real-world sense, where a supervising teacher provides a perspective from a research sense. Feedback from the supervising and cooperating teachers are often summative judgments from their observations which cannot necessarily differentiate levels of teaching effectiveness (Arends, 2006).

A performance-based observation instrument allows for better recognition of strengths and weaknesses, which is critical before pre-service teachers attempt to gain certification. Sandholtz and Shea (2012) compared supervising teachers' predictions to actual pre-service teachers' performance on the Performance Assessment for California Teachers (PACT). The inconsistencies between predictions and performance led to a follow-up study in 2015, where Sandholtz and Shea found the majority of candidates predicted to fail did not, and the majority of candidates who did fail were predicted to pass. Through observation alone, a supervising teacher does not have difficulty placing pre-service teachers at opposite ends of the spectrum of teaching ability (e.g., able or unable, high- or low-achieving) (Sandholtz & Shea, 2015). The advantage of the performance assessment (PACT) was that pre-service teachers were being evaluated using evidence from actual teaching, as opposed to their academic strengths and weaknesses (Sandholtz & Shea, 2015). The results emphasize the use of performance assessment in addition to content assessment in the development of pre-service teachers. Their findings also support the fact that high-stakes outcomes should not be determined by a single measure or single evaluator (Sandholtz & Shea, 2015).

Reliability and Validity in Performance Assessments

Teacher evaluation frameworks, such as the Marzano Teacher Evaluation Model (2013), Danielson's Framework for Teaching (2007), and Stronge's Teacher Effectiveness Performance Evaluation System (2012), allow for formative feedback and support for the classroom teacher (Tillema, 2009). Detailed analysis of the Marzano and Danielson frameworks suggests that the items and rubrics provide the intended results, whereas analysis of the Stronge framework emphasizes validity over reliability (Benjamin, 2002; Milanowski, 2004; Milanowski, 2011; Kane, et al., 2010; Stronge, 2013). These frameworks most often use a rubric to assess teachers,

rather than broad open-ended questions, which keeps in mind the importance of context when assessing (Darling-Hammond & Snyder, 2000; Roegman, et al., 2016). Supporters of these frameworks emphasize their ability to define good teaching, to guide the feedback of cooperating and supervising teachers, to act as a parallel evaluation system to that of public school teachers, and to identify and develop teaching practices (Benjamin, 2002; Danielson, 2007; Soslau & Lewis, 2014; Roegman, et al., 2016).

Performance-based teaching assessments evoke concerns of reliability and validity. Some of these assessments encompass more generic teaching skills, and often it remains unclear what processes and analyses are used to determine reliability and validity (Berry & Ginsberg, 1990; Kogan, 1989; Bergee, 1992). Within the Marzano framework, correlational studies are used to determine the validity of the evaluation model (Marzano, 2012). Correlation may be used as a starting point, but should not be the determining factor of validity (Linacre, n.d.). The Mid-continent Research for Education and Learning Teacher Evaluation System (McREL, 2013) discusses fairness, reliability, and meaning of their rubrics as they relate to specific district or state education systems, such as in districts in New Jersey. Although McREL uses pilot and field-testing to certify validity, correlation is used to determine predictive validity (Plotz, n.d.). Validity and reliability must be determined through the revalidation process, and through careful scrutiny of the instrument's items.

While teaching assessments exist and are in use, general teaching strategies do not encompass all expectations for an effective teacher. In the assessment of pre-service teachers, context must become a common thread throughout a measure of performance. In this study, those general teaching strategies put forth by the teacher evaluation frameworks are combined with previous research in effective music teaching to create a new measure. The purpose of this

study is first to determine which behaviors and skills are relevant and important in the assessment of pre-service instrumental music teachers. Additionally, this study serves to build a measure to assess pre-service instrumental music teachers' performance in the student teaching setting. The measure is intended to be used by both the supervising teacher and the cooperating teacher to allow for more specific feedback to be given to the pre-service teacher. The study is guided by the following research questions:

1. What is the psychometric quality (i.e., validity and reliability) of a rating scale used to measure the teaching ability of pre-service instrumental music teachers?
2. Which items demonstrate acceptable model fit in the development of a measure of pre-service teacher performance?
3. How well does the rating scale structure cooperate to produce meaningful measures?

Method

Item pool, Raters, and Judging Plan

The researcher-developed items (n = 41) were derived from items within the Marzano, Danielson, and Stronge teaching frameworks (Marzano, 2013; Danielson, 2007; Stronge, 2013). These items were the basis for building the rating scale, and were rephrased for clarity and length. Items were reviewed by the researcher and music content experts. In addition, these items were rephrased to express four levels of performance. For example, an original item was listed, "Teacher uses a variety of teaching methods to engage students." The item was shortened to deter any bias, "Variation of Teaching Methods." Lastly, the item was expressed at four levels of performance, from low to high achieving, (a) teacher uses an unacceptable variety of teaching methods; (b) teacher uses a slightly unacceptable variety of teaching methods; (c) teacher uses a slightly acceptable variety of teaching methods; and (d) teacher uses an acceptable variety of

teaching methods (see Figure 4.1). The item was also accompanied by a description, for example, “Teaching methods may include: Teacher-centered (lecture, demonstrations), student-centered (discussions), content-focused, and/or active learning (groups, brainstorming),” (see Figure 4.2). Items were then grouped according to four domains: (a) Classroom Management, (b) Communication, (c) Lesson Content, and (d) Teaching Strategies. Both general teaching and music-specific items were included in each of the domains (See Figure 4.1).

Videos (n = 16) were submitted by pre-service teachers from a large public university in the southeastern United States. Consent was given from each pre-service teacher to use video footage in the development of a measure of assessment for pre-service teachers. All teachers were in their fourth year of undergraduate study, with a major in music education. Pre-service teachers were also all instrumental majors, with a string (e.g., Guitar or Orchestra) or band emphasis. The pre-service teachers’ name is withheld from the videos to maintain anonymity.

Raters were chosen based on teaching position. The requirement for all raters was a background in music education, and current supervision of pre-service or student teachers. A recruitment letter was sent to music education faculty (n=8), both known and unknown to the researcher, throughout the United States. An effort was made to include equal male and female raters.

The judging plan used is classified as a balanced incomplete assessment network (Engelhard, 1997). This specific judging plan is recommended to ensure reliability and validity not only within facets (or variables), but also between facets (Wright & Stone, 1979; Linacre & Wright, 2004). Each rater evaluated four pre-service teacher lesson videos. In this plan, each video was rated by two raters. For example, Rater 1 rated videos 1, 2, 3, and 4. Rater 2 rated

videos 3, 4, 5, and 6. This overlapping pattern continued for all videos, until Rater 8, who rated videos 15, 16, 1, and 2.

The rating scale items were entered into a Google form. The raters were given instructions before and within the form as to the use of the form itself. Before the rating process, the raters were also given links to Dropbox folders containing numbered pre-service teacher videos specific to each rater. The Google form contained a statement to confirm each rater's consent in regards to the number of pre-service teacher videos and the content within the Google form. Raters were given the option to not complete the study if they did not consent to take part in the study.

Psychometric Considerations

The Rasch Measurement Model falls under the umbrella of Item Response Theory. The theory behind Rasch Measurement is that through a logistic transformation, raw scores can change from ordinal to interval level data onto a linear scale (Granger, 2008). This process allows for all information to be measured according to the same latent variable. The major benefit in using the Rasch Measurement Model is the achievement of invariant measurement when the data adequately fit the model. Engelhard and Perkins (2011) define invariant measurement using five requirements: (a) the calibration of the items must be independent of the particular persons used for calibration; (b) any person must have a better chance of success on an easy item than on a more difficult item; (c) the measurement of persons must be independent of the particular items that happen to be used for the measuring; (d) a more able person must always have a better chance of success on any item than a less able person; and (e) items must be measuring a single underlying latent variable. In the context of this study, items refer to the 41 rating scale items, and "persons" refer to the pre-service teaching episodes.

For this study, the statistics from the Rasch Measurement Model were calculated using *FACETS* (Linacre, 2014). Specifically, the Multifaceted Rasch Partial Credit Model (MFR-PC) was used to investigate the development of the performance scale. The partial credit version of the MFR model treats each rating scale category for each item independently, which allows for more precision in the analysis and interpretation of item and category usage. The Partial Credit Model is as specified as follows:

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

where

$\ln\left[\frac{P_{nijmk}}{P_{nijmk-1}}\right]$= the natural log of the probability that teaching episode *n* rated by rater *i* on item *j* receives a rating in category *k* rather than category *k*-1;

θ_n = achievement level of teaching episode *n*;

λ_i = severity of rater *i*;

δ_j = difficulty of item *j*;

τ_{ik} = the location on the logit scale where rating scale categories *k* and *k* – 1 are equally probable for Rater *i*.

In this study, each of the rating scale criteria contain four response levels within the rating scale structure.

Raters were needed to facilitate the process of assessment, as the evaluation of the teaching of pre-service teachers is a performance-based assessment. Raters may be classified on a continuum between lenient and severe (Engelhard, 1994). All raters were treated equally within the model, which allows for control of rater variability and for the determining of model-data fit for the raters. The requirements of invariant measurement as expressed in Engelhard and Perkins

(2011) can be applied to raters in the context of this particular study: (a) the measurement of pre-service teacher's lessons must be independent of the particular raters used for measuring; (b) a higher-achieving pre-service teacher's lesson must have a better chance of acquiring higher ratings than a lower-achieving lesson; (c) the calibration of raters must be independent of the pre-service teacher's lessons used for calibration; (d) a pre-service teacher's lesson must have a better chance of receiving a higher rating from lenient raters than from severe raters.; and (e) pre-service teacher's lessons and raters must be concurrently located on a single latent variable (i.e., the Wright Map) (Engelhard, 2013).

Results

Results from the analysis using the MFR-PC model are presented in the form of a pictorial Wright Map and summary statistics. These statistics include reliability components and implications of model-data fit. Additionally, each facet (e.g., teaching episode, rater, and item) will be investigated separately to discuss model-data fit and level of difficulty. Last, each item is examined to determine the functioning of each rating category, specific to the individual item.

Wright Map

The Wright Map (Figure 4.3) is a visual representation of each of the facets displayed on the same logit scale. This logit scale is the result of the logistic transformation of the raw data from ordinal level measures to interval level measures. This interval-level measurement means that an item located at logit-scale 1 is equally more difficult than an item at logit-scale 0 as that item at measure 0 compared to an item a logit-scale -1 (Bond & Fox, 2015). Pre-service teaching episodes are organized from highest achieving to lowest achieving. Teaching episode 13 is the highest achieving teaching episode, while teaching episode 9 is the lowest achieving. Raters are organized from most severe to most lenient. Raters are centered near 0, but a hierarchy

is present. Rater 2 is the most severe, while Rater 5 is the most lenient. Items are organized from most difficult to endorse to least difficult to endorse. Again, most items are centered near 0, but pre-service teachers had the most difficulty with use of technology, and the least difficulty with differentiation of instruction.

Summary Statistics

The Summary Statistics table (Table 4.1) contains mean logit-scale locations, fit statistics, reliability of separation, and Chi-square tests of significant independence for each facet. Preliminary analysis revealed overall significant differences between teaching episodes (χ^2= 367.0, p < .01), raters (χ^2= 40.2, p < .01), and items (χ^2= 155.1, p < .01).

Reliability of Separation. Reliability of separation is a statistic similarly interpreted as that of Cronbach's alpha. The higher the statistic, the more confidence can be placed that the results would be similar with another sample (Bond & Fox, 2015). This statistic may also serve as an implication that the performance measure was responsive enough to differentiate between high-achieving and low-achieving teaching episodes, severe and lenient raters, and difficult and easy items. Reliability of separation for teaching episodes (*Rel* = 0.96) is high enough to imply that teaching episodes can adequately be separated according to level of achievement. Reliability of separation for raters (*Rel* = 0.83) does not meet the high standard needed for persons (e.g., teaching episodes) or items, and does not indicate that rater severity could be separated. Reliability of separation for items (*Rel* = 0.77) is low due to the small sample size. While items were able to be separated, locations of those items on the logit-scale would be more accurate with a larger sample.

Model-data fit. Model-data fit is indicated through the use of fit statistics called Infit Mean Square Estimates (MSE) and Outfit MSE. The Infit MSE is a fit statistic that is sensitive to

unexpected behaviors of persons (e.g., teaching episodes) where items may have been improperly targeted (Linacre, n.d.b). High Infit values indicate that the items are not performing as intended, which can become a threat to validity. The Outfit MSE is a fit statistic that is sensitive to unexpected behavior on items that were too easy or too hard. Outfit MSE if sensitive to outliers, where Infit MSE is sensitive to inliers, or patterns within the distribution. High Outfit scores can indicate that the raters were not engaged with the rating scale and responded more randomly. Parameter-level mean-square fit statistics should fall within the range of 0.5-1.5 to be productive for measurement construction. Those statistics within the 1.5-2.0 range should be carefully evaluated, as they should be considered unproductive, but not detrimental to the construction process (Linacre, n.d.b). Mean Infit and Outfit MSE values should be close to a value of 1.00 to indicate good model-data fit. Mean Infit MSE values for teaching episode (1.04), raters (1.04), and items (1.00) imply overall good fit. Mean Outfit MSE values for teaching episode (1.07), raters (1.07), and items (1.07) also imply good fit.

Calibration of Teaching Episodes

Table 4.2 provides the calibration of teaching episodes. These episodes are organized from top to bottom, highest-achieving to lowest-achieving. Episode 13 was the highest-achieving teaching episode (1.55 logits) and Episode 9 was the lowest-achieving teaching episode (-2.17 logits). Based on Linacre's (n.d.b) suggestions for range of fit statistics, Episode 1 demonstrates Infit (0.46) and Outfit (0.46) MSE values just outside of the range. Because the statistic is not severely underfitting, it is likely not serving to contradict the measure (Linacre, n.d.b).

Calibration of Raters

The calibration of raters is presented in Table 4.3. Raters are ranked based on a hierarchy of leniency. Raters at the top of the table were more severe, while those at the bottom were more

lenient. Rater 2 was the most severe (observed average = 2.50, logit measure = 0.42), and Rater 5 was the most lenient (observed average = 2.70, logit measure = -0.58). Raters exhibited suitable Infit and Outfit MSE values to imply good model-data fit.

Calibration of Items

Table 4.4 displays the calibration of the preliminary items. The items are ordered based on level of difficulty, from most difficult to least difficult. This can be interpreted as difficulty for pre-service teachers to use within a lesson, or difficulty for raters to endorse in the observation process. The most difficult item was item 36, Use of Technology, (observed average = 1.20, logit measure = 2.33). The least difficult item was item 24, Differentiation of Instruction, (observed average = 2.40, logit measure = -1.00). Item 20, Alignment of Warm-up to Lesson Activity, demonstrated overfit (Infit MSE = 2.13, Outfit MSE = 2.76). Items 36, 21 and 5 exhibit noisy Outfit MSE values, which could be an indication that these items did not measure what was intended, or the anchors were incorrectly applied. This table serves to advise on the inclusion or exclusion of items. While fit is an important indicator, another statistic must be investigated. Point-biserial is a statistic that indicates if the items are all working in that same positive direction (Bond & Fox, 2015). Negative point-biserial values are an indication that an item should be reconsidered. Only item 36 was exhibited a negative point-biserial value.

Rating Scale Category Diagnostics

Once the items have been carefully evaluated for fit and intent, the structure of the rating scale must be investigated for each individual item. The structure of the rating scale may change for each item, as recommended by Linacre (2002) in his steps for the optimization of rating scale structures. The collapsing of categories into adjacent categories allows for the item to be used properly, to improve reliability, and to provide more accurate feedback regarding person (or pre-

service teacher) performance. Three analyses will be discussed, as shown in Table 4.5: category usage, observed average and expected measure, and Outfit MSE.

First, it is suggested that at least 10 observations occur in each category, based on 100 observations. Therefore, each item was evaluated not for a frequency of 10, but for a 10% usage, since the sample size was only 32. Item 1 (Category 1), Item 2 (Category 1), Item 3 (Category 1), Item 5 (Category 1), Item 7 (Category 1), Item 8 (Category 1), Item 9 (Category 1), Item 10 (Category 1), Item 11 (Category 4), Item 12 (Category 1), Item 15 (Category 1), Item 17 (Categories 1 and 2), Item 18 (Category 1), Item 19 (Category 1), Item 21 (Category 4), Item 22 (Category 1), Item 23 (Category 1), Item 24 (Categories 1 and 4), Item 26 (Categories 1 and 2), Item 27 (Category 1), Item 28 (Category 4), Item 29 (Category 4), Item 31 (Category 1), Item 32 (Category 4), Item 33 (Category 4), Item 34 (Categories 1 and 4), Item 35 (Category 4), Item 36 (Category 4), Item 37 (Category 4), Item 38 (Category 1 and 4), Item 39 (Category 1), Item 40 (Category 4), and Item 41 (Category 1) have percentage observed category usage with low frequencies that could affect a regular distribution. Categories with 9% usage were allowed because the categories provide a valuable response for raters. Those categories listed above should be collapsed into adjacent categories, for example, by combining category 1 and 2 if category 1 failed to meet the percentage usage requirement.

Next, average observed and expected measures must advance monotonically. This guideline means that from category 1 through each category to category 4, the average observed measure must continuously increase (Andrich, 1996). If one category decreases, it is in violation of the assumption of monotonicity. The second categories of Item 4, Item 19, Item 20, Item 36, and Item 37, and the third categories of Item 5 and Item 21 violate the assumption of monotonicity, and must be collapsed into adjacent categories.

Last, Outfit MSE values must be greater than or equal to 2.00. A value larger than 2.00 indicated the category was not used as anticipated. Item 4 (Category 3), Item 20 (Categories 1, 2, and 3), Item 21 (Category 3), Item 36 (Category 2), and Item 37 (Category 1) exhibit Outfit MSE values greater than or equal to 2.80 and must be collapsed.

Discussion

The first research question addresses the psychometric quality of the rating scale in its efforts to assess pre-service teachers' instruction. The psychometric quality of a scale is evaluated based on reliability, precision, and validity (Wesolowski, et al., 2017). First, evidence of reliability is presented through high reliability of separation statistics. Teaching episodes showed high reliability. Raters showed acceptable but lower reliability, which is an indication that severe raters could not be separated from lenient raters. Items also showed reliability on the lower end of acceptability, which is likely due to the sample size. These reliability statistics give evidence to support future research with this rating scale. Next, evidence of precision presented through low standard error among teaching episodes and items. Standard error below 0.3 is considered acceptable (Linacre, 1994). Both facets indicated low standard error to give evidence for strong precision. When reliability and precision are collectively considered, the evidence is that the measure was able to separate pre-service teaching episodes along the latent variable. The ordering of the teaching episodes and items supports future research involving the validity aspect of the psychometric quality of the rating scale.

The second research question addresses which items demonstrate acceptable model fit in the development of the measure. Acceptable model-data fit is expressed through the Infit and Outfit MSE values within the range specified by Linacre (n.d.b). Overall, the items demonstrated acceptable model-data fit, however, some items need careful consideration. First, Item 36, Use of

Technology not only exhibited overfit, but also a negative point-biserial. Normally, a negative point-biserial would be grounds for item dismissal. However, in the context of all teacher evaluation, the integration of technology into classroom teaching is integral, and therefore the item must remain in the measure. Item 20, Alignment of Warm-up to the Lesson Activity, likely exhibited misfit because students failed to include the warm-up into the lesson. The item did not function well because the classroom instruction in regards to the integration of a warm-up into a lesson may not have been properly introduced. Last, Item 5, Appropriate Response to Student Disruptions, exhibited overfit due to misinterpretation and lack of presence in the lessons. The item will likely be reworded, and with further research, be reviewed and potentially removed from the measure. All other items in the measure exhibited acceptable model-data fit and can be admitted as part of a pre-service teaching measure.

The third research question addresses how well the rating scale structure cooperates to produce meaningful measures. This question raises the possibility that not all items may carry an equal rating scale structure. As shown in Table 4.5 and as explained in the results, violations in the form of category usage frequency, monotonicity, and Outfit MSE cause categories to be collapsed and changed from item to item. These changes in the category structure suggest that the succeeding measure may provide different inferences. The modifications being made will require further testing, as the increase in sample size alone will alter the results.

In the field of music education, a measure of pre-service teacher performance is needed to determine how well the National Association of Schools of Music guidelines are being met. According to the NASM (2016), these six teaching competencies are listed as essential: (a) teach a variety of levels in various settings, incorporating effective classroom and rehearsal management; (b) understand the learning development of children in relation to music; (c) assess

students based on needs and backgrounds, assess content, and plan lessons with assessment in mind; (d) know repertoire and methods needed to teach various levels and students; (e) reflect on lessons and adjust teaching strategies based on various teaching situations; and (f) understand how assessment fits into the curriculum to inform of student progress. The measure being developed in this study breaks down each of these competencies into observable behaviors. By developing a uniform set of pre-service teaching goals that can be observed in the student teaching setting by the cooperating and supervising teacher, pre-service teachers can have a clearer idea of the expectations.

Not only does the measure align with the teaching competencies of the NASM, but the measure also aligns with the expectations of teacher evaluations. Instead of focusing on pre-service teaching competencies, which may be of a lower standard than that of an experienced teacher, the focus is on readying the pre-service teacher for what is to come in the real classroom. Aligning the teaching expectations at the pre-service level with those at the public teaching level will help teachers be more successful in their first years of teaching.

Future research will need to be completed with the current items in this measure, specifically in regards to sample size. Data analysis will be continued with the current set of data, but the researcher will continue to gather pre-service teaching videos and to recruit raters in order to gain a better understanding of the items and their relation to each other. The reliability among items was the most concerning aspect of this study, and it is believed that a larger sample size will relieve this concern and will provide more information to build a usable measure.

Once the items have been reviewed by a larger audience of raters, a revised rating scale will need to be transformed into a rubric. This process will involve taking each item and its four levels, and providing detailed descriptions for each level. The user will then have criteria at each

level that need to be met in order to receive that specific score. This step is most important before the integration into the music education curriculum. Ultimately, this measure would be used in music education curriculum to assist with observations of pre-service teachers. Cooperating teachers would be given the same measure as a guideline for the overall expectations. The alignment of the measure with NASM and teaching standards encourages success and readiness into the music teaching field.

References

Abeles, H. F. (1975). Student perceptions of characteristics of effective applied music instructors. *Journal of Research in Music Education, 23, 2*, 147-154.

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

Arends, R. I. (2006a). Summative performance assessments. In S. Castle & B. S. Shaklee (Eds.), *Assessing teacher performance: Performance-based assessment in teacher education* (pp. 93-123). Lanham, MD: Rowman & Littlefield Education.

Asmus, Jr., E. P. (1989). Factor analysis: A look at the technique through the data of Rainbow. *Bulletin of the Council for Research in Music Education*, *101*, 1–29.

Benjamin, W. J. (2002). Development and validation of student teaching performance assessment based on Danielson's Framework for Teaching. Paper presented at the annual meeting for the American Educational Research Association, New Orleans, LA, 1–5 April 2002.

Bergee, M. J. (1992). A scale assessing music student teachers' rehearsal effectiveness. *Journal of Research in Music Education*, *40*(1), 5–13.

Berry, B., & Ginsberg, R. (1990). Creating lead teachers: From policy to implementation. *Phi Delta Kappan, 71*, 616-621.

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

Brand, M. (1982). Effects of student teaching on the classroom management beliefs and skills of music student teachers. *Journal of Research in Music Education*, *30*, 255-265 doi: 10.2307/3345299

Brand, M. (1985). Development and validation of the home musical environment scale for use at the early elementary level. *Psychology of Music*, *13*(1), 40-48.

Cocetti, R. A. (1985, April). *Communication style or leadership: The validation and interpretation of an instrument.* Paper presented at the meeting of the Central States Speech Association, Indianapolis, IN.

Coker, H., Medley, D. M., & Soar, R. S. (1980). How valid are expert opinions about effective teaching? *Phi Delta Kappan, 62*(2), 131-134, 149.

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

Conway, C. (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*, 324-338. doi: 10.1 177/002242941

Danielson, C. (2007). *Enhancing professional practice: a framework for teaching* (2nd ed.). Alexandria VA: ASCD.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*(5), 523–545.

Engelhard Jr., G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. http://doi.org/10.1111/j.1745-3984.1994.tb00436.x

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

Engelhard Jr., G., & Perkins, A. F. (2011). Person response functions and the definition of units in the social sciences. *Measurement: Interdisciplinary research and perspectives*, *9*(1), 40–45.

Gipson, R. C. (1978). An observational analysis of wind instrument private lessons. *Dissertation Abstracts International, 39,* 2118A.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Granger, C. V. (2008). Rasch analysis is important to understand and use for measurement. *Rasch Measurement Transactions*, *21*(3), 1122–1123.

Hepler, L. E. (1968). The measurement of teacher-student interaction in private music lessons and its relation to teacher field dependence/field independence. *Dissertation Abstracts International, 47,* 2939A.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). Identifying effective classroom practices using student achievement data (NBER Working Paper 15803). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w15803.pdf.

Kogan, D. M. (1989). The cost of avoiding research. *Phi Delta Kappan, 71*, 220-224.

Kostka, M. J. (1984). An investigation of reinforcements, time use, and student attentiveness in piano lessons. *Journal of Research in Music Education, 32, 2,* 113-122.

La Paro, K. M., Scott-Little, C., Ejimofor, A., Sumrall, T., Kintner-Duffy, V. L., Pianta, R. C., Burchinal, M., Hamre, B., Downer, J., & Howes, C. (2014). Student teaching feedback and evaluation: Results from a seven-state survey. *Journal of Early Childhood Teacher Education*, *35*, 318–336. http://doi.org/10.1080/10901027.2014.968297

Linacre, J. M. (n.d.). Correlations: Point-biserial, point-measure, residual. Retrieved from http://www.winsteps.com/winman/correlations.htm

Linacre, J. M. (n.d.b). Fit diagnosis: Infit, outfit, mean-square standardized. Retrieved from https://www.winsteps.com/winman/misfitdiagnosis.htm

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,* 85–106.

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

Madsen, K. (2003). The effect of accuracy of instruction, teacher delivery, and student attentiveness on musicians' evaluation of teacher effectiveness. *Journal of Research in Music Education*, *51*(1), 38–50. http://doi.org/10.2307/3345647

Marzano Research Laboratory. (2013). *The Marzano teacher evaluation model.* Englewood, CO: Marzano Research Laboratory. Retrieved from: http://tpep-wa.org/wp-content/uploads/Marzano_Teacher_Evaluation_Model.pdfMilanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

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

Milanowski, A. T. (2011). Validity research on teacher evaluation systems based on the Framework for Teaching. Paper presented at the annual meeting of the American Education Research Association, New Orleans, L.A.

Munby, H., Russell, T., & Martin, A. K. (2001). Teachers' knowledge and how it develops. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 877–904). Washington, DC: American Educational Research Association.

National Association of Schools of Music. (2016). *National Association of Schools of Music Handbook*.

Plotz, M. (n.d.). *Teacher and principal practice rubric providers: Technical proposal application.* New York State Education Department. Retrieved from http://usny.nysed.gov/rttt/teachers-leaders/practicerubrics/Docs/mcrels-cues-framework-teacher-evaluation-instrument.pdf

O'Neill, S. A. (1993). *Identifying variations in teacher behavior during children's individual music tuition.* Paper presented, Indiana Symposium on Research in Social Psychology of Music, Indiana University, Bloomington, Indiana.

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education, 57*(1), 22-36. doi:10.1177/0022487105284045

Roegman, R., Goodwin, A. L., Reed, R., & Scott-McLaughlin, II, R. M. (2016). Unpacking the data: An analysis of the use of Danielson's (2007) Framework for Professional Practice in a teaching residency program. *Educational Assessment, Evaluation and Accountability*, *28*, 111–137. http://doi.org/10.1007/s11092-015-9228-3

Rosenshine, B. (1987). Explicit teaching and teacher training. *Journal of Teacher Education, 38*(3), 34-36.

Sandholtz, J. H., & Shea, L. (2012). Predicting performance: A comparison of university supervisors' predictions and teacher candidates' scores on a teaching performance assessment. Journal of Teacher Education, 63(1), 39-50. doi:10.1177/0022487111421175

Sandholtz, J. H., & Shea, L. M. (2015). Examining the extremes: High and low performance on a teaching performance assessment for licensure. *Teacher Education Quarterly*, *42*(2), 17–42.

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

Smith, D. T. (2009). Development and validation of a rating scale for wind jazz improvisation performance. *Journal of Research in Music Education*, (3), 217.

Sparks, G. M. (1984). Research on teacher effectiveness: What it all means. In J. Reinhartz (Ed.), *Perspectives on effective teaching and the cooperative classroom* (pp. 8-11). Washington, DC: National Education Association.

Soar, R. S., & Soar, R. (1983). Context effects in the teaching-learning process. In D. C. Smith (Ed.), *Essential knowledge for beginning educators* (pp. 65-75). Washington, DC: American Association of Colleges for Teacher Education.

Soslau, E., & Lewis, K. (2014). Leveraging data sampling and practical knowledge: Field instructors' perceptions about inter-rater reliability data. *Action in Teacher Education, 36*(1), 20–44.

Stronge & Associates. (2013). *Stronge teacher evaluation system: A validation report.* Retrieved from https://www.strongeandassociates.com/files/components/Stronge%20Validation%20Brief%2012%2015%202017.pdf

Stronge, J. H. (2012). *Teacher effectiveness performance evaluation system.* Stronge & Associates Educational Consulting, LLC. Retrieved from http://mnprek-3.wdfiles.com/local--files/teacher-effectiveness/TEPES%20-%20Stronge.pdf

Taebel, D. K. (1990). An assessment of the classroom performance of music teachers. *Journal of Research in Music Education*, *38*(1), 5–23.

Tillema, H. H. (2009). Assessment for learning to teach appraisal of practice teaching lessons by mentors, supervisors, and student teachers. *Journal of Teacher Education, 60*(2), 155–167.

Van Teijlingen, E. R., & Hundley, V. (2001). *The importance of pilot studies.* Retrieved from http://sru.soc.surrey.ac.uk/SRU35.html.

Wang, W. (2001). Verbal versus nonverbal communication in music performance instruction. *Contributions to Music Education*, *28*(1), 41–60.

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

Young, S., & Shaw, D. G. (1999). Profiles of effective college and university teachers. *Journal of Higher Education*, *70,* 670-686.

Tables

Table 4.1. *Summary Statistics*

		Facets	
	Pre-service teaching episode (θ)	Rater (λ)	Item (δ)
Measure (Logits)			
Mean	0.26	0.00	0.00
SD	0.92	0.30	0.62
N	16	8	41
Infit MSE			
Mean	1.04	1.04	1.00
SD	0.30	0.33	0.28
Std. Infit MSE			
Mean	0.10	0.10	0.00
SD	2.00	2.90	1.10
Outfit MSE			
Mean	1.07	1.07	1.07
SD	0.33	0.35	0.43
Std. Outfit MSE			
Mean	0.20	0.20	0.10
SD	1.90	2.70	1.20
Separation Statistics			
Reliability of Separation	0.96	0.83	0.77
Chi-Square	367.0*	40.2*	155.1*
Degrees of Freedom	15	7	40

*Note. * $p < 0.01$. Infit and Outfit Mean Square Error (MSE) should exist within the range of 0.5 – 1.5. Std. refers to a weighted mean square residual, where the mean is set to zero, and the variance is set to 1 (Wright & Masters, 1982). SD = Standard deviation. Reliability of separation refers to the spread of elements within the facet.*

Table 4.2. *Calibration of Pre-Service Teaching Episode*

Teaching Episode	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
13	3.10	1.55	0.19	1.22	1.40	1.21	1.20
5	3.00	1.43	0.18	0.70	-2.00	0.72	-1.80
10	3.20	1.26	0.19	1.50	2.80	1.71	3.60
3	3.00	1.23	0.18	1.13	0.80	1.10	0.60
7	2.80	0.57	0.17	1.54	2.90	1.53	2.80
2	2.80	0.50	0.17	0.82	-1.10	0.76	-1.50
8	2.70	0.48	0.17	1.17	1.00	1.10	0.60
1	2.70	0.38	0.17	0.46	-4.20	0.46	-4.10
11	2.70	0.15	0.17	0.99	0.00	0.87	-0.70
16	2.60	0.13	0.17	0.87	-0.80	0.85	-0.80
14	2.50	-0.04	0.17	0.96	-0.10	0.96	-0.10
15	2.50	-0.15	0.17	0.95	-0.20	0.96	-0.10
12	2.50	-0.23	0.17	1.18	1.10	1.12	0.70
6	2.30	-0.25	0.17	0.93	-0.40	1.31	1.60
4	2.20	-0.59	0.17	0.78	-1.50	0.96	-0.10
9	1.80	-2.17	0.19	1.50	2.80	1.51	1.80
Mean	2.60	0.26	0.18	1.04	0.10	1.07	0.20
SD	0.40	0.92	0.01	0.30	2.00	0.33	1.90

Note. Teaching episodes are ordered according to Measure, from highest value to lowest, highest-achieving to lowest-achieving.

Table 4.3. *Calibration of Raters*

Rater	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
2	2.50	0.42	0.12	0.91	-0.80	1.16	1.30
3	2.70	0.22	0.12	0.91	-0.80	0.91	-0.70
7	2.60	0.15	0.12	1.13	1.10	1.13	1.10
6	2.60	0.11	0.12	0.94	-0.50	0.86	-1.10
8	2.70	-0.10	0.12	0.59	-4.40	0.60	-3.90
1	2.70	-0.11	0.12	0.84	-1.40	0.78	-1.90
4	2.60	-0.13	0.13	1.58	4.40	1.58	3.60
5	2.70	-0.58	0.13	1.46	3.60	1.54	3.60
Mean	2.60	0.00	0.12	1.04	0.10	1.07	0.20
SD	0.10	0.30	0.00	0.33	2.90	0.35	2.70

Note. Raters are ordered according to Measure, from highest value to lowest, most severt to most lenient.

Table 4.4. *Calibration of Rating Scale Items*

Item	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
36	1.20	2.33	0.39	1.30	0.80	1.94	1.40
21	1.40	1.47	0.28	1.40	1.30	1.95	1.60
35	2.60	0.68	0.28	0.79	-0.70	0.82	-0.40
29	2.40	0.65	0.29	0.86	-0.50	0.84	-0.60
20	2.40	0.64	0.21	2.13	3.90	2.76	4.10
33	2.30	0.48	0.28	0.90	-0.30	0.86	-0.40
30	1.90	0.48	0.32	0.96	-0.10	0.95	-0.10
28	2.40	0.47	0.24	1.14	0.60	1.09	0.40
38	2.60	0.43	0.30	1.05	0.20	1.06	0.30
40	1.90	0.42	0.26	1.01	0.10	0.93	-0.20
25	2.60	0.33	0.25	0.70	-1.20	0.75	-1.00
32	2.00	0.22	0.26	1.32	1.50	1.37	1.60
9	3.00	0.18	0.31	1.03	0.20	1.03	0.10
27	3.00	0.13	0.36	1.01	0.10	1.01	0.10
13	2.60	0.11	0.25	0.80	-0.80	0.80	-0.80
14	2.60	0.09	0.23	0.90	-0.40	0.89	-0.40
5	3.10	0.04	0.25	1.23	1.10	1.67	2.40
1	3.10	0.02	0.29	0.94	-0.10	0.92	-0.20
6	2.70	-0.06	0.24	1.20	0.90	1.20	0.80
18	3.10	-0.10	0.30	1.00	0.00	0.99	0.00
41	2.70	-0.11	0.27	0.73	-1.10	0.74	-1.00
11	2.20	-0.13	0.27	1.07	0.40	1.11	0.50
26	3.10	-0.15	0.44	0.91	-0.10	0.86	-0.20
8	2.70	-0.15	0.26	0.69	-1.40	0.71	-1.30
34	2.80	-0.18	0.33	1.00	0.00	1.00	0.00
19	2.70	-0.26	0.30	1.04	0.20	1.04	0.20
4	2.90	-0.30	0.21	1.42	1.60	1.55	1.80
3	3.20	-0.30	0.31	0.97	0.00	0.96	-0.10
15	2.90	-0.33	0.27	1.28	1.00	1.34	1.20
7	2.70	-0.33	0.28	0.84	-0.60	0.84	-0.60
31	2.80	-0.37	0.28	0.78	-0.90	0.79	-0.80
12	2.80	-0.46	0.27	0.62	-1.80	0.63	-1.70
23	2.90	-0.46	0.31	0.74	-0.90	0.76	-0.80
37	2.40	-0.54	0.28	1.24	1.00	1.81	2.40
2	3.30	-0.60	0.32	1.18	0.80	1.18	0.80
39	2.90	-0.62	0.29	0.78	-0.80	0.80	-0.70
22	2.90	-0.64	0.27	0.64	-1.60	0.64	-1.60
16	3.00	-0.65	0.30	0.82	-0.50	0.82	-0.60
10	3.00	-0.68	0.30	0.83	-0.50	0.83	-0.50
17	3.10	-0.76	0.42	0.80	-0.60	0.69	-0.90
24	2.40	-1.00	0.33	0.95	-0.10	1.00	0.00
Mean	2.60	0.00	0.29	1.00	0.00	1.07	0.10
SD	0.50	0.62	0.05	0.28	1.10	0.43	1.20

Note. Rating scale items are ordered according to Measure, from highest value to lowest, most difficult to least difficult.

Table 4.5. *Item Behavior of Category Usage, Average Observed and Expected measures, and Outfit MSE*

Item	**Category Usage (%)**				**Average Observed Measure (Average Expected Measure)**				**Outfit MSE**			
	1	2	3	4	1	2	3	4	1	2	3	4
1	**-**	6(19)	17(53)	9(28)	-	-0.48 (-0.46)	0.20 (0.24)	0.82 (0.73)	-	0.90	1.00	0.90
2	**-**	3(9)	17(53)	12(38)	-	0.37 (-0.04)	0.75 (0.73)	1.14 (1.27)	-	1.20	1.30	1.10
3	**-**	4(13)	18(56)	10(31)	-	0.08 (-0.25)	0.32 (0.50)	1.20 (1.02)	-	1.10	1.00	0.80
4	4(13)	6(19)	10(31)	12(38)	0.06 (-0.50)	0.06*(0.15)	0.73 (0.64)	0.84 (1.05)	1.20	0.50	*3.10*	1.90
5	**-**	9(28)	10(31)	13(41)	-	0.01 (-0.42)	-0.08*(0.21)	0.60 (0.68)	-	1.80	1.40	1.60
6	3(9)	10(31)	13(41)	6(19)	-0.19 (-0.71)	-0.03 (0.01)	0.46 (0.53)	0.89 (0.95)	1.70	1.00	1.20	1.00
7	**1(3)**	12(38)	15(47)	4(13)	-1.71 (-0.59)	0.19 (0.22)	0.89 (0.81)	1.29 (1.24)	0.50	1.00	0.60	0.90
8	**2(6)**	11(34)	14(44)	5(16)	-1.66 (-0.69)	0.11 (0.07)	0.63 (0.63)	1.36 (1.05)	0.40	0.90	0.50	0.70
9	**-**	6(19)	19(59)	7(22)	-	-0.69 (-0.60)	0.20 (0.11)	0.46 (0.61)	-	0.90	1.20	1.10
10	**1(3)**	4(13)	21(66)	6(19)	-1.37 (-0.51)	0.71 (0.30)	0.82 (0.98)	1.91 (1.47)	0.30	1.30	0.90	0.70
11	7(22)	12(38)	13(41)	**-**	-0.26 (-0.32)	0.33 (0.34)	0.80 (0.83)	-	1.40	0.50	1.10	-
12	**1(3)**	10(31)	16(50)	5(16)	-1.59 (-0.52)	0.11 (0.28)	0.92 (0.88)	1.79 (1.33)	0.50	0.70	0.50	0.70
13	3(9)	11(34)	14(44)	4(13)	-1.44 (-0.86)	0.02 (-0.13)	0.30 (0.40)	1.20 (0.82)	0.60	0.90	1.10	0.70
14	4(13)	11(34)	11(34)	6(19)	-1.37 (-0.76)	0.12 (-0.08)	0.51 (0.42)	0.69 (0.82)	0.50	1.00	0.80	1.10
15	**2(6)**	6(19)	18(56)	6(19)	-0.39 (-0.65)	0.52 (0.11)	0.59 (0.70)	1.00 (1.16)	1.60	1.50	1.30	1.10
16	**1(3)**	5(16)	20(63)	6(19)	-1.40 (-0.50)	0.34 (0.31)	0.97 (0.97)	1.60 (1.45)	0.30	0.90	1.00	0.90
17	**1(3)**	**-**	25(78)	6(19)	-1.28 (0.09)	-	0.92 (0.94)	1.86 (1.52)	0.50	-	0.90	0.70
18	**-**	5(16)	18(56)	9(28)	-	-0.62 (-0.39)	0.45 (0.33)	0.72 (0.84)	-	0.90	1.00	1.10
19	**1(3)**	10(31)	18(56)	3(9)	0.61 (-0.71)	-0.10*(0.11)	0.76 (0.72)	1.19 (1.17)	1.90	0.80	1.00	1.00
20	10(31)	3(9)	15(47)	4(13)	-0.22 (-1.07)	-1.12*(-0.50)	-0.38 (-0.07)	-0.18 (0.31)	*3.00*	*3.60*	*3.50*	1.50
21	23(72)	5(16)	4(13)	**-**	-1.31 (-1.42)	-0.85 (-0.85)	-1.07*(-0.45)	-	1.30	0.80	*2.80*	-
22	**1(3)**	8(25)	16(50)	7(22)	-1.41 (-0.41)	0.08 (0.39)	1.15 (1.00)	1.61 (1.46)	0.40	0.50	0.60	0.80
23	**1(3)**	6(19)	21(66)	4(13)	-1.58 (-0.64)	0.03 (0.18)	0.89 (0.83)	1.49 (1.31)	0.30	0.80	0.80	0.90
24	**2(6)**	15(47)	15(47)	**-**	0.57 (0.25)	0.92 (1.05)	1.70 (1.62)	-	1.30	0.90	0.90	-
25	4(13)	9(28)	16(50)	3(9)	-1.61 (-1.03)	-0.17 (-0.32)	0.21 (0.20)	0.87 (0.62)	0.40	1.10	0.70	0.90
26	**-**	**2(6)**	26(81)	4(13)	-	-0.60 (-0.43)	0.35 (0.39)	1.30 (0.95)	-	1.00	1.00	0.80
27	**-**	4(13)	23(72)	5(16)	-	-0.45 (-0.61)	0.09 (0.15)	0.81 (0.68)	-	1.10	1.30	0.90
28	5(16)	11(34)	13(41)	3(9)	-1.04 (-1.06)	-0.38 (-0.38)	0.24 (0.11)	-0.10*(0.51)	1.00	1.00	0.70	1.50

29	3(9)	13(41)	15(47)	**1(3)**	-1.73 (1.35)	-0.61 (-0.59)	0.06 (-0.05)	0.05*(0.36)	0.70	0.80	0.90	1.10
30	7(22)	20(63)	5(16)	**-**	-0.84 (-0.83)	-0.17 (-0.13)	0.50 (0.36)	-	1.10	0.70	0.90	-
31	**1(3)**	10(31)	17(53)	4(13)	-1.67 (-0.61)	0.16 (0.21)	0.85 (0.81)	1.50 (1.26)	0.40	0.90	0.70	0.80
32	9(28)	13(41)	10(31)	**-**	-0.30 (-0.56)	0.09 (0.08)	0.31 (0.55)	-	1.70	0.90	1.20	-
33	3(9)	18(56)	9(28)	**2(6)**	-1.03 (-1.11)	-0.39 (-0.35)	0.11 (0.17)	1.12 (0.57)	1.00	0.90	1.00	0.50
34	**1(3)**	7(22)	22(69)	**2(6)**	-1.86 (-0.88)	0.27 (-0.05)	0.54 (0.60)	1.22 (1.08)	0.30	1.20	1.30	1.00
35	4(13)	7(22)	20(63)	**1(3)**	-1.96 (-1.42)	-0.35 (-0.69)	-0.19 (-0.16)	0.70 (0.27)	0.40	1.30	0.90	1.10
36	26(81)	5(16)	**1(3)**	**-**	-2.00 (-2.18)	-2.57*(-1.60)	-1.12 (-1.21)	-	1.20	*2.80*	0.70	-
37	5(16)	10(31)	17(53)	**-**	0.84 (-0.06)	0.15*(0.64)	1.19 (1.16)	-	*3.50*	0.30	0.90	-
38	**2(6)**	11(34)	18(56)	**1(3)**	-1.32 (-1.27)	-0.37 (-0.47)	0.02 (-0.10)	0.97 (0.54)	0.90	1.00	1.40	0.90
39	**1(3)**	6(19)	19(59)	6(19)	-1.42 (-0.49)	0.23 (0.32)	1.01 (0.96)	1.54 (1.44)	0.30	0.80	0.90	0.90
40	11(34)	12(38)	9(28)	-	-0.66 (-0.69)	-0.13 (-0.07)	0.43 (0.38)	-	1.00	1.00	0.80	-
41	**2(6)**	9(28)	17(53)	4(13)	-1.42 (-0.79)	-0.08 (-0.01)	0.62 (0.56)	1.22 (1.00)	0.50	0.80	0.80	0.80

Note. Categories 1 – 4 correspond to the levels of performance listed in Figure 4.2. A bold category usage indicates underuse of the category. * denotes a violation of monotonicity. Italicization denotes an Outfit MSE value outside of the necessary range.

Figures

Item	Domain	Description
1. Learning Environment	Classroom Management	The learning environment is the classroom and all students within it.
2. Respect	Classroom Management	Students should all be spoken to with appropriate language.
3. Rapport	Classroom Management	Rapport is developed through effective communication and positive expectations. Rapport is the relationship between the teacher and the students.
4. Response to Student Disruptions	Classroom Management	Student disruptions may include talking or playing out of turn, disrespecting teacher or other students. The teacher's response should occur in enough time that the lesson or classroom environment are not jeopardized.
5. Appropriate Response to Student Disruptions	Classroom Management	Response to student disruptions should not embarrass students, should given with calm voice, verbal or non-verbal cues.
6. Appropriate Balance of Playing & Talking	Classroom Management	The balance between teacher talking and group/student playing should be the most conducive to student learning.
7. Pacing of Instruction	Classroom Management	Pacing of instruction refers to the speed at which the teacher's instruction is delivered. Instruction includes directions, content delivery, and feedback.
8. Pacing in Activities	Classroom Management	Pacing of activities refers to the speed at which students are given lesson activities. Activities may include the warm-up, group work, discussion, ensemble performance, small group performance, etc.
9. Communication of Behavior Expectations	Communication	Communication can be verbal or nonverbal. Behavior expectations may include how to enter the room, answer questions, respond to student discussion.

10. Communication of Performance Expectations	Communication	Communication can be verbal or nonverbal. Performance expectations may include how to respond to director movements, how to be a good audience, how to listen.
11. Learning Objective	Communication	The learning objective (i.e., essential question, learning goal, central focus) is the main idea of the lesson. This could be written on the board, or expressed verbally throughout the lesson.
12. Verbal Communication	Communication	Verbal communication refers specifically to how the teacher delivers information using words. The information should flow easily from teacher to students. Poor verbal communication may appear in the form of student misunderstandings.
13. Non-verbal Communication	Communication	Non-verbal communication refers to the exchange of information from the teacher to the students without using words. May include: facial expressions, conducting gestures, body language, proximity, behavioral cues (snapping to get attention, hand raised).
14. Familiarity with Students	Communication	Familiarity may look like the teacher knowing students' names, knowing specific interests of the students.
15. Concise Feedback	Communication	Concise feedback refers to the length of the feedback to the individual, group, or ensemble.
16. Clear Feedback	Communication	Clear feedback refers to clarity, understandability of the feedback to the individual, group or ensemble. (Does what the teacher said make verbal sense?)
17. Appropriate Feedback	Communication	Appropriate feedback refers to the content of the feedback. It should relate to the passage just played, or about to be played. The feedback should make sense based on what was previously played.
18. Teacher Eye Contact	Communication	Eye contact between the teacher and students is important during musical performance and during feedback and content teaching.

19. Learning-Conducive Conducting	Communication	The teacher's pattern is clear. The pattern, cues, and flourishes should enhance the music and learning environment. The teacher's conducting should allow for students to have success.
20. Alignment of Assessment	Lesson Content	Assessment may be formative (throughout the lesson) or summative (at the end, final). Alignment refers to the assessment and its direct relation the the content/skills being taught in the lesson.
21. Checking for Understanding	Lesson Content	Teacher may check for understanding by asking questions, having students indicate how they feel about performance/understanding (thumbs up/down, show 1-5 on hand), moving around the room to listen to individuals.
22. Students' Interests	Lesson Content	Students' interests refers to anything familiar to the class, outside of music.
23. Use of Academic Vocabulary	Lesson Content	Academic vocabulary should include specific musical vocabulary relevant to the lesson. Students should be able to understand new words through proper use and application.
24. Differentiation of Instruction	Lesson Content	The instruction is differentiated when the teacher presents the same material in different ways. In addition, the teacher may choose to differentiate the content (some students may be at different levels in one class). Differentiation may also come in the form of groupings, either by ability level, or by mixed level to help lower achieving students.
25. Knowledge of the Score	Lesson Content	Teacher can demonstrate knowledge of the musical score through appropriate musical directions, correct conducting (cues, pattern).
26. Appropriate Tempos	Lesson Content	A tempo is appropriate when students are successful but also challenged.
27. Consistent Tempos	Lesson Content	A tempo is consistent when it does not fluctuate within sections of music.

28. Alignment of Warm-up to Lesson Activity	Lesson Content	Alignment refers to the direct relationship between the warm-up and the lesson activity.
29. Methodological Comments	Lesson Content	Methodological comments refer to how the teacher addresses specific instrumental methods. For example: embouchure, tone, fingerings, etc.
30. Questioning Type	Teaching Strategies	These questions should move beyond knowledge (recall) or comprehension, and into application, analysis, synthesis and evaluation.
31. Appropriate Questioning	Teaching Strategies	The teacher should ask questions relevant to the task at hand.
32. Engagement	Teaching Strategies	The teacher's style of teaching should encourage student engagement. Engagement looks like full participation/performance, students asking and answering thoughtful questions, students paying attention and reacting to teaching.
33. Probing	Teaching Strategies	Probing refers to the teacher asking follow-up questions when students are either incorrect or do not understand. The teacher may ask completely new questions to get to the same answer, or the teacher may rephrase the original question.
34. Variation of Teaching Methods	Teaching Strategies	May include: Teacher-centered (lecture, demonstrations), student-centered (discussions), content-focused, and/or active learning (groups, brainstorming).
35. Use of Scaffolding	Teaching Strategies	Scaffolding refers to specific support given to students within the lesson to reach the objective. May include: teacher or student modeling, guides, supporting resources (i.e., worksheet).
36. Mastery of Content	Teaching Strategies	The teacher proves to fully understand the musical concepts being taught.
37. Use of Technology	Teaching Strategies	Technology may include a SmartBoard, metronome, Harmony Director, videos/projector, recordings, iPad, etc.

38. Student Reflection	Teaching Strategies	Student reflection refers to any opportunity for students to think upon or discuss their own playing, make decisions about music.
39. Association with Previous Knowledge	Teaching Strategies	Association refers to the relation of current material to previously learned material. The teacher may discuss or remind students of the previously learned material before discussing new material. During the lesson, the teacher may reference previous material throughout.
40. Appropriate Teaching Strategies	Teaching Strategies	Teaching strategies may include technology, differentiated instruction, questioning, summarizing, practice, feedback, group-work, etc.
41. Error Detection	Teaching Strategies	Error detection refers to the ability of the teacher to hear mistakes and musical inaccuracies.

Figure 4.1. Items grouped by domain with given descriptions

Item	1	2	3	4
1. Learning Environment	Teacher establishes a very negative learning environment.	Teacher establishes a somewhat negative learning environment.	Teacher establishes a somewhat positive learning environment.	Teacher establishes a positive learning environment.
2. Respect	Teacher rarely treats students with respect.	Teacher sometimes treats students with respect.	Teacher frequently treats students with respect.	Teacher always treats students with respect.
3. Rapport	Teacher has established negative rapport with the students.	Teacher has established slightly negative rapport with the students.	Teacher has established slightly positive rapport with the students.	Teacher has established positive rapport with the students.
4. Response to Student Disruptions	Teacher did not identify disruptions.	Teacher's response to disruption detracted very much from the learning environment.	Teacher's response to disruption detracted little from the learning environment.	Teacher's response to disruption did not detract from the learning environment.
5. Appropriate Response to Student Disruptions	Teacher's response to disruption was inappropriate.	Teacher's response to disruption was slightly inappropriate.	Teacher's response to disruption was slightly appropriate.	Teacher's response to disruption was appropriate.
6. Appropriate Balance of Playing & Talking	The balance between teacher talking and student playing is inappropriate for the learning environment.	The balance between teacher talking and student playing is slightly inappropriate for the learning environment.	The balance between teacher talking and student playing is slightly appropriate for the learning environment.	The balance between teacher talking and student playing is appropriate for the learning environment.
7. Pacing of Instruction	The pacing of instruction is inappropriate.	The pacing of instruction is slightly inappropriate.	The pacing of instruction is slightly appropriate.	The pacing of instruction is appropriate.

8. Pacing in Activities	The pacing of activities is inappropriate.	The pacing of activities is slightly inappropriate.	The pacing of activities is slightly appropriate.	The pacing of activities is appropriate.
9. Communication of Behavior Expectations	Teacher communicates inappropriate behavior expectations.	Teacher communicates slightly inappropriate behavior expectations.	Teacher communicates slightly appropriate behavior expectations.	Teacher communicates appropriate behavior expectations.
10. Communication of Performance Expectations	Teacher communicates inappropriate performance expectations.	Teacher communicates slightly inappropriate performance expectations.	Teacher communicates slightly appropriate performance expectations.	Teacher communicates appropriate performance expectations.
11. Learning Objective	Learning objective is not evident within the lesson.	Learning objective is slightly evident throughout the lesson.	Learning objective is moderately evident throughout the lesson.	Learning objective is extremely evident throughout the lesson.
12. Verbal Communication	Teacher's verbal communication is unacceptable.	Teacher's verbal communication is slightly unacceptable.	Teacher's verbal communication is slightly acceptable.	Teacher's verbal communication is acceptable.
13. Non-verbal Communication	Teacher's non-verbal communication is unacceptable.	Teacher's non-verbal communication is slightly unacceptable.	Teacher's non-verbal communication is slightly unacceptable.	Teacher's non-verbal communication is acceptable.
14. Familiarity with Students	Teacher is not at all familiar with the students.	Teacher is slightly familiar with the students.	Teacher is moderately familiar with the students.	Teacher is extremely familiar with the students.
15. Concise Feedback	Teacher feedback is never concise.	Teacher feedback is rarely concise.	Teacher feedback is sometimes concise.	Teacher feedback is almost always concise.
16. Clear Feedback	Teacher feedback is never clear.	Teacher feedback is rarely clear.	Teacher feedback is sometimes clear.	Teacher feedback is almost always clear.

17. Appropriate Feedback	Teacher feedback is never appropriate.	Teacher feedback is rarely appropriate.	Teacher feedback is sometimes appropriate.	Teacher feedback is almost always appropriate.
18. Teacher Eye Contact	Eye contact is never made with students.	Eye contact is rarely made with students.	Eye contact is sometimes made with students.	Eye contact is frequently made with students.
19. Learning-Conducive Conducting	Teacher's conducting always detracts from the learning.	Teacher's conducting often detracts from the learning.	Teacher's conducting rarely detracts from the learning.	Teacher's conducting never detracts from the learning.
20. Alignment of Assessment	Assessment is not at all influenced by the lesson.	Assessment is slightly influenced by the lesson.	Assessment is somewhat influenced by the lesson.	Assessment is very influenced by the lesson.
21. Checking for Understanding	Checking for understanding is of low importance.	Checking for understanding is slightly important.	Checking for understanding is moderately important.	Checking for understanding is very important.
22. Students' Interests	Teacher rarely refers to students' interests within the lesson.	Teacher occasionally refers to students' interests within the lesson.	Teacher sometimes refers to students' interests within the lesson.	Teacher frequently refers to students' interests within the lesson.
23. Use of Academic Vocabulary	Teacher use of academic vocabulary is unacceptable.	Teacher use of academic vocabulary is slightly unacceptable.	Teacher use of academic vocabulary is slightly acceptable.	Teacher use of academic vocabulary is acceptable.
24. Differentiation of Instruction	Teacher employs unacceptable differentiation in the learning environment.	Teacher employs slightly unacceptable levels of differentiation in the learning environment.	Teacher employs slightly acceptable levels of differentiation in the learning environment.	Teacher employs acceptable levels of differentiation in the learning environment.

25. Knowledge of the Score	Teacher demonstrates poor knowledge of the musical score.	Teacher demonstrates fair knowledge of the musical score.	Teacher demonstrates satisfactory knowledge of the musical score.	Teacher demonstrates thorough knowledge of the musical score.
26. Appropriate Tempos	Teacher's tempos are never appropriate.	Teacher's tempos are rarely appropriate.	Teacher's tempos are often appropriate.	Teacher's tempos are always appropriate.
27. Consistent Tempos	Teacher's tempos are never consistent.	Teacher's tempos are rarely consistent.	Teacher's tempos are often consistent.	Teacher's tempos are always consistent.
28. Alignment of Warm-up to Lesson Activity	No warm-up is used.	Warm-up is unacceptable for the lesson activity.	Warm-up is slightly acceptable for the lesson activity.	Warm-up is acceptable for the lesson activity.
29. Methodological Comments	Teacher's methodological comments are unacceptable.	Teacher's methodological comments are slightly unacceptable.	Teacher's methodological comments are slightly acceptable.	Teacher's methodological comments are acceptable.
30. Questioning Type	Teacher never asks higher-order thinking questions.	Teacher infrequently asks higher-order thinking questions.	Teacher sometimes asks higher-order thinking questions.	Teacher always asks higher-order thinking questions.
31. Appropriate Questioning	Teacher asks inappropriate questions to the ensemble.	Teacher asks slightly inappropriate questions to the ensemble.	Teacher asks slightly appropriate questions to the ensemble.	Teacher asks appropriate questions to the ensemble.
32. Engagement	Student engagement is extremely unlikely based on the teaching style.	Student engagement is somewhat unlikely based on the teaching style.	Student engagement is somewhat likely based on the teaching style.	Student engagement is extremely likely based on the teaching style.

33. Probing	Teacher never probes incorrect answers.	Teacher rarely probes incorrect answers.	Teacher sometimes probes incorrect answers	Teacher frequently probes incorrect answers.
34. Variation of Teaching Methods	Teacher uses an unacceptable variety of teaching methods.	Teacher uses a slightly unacceptable variety of teaching methods.	Teacher uses a slightly acceptable variety of teaching methods.	Teacher uses an acceptable variety of teaching methods
35. Use of Scaffolding	Teacher employs inappropriate scaffolding.	Teacher employs slightly inappropriate scaffolding.	Teacher employs slightly appropriate scaffolding.	Teacher employs appropriate scaffolding.
36. Mastery of Content	Teacher demonstrates an unacceptable mastery of the content.	Teacher demonstrates a slightly unacceptable mastery of the content.	Teacher demonstrates a slightly acceptable mastery of the content.	Teacher demonstrates an acceptable mastery of the content.
37. Use of Technology	Teacher does not use technology in the lesson.	Connections between the lesson and technology are shallow.	Connections between the lesson and technology make sense but are not developed.	Connections between the lesson and technology are concrete and well-developed.
38. Student Reflection	Teacher never provides opportunity for student reflection.	Teacher rarely provides opportunity for student reflection.	Teacher sometimes provides opportunity for student reflection.	Teacher frequently provides opportunity for student reflection.
39. Association with Previous Knowledge	Teacher never associates the lesson with students' previous knowledge.	Teacher rarely associates the lesson with students' previous knowledge.	Teacher sometimes associates the lesson with students' previous knowledge.	Teacher frequently associates the lesson with students' previous knowledge.
40. Appropriate Teaching Strategies	Teaching strategies are completely inappropriate for the content.	Teaching strategies are slightly inappropriate for the content.	Teaching strategies are slightly appropriate for the content.	Teaching strategies are absolutely appropriate for the content.

41. Error Detection	Teacher's error detection is unacceptable.	Teacher's error detection is slightly unacceptable.	Teacher's error detection is slightly acceptable.	Teacher's error detection is acceptable.

Figure 4.2. The Pre-service Instrumental Music Teaching Rating Scale

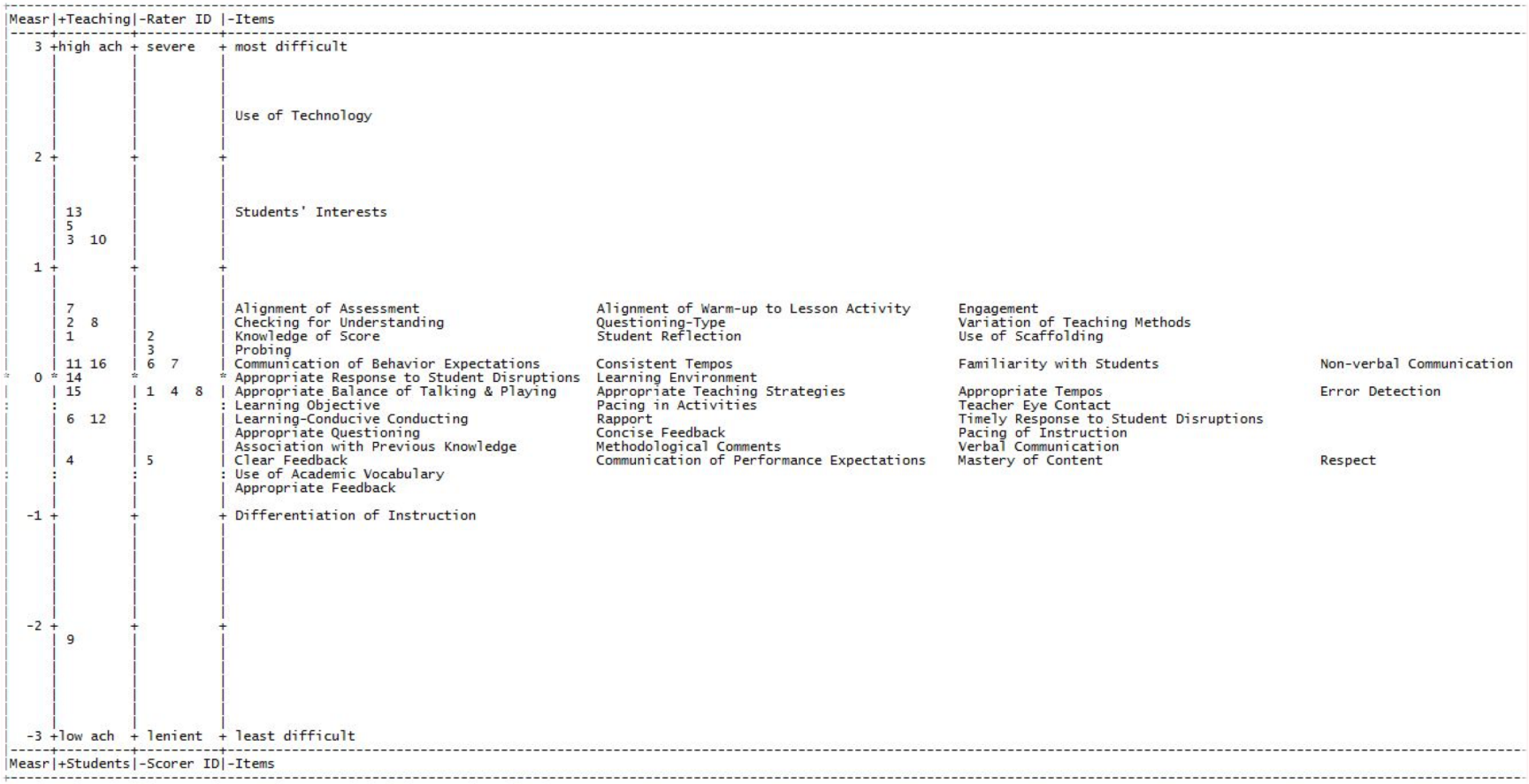

Figure 4.3. Wright Map

CHAPTER 5

CONCLUSION AND DISCUSSION

Pre-service teacher training has come under much change in the recent years with the implementation of new legislation (ESSA, 2015; ESEA, 2016). One goal of ESEA was to provide all students with teachers who were highly qualified. The quality of teachers is determined at the collegiate level through pre-serving training programs. Specifically in music, the National Association of Schools of Music (NASM) stresses the need for pre-service teachers who have mastered the necessary skills and concepts to be successful (NASM, 2016). However, the skills and concepts specified by the NASM are broad teaching competencies that cannot necessarily be observed as they currently exist. These competencies include the student's ability to teach multiple levels of music in various settings, understand learning theory and child development, plan lessons based on the background and needs of the students, understand current methodologies and repertoire, self-reflect and adjust teaching practices, and develop a thorough knowledge of assessment and applications thereof (NASM, 2016). The purpose of this research was to provide a starting point in the decision-making process of what pre-service teachers should look like at the end of their studies.

This dissertation first explores the development of a performance measure. The purpose of the first research study (Chapter 2) was to establish a methodology for the development and validation of a performance-based music assessment. First, the study addressed how raw score data can be collected from raters in a valid and meaningful way. Raw scores are gathered through the use of well-developed items as part of a rating scale or rubric. Additionally, raw data

is gathered when raters interact with said rating scale or rubric. Through the use of Rasch Measurement Theory, raw data undergoes a transformation to become interval level measures, which, in this context, allows for the comparison of pre-service teachers and of the items within the rating scales. The study also addressed how test construction and development can be handled in order to make inferences. This study formed a framework for the subsequent two chapters in the development of two new measures of performance.

The purpose of the second research study (Chapter 3) was to evaluate the validity and reliability of a rating scale to assess pre-service music teachers' lesson plan development. Specifically, lesson plans were focused on secondary-level instrumental music classrooms (i.e., middle school and high school band, or orchestra). The study addressed which items originally in the measure did not demonstrate acceptable model fit for the construct. Results indicated that five items did not adequately fit the model: activities permit student choice, learning experience connect to other disciplines, provides a variety of appropriately challenging resources that are differentiated for students in the class, engages students in authentic learning by providing real-life examples, and differentiates the instructional process to meet individual developmental needs. In addition, the study addressed the potential for change in the structure of the rating scale across items. Only four items maintained the original rating scale structure. Raters had a tendency to use the extreme categories (i.e., strongly disagree or strongly agree) least. Last, the study investigated differential severity among academic administrators and music education content specialists. Result showed that differential severity existed among only three of the items.

The results from the second study, paired with the methodology, stressed that a rating scale is never complete. In its current state, more research needs to be completed before the

measure can be used in the classroom. The lesson plan rating scale will continue to be revised as the rating scale becomes a rubric that can be used in undergraduate classrooms. Once the rubric is complete, it must continue to be revalidated in order to ensure the measure is working properly. Furthermore, the results highlighted the importance of aligning pre-service music teaching measures with current teaching expectations. It was enlightening to discover the difference between administrators' and music professors' ratings and expectations. The more that those expectations can align, the more successful pre-service teachers could be when they enter classroom teaching.

The purpose of the third research study (Chapter 4) was to determine which behaviors and skills are relevant in the assessment of pre-service music teachers. The study also built a measure to assess pre-service instrumental music teachers' performance in the student teaching setting. The first research question was aimed at evaluating the psychometric quality of the rating scale. Results showed adequate reliability for teaching episodes and raters. However, a small sample size affected the reliability of the items, and therefore more research is needed. Low standard error gave evidence of strong precision. The combination of reliability and precision gave evidence that there was separation among teaching episodes and items, which supports validity. The second research question addressed which items demonstrated acceptable model fit. Results indicated acceptable model fit for all but three items: Use of Technology, Alignment of Warm-up to Activity, and Appropriate Response to Student Disruption. Last, the rating scale structure was examined in its ability to produce meaningful measures. Results indicated violations in category usage, monotonicity, and Outfit MSE, which required categories to be collapsed and the rating scale structure to change.

Results from the third study emphasized how important sample size is when determining the fit of items. While the items demonstrate acceptable fit on their own, there was not enough separation to determine student performance on each individual item. Moreover, this study showed where improvements can be made in the pre-service teaching curriculum. One of the misfitting items was Use of Technology. The misfit of this item is an indication that pre-service teachers have not integrated technology well into their lessons, and that perhaps, not enough was done in teacher preparation to practice the integration of technology into music teaching.

Each of these studies highlights the importance of using Rasch Measurement Theory in constructing measures and measuring achievement of pre-service music educators in the context of the music classroom. Much of the research in the development of measures of music performance employs factor analysis (Miksza, 2012; Russell, 2010; Smith & Barnes, 2007; Zdzinski & Barnes, 2002, for example). This dissertation emphasizes that both analyses, factor analysis and Rasch Measurement, are capable of providing valuable information. However, Rasch Measurement Theory is more appropriate in the development of a performance assessment. Factor analysis uses raw scores, while Rasch Measurement Theory performs a logistic transformation on those raw scores to convert them from ordinal to interval data. This transformation allows persons and items to be organized on the same latent variable. Therefore, Rasch Measurement enables inferences to be made, as the findings are not sample- or test-dependent.

In the field of music education, performance assessments are becoming more essential to monitor the progress of student learning and growth. While performance assessments allow teachers to better understand their students, they also provide an accountability model for teacher evaluations. The rating scales and eventual rubrics provide concrete documentation where

students can recognize their current level of performance and can work to improve for future performance.

This dissertation also provides support for current pre-service teaching curriculum. The gap between pre-service teacher training and teacher evaluation systems can cause novice teachers to struggle to meet administrators' expectations. The more practice that pre-service teachers have with current teacher evaluation systems, the more successful they will be upon entering the public school classroom. By aligning the guidelines and expectations of the NASM and pre-service teaching with those of public teaching evaluations, the rating scales in this dissertation allow pre-service teachers to become more familiar with current evaluation practices and become better prepared for classroom teaching.

Future Research

The rating scale studies in this dissertation are both limited by sample size. First, in the development of a lesson plan rating scale, only 32 lesson plans were assessed. These lesson plans came from the students at the same university, meaning all students had similar training in music teaching. In addition, in the development of a pre-service music teacher rating scale, only 8 raters and 16 student videos were used. Again, these student videos all came from students at the same university. It is the intent of each of these studies to develop rubrics that could be used in alignment with the NASM expectations. These two studies need to be replicated using lesson plans and student videos from a larger variety of universities to account for diverse training programs.

The findings of the three studies lead to three areas for future research. These areas include exploring the components of planning and instruction more in-depth, transforming the

rating scale and pilot study into usable rubrics, and investigating current NASM-inspired pre-service teacher preparation practices at universities in the United States.

Exploring the components more in-depth

The dissertation focuses on two specific aspects of pre-service teaching: planning and instruction. The third chapter about lesson planning examines the overall lesson and how all components of the lesson plan work together. The fourth chapter about observing pre-service teachers' instruction examines a wide array of teaching elements, specifically communication, teaching strategies, and assessment.

The lesson plan rating scale and the pre-service teaching rating scale would be best used near the end of pre-service teachers' studies once those students have completed their training. These rating scales examine whole processes, and do not allow for smaller sections of the rating scale to be isolated. Through future study, there is an opportunity to break down the processes within each rating scale into smaller components, allowing for the further understanding of pre-service teachers' planning and instruction practices.

Lesson planning involves a sequential process of identifying a learning objective, planning a lesson (i.e., rehearsal) to meet that objective, and assessing students' learning in relation to that objective (Tyler, 2013; Santoyo & Zhang, 2016). In order for pre-service teachers to be most successful, each of these processes must be broken down and assessed individually. Future research would involve the development of separate rubrics for the writing of objectives, the writing of procedures, and the writing of an assessment plan. Additional components could include the use of academic language and the diagnosis of potential teaching problems with planned supports. These rubrics could be incorporated into a secondary instrumental techniques curriculum or any course where lesson planning is being introduced.

Pre-service teaching involves the executing of a lesson plan. However, a well-thought out lesson does not ensure the achievement of student learning. Teachout (1997) explored a list of 40 skills and behaviors indicative of successful music teaching. These skills and behaviors were grouped according to personal skills, musical skills, and teaching skills. Within chapter four, it was determined that musical skills should be embedded within teaching and personal skills. As in the future research with lesson planning, the future research with pre-service teaching would involve disassembling the overall teaching process into individual components. Such components could include feedback, communication, assessment, teaching strategies, and classroom management.

Transforming the rating scales into rubrics

As with all rating scales, revalidation is needed to make sure the instrument is functioning as intended. The lesson plan rating scale must be transformed from a rating scale into a rubric. The performance levels of each item will need to be rewritten using anchors (Vagias, 2006). An additional research study would be needed to investigate item and category use in relation to the new rating scale. The performance levels would then be given specific criteria needed to meet each level, which transforms the rating scale into a rubric. The rubric would then be tested to determine reliability and validity of the measure. Each time the rubric is used thereafter, reliability and validity would continue to be examined.

The pre-service teacher rating scale is in its current state after a pilot study. More research is needed involving a larger number of student videos, preferably from students from a variety of university backgrounds. Because the current study examines students from a specific university, all students have experienced extremely similar training. However, pre-service music teacher training may appear differently at other universities and colleges. Having a variety of

pre-service teachers' videos will ensure the items are representative of all pre-service instrumental music educators. Once the piloted rating scale has been tested with a new sample, the items must be rewritten in the form of a rubric. All items will be given detailed criteria for each performance level. When the rubric has been developed, it must go through the validation process to ensure the items are working properly as a rubric. Then, the rubric can begin to be implemented with current pre-service teachers.

Investigating current NASM-inspired pre-service preparation practices

The research in this dissertation has led to further interest in standards set by the NASM. First, because the standards for pre-service preparation as set forth by NASM are not well-defined or clear-cut, pre-service teacher preparation will look different at various schools of music. One future study of research will involve investigating current practices of pre-service teacher preparation in relation to observational experiences, practicum (field) experiences, and educational coursework.

Furthermore, the NASM lists desired attributes of music education students (NASM, 2016). These expectations are well communicated and specific. However, a method of assessment is not included with the NASM handbook. Future research could involve the development of an assessment (rating scale or rubric) to determine if pre-service music educators fulfill these expectations. Moreover, these attributes could be used within the music education interview process. A rubric could be developed specifically for determining potential music education candidates' possession of these necessary attributes.

The use of rating scales and rubrics within the classroom can provide valuable feedback about performance (Wesolowski, 2012). However, these tools should not be used exclusively for all assessment in education. Within the context of this dissertation, these rating scales should be

used to inform current pre-service teacher preparation practices and lead to the improvement of pre-service teachers' knowledge and skills.

References

Congress, 114th. Every student succeeds act (2015).

Elementary and Secondary Education Act of 1965 (2016).

Miksza, P. (2012). The development of a measure of self-regulated practice behavior for beginning and intermediate instrumental music students. *Journal of Research in Music Education, 59*(4), 321-338. Retrieved from http://www.jstor.org.proxy-remote.galib.uga.edu/stable/41348841.

National Association of Schools of Music. (2016). *National Association of Schools of Music Handbook*.

Russell, B. (2010). The development of a guitar performance rating scale using a facet-factorial approach. *Bulletin of the Council for Research in Music Education,* (184), 21-34. Retrieved from http://www.jstor.org.proxy-remote.galib.uga.edu/stable/27861480.

Santoyo, C., & Zhang, S. (2016). Secondary teacher candidates' lesson planning learning. *Teacher Education Quarterly*, *43*(2), 3–27.

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

Teachout, D. J. (1997). Preservice and experienced teachers' opinions of skills and behaviors important to successful music teaching. *Journal of Research in Music Education*, *45*(1), 41–50. http://doi.org/10.1080/13540600601152546

Tyler, R. W. (2013). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago Press.

Vagias, Wade M. (2006). *Likert-type scale response anchors*. Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson University

Wesolowski, B. C. (2012). Understanding and creating rubrics for the assessment of music performance, *Music Educators Journal, 98*(36), 36-42.

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, (3), 245.