THE PSYCHOMETRIC DEVELOPMENT AND REVIEW OF AN EVALUATION SYSTEM FOR STRING ENSEMBLE PERFORMANCE USING RASCH MEASUREMENT THEORY

by

KINSEY E. EDWARDS

(Under the Direction of Brian C. Wesolowski)

ABSTRACT

The purpose of these studies was to develop a valid and reliable rubric for the evaluation of large ensemble string performances using psychometric principles of invariant measurement. The three papers seek to define assessment within the music classroom, create and validate a rubric for performance evaluation, and review how the newly designed rubric operates in a live performance evaluation setting. The first portion of the study was guided by the following research questions: (a) What does Rasch Measurement analysis reveal about the psychometric quality (i.e., validity and reliability) of items, raters, and ensembles within the context of a large ensemble string performance assessment? (b) How do the items vary in difficulty, raters vary in severity, and ensembles vary in achievement? and (c) How does the rating scale structure vary across individual items? Music content experts (N = 25) were solicited to evaluate string ensemble performances. Response categories were optimized in order to increase measurement accuracy and precision. Implications for the improvement of music assessment practices are discussed.

The second part of the study was guided by the following research questions: (a) How do the numerical ratings from the condition A rating scale compare to those numerical results

yielded from the newly developed condition B rubric? (b) How do the written forms of feedback given to the directors of the ensembles from the two systems compare? and (c) How do the two forms compare in terms of overall usability for the raters? A side-by-side comparison of the condition A rating scale in relation to the condition B rubric was conducted. Music content experts (N = 3) were solicited to evaluate string ensemble performances using the condition A rating scale while three additional content experts used the condition B rubric to evaluate the performances. Results from the condition A rating scale were analyzed using both Rasch analysis and Classical Test Theory and results from the condition B rubric were analyzed using Rasch analysis. Comparisons were made to determine which method better distinguished true measurement of the actual performances. Implications for the improvement of music assessment practices are discussed.

INDEX WORDS: string ensemble, assessment, performance evaluation, invariant measurement, Rasch, rubric

THE PSYCHOMETRIC DEVELOPMENT AND REVIEW OF AN EVALUATION SYSTEM FOR STRING ENSEMBLE PERFORMANCE USING RASCH MEASUREMENT THEORY

by

KINSEY E. EDWARDS

B.M., Furman University, 2008

M.M.E., The University of Georgia, 2014

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

DOCTOR OF EDUCATION

ATHENS, GEORGIA

2017

THE PSYCHOMETRIC DEVELOPMENT AND REVIEW OF AN EVALUATION SYSTEM FOR STRING ENSEMBLE PERFORMANCE USING RASCH MEASUREMENT THEORY

by

KINSEY E. EDWARDS

Major Professor: Brian Wesolowski

Committee: Clinton Taylor
Emily Gertsch

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2017

DEDICATION

To God be the glory! It is only through Him and His gracious and masterful workings that I have been able to complete the course work and dissertation in order to earn this degree. I thank Him for the many blessings He has given me, and for the strength and persistence He gave me to complete the work. I would also like to thank my husband, Andy Edwards, who has provided unending amounts of love and support throughout this process. He is always my voice of reason, and my encourager and I simply would not be where I am today without him. I thank God for putting him in my life during our undergraduate days, and I look forward to spending the rest of my life with him as we make more wonderful memories in our next phases of life. I would like to also thank my parents, George and Nancy Bagwell, for praying for me always, for teaching me everything I know about what it means to work hard to accomplish your dreams, for providing everything I ever needed plus more, and most importantly, for loving me and being the best parents a girl could ask for. I also want to thank the Edwards family for the love, continued support, and unending prayers that undoubtedly helped to carry both Andy and I through the process.

ACKNOWLEDGEMENTS

I would like to thank Dr. Brian Wesolowski who has taught me the true meaning of what it means to put forth ten times the effort to in order to meet your goals. He has motivated me to think critically about what it is I choose to study, and to write so that others in the field will have access to the meaningful research I have chosen to pursue. Dr. Wesolowski has invested so many hours to teach me to never settle and to always aim to better the profession. I feel that I am a better music educator after having spent time studying with him and the many lessons I learned from him will undoubtedly continue to shape my career in the future. I also want to thank Dr. Taylor who taught me the importance of teaching outside of the box, exposing students to new string concepts that will take them outside of their comfort zones, and to always fight the battles that matter. Lastly, I want to thank Dr. Gertsch for her involvement in both my undergraduate and my graduate experiences. She is always so patient, so encouraging, and has always been willing to put forth the effort to help me become a better musician.

I want to thank Sarah Ball and Amy Clement, my slightly older sisters who have taught me everything I know about being in the string classroom. They have provided so much wisdom and encouragement over the years and I am truly thankful to have them as friends. I also want to thank Suzanne Pearson, my co-teacher, who has kept me sane throughout my entire graduate school experience. She is always so encouraging, supportive, and it makes me truly happy to share an office with her each day. I would also like to thank Dr. Bernadette Scruggs for the many advice tidbits and strategic planning thoughts that she shared in order to help with the data

collection. I want to extend an extreme thank you to the orchestra directors that agreed to participate in my studies in order to allow for an authentic evaluation experience to be studied.

Lastly, I want to thank my teachers who inspired me to enter the profession. Sydney Mellard, who helped me to first discover my love for the violin, and Dana Ballard and Dr. Peter Lemonds who continued to foster that love of music and inspired me to become a teacher. I am so thankful that they are my teachers, and I am proud to say that I am where I am today because of them. I also would like to thank Dr. Thomas Joiner and Dr. Anna Joiner for investing countless hours in my studies at both the Brevard Music Center and at Furman University. I strive each day to live up to the example that they set for me so many years ago.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

As educational reform continues to be in the spotlight, the data-driven approach is one method being pursued by politicians, policy makers, and administrators in an effort to improve teaching and learning (Brookhart, 2013). In the music education setting, this approach can be considered somewhat problematic. As with commonly tested subjects, such as math or science, there has been a push for the development of written assessments that can gather data on student learning in the music education classroom (Hope and Wait, 2013). A conflict arises when the performance nature of the class is considered. The use of written exams for gathering data may not completely and accurately measure the overall effectiveness of music education programs. Though written comprehension of music concepts is important, at least part of the assessment should take into consideration the performance skills that are learned in conjunction with those music concepts in order to consider how those parts work together to develop a complete understanding (Hope and Wait, 2013). Evaluating performance with a possible goal of determining overall teaching and learning effectiveness adds an additional challenge in the performing arts classrooms. There are forms of assessment already in place for rating performing arts ensembles, but the rater-mediated nature of the current performance assessments poses questions of validity and reliability because of the subjectivity present in raters' evaluations (Wesolowski, 2016a). The use of rater-mediated assessment results can be particularly problematic when claims of achievement, or statements of how performing arts programs

contribute to the overall education of students are made based on those performance evaluation results.

Determining the best method for performing arts assessment is important and should be a priority for music educators. Band, Choir, and Orchestra classrooms are, for the most part, performance-based in that the majority of time is spent developing performance skills (Hope and Wait, 2013). There are undoubtedly many variations in terms of the amount of time that is spent on teaching and practicing written concepts that are a necessary part of the discipline, but most music educators want their ensembles to be capable of giving strong performances. The amount of time spent on teaching and rehearsing is therefore in an effort to help students improve their playing ability. The desire is for the students to sound good, to understand the concepts behind making aesthetic decisions and adjustments in performance, to feel proud of how they sound when they perform, and to develop a lifelong love for music. Furthermore, students that are a part of a music performance program gain access to an outlet that allows for them to engage in creative expression that requires no words (Oxley, 1996). They certainly have the ability to develop and exercise creativity in other areas throughout the school day, but many choose music performance classes as their creative venture of choice.

The aesthetic, or non-tested nature of the music performance classes, poses two primary challenges when it comes to a data driven focus: how we assess student achievement, and how we show evidence of the benefit that students receive when they are a part of a music program during the school day. It is important that we develop a method of authentic assessment that will determine levels of student achievement, but because of the different nature of the classroom, the assessment must also be different from those assessments that are developed and implemented in tested subject classrooms, such as math and science (Zaleski, 2014). The nature of the

assessment should match the nature of the instruction and more importantly, it should be considered authentic in that it provides opportunities for students to perform tasks that are considered to be relevant to real world experiences (Asmus, 1999). If the primary focus is performance skills in the music classroom, then the assessment approach should also be primarily performance based as well (Brewer et al., 2014). This does not mean there should be an absence of written assessment, however, there might need to be a balanced combination of the two in order to fully determine overall effectiveness in the music classroom. The performance-based aspect of assessment has led some music educators to claim that music performance classrooms should not be tested (Fisher, 2008). Students should be given the opportunity to show what they have learned, this just means that time, energy, and resources need to be used in order to collectively develop authentic assessment measures that adequately measure student achievement levels, and that can be used to communicate and confirm evidence of the importance of these programs in relation to student achievement results. A change in assessment measures will allow for the use of a common language that can help music educators to confidently report assessment results that will show the benefits of music education programs (Pellegrino et al., 2015). This does not mean that we change what we do as a discipline; rather, we just develop an authentic assessment that can clearly assess and communicate evidence of student achievement levels and teacher effectiveness.

Topics of consideration for the development of an authentic music performance assessment must include a discussion of the importance of the results of the assessment, how the results are calculated and analyzed, and how the assessment is used to inform teaching in the classroom and for large ensemble performance evaluations. For these studies, overarching ideas of the data driven era, and specific relation of those ideas to the music education classroom were

gleaned by reading Brewer's (2014) *Consequential Validity of Accountability Policy: Public Understanding of Assessments,* Brookhart's (2013) *The Public Understanding of Assessment in the Educational Reform*, Hope and Wait's (2013) *Assessment on Our Own Terms*, and Pellegrino's (2015) *Assessment in Performance-Based Secondary Music Classes*. After understanding why authentic assessment is needed and the challenge that is presented in trying to develop music performance assessment, a more focused look revealed that music performance assessments must be valid and reliable. This refers to the ideas that an assessment must measure what it intends to measure, and the results must be consistent with each administration of the assessment. These aspects of assessment were researched and explored primarily with Messick's (1989) *Validity* and Asmus' (1999) *Music Assessment Concepts.*

Current evaluation practices have limitations that provide evidence that it may be worth investigating new ways to develop meaningful, valid, and reliable assessments. This allows for the investigation of Classical Test Theory (CTT) versus Item Response Theory (IRT), and more specifically, the Rasch Measurement Model as a type of IRT. The difference between CTT and IRT are explored in order to determine how the data should be best collected, treated, and analyzed. Advocacy for the implementation of assessments supported by an analysis using IRT is presented. The Rasch Measurement Model is used as a method for the development of music performance assessment due to its ability to support partial credit responses and analyze multiple facets, including rater, item and performance function. The primary bodies of literature used to inform decisions made based on the Rasch Measurement Model were *Constructing Measures* by Wilson (2005), *Optimizing Rating Scale Category Effectiveness* by Linacre (2002), *Invariant Measurement* by Engelhard (2013), and *Applying the Rasch Model* by Bond and Fox (2015). The Rasch model was used to develop a rating scale that was eventually converted into a rubric for

music performance assessment in order to allow for a method of communication between performers and raters. The use of the Rasch Measurement Model also allowed for the discussion and analyses of how the raters functioned when rating music performance. Both aspects of rubric development and rater functioning in music performance assessment were primarily informed through the use of *Understanding and Developing Rubrics for Music Performance Assessment* by Wesolowski (2012), *Documenting Student Learning in Music Performance: A Framework* by Wesolowski (2014), and *Rater Analyses in Music Performance Assessment* by Wesolowski (2016a). The bodies of literature mentioned here merely represent an overview of the research that was used. The sampling listed above included thought-provoking examples that eventually led to the investigation of a larger body of literature that focused on supporting topics and ideas.

The second chapter, previously published in *Georgia Music News* in the spring of 2016, aims to answer the questions: (a) What is assessment? (b) How should assessment be developed and used? and (c) How should we taper our approach to advocacy if assessment is to only be used to show evidence of student growth? This paper serves primarily as a practitioner manuscript in which a philosophy is presented in order to support the remaining two papers. The first paper does not contain a methodology for this reason.

The third chapter, currently in press for the *Bulletin of the Council for Research in Music Education*, is a research study that aims to answer the questions: (a) What does Rasch Measurement Analysis reveal about the psychometric quality (i.e., validity and reliability) of items, raters, and ensembles within the context of a large ensemble string performance assessment? (b) How do the items vary in difficulty, raters in severity, and ensembles vary in achievement? and (c) How does the rating scale structure vary across individual items? For this portion of the research, a rubric was developed to serve as a large ensemble music performance

assessment. Twenty-five content experts listened to fifty-two recordings from a previous large ensemble performance evaluation event and rated those performances using the newly developed rating scale. The rating scale was developed using an item pool from a previous study (Zdzinski and Barnes, 2002). An incomplete rater assessment network was used to ensure that two raters rated each performance. The resulting rating scale was then developed into a rubric using Vagias' (2006) *Likert-type Scale Response Anchors* in order to allow for more specific forms of feedback.

Chapter four is a second research study that outlines the process used to investigate a side-by-side comparison of the current large ensemble performance evaluation system, referred to as condition A, in contrast to the newly proposed rubric that was developed in the previous study and that is referred to as condition B. The research study aims to answer the questions: (a) How do the numerical ratings from the condition A rating scale compare to those numerical results yielded from the condition B rubric? (b) How do the written forms of feedback from the two systems given to the directors of the ensembles compare? and (c) How do the two forms compare in terms of overall usability for the raters? Thirty-four large string performance ensembles participated in the study. Each ensemble performed three selections and received ratings from the six raters: three condition A raters (using the condition A rating scale) and three condition B raters (using the condition B rubric). Ratings were compared for analyses and an informal survey was given to the condition B raters and participants in order to provide further thought and comparison on the usability of the two systems.

The current approach to performance evaluation systems was developed sometime in the early to mid-twentieth century and has remained consistent throughout most of its administration (Colwell, 1970; Mark and Gary, 2007). As educational trends continue to evolve, the time is

right for music educators to implement new assessment measures. Students can undoubtedly benefit from being involved in music education programs, however, we do need to be able to show how they benefit (Morrison, 1994). Assessment results from this newly developed performance evaluation rubric have the potential to confirm the high student achievement levels that are resulting in performance-based classrooms.

CHAPTER 2

ASSESSMENT IN MUSIC EDUCATION: SHOWING GROWTH, NOT ADVOCACY[1]

Assessment in the music classroom can be a valuable tool for tapering instruction in a way that will most benefit students. In order for assessment to be helpful for students and educators, it must be developed and used appropriately. This process entails developing valid, reliable, and useful tests that will accurately measure what students know in an effort to redirect their learning. Unfortunately, assessment is used inappropriately in the classroom each day. For those that have a limited understanding of how assessment works and of how assessment results should be used, the process can actually negatively impact student learning. As music educators, the ultimate goal should be to develop assessments that will correctly demonstrate students' level of understanding. These results can then be used to help us guide instruction and to provide valuable feedback for students in order to improve understanding of content and performance in the music classroom.

In his article, *The Status of Arts Assessment: Examples from Music*, Richard Colwell (2003) warns, "Without intellectual honesty and a deep understanding of assessment, artists and advocates may be led to assess the wrong experiences…they may fail to relate assessment to what is to be learned, they may use inadequate instruments, and, most important, they may be unequipped to deal with the intricacies of interpretation and dissemination" (Colwell, 2003, p.12). The validity of this point lies in the notion that assessment is very easily used incorrectly. And when used incorrectly, learning can become meaningless. Examples of such can be seen in

[1] Edwards, K. (2016). Assessment in music education: showing growth, not advocacy. *Georgia Music News*. 76(3).

recent federal legislation that provided mandates for the uses of assessment. *No Child Left Behind* (2001) was passed during George W. Bush's administration and was intended to provide aid for schools with low achievement rates (Abbott, 2013). Unfortunately, financial help was only provided to schools that were able to show Adequate Yearly Progress (Abbott, 2013). This stipulation undermined the intentions of the legislation because assessment results were used ineffectively to allocate funding instead of helping to drive instruction within the classroom. Following *No Child Left Behind* (2001), *Race to the Top* (2009) was passed during Barack Obama's administration (Abbott, 2013). *Race to the Top* (2009) was far more specific and directed in terms of specific needs that each state needed to meet, however once again, only states with the highest level of achievement were awarded with additional funds for educational use (Hill, 2014). The punitive nature of both *No Child Left Behind* (2001) and *Race to the Top* (2009) are primary examples of assessment results being used incorrectly and inappropriately as a method to monitor and improve student learning. When assessment is used for political advancement and to manipulate the distribution of funds, students do not benefit and no real advancement in student achievement will result (Hill, 2014). This notion is further confounded when assessments are created to measure growth on a national level, regardless of the nature of the discipline or of needs that can only be determined on a local basis.

Developing the right assessment for a music classroom can be very troubling. Politicians, test developers, administrators, community members, and teachers typically feel that they have a thorough understanding of assessment, however, most often they actually have a limited and often superficial understanding of assessment. Adopting consistent terminology in relation to assessment is therefore a necessary first step in the process of developing effective forms of assessment. The manner in which we assess students should always be relative to the type of

learning that is taking place. Formative assessment is most commonly used to guide teaching throughout the learning process because students receive periodic feedback from teachers as to what they comprehend versus what they need to reconsider. Scott (2012) describes formative assessment as "Assessment for Learning" (Scott, 2012, p. 4). This type of assessment is often referenced to a constructivist approach in which students receive feedback as a way to expand upon their current understanding. According to Scott (2012), formative assessment is highly effective in improving student learning because it is an opportunity for students to learn from their mistakes and to take chances in learning without focusing too heavily on grades or on the final judgment of comprehension.

Summative assessment, on the other hand, provides the opportunity for a final judgment to be implemented within a specific unit. This final assessment might result in assigning students a grade at the end of a grading period or class. Scott (2012) defines summative learning as "Assessment of Learning" (Scott, 2012, pg. 3). This is the form of assessment that students are most accustomed to because it is the basis upon which teachers form a final evaluation of the students' level of understanding in relation to the objectives outlined at the beginning of the unit or course (Scott, 2012). For assessment to be most valuable, formative results should be used to guide learning that occurs prior to summative assessment. Students should have the opportunity to continuously receive feedback and redirection in terms of their thought processes and understanding of content prior to receiving a final evaluation. With this structure in mind, learning can be thorough and focused in the classroom.

In addition to deciding whether to use formative or summative assessment, teachers must also decide what type of assessment best fits the type of instruction that is being delivered. In academic classrooms, teachers most often assess student understanding of certain knowledge and

concepts with paper and pencil tests. Knowledge and concepts are a component of the music discipline, however performance skills are also a large component of music education. In the music classroom, performance skills refer to those things that students can do: sing, play, create, and listen, for example. Herein lies the issue with assessment in the music classroom. If we are teaching students to sing and play in the music classroom, then we need to also be able to assess how well they are able to execute those skills, and perhaps more importantly, that students know when to apply and use those skills. According to Wesolowski (2014), "Learning objectives should be clear, specific, and measurable and describe the most important learning that is taking place in the classroom" (p. 3). The application of such skills requires that we are assessing conceptual understanding in the classroom. Though the skill is performance based, teaching students how to decipher what skills should be employed and how to execute those skills relies on the fact that they have a thorough and accurate conceptual understanding of the skills in relation to the bigger musical picture.

Valid artistic assessment therefore entails evaluating performance skills separately, but then also assessing how those parts work together in order to form a complete musical understanding (Hope and Wait, 2013). Though we can assess conceptual knowledge and understanding with paper and pencil tests, to assess skills in relation to those concepts, Hope and Wait (2013) insist that we must think differently about how to show this type of progress: assessment should help to determine if an individual can reach a "meaningful interpretation" that only somewhat relies on his or her understanding of and ability to demonstrate technical skill (Hope and Wait, 2013, p. 7). This means that one form of assessment (paper and pencil for instance) will not suffice to assess everything that students learn during a music class: assessment needs to show how concepts and skills are used together in the music discipline in

order to ultimately result in a complete product. Hope and Wait (2013) explain, "scientific kinds of evaluations can never do the entire job of evaluation in the arts disciplines…science is looking for single answers: the arts, for multiple answers conceived by individual creators as they set their particular goals" (Hope and Wait, 2013, p. 5). If our classroom instruction is based heavily on the acquiring of performance skills and the application of those skills, the assessment needs to assess the performance in the same way. Thus, we need to assess understanding of such in a way that is similar to instruction delivery methods. For this reason, a dual approach to assessment in the music classroom is favored: one in which we are able to assess student understanding of knowledge and concepts through paper and pencil tests, in combination with a specific form of assessment that assesses students' ability to demonstrate performance skills.

Wesolowski (2014) explains that the best manner to assess students in terms of performance-based music tasks is to use checklists, rating scales, and rubrics. Checklists simply determine if a certain behavior or skill is being demonstrated, but rating scales provide a bit more information that can be used to direct student learning. Though rating scales may be considered to be a subjective form of assessment, the important aspect of this type of assessment is that students and teachers have a common understanding of what is being communicated. If used appropriately, this information can then be used to address the specific needs of each individual student. For this reason, rating scales must always contain three aspects that describe the performance: the content expected, description of each action in relation to the content, and a scale that describes the level at which students perform each task (Wesolowski, 2014). As was mentioned earlier, the primary reason to assess student learning, especially in terms of the formative level, is to be able to provide students with feedback in order to strengthen understanding. Ratings can serve this purpose in the music classroom, especially in terms of

performance skills. As long as goals are communicated clearly and each level of the rating scale explains the standard to which the skill is performed, students and teachers alike will be able to understand current progress in relation to the completion of the final goal. This idea therefore operates on the understanding that the rating scales are used for the classroom with the intent of improving student progress – thus, results should not be used for any ulterior purposes outside of the classroom.

Checklists, rating scales, and rubrics serve as formal formative assessments. These are assessments that students complete individually in the classroom in order to determine levels of understanding. In addition to formal formative assessments, the music classroom is unique to education in that informal formative assessment occurs almost every minute in the classroom. Teachers (as conductors) are continuously providing feedback for students in order to perfect and improve student performance within a group setting. Teachers use what they are hearing to assess how students are performing and to communicate what changes need to be made in order to adjust the performance in an effective way. Assessment of individual performance, as well as group performance, is continuously occurring in the music classroom. Taking time to listen to every performance and to determine what needs to be improved upon is an opportunity for teachers and students to constantly redirect practice and learning. This process can even be implemented following a final performance in order to determine if objectives and goals were met, and to what level of proficiency those were achieved (Hope and Wait, 2013). As musicians, these objectives and goals are based on artistry and the application of individual or group-based aesthetic decisions that may be open to differences in interpretation. When informal assessment of technical skills is used in combination with a sense of ambiguity in terms of aesthetic decisions, a conceptual approach to the acquisition of musical knowledge will help to enrich the

learning environment for students (Hope and Wait, 2013). In learning environments such as this, students are able to engage in informal assessment while providing opinions about how the group performed, or perhaps in relation to how another individual performed. This results in an aesthetic approach to conceptually applying technical skill to both past and future performances, therefore resulting in music educators and students developing an appreciation for "personal aesthetic preference" (Hope and Wait, 2013, p. 7). As music educators, our goal should be to better educate students on how to assess themselves in such a way that they can learn how to listen during performance, to have an understanding about what sounds good and what could be changed, and then to implement those changes in subsequent performances. Hale and Green (2009) refer to a cyclical process of learning in which students are given feedback, provided strategies to fix misconceptions, and then allowed time to practice until mastery of the skill occurs. Feedback, discussion, and practice should continue until students are able to show growth and perhaps even mastery in terms of both executing technical skill and in making aesthetic decisions. As was mentioned earlier, our primary goal should be to help students learn to assess themselves. Thus, taking time to provide feedback in terms of informal and formal formative assessment is the best way to help students understand and glean the most from the learning process (Hale and Green, 2009).

To academic educators, administrators, or parents, the strategies mentioned above may not seem like formative assessment, but Hope and Wait (2013) contest that it absolutely is. Music educators recognize that assessment occurs continuously in the classroom, but perhaps this idea is not common knowledge to most. Hope and Wait (2013) clarify, "Our problem is not that we do not know how to make assessments and evaluation, but rather that we are not as adept as we need to be in explaining to others what we do, how it works, and why it works" (Hope and

Wait, 2013, p. 3). We are accustomed to using assessment for the sake of improving student conceptual understanding as it relates to musical performance; however, those that are not involved in music education may not understand what assessment looks like in a performance-based classroom. As music educators, it is our job to convey to others what assessment looks like in the music classroom. Taking time to explain assessment in the music classroom to fellow educators, administrators, politicians, community members, and to parents is certainly a more effective approach in contrast to being defensive when other educators or the public show a lack of understanding or disbelief that assessment does in fact take place in the music classroom. Agreeing on consistent terminology to describe and discuss assessment is the first step in being able to discuss what this process looks like in the music classroom. Hope and Wait (2013) challenge music educators to "think more deeply about communication, with the goal of maintaining assessment on terms useful and productive to the music profession" (Hope and Wait, 2013, p. 3). Taking the time to become educated in the ways of assessment can help to ensure that we are able to adequately describe assessment in the music classroom without denying music's integrity as a discipline.

It is important to remember that the ultimate goal of assessment should be to direct continued growth and learning. Messick (1989) affirms this concept with his idea of consequential validity. Validity, in its most simplest form is generally used to confirm that the test accurately and fairly assess what it is supposed to test. Consequential validity refers to the manner in which data resulting from the assessment is used to form conclusions (Brewer et al., 2014). There are two components to consequential validity that specifically relate to the use of assessment in the classroom. The first of these is the fact that the label of the assessment should directly represent what knowledge and skills are being assessed (Brewer et al., 2014).

Assessment should ultimately provide opportunities to redirect learning in order to help classroom objectives to be reached (Crochet and Green, 2012). This involves diagnosing what students do not understand and then developing instruction based on redirecting misconceptions commonly held by students in the classroom. Consequential validity also refers to the manner in which the information derived from the assessments is used to inform decision making or more specifically, policy making (Brewer et al., 2014). Once assessment results are gathered, the decision of how to use those results is of upmost importance. We have already discussed the fact that student learning should be the primary purpose, and this also means that students should never be assessed just for the sake of doing so. Thus, if a test is being used to incorrectly assess what is actually being learned or if the results from a test are being improperly used, the educational outcomes will be negative (Brewer et al., 2014). Educators need to be certain to view assessment results, to study those results, and to ultimately make adjustments to practice in order to assist students in achieving comprehension. We assess students so that we can modify the manner in which we deliver instruction, hopefully resulting in students experiencing growth and development. Using assessment for any other reason is wasteful and inconsiderate of students' time and energy. For this reason, assessment should always be considered in terms of what it accomplishes in the classroom for student understanding. What happens outside of the classroom should not affect assessment results, nor should assessment results be used to affect any happenings outside of the classroom. In order for assessment to be effective and not an abusive process, we should continually use the results in the classroom for the benefit of individual students.

Inappropriate uses of assessment outside of the classroom include teacher evaluation initiatives, promotion of political agendas, or most astonishing, advocacy for music education.

Unfortunately, most of the recent educational trends falsely use assessment results to provide evidence of the overall effectiveness of political education initiatives (Brookhart, 2013). In contrast to this purpose of assessment, what matters most is the ultimate growth of students and the use of assessment results outside of the classroom serves no purpose in accomplishing this goal. One of the most inappropriate uses of assessment outside of the classroom is the use of assessment for advocacy efforts. The biggest issue in trying to use assessment as a component of advocacy is the fact that assessment results from the music classroom will be used to compare music classrooms to the traditional classroom (Hope and Wait, 2013). This is problematic due to the fact that assessment in the music classroom is very different from traditional classroom settings. Assessing musical knowledge and understanding in the same manner as that of academic classes is unrealistic primarily due to the subjective and aesthetic nature of the music discipline (Blakeslee, 2004). When academic achievement methods are implemented to describe the level of learning taking place in the music classroom, faulty and inappropriate results will only lead to confusion and outrage due to society's lack of understanding of true arts assessment (Brookhart, 2013). If we instead continue to focus on how assessment can indicate student learning, students will undoubtedly continue to benefit from guided instruction in the classroom. In contrast, resorting to using test results as a form of advocacy is dangerous to the integrity of music as a discipline and only works to undermine music's contribution to student growth and learning (Colwell, 2003). Therein lies the overarching problem with the (misinformed) common idea of using assessment as a method of advocacy for music education.

Assessment and advocacy are two separate entities and should not be used in combination with one another. Richard Colwell (2003) warns "if the arts should become a core subject based on the wrong foundation, however, with standards and other riggings modeled on

math and science, then the use of needs assessments, ability measures, diagnostic tools, and formative evaluation all will focus on the wrong objectives" (Colwell, 2003, p. 12). When used for advocacy, assessment becomes ineffective and draws the attention away from students. The primary reason for this is current assessment tools do not accurately measure what students actually know about the content, nor does it show the benefits music can provide for students outside of the classroom (Colwell, 2003). Ultimately, music is an aesthetic discipline and a form of artistic expression that cannot fully and accurately be measured based on academic test results (Fisher, 2008). Furthermore, if testing is used for advocacy, this (in some ways) implies that testing should serve as a national gauge for student achievement and evaluation of music programs (Fisher, 2008). This is an inaccurate view because the effectiveness of an arts program should be measured on a local level (Welenc, 2010). Each community needs something different from their arts programs and each school serves a completely diverse set of students. Why then, should we try to assess students on a national level when goals and objectives should be set to best serve students at the local level? When asking school board members what constitutes an effective music program in their schools, Orzolek (2006) found that officials perceived success in such programs related directly to performances at school events, good quality concerts, numbers of students involved in the programs, a small number of parental complaints, and awards for these programs (Orzolek, 2006). With these aspects in mind, music educators can ensure that they have a successful program and can use assessment to make sure high standards are being achieved in the classroom, all in an effort to ensure a quality education for students. When assessment is used to guide learning and ensure understanding, the rest of the successful components of a music program will fall into place. Herein lies the argument as to why assessment should not be used for advocacy efforts: when assessment is used for advocacy, short

fallings will be reported because no academic test can accurately show what the discipline can do for students. However, when assessment is used appropriately in order to direct learning and to provide measures of growth for students, individual students will be successful. This will in turn result in the development of successful music education programs that will positively affect surrounding communities.

Assessment is best used to guide learning in the classroom, therefore, the music education community must decide how to advocate without using assessment results. In my opinion, the answer is simple: the most effective way in which we can advocate for music education is to simply not advocate at all. Advocacy in itself seems to suggest that the discipline needs rescuing and that we must defend something that is failing in the eyes of the general public. By speaking out and scrambling over our words while showing only a slightly united front (due to the fact that anyone is able to present their feelings on behalf of the discipline), our relevance and credibility is completely limited. Once an advocacy approach is attempted, whether it is affective or not, it will undoubtedly permeate the educational community and could potentially do more harm than good (Blakeslee, 2004). It is for this reason that we must carefully consider our approach to advocacy and refrain from using assessment inappropriately for advocacy efforts. By using assessment in the classroom for the appropriate reasons, we can instead show the benefits we provide students. If we as music educators take a different approach by allowing the true benefits and unique ways in which we affect so many lives to resonate with those around us, music as an aesthetic discipline will work to prove its own worth. Ebie (2005) found that students chose to enroll in music and arts programs in order to attain personal growth and satisfaction; in order to learn how to express music, ideas, and feelings to audiences; and to have the opportunity to participate in “spiritual experiences or ways of expressing spiritual

thoughts or feelings" (p. 290). According to Abril and Gault (2008), principals value music programs in the schools due to the ability of such programs to help students be creative and to reach a broad array of personal and educational goals. In corroboration, Weerts and Greenwood (1993) posits that principals appreciate music programs for the opportunities that are provided in terms of self-expression, acquiring performance skills, school spirit, teaching students to learn how to cooperate with one other, and promoting a sense of community and school pride among students. Glenn (1928) describes the musical experience for students by stating, "We must see in public school music as a means of feeding man's need for beauty" (Glenn, 1928, p. 19). According to Gambler (2003), J. Terry Gates states that music benefits individuals in society by helping people to communicate, create, derive meaning from musical experiences, empower emotional experiences, contribute to personal growth, encourage and promote self-discipline and the meeting of personal goals, embracing diversity, promoting a sense of community and helping to advance cultural values within communities. Gambler (2003) also quotes Jensen (2001) in a way that perhaps best summarizes the benefits of students in music programs: "the arts enable students to focus on the things that matter most in the world: order, integrity, thinking skills, a sense of wonder, truth, flexibility, fairness, dignity, contribution, justice, creativity, and cooperation" (Gambler, 2003, p. 12). This list indicates many benefits that may result when students participate in music. However, these benefits are unquantifiable, are indefinable, and are too numerous to try and list in the scope of this (or any other) single document. It is for this reason that we should not attempt to define any one aspect or benefit that results from students being exposed to an education in music. "Advocacy" for music education, then, can best be accomplished by discontinuing our efforts to verbally advocate for our discipline, but to instead use a new approach that we might refer to as "silent advocacy." "Silent Advocacy" involves

maintaining our integrity as a discipline while ensuring that valuable learning is taking place in the classroom each day; by providing performances in surrounding communities; by exposing administrators, parents, and community members to the successes of our programs; by keeping parents and students happy; and by ensuring that students feel successful while developing a pride for their learning and level of performance, both as individuals and as an integral part of an ensemble. By focusing on what we are doing in the classroom and allowing that work to pervade to other classrooms in the school, to the principal's office, to our colleagues' environments, and to surrounding communities, music education will thrive. If we maintain our focus on the development of student learning through formative and summative assessment, improvement and success will permeate our classrooms and communities. Our overarching mission should be to leave our classrooms and surrounding communities better than we found them when we first arrived. As long as we use assessment to improve the conditions within the music discipline, the larger aspects of credibility and application to society will follow. Hope and Wait (2013) best summarize this by stating, "the most meaningful improvement comes from within a discipline, not from outside it…it seeks not to measure, but to make better" (p.11).

References

Abbott, C. (2013). The “Race to the Top” and the inevitable fall to the bottom: How the principles of the “Campaign for Fiscal Equity” and economic integration can help close the achievement gap. *Brigham Young University Education and Law Journal*, 93–123.

Abril, C. R., & Gault, B. M. (2008). The state of music in secondary schools: The principal's perspective. *Journal Of Research In Music Education*, *56*(1), 68–81.

Blakeslee, M. (2004). Assembling the arts education jigsaw. *Arts Education Policy Review, 105*(4), 31.

Brewer, C., Knoeppel, R. C., & Lindle, J. C. (2014). Consequential validity of accountability policy: Public understanding of assessments. *Educational Policy*, *29*(5), 711–745.

Brookhart, S. M. (2013). The public understanding of assessment in educational reform in the United States. *Oxford Review of Education*, *39*(1), 52–71.

Colwell, R. (2003). The status of arts assessment: Examples from music. *Arts Education Policy Review*, *105*(2), 11–18.

Crochet, L. S., & Green, S. K. (2012). Examining progress across time with practical assessments in ensemble settings. *Music Educators Journal*, *98*(3), 49–54. doi:10.1177/0027432111435276

Ebie, B. D. (2005). An investigation of secondary school students' self-reported reasons for participation in extracurricular musical and athletic activities. *Research And Issues In Music Education*, *3*(1),

Fisher, R. (2008). Debating assessment in music education. *Research And Issues In Music Education*, 6(1),

Gambler, M. B. (2003). *The importance of music education and reasons why administrators should develop curriculum, schedules, budgets, and staffing to meet the needs of the music program and its students* (Order No. EP21295). Available from ProQuest Dissertations & Theses A&I. (305265072). Retrieved from http://search.proquest.com/docview/305265072?accountid=14537

Glenn, M. (1928). The school administrator and the music program. *Music Supervisors' Journal*, (2). 11.

Hale, C. L., & Green, S. K. (2009). Six key principles for music assessment. *Music Educators Journal*, (4). 27.

Hill, B. L. (2014). A call to congress: amend education legislation and ensure that president Obama's "Race to the Top" leaves no child behind. *Houston Law Review*, *51*(4), 1177–1205.

Hope, S., & Wait, M. (2013). Assessment on our own terms. *Arts Education Policy Review*, *114*(1), 2–12.

Jensen, E. (2001). *Arts with the brain in mind*. Alexandria, Va.: Association for Supervision and Curriculum Development, c2001.

Messick, S. (1989). Validity. In R.L. Linn (Eds.), *Educational Measurement*, 3rd ed. (pp. 13–103). New York, NY, England: Macmillan Publishing Co, Inc.

No Child Left Behind (2001). Act of 2001, Pub. L. No. 107-110, § 115. *Stat*, *1425*, 107–110.

Orzolek, D. C. (2006). The paradox of assessment: Assessment as paradox. *Research And Issues In Music Education*, *4*(1), 1–5.

Race to the top: U.S. Department of Education. (2009). Washington, DC U.S. Dept. of Education, [2010].

Scott, S. J. (2012). Rethinking the roles of assessment in music education. *Music Educators Journal*, *98*(3), 31–35.

Weerts, R. K., & Greenwood, R.A. (1993). Dissertation reviews: Secondary school administrators' attitudes and perceptions on the role of music and school bands. *Bulletin Of The Council For Research In Music Education*, (118), 52–54.

Welenc, J. (2010). An ugly truth about music advocacy. *Triad*, *78*(1), 123–125.

Wesolowski, B. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal*, *101*(1), 77. doi:10.1177/0027432114540475

CHAPTER 3

EVALUATION OF A STRING PERFORMANCE RATING SCALE USING THE RASCH MEASUREMENT MODEL[2]

As the educational setting becomes more data driven, valid, reliable, and fair empirical evidence is needed to demonstrate growth in student achievement (Brookhart, 2013). Additionally, the trend of measuring student achievement using valid and reliable empirical data continues to become more prominent with continued focus on teacher effectiveness (Brookhart, 2013). In the field of music, however, the resulting data that is used for these purposes is often psychometrically flawed and therefore misleading due to the current misalignment between instructional focus and corresponding assessment methods (Colwell, 2003).

The results from evaluations in academic classroom settings (such as math and science) are typically student interaction measures that can be enumerated easily through the use of cognitive tests including multiple-choice, true/false, and other selected-response-type examinations (Blakeslee, 2004). In contrast, assessment results in music performance settings are often given in the form of a qualitative narrative that addresses how well parts of a whole work together to contribute to the final performance (Hope and Wait, 2013). Assessment measures in a music classroom are best achieved through the use of rater-mediated evaluations that allow for such narrative and critique to be shared (Wesolowski, 2012). Most often, in order to cater to the data-driven focus in non-performance-based classrooms, the approach implemented by music

[2] Edwards, K., Edwards, A., Wesolowski, B. (in press). Evaluation of a string performance rating scale using the Rasch Measurement Model. *Bulletin of the Council for Research in Music Education*.

teachers is the use of selected-response testing similar to academic classrooms (Hope and Wait, 2013). These methods allow for only one clear answer, thereby allowing empirical data to be provided more easily. Assessment practices should instead be considerate of authentic behaviors in the context of music teaching and learning, where students demonstrate performance-based tasks that are relevant to the content area, and the manner in which the content is being delivered (Zaleski, 2014).

Regardless, music educators must present empirical evidence to document levels of student achievement despite the performance-based nature of music (Hope and Wait, 2013). In the music classroom, the instructional focus requires spending time to develop performance skills while also developing the ability to decipher how to execute and make artistic decisions based on those performance skills (Hope and Wait, 2013). Evidence of student learning therefore needs to be gathered by assessing those performance knowledge and skills that are being taught (Brewer et al., 2014).

Examples of current music performance evaluations that are used to evaluate student performance achievement include juries, auditions, chair placements, large ensemble performance competitions, or community and public performances (Hope and Wait, 2013). One particular challenge arises in the use of student performance opportunities for empirical evidence of achievement due to the fact that such observations of student performance are rater mediated (Wesolowski et al., 2015). The rater interaction in such situations allow for multiple perspectives and opinions to taint the possibility of completely consistent feedback (Hope and Wait, 2013). Unlike selected response standardized assessments used in more academic type classrooms, validity issues in music assessment practices occur because these measures are not indicative of direct student interaction, but rater interaction instead. Thus, even if the performer and

performance remains constant, a different evaluator may very easily yield entirely different results each time. The nature of the performance-based, rater-mediated assessment unfortunately allows for subjectivity to occur on the part of the rater (Wesolowski et al., 2016a). For this reason, the field of music education must continue working to develop valid, reliable, fair, and more importantly, authentic, performance evaluation tools in order to replace the current measures that are being used (McMillan, 2003). The purpose of this study is to help meet this immediate need by developing a performance measure that is authentic to the teaching and learning in the classroom, and that has been tested for validity and reliability the same way that high stakes student interaction measures (such as open response writing assessments) are validated. The development of such an assessment measure that can be used in the performance-based music classroom will require empirically investigating the process involved in developing measurement tools that can be used both in the classroom and in evaluating large ensemble performance evaluation.

The use of rubrics in music teaching and learning

One solution to the misalignment between teaching and assessment practices in the music profession can be achieved through the use of valid and reliable rubrics, as rubrics can accurately account for multiple technical and expressive aspects of music performance (DeLuca and Bolden, 2014). An understanding of what a rubric is and why it is an effective measure of music performance is required in order to entertain a discussion on assessment practices in the music discipline. According to Asmus (1999), rubrics can be defined as "a set of scoring criteria used to determine value of a student's performance on assigned tasks; the criteria are written so students are able to learn what must be done to improve their performance in the future" (p. 21). The information and criteria presented in a rubric helps to ensure that both the student and teacher are

informed as to the direction and expectations that will materialize in the classroom, thereby providing a better established communication link between the teacher and student (Whitcomb, 1999). Rubrics also provide a method for documentation of student achievement levels wherein specific written feedback can be shared with teachers, parents, and students in order to provide evidence of student performance achievement in the music classroom (Wesolowski, 2012). The effective use of rubrics in music performance settings can help students to develop ownership and to take control of their learning efforts (DeLuca and Bolden, 2014).

Research efforts in the music education community have contributed to an increased understanding of how rubrics can improve scoring reliability and consistent grading methods for the music classroom (DeLuca and Bolden, 2014). However, there is still an urgent need for the development of valid and reliable rubrics that can be used to consistently yield results that adequately measure student performance in large ensemble performance evaluation settings, and that can also help to inform preparations for such in the music classroom (Colwell, 2003).

The misalignment between musical performance abilities and resulting assessment data is also part of a larger problem that stems from a lack of teacher training focused on understanding true assessment methods (Colwell, 2003). Parkes (2010) states “Few educators received any formal training in assigning marks to students’ work or in grading students’ performance and achievement” (p. 98). As a result, many music educators do not fully understand the true functions and possibilities that can result from using valid and reliable rubrics in the music classroom (Wesolowski, 2012). The unfortunate reality that further confounds this issue is that teachers are often given the task of developing rubrics with very little training (Pellegrino et al., 2015). This greatly limits the ability of music educators to accurately measure student

achievement in a valid and reliable way that can be used to help inform teaching and learning practices (DeLuca and Bolden, 2014).

Due to the current focus on assessment and accountability at the national level, the notion of developing a system of more authentic measures for the music classroom is gaining thought and attention (Model Cornerstone Assessments, 2015). A change in the way in which the field of music assesses musical behaviors in favor of a more validated approach will allow for music educators to accurately and confidently report evidence of student learning to students, parents, colleagues, and administrators (Pellegrino et al., 2015). Teachers should confirm that the rubrics are valid in that they measure what is intended to be measured, and also that they are reliable in that they provide consistent results throughout the assessment process (Pellegrino et al., 2015).

Future implications of this measurement system are specific to the validation of a new system that can be used by classroom and university music educators to rate string large ensemble performances. Achievement parameters and intended goals of string performance need to be constructed and agreed upon prior to the development and implementation of performance evaluations. This contributes to the idea that rubrics should only be used in the classroom for performance preparations if they are first rigorously tested. This is especially true if the data from these assessment measures is going to be used as a means to infer teacher effectiveness. Messick (1989) refers to this consideration as consequential validity. In this instance, the results that are used from performance evaluations have important implications and social consequences that must be considered prior to implementing such high level forms of assessment.

In order for the results of the rubrics to maintain meaning as a part of the instructional process, focused efforts must be directed toward the development and validation of rubrics (DeLuca and Bolden, 2014). Though there is an increased focus on the use of theoretically

informed rubrics in student performance assessment, there is still a need for the development of empirically supported rubrics (DeLuca and Bolden, 2014). The challenge in using rubrics as an assessment mechanism for string ensemble performance lies in the notion that raters mediate current assessments in music as a mechanism for providing the empirical data. As such, these measures do not take into consideration various rater errors such as severity/leniency and raters' specific use of the rating scale structure. This leads to subjectivity on the part of the rater, thereby providing data that does not accurately measure the true level of student achievement (Wesolowski et al., 2016a, 2106b).

The proposed method therefore favors the consideration of a psychometric approach to developing such rubrics. An advantage of using such an analysis approach is that the metrics allow us to gain a better understanding for, and infer, local independence prior to the use of the developed rubric (Linacre, 2010). There is strong evidence that a measure that is put through rigorous testing outside of the classroom (i.e., in a large ensemble performance evaluation setting) will hold onto its validity when used in the true testing environment (i.e., the performance-based music classroom).

Psychometric considerations

Arguably, the most significant limitation of performance assessments is measurement variance attributed to raters. Rater scores are less associated with the performances themselves and more associated with the perceptive lens of the rater (Brunswik, 1952; Engelhard, 2002). This is true in any rater-mediated situations, however, this specific scenario requires that the nature in which individuals rate string ensemble performance is examined. Traditional methods of evaluating rater behavior in music include consistency and consensus estimates. These methods do not adequately estimate true scores of performances (Wesolowski et al., 2015). More

specifically, raters can consistently over- or underestimate true scores but demonstrate high consistency and consensus estimates. Inferences drawn from such instances can therefore be misleading. In order for rater-mediated assessment processes to be more fair, rater errors such as severity/leniency, need to be investigated as part of the measurement process (Wesolowski et al., 2016a).

The development of a valid performance evaluation therefore requires consideration of the psychometric process in which a rating system can be constructed in a way that will allow for it to be applied to future performance scenarios. The Multifaceted Rasch Partial Credit Measurement Model (Linacre, 1989/1994) was used in this study to investigate the psychometric properties (i.e., validity and reliability) of the original rating scale because of the properties of invariant measurement underscoring the Rasch family of measurement models (Engelhard, 2013). Content validity is considered through the discarding of items that are not considered to be useful in the measurement of string ensemble performance. This notion is specifically referred to as data-model fit when using the Rasch model (Wesolowski et al., 2016a, 2016b). When using the Rasch model, instead of the model of a normal bell curve being mapped to the data, the data is instead compared to an already existing and consistent model. Any data that does not adequately fit the model based upon the properties of invariant measurement is discarded. Data-model fit will be determined by evaluating fit indices for all items.

When using the Rasch model, data-model fit is determined based on the degree to which invariant measurement is met. Adequate fit to the model results when the five requirements for invariant measurement are met (Engelhard and Perkins, 2011): (a) the items must be independent of the persons used for measurement (i.e., person-invariant calibration of items); (b) persons must have a higher probability of success on easy items in comparison to the more difficult items

(i.e., non-crossing person response functions); (c) the persons must be independent of the items used for measurement (i.e., item-invariant calibration of persons); (d) a person who is more able must have a higher probability on succeeding on more difficult items than that of a less able person (i.e., non-crossing person responses); and (e) items must measure a single underlying latent variable (i.e., Engelhard, 2013). When adequate fit to the model is obtained, invariant measurement is achieved.

In contrast to factor analysis methods, when using Rasch, there is no conflict between the observed data and the future use of the model. Due to the independent measurement of raters, performances, and items, the model is sample independent and can therefore be applied to future assessments (Meredith, 1993). This aspect of the Rasch model accounts for reliability. The developed rating scale can be applied to future performances and will yield consistent results because invariant measurement was used.

The partial credit component of the model (Masters, 1982) allows for the additional parameter of rating scale structure to be explored across each item. In the context of music performance assessment, evidence exists that each of the categories (e.g., strongly agree, agree, disagree, strongly disagree) for each item are not equidistant because they vary in difficulty in terms of ability to endorse (Wesolowski, 2016b). The partial credit aspect of the Rasch model allows for the investigation of the difficulty level across each rating scale category. For instance, a strongly agree is harder to achieve than agree. The partial credit allots for this varying degree of difficulty because a higher level of achievement should be earned with strongly agree, as opposed to agree. To investigate the partial credit aspect of the model, consideration of monotonicity (e.g., proper ordering of rating scale categories) between categories, appropriate distinction made between performances, frequency of use by raters, and probability measures

was taken into account in order to determine the most optimized structure for each of the rating scale categories (Linacre, 2002). If the logit measurements show that agree was harder to earn than strongly agree, that would result in a violation of monotonicity. Analysis of the rating data was conducted using the computer program *FACETS* (Linacre, 2014).

In this data-driven educational climate there is a critical need for the development of valid, reliable, and fair measures that can be used to measure student achievement in string ensemble performance settings. Such assessment measures need to also be applied in performance-based music classrooms by providing information that can help to guide instructional decisions that are implemented as a component of preparations for such large ensemble evaluations. The purpose of this study was to develop a valid and reliable rubric for the evaluation of large ensemble string performance. This study was guided by the following research questions:

1. What does Rasch Measurement Analysis reveal about the psychometric quality (i.e., validity and reliability) of items, raters, and ensembles within the context of a large ensemble string performance assessment?
2. How do the items vary in difficulty, raters vary in severity, and ensembles vary in achievement?
3. How does the rating scale structure vary across individual items? (The null hypothesis states that the final items on the rating scale will share identical response structure).

Method

Rater cohort of content experts. Twenty-five content experts participated in this study by agreeing to listen and evaluate four full ensemble orchestra recordings each. Fifteen females and ten males participated in the study, twenty of which attained a Bachelor, Masters, or

Specialist degree and five of which attained their Doctoral degree. Of the twenty-five content experts, fifteen teach in a middle school string setting, and ten teach in a high school string setting. These content experts will benefit from the developed rubric in that the resulting finds can be used to evaluate performances by their programs. Each rater was chosen based on their availability, experience, influence in the field, and willingness to listen to the recordings and rate the performances. The selection of the content experts was based on the assumption that "Best practice in the selection and utilization of adjudicators in the field of music performance suggests that expert teachers and performers offer the best chance for providing a fair and equitable assessment" (Wesolowski et al., 2015, p. 165).

Development of initial item pool. Thirty-eight item stems were extracted from an original item pool previously developed by Zdzinski and Barnes (2002). Zdzinski and Barnes (2002) used some stems from the same item pool to develop an earlier rating scale for string performance. The treatment of the item stems from this pool was different in the original study from which they were extracted because factor analysis was used. When using the factor analysis method, individual characteristics are not independent of one another and therefore the resulting rating scales cannot be applied to future situations.

The descriptive statements were organized into four a priori categories based upon the performing dimension of the National Association for Music Education Model Cornerstone Assessment: (a) tone production, (b) rhythm and pulse accuracy, (c) pitch and intonation accuracy, and (d) expressive qualities/stylistic interpretation (National Association for Music Education, 2015). Three content experts reviewed each of the original item stems in order to evaluate the manner in which the stem was able to accurately describe the music concept. Discussions resulted in the editing and adapting of stems that were not considered to be clear and

appropriate for the study. Agreement was reached in the directionality of the items, resulting in 20 positively phrased items and 18 negatively phrased items. The 38 items were randomized and paired with a four-point Likert-type scale (see Figure 3.3).

Performance stimuli. The content experts evaluated a total 52 recordings from a formal district music performance assessment, in a large southern state, that occurred in the previous year. These recordings included string ensemble performances from both middle school and high school groups of various ability levels. These performances were representative of the population that will benefit from the development of the rubric. All recordings were professionally created and matched in sound quality.

Rater assessment network. An incomplete assessment network was used where a total of fifty-two performances were evaluated. Each rater listened to and evaluated a total of four performances, but two of those performances overlapped with the subsequent rater. The last rater and first rater were overlapped in order to account for all performances (Engelhard, 1997). Performances were randomly assigned to raters and were shared with individual Dropbox links. Raters used a separate randomized Google form in order to submit evaluations of each performance. Item stems were randomly presented on each Google form in order to control for rater fatigue. Once completed, negatively phrased items were reverse coded prior to analysis (See note at the bottom of Table 3.3 for stems that were reverse coded).

Results

Variable map. The variable map is a visual representation of the latent construct (eg., large ensemble string performance) (See Figure 3.1). Each of the facets included in the study are displayed in each of the columns on the variable map. The first column shows the logit scale that serves as a "ruler" in order to allow for the measurement of each facet to be shown on a common

map. The second column shows the performances, notated through the use of an asterisk for each performance. The performances near the top are considered to be the highest achieving performances and those closer to the bottom are the lower achieving performances. The measures ranged from -1.81 logits to 2.56 logits with a demonstrated range of 4.37 logits (M = -0.02, SD = 1.01, N = 50). The third column represents the severity of raters. Severity and leniency ranged from -2.19 logits to 2.07 logits with a demonstrated range of 4.26 logits (M = 0.00, SD = 0.88, N = 25). The raters closest to the bottom of the map are considered to be more lenient and raters closer to the top are considered to be more severe in their measurement practices. The fourth column shows the difficulty to endorse each item. Difficulty ranged from -1.18 logits to 1.68 logits with a demonstrated range of 2.86 logits (M = 0.00, SD = 0.62, N = 38). The items closer to the bottom of the map are considered to be easier to endorse and items closer to the top are considered to be more difficult to endorse. The measurement of these three facets will be used to infer measurement on the latent construct of large ensemble string performance.

The variable map provides a visual representation of the information that is needed to answer the second research question. Psychometric aspects of the model allow for the investigation as to how well the items, raters, and ensembles fit the model. In this particular investigation, any items that did not fit the model were discarded. Items that were considered too easy or too difficult to endorse will not be included in the final validated scale, and will therefore also not be included as a part of the final rubric. The use of the Rasch model allows for any unfit items to be removed in order to aid in the creation of a valid and reliable final rating scale. The final rating scale was then translated into a rubric that will be useable to professionals in the music education setting hoping to rate string ensemble performances to show levels of student achievement.

The following calibration details explain the intricacies of how items were either considered to fit the model, or considered as being misfit. Infit MSE statistics considered to fit the model are within the range of 0.80 and 1.20 logits as indicated by Wright and Linacre (1994) and Engelhard (2009). Measurements below 0.80 are considered to be underfit, and any measurements above 1.20 are considered to be overfit. Underfit and overfit items are thusly considered to be misfit when applied to the model.

Calibration of ensemble performances. The calibration of student performances is provided in Table 3.4. Higher numbers represent higher performance achievement and lower measures represent lower performance achievement. Performance 5 represented the highest performance achievement (2.56 logits) and performance 3 represented the lowest performance achievement (-1.81 logits). Misfitting performances is based upon Infit MSE statistics that fall outside of the ranges of 0.80 and 1.20 logits as indicated by Wright and Linacre (1994) and Engelhard (2009). Over-fitting performances include performances 2, 3, 10, 15, 17, 18, 21, 26, 31, 41, and 47. Under-fitting performances include performances 4, 6, 9, 11, 14, 19, 23, 24, 25, 27, 28, 30, 32, 36, 39, and 45.

Calibration of raters. The calibration of raters is provided in Table 3.5. The table demonstrates a ranking of the raters in terms of severity and leniency. Rater 6 was the most severe (observed average = 1.77, logit measure = 2.07) and rater 9 was the least severe (observed average = 2.96, logit measure = -2.19). Raters 2, 4, 7, 8, 10, 12, 13, 14, 15, 20, 21 and 22 were considered to demonstrate muted patterns with Infit MSE less that 0.80. Raters 1, 5, 9, 16 and 25 were considered to demonstrate sporadic patterns with Infit MSE greater that 1.20. This aspect of the model accounts for rater behaviors, which will provide pertinent information for future rater training.

Calibration of items. The calibration of items is presented in Table 3.2. The calibration of items displays the difficulty of each item. The more difficult items are evident in the larger logit measures, and the easier items are evident in the smaller logit measures. The most difficult item was item 23 (Ensemble performs with consistently good intonation in all registers) (observed average = 1.95, logit measure = 1.68) and the easiest item was item 10 (Tempi are appropriate for style of composition)(observed average = 3.14, logit measure = -1.18). Items that demonstrated overfit included items 2, 10, 15, and 29. Items that demonstrated underfit included items 1, 3, 4, 5, 22, 33, and 34. This provided grounds for removal from the final rating scale, which also meant that these items would not be used in the development of the rubric. Misfit items do not adequately contribute to the rating of string performance evaluation, so they should not be kept in the rubric as a means of defining exemplary string ensemble performance evaluation. Further analysis could provide the opportunity to investigate additional stems that would replace the gaps represented by the removal of these stems.

Summary statistics. Summary statistics are provided in Table 3.1. Analysis indicates significant differences between performances (χ^2=1383.0, $p < .01$), raters (χ^2=1030.1, $p < .01$), and item stems (χ^2 = 542.6, $p < .01$). Good data fit is evident in that the mean square fit values (Infit MSE and Outfit MSE) are close to the expected value of 1.00. Acceptable range for productive parameter-level mean square statistics is between 0.80 and 1.20 according to Wright and Linacre (1994) and Engelhard (2009). Therefore, the reliability of separation for performances ($Rel_{performances}$ = .97), raters (Rel_{raters} = .98), and items (Rel_{items} = .93), shows an adequate amount of separation to confirm the construct validity of the measurement instrument. This might be more clearly understood by saying that there is 97% reliability that this assessment tool distinguishes the level of achievement of each of these performances. Thus, the final rating

scale can be considered valid because the results are independent from the performances, raters, and items used to construct the rating scale. Table 3.1 provides information that can be used to provide the analysis necessary to answer the first research question.

Rating scale category diagnostics. The original thirty-eight item stems were extracted from an original item pool previously developed by Zdzinski and Barnes (2002). Following the study, misfit items were removed from the item pool and the rating scale was closely studied in order to determine the best structure for the remaining items (Linacre, 2002). In order to improve validity of the rating scale, modification of the structuring was made to provide for a more exact description of the performances. This was completed under the assumption that each category is not considered to be equal distance from the previous or subsequent categories. Making such changes will improve the ability and ease associated with the use of the model in future applications, as well as, the validity and reliability. Table 3.3 provides the data that was taken into consideration when collapsing the rating scale structure.

Frequency counts were investigated based on Linacre's (2002) recommendation of 10 uses per category. Any categories with less than 10 uses for certain items were collapsed in order to represent the best possible structure for those specific items and to avoid skewed distribution of item usage. Item 7 (category 1), item 8 (category 4), item 9 (categories 1 and 4), item 11 (categories 1 and 4), item 12 (categories 1 and 4), item 13 (category 4), item 14 (category 1), item 16 (category 1), item 17 (category 1), item 18 (category 4), item 19 (category 1), item 20 (category 4), item 21 (category 4), item 23 (category 4), item 24 (category 4), item 25 (category 4), item 26 (category 4), item 30 (category 1), item 31 (category 1), item 32 (category 1), item 35 (categories 1 and 4), and item 38 (category 1) were collapsed into adjacent categories (based on frequency counts) in order to better serve the rating scale structure. Outfit mean squares (MSE)

were examined for values >/= 2.0 because such values would indicate excessive sporadic measures in the ratings. Items 21 and 25 (category 4) was collapsed into adjacent categories in order to better serve the rating scale structure. Lastly, average observed logit measures were examined for violations of monotonicity. Monotonicity is the continuous advancement of step calibrations (Andrich, 1996). Agreement of monotonicity operates under the assumption that strongly agree is more difficult to endorse than agree and so forth. Therefore, if an item showed a violation in this monotonicity in the difficulty to endorse, the structure was collapsed. Item 10 was the only item that demonstrated violations of monotonicity and was thusly collapsed. Item 10 had already been discarded due to overfit, but if not, this would mean that only disagree or agree options were needed, as opposed to four separate Likert scale categories in the final rating scale. Without a qualitative investigation with the raters, it is hard to determine what might cause this result. Only an assumption can be made because an investigation such as this is outside of the scope of this study. In this study, the quantitative results of the rating scale category structure and optimizing it based upon the analytics is the primary focus. Collapsing this item prior to developing the final rating scale would contribute to the usability of the final rubric in that raters would more easily be able to rate tempi by either disagreeing or agreeing that the tempi were appropriate for the style of the composition.

A finalized version of the String Performance Rating Scale is shown in Figure 3.4. The final rating scale reflects the absence of discarded items, as well as, the modifications made to the structure of each item. These modifications were decided based on frequency of use, outfit measures, and monotonicity. The information provided in Table 3.3 and Figure 3.4 can be used to answer the third research question.

Rubric development and defining performance criteria descriptors. Following the investigation of quantitative results for the rating scale, three content experts engaged in ex post facto qualitative analyses through the development of descriptors for each of the rating scale categories. This step was necessary in order to develop a rubric that would be considered useable in future string large ensemble performance evaluation situations. The resulting rubric will allow for feedback to be shared with performers following the evaluation process and will be easier for raters to use in future evaluation settings.

The content experts provided the expertise needed to develop a rubric that reflected the appropriate wording and descriptions for each item in a way that would be meaningful to the middle school and high school string performance community. Careful consideration was taken to eliminate repetition, and to ensure clarity and precision for each of the items in the rating scale. The original directionality within each item was removed in order to maintain a content-specific and non-directional learning outcome.

Pre-established anchors were used to develop statements for each criterion performance level descriptors (Vagias, 2006). Anchor selection included the categories of quality (item 6), detraction (items 7, 9, 17, 19 and 27), effectiveness (items 8, 16, and 35), appropriateness (item 11), frequency (items 13, 14, 20, 21, 24, 26, 32, 37, and 38), problem (items 18, 23 and 28), acceptability (item 25), influence (item 30), desirability (item 31), and satisfaction (item 36). The finalized rubric is shown in Figure 3.2.

Conclusions and future research

The first research question addresses how Rasch measurement analysis is used to reveal the psychometric quality (the validity and reliability) of the assessment used to evaluate large ensemble string performance. The item stems, raters, and performances were measured

independent of one another and the resulting metrics for each were used to determine which of each of the items, raters, and performances fit the model. This discernment between fit and misfit items, raters, and performances addresses the validity of the rating scale. Individual items that were considered to be misfit (outside of 0.80 and 1.20 logits) were discarded in favor of the items that adhered to the model.

The resulting rating scale is also considered to be reliable because a high reliability of separation is evident for persons, items, and performances. This reliability of separation ($Rel_{performances}$ = .97, Rel_{raters} = .98, and Rel_{items} = .93) increases the confidence that each of the measures accurately separates the facets that were measured. Thus, the high reliability of separation confirms that the characteristics were measured independent of one another within the context of the assessment.

The material was carefully examined in order to verify that the items accurately represented the construct that was being measured and any impeding material was omitted. Three content experts accounted for the validity of the items during initial reading, collection of materials, and discussion of the nature of the items. As a result of the psychometric analysis, 11 of the 38 items were considered misfit based on infit and outfit metrics. The 11 items that were removed only increase the future functionality of the measurement. Any items considered underfitting were too muted and did not provide enough variety in order to be considered valid, whereas the over fitting items were considered to be too sporadic and failed to adequately contribute to the scoring process. Items 6,7, 8, 9, 11, 12, 13,14, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 30, 31, 32, 35, 36, 37, and 38 were considered to be valid because they fit the model (Figure 3.4).

There are small standard errors associated with each rater and item that contributes to the degree in which the newly developed rating scale can be used for further investigation. This also

contributes to evidence of strong precision. The precision and reliability evidence demonstrates that the measure was able to adequately distinguish between high and low performances while using the logit continuum. Further implications include the possibility of predicting the level of difficulty for items prior to data collection due to the fact that the sample is independent from the data to model fit. This particular aspect of the results could help in preparation efforts for ensemble rehearsals in the music classroom.

The second research question address how the items vary in difficulty, the raters in severity, and the ensembles in performance achievement. The variable map shows how each of these facets vary. In terms of items, the more difficult items to endorse are closer to the top, and the easier items to endorse at the bottom (Figure 3.1). Item 23 is considered to be the most difficult item to endorse (ensemble performs with consistently good intonation in all registers), and item 32 is considered to be the easiest item that fit the model (stylistically appropriate articulations). Item 10 cannot be considered the easiest item to endorse because the item was considered misfit (1.47 logits) and was therefore discarded.

Items that did not fit the model were considered invalid and were therefore discarded. Each of the four a priori categories contained item stems that were discarded. Five of the nine item stems from the tone production category were discarded, and each of those five item stems discarded were positively worded (items 1, 2, 3, 4, and 5). Item 2 (Players use sufficient bow weight) was overfit, but the remaining discarded tone production items were underfit. This implies that raters might not be able to adequately describe tone production in a positive manner, or perhaps the raters have unrealistic expectations for what a characteristic tone might sound like in an exemplar string ensemble performance. Replication of this study might allow for more item

stems relating to tone production to be tested in order to hopefully be able to provide more descriptors for this category in the final rating scale.

Two stems from the rhythm and pulse accuracy category were discarded (items 10 and 15). Both of these items were considered overfit, item 10 was positively worded and item 15 was negatively worded. In contrast to the tone production category wherein 50% of the items were discarded, only 20% of the rhythm and pulse accuracy stems were discarded. The remaining items that were discarded were considered underfit. Two items from the pitch and intonation accuracy was discarded (items 22 and 29). Item 22 was positively worded and item 29 was negatively worded. Two items were discarded from the expressive qualities/stylistic interpretation category (items 33 and 34). Both of these items were positively worded. Aside from the five tone production item stems that were discarded, it seems as if the other stems that were discarded could perhaps be hard to hear and discern a true rating with the absence of a score during string performance.

Though modification is a possibility for future studies, the adaptation of the misfit items was outside of the scope of this study. For this reason, any items that did not fit the model were discarded and not included in the final rating scale, or in the rubric. Future replications of the study would allow for the introduction of more stems in order to provide opportunities to further discover the unidimensionality and possible modification of such stems in order to counteract multidimensionality.

The third research question addresses the structure of the rating scale. The researcher investigated the null hypothesis that the original items share an identical response structure. This consideration provided the opportunity to show that the item stems do not all share the same structure. Inconsistencies in terms of monotonicity, frequency of use, and aspects of

predictability provide evidence of the notion that item stems require different levels of answer responses according to the difficulty to endorse. The null hypothesis was rejected. The modifications that were made through collapsing the structure of certain item stems helped to improve the data to model fit. This collapsing also contributes to the overall usability of the rubric. For certain stems, it will suffice for future raters to either agree or disagree, not having to choose between strongly agree and strongly disagree for what are considered to be items that are more dichotomous in nature. Table 3.3 and Figure 3.4 provide information that can be used to investigate the third research question.

The third research question also contributes to the overall usability of the final rubric. The changes in structure contributed to usability in that future raters will be able to better use the rubric in order to evaluate live performances of large string ensembles under specific time constraints that are typically characteristic of these situations. The overwhelming number of 38 original items, each with four levels of rating was greatly reduced through the course of the study, in order to create a valid, reliable, and more usable rubric for the intended application.

The Rasch Partial Credit Measurement Model is ideal for this situation because the data can be treated differently in order to provide an accurate reading as to specific performance levels. Previous scoring practices operate under the assumption that rubric data is considered to be interval-level data, but in this instance, the data received is actually ordinal-level data. The Rasch Measurement Model transforms the data from ordinal to interval data through metrics. Receiving data at a higher level means that the data is more meaningful in terms of how it relates to other and future performances. In developing this model, the data can be seen as counts that are directly correlated to a common measure as opposed to isolated measures that cannot be aligned with the overall relationship between the facets that were measured. This adjustment in

data increases the possibility of the development of a validated assessment measure that should and can ideally be developed in relation to the pre-determined standards for performance.

With its predictive nature, perhaps the most important future implication of a validated rating scale lies in the capabilities of such to be used in training and aligning those practices of individuals who rate performances. Rater severity measures can be used to help raters understand how performances should be evaluated in an effort to adequately align the opinions of those who serve as raters in such performance evaluation situations. The development of the rubric is an important process and can be revealing. However, future replications and uses of the scale itself can be used in order to help develop a more valid, reliable and meaningful rating process for students, teachers, administrators, community members, and teacher candidates. This newly developed rubric can be used to provide valid and reliable empirical evidence of student performance achievement. Specifically speaking, in a performance-based classroom setting that primarily utilizes rater-mediated assessments, this rubric provides a way for music educators to objectively rate and evaluate string ensemble performances.

As data driven efforts continue to progress in the educational setting, the use of such rubrics in music performance evaluation will become a necessity. In moving forward, this particular rubric needs to be retested using a different population, including classroom teachers, in order to confirm that a distinction between expert raters and classroom teachers would yield positive results. As was evident in the pilot test of the Model Cornerstone Assessments, teachers evaluating their own student work and outside expert evaluators evaluating the same work did not demonstrate any form of differential rater functioning (i.e., bias) based upon their population grouping (Model Cornerstone Assessments, 2015). There is a high level of confidence that a continuation of the research and development of this rubric would yield the same positive results

in future retesting. As such, a continued focus is needed to provide such performance assessment measures for the music education community.

This continued development of these assessment measures needs to involve the rigorous and concentrated efforts of content experts, expert raters, psychometric analysts, and influential political and community individuals as well. Once established and agreed upon, content experts and stakeholders could work to establish a valid and reliable assessment system that would provide achievement levels and descriptive forms of feedback for those performances that are either in need of improvement, or should be commended for serving as an adequate model for future exemplary performance status.

The development of a valid rating scale also provides pedagogical speaking points for music teacher training programs. Specifically, the ranking of items in terms of difficulty presents the opportunity for pre-service teachers to engage in meaningful conversations about the components of string performance and which items are going to be the most difficult. Helping teachers to discover and discuss the difficult aspects of rating string performance can help teacher candidates to formulate teaching strategies and approaches that will be effective. Furthermore, assuming a good data to model fit will provide the opportunity for stellar performances to be identified and will therefore supply pre-service teachers and novice teachers with the ability to listen to exemplars as they learn how to achieve the same performance standard. Essentially, a valid and reliable string performance rating scale will help string music educators to better understand what distinguishes a great performance from a mediocre performance, therefore assisting teachers in being able to better align instruction in the classroom with valid and meaningful assessment processes.

References

Andrich, D. (1996). Measurement criteria for choosing among models with graded responses. In Eye, A. V., & Clogg, C. C. (Eds.), *Categorical Variables in Developmental Research* (pp. 3–35). San Diego: Academic Press.

Asmus, E. P. (1999). Music assessment concepts. *Music Educators Journal*, *86*(2), 19–24.

Blakeslee, M. (2004). Assembling the arts education jigsaw. *Arts Education Policy Review, 105*(4), 31.

Brewer, C., Knoeppel, R. C., & Lindle, J. C. (2014). Consequential validity of accountability policy: Public understanding of assessments. *Educational Policy*, *29*(5), 711–745.

Brookhart, S. M. (2013). The public understanding of assessment in educational reform in the United States. *Oxford Review of Education*, *39*(1), 52–71.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago, IL: Chicago University Press.

Colwell, R. (2003). The status of arts assessment: examples from music. *Arts Education Policy Review*, *105*(2), 11–18.

DeLuca, C., & Bolden, B. (2014). Music performance assessment: Exploring three approaches for quality rubric construction. *Music Educators Journal*, *101*(1), 70–76.

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

Engelhard, G. J. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: development, implementation, and analysis* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*(4), 585–602.

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

Hope, S., & Wait, M. (2013). Assessment on our own terms. *Arts Education Policy Review*, *114*(1), 2–12.

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

Linacre, J. M. (1994). *Many-facets Rasch measurement, 2nd ed.* Chicago: MESA Press.

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

Linacre, J. M. (2010). More objections to the Rasch Model. *Rasch Measurement Transactions, 24*(3), 1298–1299.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, *22*(4), 34–43.

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Messick, S. (1989). Validity. In R.L. Linn (Eds.), *Educational Measurement*, 3rd ed. (pp. 13–103). New York, NY, England: Macmillan Publishing Co, Inc.

Model Cornerstone Assessments (2015). National Association for Music Education. Retrieved from http://www.nafme.org/my-classroom/standards/mcas-information-on-taking-part-in-the-field-testing/ October, 22 2015.

National Association for Music Education. (2015) Music model cornerstone assessment: Performing ensemble proficient, (August), 9.

Parkes, K. (2010). Performance assessment: Lessons from performers. *International Journal of Teaching and Learning in Higher Education*, *22*(1), 98–106.

Pellegrino, K., Conway, C. M., & Russell, J. A. (2015). Assessment in performance-based secondary music classes. *Music Educators Journal*, *102*(1), 48–55.

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

Wesolowski, B. C. (2012). Understanding and developing rubrics for music performance assessment. *Music Educators Journal*, *98*(3), 36–42.

Wesolowski, B. C., Wind, S. A., & Engelhard, J. G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147–170.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016a). Rater analyses in music performance assessment : Application of the many facet Rasch model. In *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education* (pp. 335–356).

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

Whitcomb, R. (1999). Writing rubrics for the music classroom. *Music Educators Journal, 85*(6), 26.

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

Zaleski, D. J. (2014). An introduction to classroom assessment for today's music educator. *Illinois Music Educator*, *75*(1), 58.

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, *50*(3), 245.

Tables

Table 3.1

Summary Statistics from the PC-MFR Model

	Facets		
	Performance (θ)	**Rater (λ)**	**Item (δ)**
Measure (Logits)			
Mean	-0.02	0.00	0.00
SD	1.01	0.88	0.62
N	50	25	38
Infit *MSE*			
Mean	0.99	1.00	0.99
SD	0.38	0.44	0.19
Std. Infit *MSE*			
Mean	-0.30	-0.40	-0.10
SD	2.30	3.50	1.30
Outfit *MSE*			
Mean	1.00	1.00	1.00
SD	0.39	0.43	0.21
Std. Outfit *MSE*			
Mean	-0.20	-0.30	0.0
SD	2.30	3.50	1.40
Separation Statistics			
Reliability of Separation	0.97	0.98	0.93
Chi-Square	1383.0*	1030.1*	542.6*
Degrees of Freedom	49	24	37

* $p < 0.01$

Table 3.2

Calibration of the Item Facet

Item Number	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
23	1.95	1.68	0.16	0.84	-1.30	0.89	-0.70
29	1.91	1.18	0.16	1.21	1.40	1.29	1.80
26	2.06	0.98	0.15	0.98	-0.10	1.00	0.00
25	2.11	0.85	0.15	1.16	1.20	1.23	0.60
21	2.14	0.84	0.15	0.80	-1.60	0.84	-1.10
24	2.16	0.81	0.16	0.97	-0.10	1.00	0.00
20	2.28	0.61	0.16	0.87	-1.00	0.87	-1.00
5	2.28	0.54	17.00	0.75	-1.90	0.75	-1.90
8	2.41	0.53	0.17	0.99	0.00	0.98	0.00
33	2.52	0.27	0.10	0.74	-1.90	0.71	-2.10
27	2.36	0.23	0.13	1.07	0.50	1.11	0.80
18	2.44	0.23	0.17	1.10	0.70	1.11	0.70
13	2.46	0.21	0.16	0.91	-0.60	0.96	-0.20
22	2.48	0.20	0.16	0.67	-2.70	0.67	-2.60
4	2.48	0.17	0.16	0.75	-1.90	0.74	-2.00
9	2.49	0.09	0.17	0.94	-0.40	0.94	-0.40
37	2.50	0.09	0.15	1.14	1.00	1.18	1.20
36	2.45	0.06	0.15	0.86	-1.00	0.88	-0.80
6	2.52	0.06	0.15	0.93	-0.50	0.94	-0.40
15	2.58	-0.08	0.18	1.34	2.20	1.39	2.40
11	2.60	-0.18	0.17	1.14	1.00	1.20	1.30
28	2.67	-0.25	0.15	0.97	-0.10	0.94	-0.30
30	2.62	-0.27	0.16	1.11	0.80	1.13	0.90
12	2.60	-0.30	0.17	0.94	-0.30	0.94	-0.30
7	2.65	-0.35	0.17	0.93	-0.50	0.93	-0.40
19	2.71	-0.35	0.17	1.20	1.30	1.23	1.50
1	2.68	-0.36	0.16	0.71	-2.30	0.71	-2.30
14	2.67	-0.38	0.16	1.17	1.20	1.21	1.40
35	2.56	-0.53	0.10	0.92	-0.50	0.94	-0.40
3	2.70	-0.53	0.18	0.77	-1.70	0.73	-1.90
34	2.75	-0.55	0.18	0.77	-1.60	0.74	-1.80
31	2.76	-0.57	0.17	1.07	0.40	1.12	0.80
38	2.77	-0.61	0.17	1.15	1.00	1.12	0.80
16	2.89	-0.72	0.18	1.06	0.40	1.06	0.40
2	2.82	-0.73	0.17	1.38	2.30	1.37	2.20
17	2.87	-0.83	0.17	0.96	-0.20	0.90	-0.60
32	2.85	-0.83	0.17	0.93	-0.40	0.91	-0.60
10	3.14	-1.18	0.17	1.47	2.20	1.54	2.60
Mean	2.52	0.00	0.16	0.99	-0.10	1.00	0.00
SD	0.27	0.62	0.01	0.19	1.30	0.21	1.40

Note. The items are presented in measure order from most difficult to least difficult.

Table 3.3

Rating Scale Structure Analysis

Item	Category Usage (%)				Average Observed Measure (Average Expected Measure)				Outfit *MSE*			
	1	2	3	4	1	2	3	4	1	2	3	4
1	8(8)	30(30)	48(48)	14(14)	-1.36(-1.14)	-.53(-.31)	.68(.60)	2.02(1.72)	0.80	0.70	0.60	0.80
2	4(4)	24(24)	58(58)	14(14)	-.32(-1.01)	.09(-.10)	.88(.85)	1.36(2.03)	1.90	1.30	0.90	1.50
3	4(4)	29(29)	57(57)	10(10)	-1.44(-1.16)	-.44(-.23)	.82(.76)	2.31(1.95)	0.80	0.70	0.70	0.80
4	11(11)	37(37)	45(45)	7(7)	-1.98(-1.53)	-.71(-.69)	.40(.30)	1.59(1.42)	0.60	0.70	0.80	0.90
5	13(13)	51(51)	31(31)	5(5)	-2.12(-1.78)	-.89(-.87)	.27(.18)	1.73(1.23)	0.80	0.80	0.80	0.60
†6	13(13)	32(32)	45(45)	10(10)	-1.65(-1.40)	-.57(-.60)	.46(.33)	1.08(1.43)	0.70	1.10	0.70	1.30
†7	6(6)	34(34)	49(49)	11(11)	-1.83(-1.21)	-.28(-.32)	.75(.64)	1.51(1.79)	0.60	1.00	0.90	1.20
†8	11(11)	40(40)	46(46)	3(3)	-2.14(-1.88)	-.82(-1.00)	-.10(.05)	2.04(1.19)	0.80	1.10	1.20	0.70
†9	9(9)	40(40)	44(44)	7(7)	-1.48(-1.50)	-.64(-.62)	.32(.38)	1.92(1.50)	1.10	1.10	0.90	0.80
10	5(5)	4(4)	63(63)	28(28)	.89(-.72)	-0.20* (0.1)	.88(.95)	1.97(2.12)	3.30	1.00	0.80	1.00
11	7(7)	35(35)	49(49)	9(9)	-.87(-1.33)	-.32(-.45)	.30(.53)	2.08(1.68)	1.80	1.20	1.10	1.00
12	5(5)	39(39)	47(47)	9(9)	-.82(-1.27)	-.51(-.34)	.74(.66)	1.86(1.81)	1.20	0.80	0.90	0.90
13	12(12)	37(37)	44(44)	7(7)	-1.54(-1.55)	-.78(-.71)	.29(.27)	1.54(1.38)	1.30	0.80	0.80	0.90
14	8(8)	32(32)	45(45)	15(15)	-1.06(-1.11)	-.11(-.28)	.53(.62)	1.57(1.72)	1.00	1.70	0.90	1.20
†15	6(6)	36(36)	52(52)	6(6)	-1.05(-1.47)	-.20(-.56)	.16(.47)	1.79(1.65)	1.30	1.60	1.60	0.90
†16	5(5)	15(15)	66(66)	14(14)	-1.06(-1.04)	-.08(-.18)	.78(.76)	1.78(1.98)	1.30	1.10	0.90	1.00
†17	4(4)	21(21)	59(59)	16(16)	-1.33(-.95)	-.02(-.05)	.89(.88)	2.09(2.06)	0.70	0.90	0.90	1.00
†18	10(10)	42(42)	42(42)	6(6)	-1.48(-1.60)	-.63(-.72)	.16(.30)	1.59(1.41)	1.10	1.20	1.10	0.90
†19	7(7)	26(26)	56(56)	11(11)	-1.18(-1.22)	.06(-.37)	.33(.57)	1.93(1.75)	1.30	1.70	1.10	0.90
20	17(17)	43(43)	35(35)	5(5)	-1.74(-1.80)	-1.08(-.95)	.13(.06)	1.67(1.12)	1.00	0.90	0.80	0.70
21	30(30)	32(32)	32(32)	6(6)	-1.97(-1.84)	-.98(-1.03)	-.13(-.10)	1.38(.89)	0.70	1.30	0.80	0.70
22	12(12)	35(35)	46(46)	7(7)	-1.79(-1.54)	-.90(-.71)	.38(.26)	1.94(1.39)	0.70	0.70	0.60	0.70
23	31(31)	44(44)	24(24)	1(1)	-2.66(-2.61)	-1.78(-1.69)	-.42(-.61)	.72(.32)	1.10	0.80	0.70	0.90
24	22(22)	45(45)	28(28)	5(5)	-2.07(-1.89)	-.95(-1.02)	.12(-.01)	.36(.99)	0.80	1.10	0.70	2.00
†25	27(27)	41(41)	26(26)	6(6)	-2.02(-1.86)	-.78(-1.01)	.01(-.04)	-0.141	0.80	1.40	0.80	2.70
†26	28(28)	43(43)	24(24)	5(5)	-2.10(-1.96)	-.92(-1.09)	-.26(-.09)	1.03(.87)	0.80	1.20	1.00	1.10

†27	25(25)	30(30)	29(29)	16(16)	-1.50(-1.34)	-.42(-.59)	.45(.23)	.74(1.19)	0.70	1.50	0.60	1.80
†28	10(10)	26(26)	51(51)	13(13)	-1.32(-1.21)	-.30(-.42)	.39(.49)	1.88(1.62)	0.90	1.10	0.90	0.90
†29	31(31)	52(52)	12(12)	5(5)	-2.15(-2.08)	-.99(-1.13)	-.02(-.08)	-0.2673	0.90	1.10	1.10	2.80
30	7(7)	35(35)	47(47)	11(11)	-1.04(-1.24)	-.23(-.37)	.41(.59)	1.90(1.72)	1.30	1.30	1.00	0.90
31	5(5)	27(27)	55(55)	13(13)	-.98(-1.09)	-.10(-.20)	.69(.75)	1.93(1.92)	1.40	1.10	1.10	1.00
32	4(4)	24(24)	55(55)	17(17)	-.79(-.92)	-.23(-.02)	.99(.90)	2.04(2.06)	1.20	0.70	0.80	1.00
33	7(7)	37(37)	53(53)	3(3)	-2.29(-1.77)	-.95(-.86)	.29(.20)	2.23(1.40)	0.60	0.70	0.80	0.80
34	4(4)	27(27)	59(59)	10(10)	-1.61(-1.16)	-.41(-.23)	.80(.75)	2.32(1.96)	0.60	0.70	0.90	0.80
†35	2(2)	48(48)	42(42)	8(8)	-1.57(-1.14)	-.04(-.08)	.85(.97)	2.58(2.11)	0.80	1.00	1.10	0.70
†36	12(12)	42(42)	35(35)	11(11)	-1.64(-1.38)	-.48(-.53)	.34(.41)	1.75(1.46)	0.80	1.20	0.80	0.80
†37	15(15)	31(31)	43(43)	11(11)	-1.2(-1.39)	-.51(-.60)	.14(.31)	1.51(1.39)	1.40	1.30	1.00	0.90
†38	5(5)	27(27)	54(54)	14(14)	-.72(-1.05)	-.15(-.17)	.78(.77)	1.74(1.93)	1.40	1.	0.80	1.2

Note. Category 1 = “strongly disagree;” Category 2 = “disagree;” Category 3 = “agree;” Category 4 = “strongly agree”

† Category 1 = “strongly agree;” Category 2 = “agree;” Category 3 = “disagree;” Category 4 = “strongly disagree”

* Violation of monotonicity

Table 3.4

Calibration of the Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
5	3.26	2.56	0.20	0.92	-0.40	0.92	-0.40
14	3.18	2.01	0.20	1.18	1.10	1.20	1.20
50	2.95	1.96	0.19	1.00	0.00	1.06	0.40
39	3.07	1.29	0.20	0.70	-1.90	0.69	-2.00
45	2.92	1.26	0.19	0.53	-3.10	0.53	-3.10
49	3.04	1.25	0.20	1.08	0.50	1.10	0.60
41	2.99	1.10	0.20	1.79	3.70	1.88	4.10
34	2.96	1.03	0.20	0.88	-0.60	0.93	-0.30
35	3.03	1.02	0.20	0.90	-0.60	0.95	-0.30
19	2.91	0.90	0.19	0.46	-3.90	0.45	-4.00
48	2.93	0.82	0.19	1.00	0.00	1.11	0.60
23	2.88	0.76	0.19	0.34	-5.10	0.34	-5.00
18	3.07	0.67	0.20	1.73	3.70	1.88	4.40
42	2.84	0.66	0.19	0.88	-0.60	0.83	-0.90
25	2.80	0.61	0.19	0.41	-4.40	0.45	-4.00
47	3.05	0.58	0.20	1.45	2.50	1.47	2.60
31	3.03	0.56	0.20	2.02	4.90	2.08	5.20
29	2.47	0.44	0.18	0.86	-0.80	0.86	-0.80
13	2.72	0.41	0.19	0.84	-0.90	0.91	-0.40
14	2.33	0.23	0.19	0.56	-3.20	0.55	-3.20
28	2.47	0.18	0.18	0.69	-2.10	0.72	-1.80
38	2.64	0.15	0.19	1.04	0.20	1.09	0.50
26	2.42	0.05	0.18	1.26	1.50	1.19	1.10
24	2.50	0.01	0.18	0.66	-2.30	0.66	-2.20
46	2.64	-0.04	0.19	1.11	0.70	1.00	0.60
6	2.47	-0.11	0.18	0.79	-1.30	0.76	-1.50
15	2.24	-0.36	0.18	1.87	4.50	1.78	4.10
4	2.36	-0.37	0.18	0.73	-1.80	0.69	-2.10
22	2.30	-0.40	0.18	0.87	-0.70	0.87	-0.80
17	2.04	-0.47	0.19	1.33	1.90	1.45	2.40
9	2.45	-0.49	0.18	0.77	-1.50	0.74	-1.70
36	2.32	-0.50	0.18	0.61	-2.90	0.60	-2.90
21	2.18	-0.53	0.18	1.69	3.80	1.58	3.20
20	2.08	-0.54	0.18	1.04	0.30	1.01	0.00
32	2.34	-0.63	0.18	0.71	-2.00	0.73	-1.80
37	2.18	-0.65	0.18	0.87	-0.80	0.89	-0.60
10	2.12	-0.65	0.18	1.74	4.00	1.69	3.70
7	2.32	-0.71	0.18	0.86	-0.90	0.93	-0.40
12	2.18	-0.79	0.18	1.11	0.70	1.09	0.60
11	2.26	-0.85	0.19	0.64	-2.50	0.60	-2.70
43	2.14	-1.05	0.18	0.95	-0.20	1.01	0.10

33	1.95	-1.11	0.18	0.85	-1.00	0.85	-0.90
2	2.03	-1.14	0.18	1.27	1.70	1.32	2.00
30	2.12	-1.17	0.18	0.60	-3.10	0.63	-2.80
8	2.11	-1.19	0.18	0.89	-0.70	0.90	-0.60
27	1.76	-1.39	0.18	0.77	-1.60	0.81	-1.30
16	2.30	-1.45	0.19	1.02	0.10	1.03	0.20
1	1.70	-1.56	0.21	1.07	0.40	0.99	0.00
40	2.61	-1.64	0.19	0.89	-0.60	0.90	-0.50
3	2.54	-1.81	0.18	1.41	2.20	1.46	2.40
Mean	2.52	-0.02	0.19	0.99	-0.30	1.00	-0.20
SD	0.39	1.01	0.01	0.38	2.30	0.39	2.30

Note. The performances are presented in measure order from highest achievement to lowest achievement.

Table 3.5

Calibration of the Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
6	1.77	2.07	0.14	0.83	-1.50	0.87	-0.90
11	2.15	1.08	0.13	1.09	0.80	1.05	0.40
24	2.39	1.00	0.13	0.86	-1.20	0.87	-1.10
16	2.02	0.93	0.13	2.24	8.80	2.50	8.20
19	2.24	0.76	0.13	0.90	-0.90	0.91	-0.70
8	2.46	0.39	0.13	0.66	-3.20	0.65	-3.30
18	2.38	0.34	0.13	0.97	-0.20	0.97	-0.20
22	2.46	0.34	0.13	0.77	-2.10	0.77	-2.00
4	2.51	0.24	0.13	0.72	-2.60	0.76	-2.10
21	2.46	0.21	0.13	0.76	-2.30	0.73	-2.50
1	2.49	0.16	0.13	1.26	2.10	1.37	2.80
12	2.30	0.15	0.13	0.64	-3.00	0.65	-3.50
2	2.36	0.13	0.13	0.75	-2.40	0.76	-2.30
20	2.39	0.01	0.13	0.53	-5.10	0.54	-4.80
14	2.80	-0.08	0.13	0.73	-2.40	0.73	-2.40
17	2.37	-0.08	0.13	1.11	1.00	1.07	0.60
15	2.75	-0.16	0.13	0.49	-5.30	0.47	-5.50
3	2.58	-0.31	0.13	1.10	0.80	1.11	0.90
13	2.54	-0.35	0.13	0.79	-1.90	0.86	-1.20
5	2.93	-0.39	0.14	1.80	5.60	1.85	5.90
7	2.76	-0.54	0.13	0.67	-3.10	0.69	-2.90
23	2.88	-0.56	0.14	1.03	0.20	1.06	0.50
10	2.80	-1.45	0.13	0.73	-2.50	0.75	-2.40
25	3.34	-1.72	0.15	1.89	6.90	1.87	6.80
9	2.96	-2.19	0.14	1.58	4.10	1.62	4.40
Mean	2.52	0.00	0.13	1.00	-0.40	1.00	-0.30
SD	0.32	0.88	0.00	0.44	3.50	0.43	3.50

Note. The raters are presented in measure order from most severe to least severe.

Figures

Figure 3.1 – Variable Map

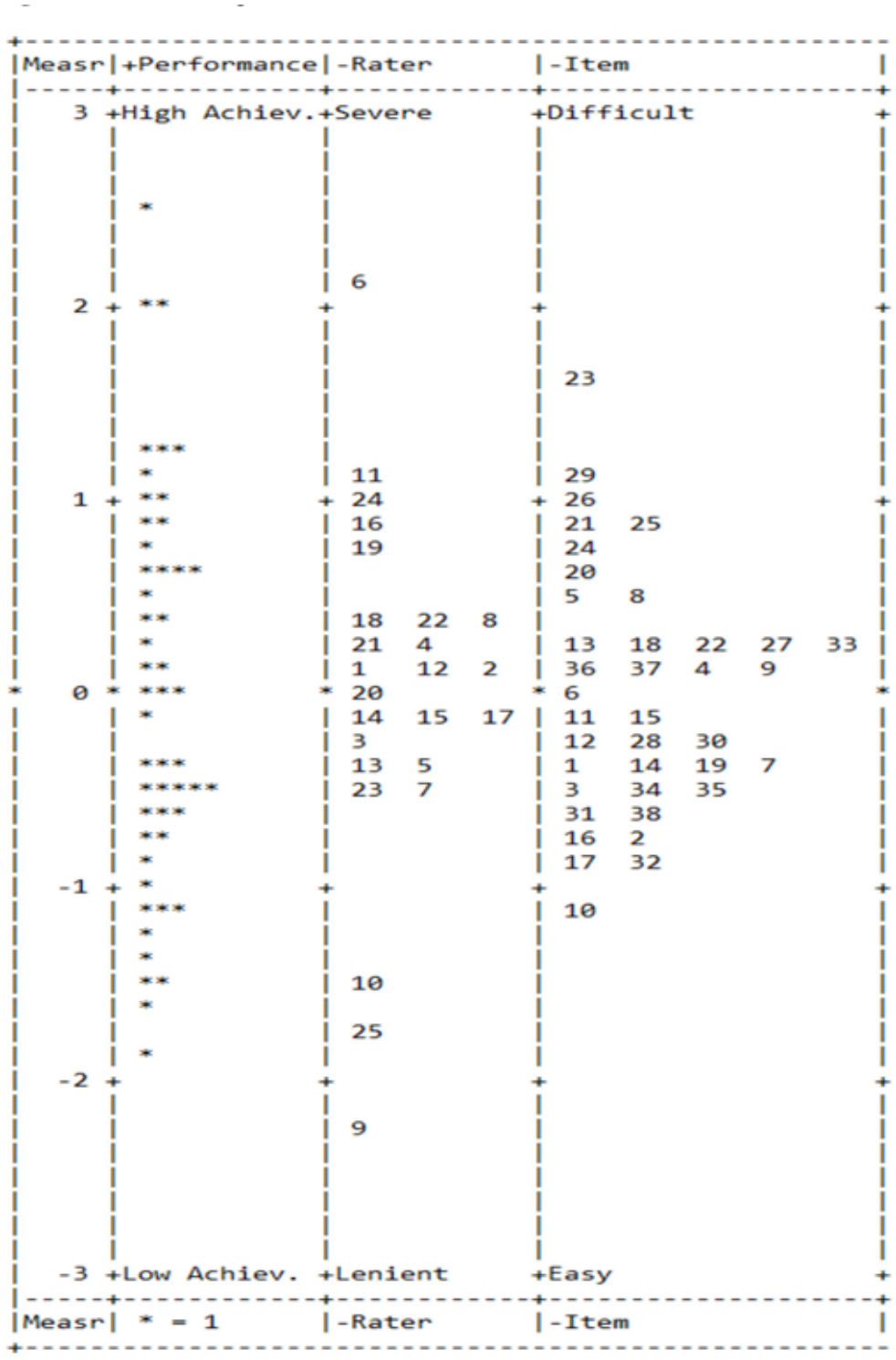


Figure 3.2 – Music Performance Rubric for String Orchestra

Tone Production				
6. Tone quality in varying registers	Tone quality is poor	Tone quality is fair	Tone quality is good	Tone quality is very good
7. Consistency of attacks	Unclear attacks always detract from performance	Unclear attacks sometimes detract from the performance	Unclear attacks never detract from the performance	
8. Tone while executing expressive gestures	The execution of expressive gestures has a major negative effect on tone quality	The execution of expressive gestures has a moderate negative effect on tone quality	The execution of expressive gestures does not have a negative effect on tone quality	
9. Consistency of tone across sections	Tone quality across sections detracts very much from the performance	Tone quality across sections detracts very little from the performance		
Rhythm and Pulse Accuracy				
11. Expressive pulse and tempo fluctuations	Expressive changes in tempo and pulse are inappropriate for the style	Expressive changes in tempo and pulse are appropriate for the style		
12. Sustained notes	Notes are not consistently held for full value	Notes are consistently held for full value		
13. Precision of attacks	Attacks are rarely executed with precision across the ensemble	Attacks are sometimes executed with precision across the ensemble	Attacks are consistently executed with precision across the ensemble	
14. Consistency of articulation	Rhythmic articulations are often inconsistent with the style of music and consistently lack ensemble uniformity	Rhythmic articulations are occasionally inconsistent with the style of music and sometimes lack ensemble uniformity	Rhythmic articulations are consistent with style of music and maintain ensemble uniformity	
16. Consistency of rhythmic stress	Rhythmic stress does not effectively communicate proper musical style	Rhythmic stress somewhat effectively communicates proper musical style	Rhythmic stress effectively communicates proper musical style	
17. Steadiness of pulse	A lack of steady pulse detracts much from the continuous flow of the music	Wavering steady pulse sometimes detracts from the continuous flow of the music	Control of steady pulse does not detract from the continuous flow of the music	

18. Appropriateness of tempo in technical passages	Tempo fluctuations during technical passages are a serious problem	Tempo fluctuations during technical passages are a moderate problem	Tempo fluctuations during technical passages are not at all a problem
19. Subdivision of the rhythm	Inaccurate performance of subdivisions frequently detracts from solidly communicated tempo and meter	Inaccurate performance of subdivisions occasionally detracts from solidly communicated tempo and meter	Accurate performance of subdivisions contribute to solidly communicated tempo and meter
Intonation Accuracy			
20. Intonation of cadential points	Cadential points are rarely in tune	Cadential points are occasionally in tune	Cadential points are consistently in tune
21. Centered pitch	The pitch is rarely centered	The pitch is occasionally centered	The pitch is centered a great deal of the time
23. Overall intonation accuracy	Maintaining consistently good intonation in all registers is a serious problem during performance	Maintaining consistently good intonation in all registers is a moderate problem during performance	Maintaining consistently good intonation in all registers is not a problem during performance
24. Pitch adjustments	It is rarely evident that players are able to accurately and quickly adjust pitch when necessary	It is sometimes evident that players are able to accurately and quickly adjust pitch when necessary	It is frequently evident that players are able to accurately and quickly adjust pitch when necessary
25. Half step intonation	Half step intonation is unacceptable	Half step intonation is slightly unacceptable	Half step intonation is perfectly acceptable
26. Chromatic alterations intonation	Chromatic alterations are rarely in tune	Chromatic alterations are sometimes in tune	Chromatic alterations are consistently in tune
27. Presence of wrong notes	Wrong notes detract from the performance a great deal	Wrong notes occasionally detract from the performance	Wrong notes do not detract from the performance
28. Open string intonation	Out of tune open strings is a serious problem	Out of tune open strings is a moderate problem	Out of tune open strings is a minor problem	Out of tune open strings is not at all a problem
29. Intonation in technical passages	Intonation fluctuations during technical passages are a serious problem	Intonation fluctuations during technical passages are a moderate problem	Intonation fluctuations during technical passages are not at all a problem

Expressive Qualities/Stylistic Interpretation				
30. Presence of crescendo and diminuendo	Crescendo and diminuendo are not at all influential on effective expression		Crescendo and diminuendo are somewhat influential on effective expression	Crescendo and diminuendo are extremely influential on effective expression
31. Balance between melody and accompaniment	Balance between melody and accompaniment is undesirable		Balance between melody and accompaniment is desirable	Balance between melody and accompaniment is very desirable
32. Stylistically appropriate articulations	Stylistically appropriate articulations are never evident		Stylistically appropriate articulations are sometimes evident	Stylistically appropriate articulations are always evident
35. Connection of phrases	Ensemble does not meaningfully connect phrases		Ensemble meaningfully connects phrases	
36. Articulation	Articulations are inconsistent in passages with notes of a similar style, resulting in a very dissatisfactory performance	Articulations are often inconsistent in passages with notes of a similar style, resulting in a dissatisfactory performance	Articulations are sometimes inconsistent in passages with notes of a similar style, resulting in a satisfactory performance	Articulations are consistent in passages with notes of a similar style, resulting in a highly satisfactory performance
37. Contrast in dynamics	Dynamic contrasts are never evident	Dynamic contrasts are almost never evident	Dynamic contrasts are sometimes evident	Dynamic contrasts are frequently evident
38. Expressive modifications (>, sfz., rit., ten., cantabile)	Stylistic or expressive modifications are rarely appropriate or present in performance		Stylistic or expressive modifications are typically appropriate and somewhat present in performance	Stylistic or expressive modifications are appropriate and consistently present in performance.

Figure 3.3

Original 38-Item Likert-Type Rating Scale

1. Characteristic tone is used throughout performance	SD D A SA
2. Players use sufficient bow weight	SD D A SA
3. Appropriate control of bow speed is evident	SD D A SA
4. Players use even bow strokes	SD D A SA
5. Tone quality is consistently rich in all registers	SD D A SA
6. Tone is fuzzy	SD D A SA
7. Attacks are not clearly defined	SD D A SA
8. Tone is compromised while executing expressive gestures	SD D A SA
9. Inconsistent tone across sections	SD D A SA
10. Tempi are appropriate for style of composition	SD D A SA
11. Tempo fluctuations are stylistically characteristic	SD D A SA
12. Full value is given to sustained notes	SD D A SA
13. Attacks are executed with precision across the ensemble	SD D A SA
14. Rhythms are clearly articulated across ensemble	SD D A SA
15. Ensemble tends to rush	SD D A SA
16. Rhythmic stress of strong and weak beats are uncharacteristic	SD D A SA
17. Steady pulse is unclear in performance	SD D A SA
18. Tempo fluctuates during technical passages	SD D A SA
19. Rhythmic figures are subdivided inaccurately	SD D A SA
20. Cadential points are in tune	SD D A SA
21. Pitch is centered	SD D A SA
22. Key signatures are accurately performed	SD D A SA
23. Ensemble performs with consistently good intonation in all registers	SD D A SA
24. Players are able to accurately and quickly adjust pitch when necessary	SD D A SA
25. Half steps are not performed with accurate intonation	SD D A SA
26. Chromatic alterations are not in tune	SD D A SA
27. Ensemble performs enough wrong notes to detract from performance	SD D A SA
28. Open strings are not tuned properly	SD D A SA
29. Intonation fluctuates during technical passages	SD D A SA
30. Crescendo and diminuendo contribute to effective expression	SD D A SA
31. Balance is well maintained between the melody and accompaniment	SD D A SA
32. Stylistically appropriate articulations	SD D A SA
33. Individual phrases are well controlled	SD D A SA
34. Appropriate inflection at cadential points	SD D A SA
35. Inconsistent connection of phrases	SD D A SA
36. Articulations lack consistency in performance	SD D A SA
37. Dynamic contrasts are insufficient	SD D A SA
38. Stylistic or expressive modifications (such as >, sfz, rit., ten., cantabile) are not evident in performance	SD D A SA

Figure 3.4

Final Rating Scale

6	Tone is fuzzy	Strongly Disagree	Disagree	Agree	Strongly Agree
7	Attacks are not clearly defined	Strongly Disagree	Disagree	Agree	
8	Tone is compromised while executing expressive gestures	Disagree		Agree	Strongly Agree
9	Inconsistent tone across sections	Disagree		Agree	
11	Tempo fluctuations are stylistically characteristic	Disagree		Agree	
12	Full value is given to sustained notes	Disagree		Agree	
13	Attacks are executed with precision across the ensemble	Strongly Disagree	Disagree	Agree	
14	Rhythms are clearly articulated across ensemble	Disagree		Agree	Strongly Agree
16	Rhythmic stress of strong and weak beats are uncharacteristic	Strongly Disagree	Disagree	Agree	
17	Steady pulse is unclear in performance	Strongly Disagree	Disagree	Agree	
18	Tempo fluctuates during technical passages	Disagree		Agree	Strongly Agree
19	Rhythmic figures are subdivided inaccurately	Strongly Disagree	Disagree	Agree	
20	Cadential points are in tune	Strongly Disagree	Disagree	Agree	
21	Pitch is centered	Strongly Disagree	Disagree	Agree	
23	Ensemble performs with consistently good intonation in all registers	Strongly Disagree	Disagree	Agree	
24	Players are able to accurately and quickly adjust pitch when necessary	Strongly Disagree	Disagree	Agree	
25	Half steps are not performed with accurate intonation	Disagree		Agree	Strongly Agree
26	Chromatic alterations are not in tune	Disagree		Agree	Strongly Agree
27	Ensemble performs enough wrong notes to detract from performance	Disagree		Agree	Strongly Agree
28	Open strings are not tuned properly	Strongly Disagree	Disagree	Agree	Strongly Agree
29	Intonation fluctuates during technical passages	Disagree		Agree	Strongly Agree
30	Crescendo and diminuendo contribute to effective expression	Disagree		Agree	Strongly Agree
31	Balance is well maintained between melody and accompaniment.	Disagree		Agree	Strongly Agree
32	Stylistically appropriate articulations	Disagree		Agree	Strongly Agree
35	Inconsistent connection of phrases	Disagree		Agree	

36	Articulations lack consistency in performance	Strongly Disagree	Disagree	Agree	Strongly Agree
37	Dynamic contrasts are insufficient	Strongly Disagree	Disagree	Agree	Strongly Agree
38	Stylistic or Expressive modifications (such as >, sfz., rit., ten., cantabile) are not evident in performance.	Strongly Disagree	Disagree	Agree	

CHAPTER 4

PSYCHOMETRIC COMPARISON OF A STRING EVALUATION SYSTEM USING THE MULTIFACETED RASCH PARTIAL CREDIT MEASUREMENT MODEL AND CLASSICAL TEST THEORY

The current educational trend is to implement decisions that are informed by data results derived from a partial focus on student achievement scores (Swan & Mazur, 2011). Since the late 1980s, education policy has primarily focused on the attempt to improve schools through intensifying curriculum standards, focusing on more refined teacher credentials, and empowering other levels of government involvement to make educational decisions (Brewer et al., 2014). Most recently, there is a renewed focus on the implications of student assessment results particularly with how they relate to teacher effectiveness (Bond and Bond, 2010). Using assessment results in a tested subject classroom, such as math or science, is more commonly done than in comparison to a more non-tested subject classroom, such as art or music. Herein lies one of the weaknesses with the assessment driven approach. There is a shortage of measures being implemented that can provide empirically based data that will serve as true evidence of student achievement in the more aesthetic disciplines (Fisher, 2008). A lack of valid and reliable measures therefore means that the claims and decisions that are made in response to such assessment results could be null and void (Brewer et al., 2014).

This lack of focus on correctly implementing assessment measures that will provide valid and empirical evidence means that music educators are currently unable to accurately and precisely communicate levels of student achievement to administrators, community members,

parents, students, policy makers, and other stakeholders (Hope and Wait, 2013). Music educators can attempt to quantify levels of student achievement in their performance-based classrooms, but the documentation of teaching and learning needs to be more on par with standardized testing results. This means that a more focused method of gathering empirical evidence is needed in order to provide data that can show the effectiveness of instructional planning and implementation that is occurring in the classrooms (Bond and Bond, 2010).

In the current educational setting, the reporting of student achievement levels is partly intended for policy makers and community members to have access to numbers that will allow them to gain a general idea of the overall quality of both the teachers and the school (Brewer et al., 2014). This accentuates the need for valid, reliable, and fair assessments that can provide more accurate and precise student achievement data in a music performance classroom. The development of specific measures to rate music performance setting would provide the means by which similar terminology could be used to communicate student performance results as is done with the more commonly tested subjects, such as math or science (Fisher, 2008). Contrary to popular misconception, this does not mean there needs to be a change in how teachers teach or what they teach in music performance based classrooms, rather, there needs to be a method for accurately measuring and communicating the effectiveness of current teaching practices (Hope and Wait, 2013). The development of such assessments is not an easy task and requires the expertise and involvement of musicians, teachers, and statisticians in order to ensure that the assessments that are used are considered to be measures that will actual yield valid and reliable results (Brewer et al., 2014).

Assessment in the classroom should provide the opportunity for students to complete tasks that show what they have learned in a way that is compatible and relevant to the content

area (Zaleski, 2014). The nature of instruction in music classes, specifically speaking about band, choir, and orchestra classrooms, is often performance-based. Therefore, the assessment methods used to determine if valuable learning is taking place in these non-tested subjects should also be performance-based so that the instruction and assessment methods are properly aligned, and will therefore provide methods of authentic assessment (Asmus, 1999). Authentic assessment seeks to gather evidence as to if students are able to perform tasks that will ultimately help them to thrive in real-world situations (Asmus, 1999). The goal of assessment development should be to create assessments that adequately represent the high level of learning that is taking place in the classroom, while also showing that students comprehend what they are learning to a point that they can apply their understanding. In the music classroom, authentic assessment would focus on how students learn to accurately apply music concepts in performance (Asmus, 1999).

Aside from providing opportunities for authentic assessment, there is also a need to determine if measurement is truly achieved from the implementation of developed assessments (Messick, 1989). Unfortunately, the perception is sometimes that if numbers are reported, those numbers are accurate, regardless of the methods of collection and types of analyses that went into reporting that data (Brookhart, 2013). Kaplan (1964) posits, "the validity of a measurement consists in what it is able to accomplish, or more accurately, in what *we* are able to do with it…the basic question is always whether the measures have been so arrived at that they can serve effectively as means to the given end" (p. 198). Measurement is sometimes defined as the assigning of numbers to objects or observations (Wilson, 2005). However, reporting raw data, or counts from assessments, does not mean that anything was measured. Furthermore, applying statistical analyses on such numbers does not necessarily mean that a measure was derived (Bond and Fox, 2015). Classification and assigning numbers to a specific object is only the beginning

because a comparison between two objects must be gleaned in order to actually achieve measurement. In actual measurement, a set of rules must be followed in order to ensure that the values presented remain constant and maintain their meaning along the entire scale. Once that scale is developed and a comparison can be made, measurement is achieved (Bond and Fox, 2015).

Measurement in the music performance setting is, however, more cumbersome due to the non-tested, performance nature of those classrooms. Music performance requires that aesthetic decisions be made in order to improve personal performance. This also includes having the ability and understanding of how to make specific aesthetic decisions and adjustments during performance (Hope and Wait, 2013). Many parts work together as a whole in order to contribute to an exemplary final product. Helping students to reach a point where they adequately understand the concepts behind what they are performing is a necessity in the music classroom. If music educators teach by relating music concepts directly to the performance skills, students will gain a more thorough understanding of the art form and achievement results from an authentic assessment will show that overall level of effectiveness. With that being said, there is a need for instruments that can accurately evaluate student understanding of how the parts fit together to contribute to a final music performance product (Hope and Wait, 2013).

Though there are current music performance assessments being implemented, these approaches could be improved in order to ensure that they are adequately measuring what they intend to measure, in an authentic way. Examples of such include all-state auditions, concert festivals (large ensemble performance evaluations), juries, playing quizzes, and tests in the classroom setting (Zdzinski and Barnes, 2002; Hope and Wait, 2013). Questions of validity and reliability in terms of how well these assessments measure student achievement arise due to the

fact that these are all rater-mediated assessments. Even if judges are considered to be experienced in evaluating performances, variability and inconsistencies can emerge among panels (Hash, 2013). Different performances, raters, circumstances, or other variable conditions can affect or alter results from such measures in many different ways. As mentioned above, the reporting of numbers for the observations does not necessarily guarantee measurement. A comparison must be made on a common scale in order for measurement to actually occur (Bond and Fox, 2015). It is for this reason that results of such performance assessment measures may not currently be agreed upon as having true empirical value or merit (Hope and Wait, 2013).

Because of the commonly used rater-mediated assessments and aesthetic nature of the music discipline, student achievement results may be skewed. The future implementation of assessment measures must be fair, valid, and reliable in order to both satisfy the high stakes testing requirements that are becoming more apparent in the education community, while also holding true to the nature of the discipline. Wesolowski et al. (2016a) outlines four aspects that must be taken into consideration when using performance evaluations as measures for student achievement, "(a) the ability of the ensemble, (b) the difficulty of the task, (c) variability in rater judgments, and (d) the manner in which the rater applies the measurement instrument" (p. 336).

The current implementation of performance evaluation for large ensemble performances involves using a form to facilitate the overall rating of a performance on a scale of I-V, with one being the best and five being the lowest (Hash, 2013). This traditional approach to rating performance evaluations is driven by concepts of Classical Test Theory. Variability in rater judgments can be present because three judges (raters) evaluate student performances using an evaluation form. These raters are not necessarily trained for the event. Each rater is potentially more concerned with his/her personal characteristics or perceptive lens more so than the actual

performance (Engelhard, 2013). Consensus estimates can be calculated, but these numbers may not actually ensure that the raters are compatible in their scoring practices (Engelhard, 2002). Regardless of the above-mentioned issues, results from these evaluations are still used to report achievement levels and to determine the quality of teaching that is taking place in music classrooms. Comparisons from one program to another are also falsely considered based upon the results of these performance ratings (Hash, 2013). It is understandable that administrators and policy makers may not accept results from such events as true evidence of learning because the rater-mediated process is not considered to be a valid or unbiased method of passing judgment on the quality of teaching occurring to prepare ensembles for performance evaluations (Hash, 2013; Wesolowski et al., 2016a).

Large ensemble performance evaluation should be strongly considered as a possible approach for determining if quality teaching and learning is occurring in the performance-based classrooms. However, in order for the evaluation to be able to meet the demands of the data driven focus, certain changes may need to be implemented to improve the process by addressing any possible limitations of the current system. The setting of the assessment does not have to be altered; instead, the method of assessment within that setting can be adjusted in order to provide a valid and reliable method to rate performances.

The purpose of this study was to do a side-by-side comparison of the current large ensemble performance evaluation system using foundations based on CTT, simultaneously with a previously developed and validated rubric that was created using the Multifaceted Partial Credit Rasch Measurement Model (using foundations based on IRT) (Linacre, 1989/1994). The current large ensemble performance evaluation system will be referred to as condition A, and the newly developed rubric will be referred to as condition B. The research questions are as follows:

1. How do the numerical ratings from the condition A rating scale compare to those numerical results yielded from the newly developed condition B rubric?

2. How do the written forms of feedback given to the directors of the ensembles from the two systems compare?

3. How do the two forms compare in terms of overall usability for the raters?

Review of literature

The current data driven educational focus in classrooms is being implemented for two main purposes. The first of these is to improve teaching and learning by extensively collecting student achievement data and focusing on trends represented within that data (Zaleski, 2014). Collecting such data has resulted in the development and implementation of new forms of assessment measures for all classrooms. This has been accomplished easier for tested subjects because as student achievement in these settings can be more closely monitored through the use of written tests (Brookhart, 2013). There are two settings in which written tests do not work as well to provide adequate student achievement data: special education classrooms and the non-tested subject classrooms that have more of an aesthetic focus, such as music and art (Prince et al., 2009). The implementation of newly developed assessments in the general education tested subject classrooms has allowed for the collection of valid and reliable student achievement data in those classrooms due to the nature of the classroom setting. In contrast, certain limitations for aesthetic-based and special education classroom environments have somewhat prevented the collection of empirical evidence for student achievement. Measuring student achievement in these two settings has faced a set back because of the challenges that arise when attempting to measure student performance in a clear, fair, and valid manner (Prince et al., 2009).

The second purpose for pursuing a more data driven approach in education is to fulfill claims of accountability (Brookhart, 2013). Approaches to teacher evaluation have typically been determined by each state government and in the past have merely included a certain number of observations during the school year. With the introduction of more intensified federal legislations, states have begun to try and accommodate political and economical demands by attempting to provide evidence of teacher accountability (Brewer et al., 2014). Brewer et al. (2014) states, "Accountability systems in most states are based on the premise that 'schools and school systems should be held accountable for their contribution to student learning'" (p. 719). Due to the potential high stake nature of accountability claims, the need arises for a more sound method for determining what level of performance either has or has not been met.

One of the primary approaches that has evolved from these demands is to add student achievement results directly to teacher evaluation systems. This aspect of the assessment driven era is intended to determine overall teacher effectiveness by linking levels of student achievement in the hopes of showing the assumed impact of specific teachers that have influence over those particular students (Springer, 2009). Two main issues have risen out of this implementation. The first is that each teacher has a different set of students in his/her classroom and using student achievement results for teachers provides an extra challenge in that some teachers are teaching students with certain circumstances that may or may not impact their ability and resolve to perform well on student achievement measures (Springer, 2009). Unfortunately, there is no real way to determine that teacher influence and curricular decisions are greater than other outside factors that may contribute to levels of student achievement (Brewer et al., 2014). The second issue is that teacher pay may soon be connected to student achievement levels, thus complicating the possible implications (Springer, 2009).

Using assessment results to measure student achievement with the hopes of placing accountability measures on teachers provides for high stakes assessment situations. Messick (1989) refers to this as consequential validity wherein the results of the assessment hold great consequences. Therefore, much care should be taken in order to ensure that what is intended to be measured is actually measured by the assessments that are implemented. Otherwise, the results may profoundly affect certain outcomes in a falsely and highly negative manner.

In music performance situations, a particular challenge arises in terms of consequential validity, or the resulting interpretation from implementation of a particular assessment (Brewer et al., 2014). This can be best explained through entertaining a discussion about methods of measurement. Fundamental measurement (or conjoint measurement) occurs when the item that you are trying to measure can directly be measured because it is tangible or physical (Ferrero et al., 2015). An example of such would be weight, height, or length. Music performance on the other hand, is a latent construct, meaning that it cannot be physically measured as in other assessment situations. This can be defined as derived measurement, which is not considered to be concrete. Derived measurement must be used when the item that you are attempting to measure can only be measured by first measuring other factors that relate to that item, but that actually have observable characteristics (Ferrero et al., 2015). Derived measurement is defined by the interaction of multiple, fundamental measures that are simultaneously measured in order to infer measurement on the larger latent construct. This approach is more commonly used in psychological measurement (Ferrero et al., 2015). Examples of such may include music performance, musical aptitude, or affect (emotional response to music). In this particular instance, music performance is the latent construct that is being measured and in order to

measure music performance, you must first measure observable behaviors such as tone, technique, and intonation.

More specifically speaking, fundamental versus derived measurement involves a difference in how the data is treated for measurement. There are two approaches Classical Test Theory (CTT) and Item Response Theory (IRT) (Gruijter and van der Kamp, 1984). CTT is a true score theory wherein the true score is equal to observed (raw) scores with some error (Wesolowski et al., 2016a). Though it has been the predominant method for developing measurement instruments in the last century, there are limitations to using CTT (Raykov and Marcoulides, 2016). Most importantly, sum scores are used by determining the proportion correct, which means that the results are considered to be sample dependent and cannot be applied to future situations (Raykov and Marcoulides, 2016). On the contrary, IRT uses probabilistic responses based upon the difficulty of the item and the ability of the performer. The probabilistic relationship between item difficulty and performance ability accounts for the limitation of the sum scores being used in CTT, which then means that the results are no longer sample dependent and can be applied to future situations (Rusch et al., 2017).

When considering the development of new assessment measures that will greatly impact the educational world, CTT shows some weakness in terms of meeting the high stakes testing needs in these particular situations. As opposed to using CTT as is currently done in rater-mediated performance evaluation settings, IRT would be better suited to aid in the development of a valid measurement tool because it can account for the limitations of sample dependence (Wesolowski et al., 2016a). The Rasch Measurement Model is a specific IRT model that is considered superior to other methods for measurement because of certain practical and theoretical benefits (Bond and Bond, 2010). Rasch is often the preferred measurement theory for

behavioral, health, and social sciences (Engelhard, 2013). Perhaps the most notable benefit is the invariant measurement aspect of the model (Wesolowski et al., 2016a). If assessment measures are developed using Rasch principles, the resulting rubrics will be sample independent and can be used to accurately measure student achievement levels in large ensemble performance evaluation.

Rasch would be an ideal model to take into consideration for several reasons. The first of these addresses how the data is represented on the measurement model. CTT operates with the idea of model to data fit. The normal bell curve is forced upon the data, which means that the data is sample dependent (Crocker and Algina, 1986). In contrast, the Rasch Measurement Model allows for the data to model fit, wherein the model does not change with each new set of data; rather, any data that does not fit the model is discarded because it does not accurately contribute to the future application of such assessments (Wesolowski et al., 2016a). In addition to the measurement being sample independent, the resulting measurement is unidemensional, meaning it measures one latent construct by preventing conflict from any other factors that might interfere with the purity of measurement (Bond and Fox, 2015). When the data fits the requirements of the model, invariant measurement has been achieved, and perhaps more importantly, rater-invariant measurement of performances can be confidently reported (Engelhard, 2013). CTT directly uses raw scores for analyses, but the Rasch Measurement Model converts those raw scores to log-odds that allows for one dependent variable (latent construct) to be measured with multiple independent variables, or facets (Wesolowski et al., 2016a). CTT deals with discrete data which can lead to faulty assumptions, as opposed to IRT measurement models that convert the data to interval level data, thus allowing for a higher level of statistical testing.

Furthermore, difficulty thresholds are not generally taken into account when measuring item responses using CTT (Bond and Fox, 2015). If analyzing responses received from a Likert-type scale, the assumption is generally that each response is equally difficult to endorse (Wesolowski et al., 2016b). There is fault in this assumption because each response is not equally difficult to endorse. It is harder for a rater to commit to strongly agree as opposed to agree. Using the Rasch Measurement Model means that difficulty thresholds are not equal so credit for achieving certain categories is given the proper weight in the final measurement report (Bond and Bond, 2010). Ability is therefore taken into account by turning raw scores (what would be sum scores in classical test theory) into actual measures. When considering ability in relation to difficulty, more thorough and sound conclusions can be drawn from the data.

The Rasch model can help to compensate for issues and aspects of variance that arise with the implementation of rater-mediated assessment. When considering music performance, rater-mediated assessment calls into question a conflict between true performance versus estimated performance. There is no real way to determine if the raters are rating the performance in a way that accurately measures the actual level of performance. Rater error is one aspect of rater-mediated assessment for which the Rasch model compensates. Rater error refers to the severity or leniency in how raters are considered to either rate too difficult, or too easy. This can also address the tendency for raters to either be muted by supplying the same responses continuously, or sporadic in which the raters have no consistent responses. Such responses can affect the alignment between the true performance and estimated performance results (Engelhard, 2013). Other examples include differential rater functioning (bias) in which rater pre-developed opinions affect how items are used to evaluate performances; rater fatigue in

which rater exhaustion affects how the items are scored; or the halo effect in which no real distinctions are made between performances (Johnson et al., 2009; Engelhard 2013).

The Rasch Measurement Model is set apart from other IRT models because invariant measurement can be achieved. There are five requirements that must be met in order to achieve invariant measurement (Engelhard and Perkins, 2011): (a) the items must be independent of the persons used for measurement (i.e., person-invariant calibration of items); (b) persons must have a higher probability of success on easy items in comparison to the more difficult items (i.e., non-crossing person response functions); (c) the persons must be independent of the items used for measurement (i.e., item-invariant calibration of persons); (d) a person who is more able must have a higher probability on succeeding on more difficult items than that of a less able person (i.e., non-crossing person responses); and (e) items must measure a single underlying latent variable (Engelhard, 2013). Invariant measurement is achieved when the final results are considered to be sample independent. Thus, the developed model can be used in future situations to yield the same results (Hambleton et al., 1991). In a more aesthetic setting, the Rasch model allows for the development of measures that can be used to better evaluate and report evidence of student achievement in the music performance environment.

Rasch allows for the consideration of data to model fit, undimensionality in measurement, interval level data, task difficulty in relation to performance ability, considerations of rater error involved in rater-mediated assessment situations, and invariant measurement. The Rasch model thereby addresses the confounding issues that could potentially impact unreliable measures of student performance achievement (Wesolowski et al., 2016a). This model can be used to develop a rubric to rate such performances. Rubrics allow for common language to be used in order to develop a concrete understanding between teachers and students, teachers and

parents, students and parents, teachers and administrators, administrators and supervisors, parents and administrators, and teachers with other teachers (Pellegrino et al., 2015). Recent research shows that the development and use of rubrics in the classroom could help to improve not only teaching and learning, but also how we communicate evidence of student achievement (Whitcomb, 1999). This allows for evidence of student achievement to be shared equally with those outside of the classroom. Valid and reliable rubrics provide specific qualitative descriptors that can be used to delineate exact information as to what was done well and what needs to be improved in performance (Pellegrino et al., 2015). If the rubrics are developed in a way that clearly outlines differing levels of achievement, there is no room for misinformation or perhaps more importantly, miscommunication about student achievement.

The use of rubrics does not have to be limited to inside the four walls of the classroom. Though rubrics in the classroom can help to inform, guide, evaluate, and verify instructional planning, rubrics can also be used outside of the classroom to evaluate music performance. Specifically speaking, research is beginning to show that evidence of student achievement in the music performance classroom could be best represented by looking at the large ensemble performance results (Parkes, 2010). The daily learning and instruction that occurs in the classroom is intended to develop and improve the overall final product that results when parts of the ensemble work together for a unified performance (Scott, 2012). For this reason, rubrics can and perhaps should be used for the evaluation of large ensemble music performances.

This particular study is geared toward the use of IRT in the development of rubrics for use during large ensemble performance evaluation. Though the specific use of these rubrics is intended for outside of the music performance classroom, the understanding is that if these rubrics are developed under a more restrictive and strenuous setting such as large ensemble

performance evaluation, then the rubrics will also be usable in the classroom in order to evaluate the results of instructional methods that are used to prepare for large ensemble performance evaluation. If music educators are indeed teaching conceptually in the classroom, students' understanding of musical concepts and how those concepts are manipulated during performance will be evident during large ensemble performances: it is the true intersection of knowledge, technical skill, and artistry (Hope and Wait, 2013).

In order for rubrics to be considered valid and reliable for both classroom testing and large ensemble performance evaluation settings, empirical support is needed. This is partly due to the fact that we must meet the demands that continue to emerge as a part of the data driven era. Most importantly, however, the development of valid and reliable measures for student music performance settings means that stronger curricular decisions can be made in order to improve teaching and learning in music education classrooms abroad. Future replication and specific development of classroom rubrics will be important in order to guarantee an alignment between rubrics used in the classroom and those used for large ensemble performance evaluation. Such practices will help to strengthen instructional reflection, planning, and implementation by providing more information on student achievement levels that can be used to better inform the demands of the educational data driven era.

The current rating sheet used in large ensemble performance evaluation utilizes assumptions that are in line with CTT, meaning that there are certain limitations when considering the nature of rater-mediated assessment. In contrast, the newly proposed system uses the Rasch Measurement Model that operates under IRT assumptions, which can help to compensate for some of the limitations of the current system.

Methodology

Participants. 34 String Ensembles (*N=34)* from 14 different schools participated in this study. Ensembles represented 8 middle schools and 6 high schools. There were a total of 16 middle school performances and 18 high school performances. These schools belong to a suburban school district in a southern state and were participating in the state's performance evaluation event. The schools that performed at this particular site were assigned according to their location in the district. Each ensemble performed three pieces for raters during their performances.

Raters. The large ensemble performance evaluation event had two sets of raters: one set of raters ($N = 3$) was hired by the state's music educators association as is typically done (condition A raters), and the other set ($N = 3$) was hired by the researcher in order to use the condition B rubric (condition B raters). Each rater (in both sets) is on the approved adjudicators list for the state music educators association. The condition A raters were two males and one female rater with an average of 26.7 years of teaching experience ($SD = 1.7$) in string education. Each rater is experienced and has served as a rater for many of the state adjudicated large ensemble performance evaluation events. Two of the raters have earned master's degrees in educational fields, and one has earned a doctorate degree as their highest degree earned. Each rater was given the rubric ahead of time in order to familiarize themselves with the wording and organization of the document.

The condition B raters were three female raters with an average of 23.3 years of teaching experience ($SD = 4.71$) in the string classroom, and an average of 15.7 years experience in rating ($SD = .94$) large ensemble performance evaluations. Each rater is experienced and has served as a rater for many of the state adjudicated large ensemble performance evaluation events. All three

of the raters have earned master's degrees in educational fields as their highest degree earned. Each rater was given the rubric ahead of time in order to familiarize themselves with the wording and organization of the document. Selection of the raters as content experts was based on the assumption that selecting raters that are expert teachers and musicians offers the best chance for developing a fair and reasonable assessment (Wesolowski et al., 2015).

Instrument. Two rating sheets were used in order to provide for a comparison between the current method, condition A, and the new method, condition B. The state-hired raters used the current string performance evaluation rating sheet in order to provide a numerical rating for each of the seven categories for each piece (Figure 4.1) (SMEA, 2017). The categories for the current rating sheet are: tone, intonation, technique, balance, interpretation, musical effect, and general effect. Each of those categories is rated on a scale of I through V. For the purpose of this study, five was recorded as the best possible score and one was recorded as the lowest. The ratings for each category and each piece is then added in order to achieve an overall numeric score. In addition to completing the current rating sheet, the state-hired raters also recorded verbal comments during the performance and wrote summative feedback on the final rating as is typically done in the current large ensemble performance evaluation setting.

The condition B rubric used for the event was developed in a previous study by the researcher (in press) using the Rasch Multifaceted Partial Credit Measurement Model (Linacre 1989/1994)(see Figure 4.2). The condition B rubric operates with assumptions of IRT, in contrast to the condition A rating scale that utilizes concepts of CTT. The raters hired for the purpose of this study completed one condition B rubric for each ensemble. The condition B rubric contains four a priori categories: tone production, rhythm and pulse accuracy, intonation accuracy, and expressive qualities/stylistic interpretation. Each category has a different number of item stems.

Tone production contains four items, rhythm and pulse accuracy contains eight items, intonation accuracy contains nine items, and expressive qualities/stylistic interpretation contains seven items. Each of the items contains anywhere from two to four categories that can be used to rate the performance. The items and number of categories per each item were validated as a part of the previous study (in press). Item stems were validated as a part of the original rating scale and then a rubric was developed for the purpose of this study. The computer program *FACETS* (Linacre, 2014) was used to run analysis of the rating data.

Procedures. Each ensemble performed three pieces for the raters in an auditorium. One rater from each group was positioned on the left side of the auditorium, in the middle, and on the right side. Both sets of raters had musical scores for each selection to use during the rating process. Following the performances, the ratings for each performance were collected from both the condition A raters and the condition B raters.

Results

Classical Test Theory item analyses were used to investigate how the condition A rating scale worked within the realm of its primary theoretical framework. Rater agreement for the condition A system was determined using Fleece's Kappa, and further item analyses that included Difficulty, Discrimination, and Correlation were used to draw conclusions on the overall effectiveness of the condition A rating scale. For an additional level of analysis, the results from the condition A rating scale were also subjected to Rasch analysis. Rasch analysis was used to determine the effectiveness of the condition B rubric. CTT item analyses could not be reciprocated to investigate the condition B results further because there was missing data that prevented that specific type of analysis. Within the guidelines of the two opposing theoretical

frameworks, the results from the Rasch analyses are sample independent, but the results from the CTT item analyses are sample dependent.

Condition A rating scale

Performance results. The highest achieving performances were performances 36 and 38 and the lowest achieving performance was performance 24. When ordered according to highest and lowest ratings, the condition A method showed ties for performances that were ranked first, second, eighth (three way tie), ninth, twelfth, seventeenth (three way tie), and twenty-first (it is important to note that these are not the performance numbers, but the rankings from highest achieving to lowest according to the condition A method). The calibration of performances that resulted from the Rasch analysis performed on the data from the condition A rating scale is presented in Table 4.7. The highest achieving performances were still 36 and 38 (4.24 logits) and the lowest was performance 24 (-0.33 logits). This demonstrates a range of 4.57 logits ($M = 2.41$, $SD = 1.03$, $N = 34$).

Rater results. A mean of each rating given by the condition A raters was calculated. The overall mean for all of the ratings for all three raters was 4.85. The first rater had a mean score of 4.85 (Difficulty = 0.97), the second rater had a mean score of 4.74 (Difficulty = 0.95), and the third rater had a mean score of 4.82 (Difficulty = 0.96). Table 4.9 contains the information for the condition A raters when analyzed using Rasch. The second condition A rater was the most severe (0.37 logits), the third condition A rater was in the middle (-0.17 logits), and the first condition A rater was the most lenient (-0.20 logits). This demonstrates a range of 0.57 logits ($M = 0.0$, $SD = 0.26$, $N = 3$). Condition A raters one and two were fit, and condition A rater three was misfit based upon Infit MSE statistics that fall outside of the 0.8-1.2 range as indicated by Wright and Linacre (1994) and Engelhard (2009). Item Analysis using Fleece's Kappa was

calculated based on the condition A raters' responses for each category on each of the three selections. Technique 1 (technique on the first selection), Balance 1, General Effect 1, Tone 2, Balance 2, and General Effect 3 were above 0.2 and were the only categories considered to be in the acceptable range for general agreement between judges (See Table 4.11).

Category results. Averages were collected for each of the seven items on the condition A rating sheet (tone quality, intonation, technique, balance, interpretation, musical effect, and general effect). Of those categories, item 2 (*intonation*) had the lowest average ($M = 4.2$), so it was the most difficult item to endorse. Item 7 (*general effect*) had the highest average ($M = 4.8$) so it was the easiest item to endorse. Identical results were found when Difficulty for each item was calculated. General Effect 1 (0.976) was the easiest, and Intonation 3 (0.825) was the most difficult. Discrimination was also calculated in order to determine the ability to demonstrate between the high level scores and the low level scores. The highest discrimination score was Intonation 1 (0.184) and the lowest was Balance 1 (0.049)(See Table 4.11).

Summary statistics. Frequency counts for each of the items was calculated in order to decipher how many of each rating was given to ensembles within the categories. For a total of 34 ensembles, three pieces for each ensemble, 3 raters for each ensemble, and 7 categories for each piece, 2,142 total ratings were given in each category using the condition A sheet. The frequency counts were as follows: 1,415 fives were given, 653 fours were given, 71 threes were given, and 3 twos were given. No ones were given. When subjected to Rasch analysis, performances ($Rel_{performances} = .91$) had the highest reliability of separation followed by raters ($Rel_{raters} = .89$) and items ($Rel_{items} = .87$) (See Table 4.1).

Condition B rubric

Summary statistics. Table 4.2 shows the Summary Statistics for the Rasch Measurement Model. Chi square statistics for each of the three facets show significant levels of difference between performances (χ^2 =411.9, $p < .01$), raters (χ^2 = 173.5, $p < .01$), and items (χ^2=437.6, $p < .01$). High reliability of separation (as demonstrated by .9 or higher) further provides confidence that each of the facets was accurately measured independent of the others. Raters (Rel_{raters} = .98) had the highest reliability of separation while performances ($Rel_{performances}$ = .94) and items (Rel_{items} = .94) had the same reliability of separation. Infit measures determine how well the data fits the model. According to Wright and Linacre (1994) and Engelhard (2009), the high-stakes testing threshold for the Infit statistic considered to appropriately fit the model are those that fall within the range of 0.80 and 1.20.

Variable map. The variable map serves as an operational definition of the latent construct, which in this case is string large ensemble performance achievement (See Figure 4.3). The three facets (performances, items, and raters) are demonstrated on the three columns so that they may be compared to a common "ruler", otherwise known as the logit scale (shown on the far left). The first column following the ruler shows the performances in relation to the logit scale. The highest achieving performances are near the top and the lowest achieving performances are closer to the bottom. The second column shows the raters, with the most severe rater being closer to the top and the most lenient rater being closer to the bottom. The third column shows the items according to how difficult each item is to endorse. The items that are more difficult to endorse are closer to the top and the items that are less difficult to endorse are closer to the bottom.

Calibration of ensemble performances. The calibration of performances is provided in Table 4.8. Performances considered to have higher achievement are closer to the top and performances with lower achievement are listed closer to the bottom. The highest achieving performance was performance 36 (5.42 logits), and the lowest achieving was performance number 24 (-0.48 logits). This demonstrates a range of 5.90 logits ($M = 1.60$, $SD = 1.28$, $N = 34$). There were four ties for performance measures: performances 20 and 30 (1.33 logits), performances 7, 18, and 34 (1.14 logits), performances 10 and 23 (0.84 logits), and performances 11 and 31 (0.77 logits). Misfitting performances are based upon Infit MSE statistics that fall outside of the 0.8-1.2 range as indicated by Wright and Linacre (1994) and Engelhard (2009). Overfitting performances included 27, 33, 36, 37, and 38 and underfitting performances included 3, 9, 14, 17, and 31. These performances should be reviewed for anomalies in the rating process.

Calibration of raters. The calibration of raters is presented in Table 4.10. Raters are listed according to severity and leniency. Rater 1 (0.76 logits) was considered to be the most severe rater, followed by Rater 2 (.00 logits), and Rater 3 (-0.76 logits) was considered to be the most lenient rater. This demonstrates a range of 1.52 logits ($M = 0.0$, $SD = 0.62$, $N = 3$). According to the Infit MSE statistics, all three raters fit the model. None of the raters were outside the 0.8-1.20 range, so none were considered to be too muted or sporadic.

Calibration of items. The calibration of items is presented in Table 4.4. Items are listed according to the difficulty. The most difficult item was item 21 (*intonation in technical passages*, 2.05 logits). The easiest item was item 25 (*connection of phrases*, -1.46 logits). This demonstrates a range of 3.51 logits ($M = 0.0$, $SD = 1.02$, $N = 26$). Items that were considered to be overfit were 10, 12, 24, and 28. Items that were considered to be underfit were 13, 15, 19, and 22. These items should be reviewed for anomalies in the rating process.

Rating scale category diagnostics. Table 4.6 provides information that was used to investigate the usage of each item in the rubric. Frequency counts shown in the first column were investigated based on Linacre's (2002) recommendation of ten uses per category. Any categories with less than ten uses would typically be collapsed in order to better fit the model. Item 10 (category 1), Item 13 (category 1), Item 14 (category 1), Item 15 (category 1), Item 17 (category 1), Item 18 (category 1), Item 19 (category 1), Item 23 (category 1), Item 25 (category 1), Item 26 (category 2), Item 27 (category 2), and item 28 (category 1) contained fewer than ten uses. Additionally, for items 4 and 5, all three raters chose the same category for every performance, showing no variance in the usage of these two items.

Each number listed in the second column is the specific logit location for each category of each item. Monotonicity refers to the continuous advancement of step calibrations (Andrich, 1996). Assumptions dictate that category 4 is more difficult to endorse than category 3, category 3 is more difficult to endorse than category 2 and so forth. If an item showed a violation in monotonicity, the item might be collapsed in order to contain fewer categories. If there were any violations of monotonicty, that would be apparent in the average observed logit column. No items showed a violation in monotonicity, meaning difficulty thresholds were as expected.

The third column shows Outfit mean squares (MSE). This data was examined for values >/= 2.0 because high values would indicate sporadic measures in the ratings. Item 7 (*attacks are sometimes executed with precision across the ensemble*), item 10 (*wavering steady pulse sometimes detracts from the continuous flow of the music*), and item 12 (*inaccurate performance of subdivisions occasionally detracts from solidly communicated tempo and meter*) were considered to be overfit. Replication of uses with this particular rubric could determine if these item stems should be collapsed for future use.

Conclusions and further research

Comparison will begin by taking a look at how each of the two systems handled the actual evaluation of each performance, how the raters operated within each of their groups (condition A raters versus condition B raters), and the manner in which each item was used in the evaluation process. The numerical reports and comparisons of each of the performances, raters, and items will be used to answer the first research question.

In addition to the statistical tests, the results from both sets of raters was given to directors, along with an informal survey in order to glean information as to if the newly proposed rubric can provide adequate information that might be able to inform instructional planning and curricular decisions for string ensemble directors. The informal survey was also given to the condition B raters. The survey results therefore address both the directors' and raters' perception of usability of the condition B rubric. Responses from the surveys will be used to help answer the second and third research questions.

The condition B results demonstrated very high reliability of separation for performances. 0.94 reliability of separation provides a high level of confidence that each individual performance was measured independent of the items and raters. The reliability of separation for performances was .91 when Rasch analysis was applied to the ratings from the condition A rating scale. This number is lower, therefore suggesting that the condition A rating scale is generally less reliable than the condition B rubric. In contrast, the higher reliability of separation for the condition B rubric shows that the rating given to each performance serves as an accurate measure that can be used to compare the performances. Representation of this can be seen in the variable map as each performance is demonstrated on a visual representation of the latent construct. The best performance was performance 36 with a measure of 5.42 logits. The next

highest performance was performance 38 with 4.32 logits, and so forth with each performance receiving a unique measure on the logit continuum.

The summary statistics show that the mean measure for performances using the condition B rubric was 1.60 logits (*SD* = 1.28) (See Table 4.2) and the measure for performances using the condition A rating scale was 2.41 (*SD* = 1.03) (See Table 4.1). The higher measure apparent in the analysis of the condition A rating scale shows that the average performance rating was higher, thus signifying possible inflation. This might suggest that it was easier to score on the higher end when the condition A rating scale was used. Furthermore, the smaller standard deviation for the condition A rating scale further confirms that there was less differentiation between each performance.

In terms of overall ranking, the two systems ranked the four highest achieving performances in a somewhat similar manner (See Table 4.12). The highest achieving performance for both systems was performance 36 and the next three highest achieving performances were the same for both systems as well (38, 37, and 15). Similarly, the lowest achieving performance for both the condition B rubric and the condition A rating scale was performance 24. Aside from that detail, there were some significant deviations. Performances 1, 2, 3, 4, 6, 11, 12, 13, 14, 15, 18, 19, 22, 23, 24, 26, 27, 34, 36, 37, and 38 were all ranked within 4 slots of one another between the two rating systems. The remaining 13 performances, however, were more than 5 slots different in rankings between the two systems: performances 7 (17^{th} tie with condition B and 9^{th} tie with condition A), 9 (16^{th} with condition B and 8^{th} tie with condition A), 10 (19^{th} tie with condition B and 12^{th} tie with condition A), 16 (26^{th} with condition B and 21^{st} tie with condition A), 17 (25^{th} with condition B and 18^{th} with condition A), 20 (15^{th} with condition B and 3^{rd} with condition A), 21 (27^{th} with condition B and 14^{th} with condition A), 28

(22nd with condition B and 13th with condition A), 29 (12th with condition B and 4th with condition A), 30 (15th tie with condition B and 5th with condition A), 31 (20th tie with condition B and 11th with condition A), 32 (23rd with condition B and 17th tie with condition A), and 33 (8th with condition B and 20th with condition A). The most concerning of these are performances 7, 20, and 33 with a difference of 12 between the two rankings. These differences should alert music educators that there is a problem with current performance evaluation results.

Table 4.12 shows the ranking of each performance for the condition A rating scale and for the condition B rubric side by side. The results from the condition A rating scale are shown in raw scores (first column), and also in logits that resulted from Rasch analysis (second column). The third column shows the Rasch results from the condition B rubric. Z-Scores are also included for each set of data. Performance 36 was the overall top performance, but it is interesting to note how differently this performance is represented with each analyses. The Z-Score for performance 36 for the condition A rating scale is 1.21 when using raw scores, 1.78 when using logits, and 2.99 when using logits for the condition B rubric. The condition B rubric shows that performance 36 was a much higher achieving performance when compared to the remaining performances, whereas the results from the condition A rating scale do not seem to show how much better performance 36 actually was in comparison to the others. The overall range represented by the condition B rubric (range = 4.61) demonstrated a much more realistic spread of the performances as opposed to the condition A rating scale (range = 4.52) or the condition A rating scale with Rasch scores (range = 4.43). This is more clearly demonstrated by the graphic representation shown in Figure 4.4. The increasing differentiation is evident when looking at the three box and whisker plots because the condition B rubric shows a bigger spread for the performance ratings, thus providing the opportunity for a more refined rating for each

performance. With this being said, the box and whisker plot for the condition B rubric shows a higher concentration for the lower achieving performances, suggesting the need for items that better distinguish ratings for the lower achieving performances.

The presence of concentrated ties for performance measures that are apparent with both the condition A rating scale and the condition B rubric means that certain performances were not given a unique and individual score on the continuum. Rather, several performances shared the same score. The condition A rating scale contained seven ties (two of which were three way ties), and the condition B rubric contained four ties (one of which was a three way tie). This results in difficulty when trying to accurately identify which performances were better than others. Due to the larger number of ties with the condition A rating scale, it can be concluded that the condition A method did a worse job of adequately distinguishing between true performance and estimated performance. There is little confidence that each performance received a true rating of the estimated performance that will allow for it to be compared with others in a valid manner.

It is interesting to note that only two performances scored in the negative logits when looking at the Rasch results. This provides strong evidence that most of the performances at this particular site were considered to be high quality performances, which holds true for the reputation of this particular district. Further replication of the rubric in a different large ensemble performance evaluation setting might provide for a different range in performances, with some possible lower scores in the negative logit measures.

The Rasch summary statistics showed a high reliability of separation for the condition B raters. Of the three facets (performances, items, and raters), the raters had the highest reliability of separation at 0.98. There is a high level of confidence that each rater was accurately measured

independently of the other two facets, and Infit statistics show they were inside the normal range (0.8-1.2), accurately fitting the model. Notions of severity and leniency can be taken into account in order to realize that future training might help to decrease the range in measures from the current range of 1.52 logits. This is a strong range to begin with, but one of the advantages of using the condition B rubric is that raters can use the statistics in order to provide training on how to better align rater practices.

The first rater using the condition A rating scale was considered the most lenient with an average of 4.85 (-0.20 logits in Rasch)(See Table 4.9). The third rater using the condition A ratings scale was the second most lenient with an average of 4.82 (-0.17 logits in Rasch), and the second rater using the condition A rating scale was the most severe with an overall average of 4.74 (0.37 logits in Rasch). Severity and leniency trends for the condition A raters were identical both when raw scores were used and when results were analyzed using Rasch. Additionally, Fleece's Kappa was applied to analyze the amount of general agreement between the condition A raters on individual items (See Table 4.11). Kappa results 0.2 and up is acceptable, with results improving the closer they get to 1.0. Negative numbers are considered poor in nature. Kappa results for the condition A raters ranged from -0.156 to 0.290. Acceptable agreement was only reached on five individual items. Those individual items were Technique 1, Balance 1, General Effect 1, Tone 2, and Balance 2. Fleece's Kappa results suggest that there was very little agreement amongst the three condition A raters on individual items. Rasch analysis of the condition A rating scale shows that rater 3 (Infit MSE = 1.25) was misfit and rater 1 (Infit MSE = .81) was almost misfit. Though the condition A raters were fairly close in terms of ratings assigned to performances with a range of 0.11, the calculated averages show that each of the raters using the condition A rating scale tended to err on the higher side, awarding more fives

than any other rating. If only fives are going to be given, perhaps the inclusion of other ratings in the system should be considered superfluous.

Rasch statistics for the condition B rubric showed a high reliability of separation for the items, providing confidence that the items were measured independently from the other facets. Difficulty for the condition A rating scale was calculated by dividing the average rating by the maximum possible score. The goal is 0.5 and the closer to 1.0, the easier the score. Difficulty results ranged from 0.825 (Intonation 3) to 0.976 (General Effect 1) (See Table 4.11). The individual items were extremely close to 1 and represented a range of only 0.151, thus demonstrating that each of the items were very easy to endorse with relatively little variability amongst each of the items. The hardest item on the condition B rubric was item 21 (*intonation in technical passages*) and the hardest category on the condition A sheet was Intonation 3. This shows that intonation is the hardest area for string ensembles to execute in an exemplary fashion. The easiest item on the condition B rubric was item 25 (*connection of phrases*) and the easiest category on the condition A sheet was General Effect 1. It makes sense that general effect was the easiest for the condition A rating scale because this category does not actually include any aspects of musical performances. The sub categories listed within this category are choice of music, discipline, instrumentation, and appearance. Perhaps the most concerning part of this category is that a string large ensemble could play horribly and give a low achieving performance, but score a five in this category if the rater felt that each of these sub categories listed was adequately met. It seems problematic that this category, which contains no aspects of musical performance, is included in the final overall rating. This detail of the condition A rating scale seems to contribute to the possible invalidation of performance ratings.

Frequency counts of the items used in the condition B rubric showed that many items would be collapsed because certain categories were used less than ten times. This is a bit concerning, but also something that replication of the study might address. It seems as if there might have been a halo effect with the raters that resulted in the raters sometime hovering in the middle two categories, as opposed to using categories one and four in order to truly distinguish between certain aspects of each performance (Wesolowski, 2015). This was particularly problematic with items 4 and 5 in which only one category was selected for every performance. This might partly explain why there was not a larger range demonstrated among the highest and lowest achieving performances. A possible limitation of the study lies in the notion that the string large ensembles that performed at this particular location are considered to be in an area that is known for having strong string programs. There might not have been enough distinction between the categories that were used to rate the performances because the performances were all generally strong. Replication of the study in a more rural school district also might provide more extended use of the items for rating purposes. The continued usage of the condition B rubric in diverse environments could provide more metrics that can be used to shape and improve the precision and validity of this measurement instrument.

Frequency counts for each of the items used in the condition A rating scale revealed that a large surplus of fives were given for individual categories, whereas hardly any ones or twos were given (See Table 4.5). Figure 4.5 provides a graphic representation of the categorical usage for the condition A rating scale wherein it is visibly obvious that the ratings are skewed to the higher end. These showings are somewhat similar to the findings in the Rasch analysis in that results seemed to err on the higher side as opposed to a wider spread of ratings that could account for each possible level. The five ratings (1,415) given for categories using the condition

A rating scale were over double of the next highest frequency count, which was the four rating category at 653 total uses. Another very large jump is evident in the next highest category usage, which was 71 uses of threes, and then again, only three uses of twos. These frequency counts show that there is not an even amount of each rating given in the categories that make up each overall rating. Perhaps the performances were all very good, however, there should be more of a distinguishing factor between each performance in order to ensure that the true versus estimated performance measures are closer in proximity. This also begs the question as to if the bottom two overall ratings of fair and poor re needed if they are not going to be used to help distinguish between performance levels. Discrimination statistics was run on the condition A results in order to determine the ability of the condition A rating scale to differentiate between high-level scores and low-level scores. It is considered acceptable 0.2 and up on a range of -1 to 1. Not a single item received a discrimination rating that was acceptable, thus showing that ensembles with higher ability did not always out perform lower ability groups. This means that the items did not reliably distinguish between upper and lower level groups in the condition A system.

The purpose of performance evaluation is to determine the overall level of performance achievement. Our goal should therefore be to rate performances in a way that is considered valid and reliable. Items on a rubric need to measure the same latent construct. For the condition A rating scale, correlation analyses were used to determine the amount of measurement error associated with the score. High reliability would indicate that the items were all measuring the same latent construct. For the correlation statistics, 0.6 and higher is considered to be acceptable. Only four categories received an acceptable correlation statistic. The items that met these criteria were Tone 1, General Effect 1, Tone 2, Musical Effect 2, and Tone 3. These results show that tone is perhaps considered to be the most important by the three condition A raters. The lack of

more categories with acceptable correlation suggests that the condition A rating scale is flawed, as it does not contain items that accurately and consistently measure the performance achievement construct.

Though the reliability of separation was high on the condition B rubric, there were still a few issues with items that were considered to not fit the model. Items that were considered to not fit the model for either being underfit or overfit were 10, 12, 13, 15,19, 22, 24, and 28. Though these items would typically be removed because they did not fit the model, such decisions are outside the scope of this study. Future replication would allow for the researcher to investigate why those items did not fit the model, and additional items might be added in order to fill the void that was created by these items being removed. Continued research into the creation of more (and better) items stems would serve the purpose of filling any gaps along the logit continuum: thus, a more comprehensive collection of item stems would help raters to be able to better differentiate between performances in order to compensate for the ties between performance ratings mentioned above.

When looking at an overall comparison, one must consider how each of the systems treats the results for items, raters, and performances. In answering the first research question, the two systems seemed somewhat similar in terms of how the items were handled during the rating process. The hardest item to endorse for both systems was playing with good intonation, which seems to maintain the expected rigor in defining what constitutes an exemplary performance. The reliability of separation was much stronger for the condition B rubric, meaning that the condition B rubric serves as a more adequate measurement for performance achievement. Though both systems provided evidence that only certain categories of items were used and that

findings sometimes hovered around the easier side, category usage with the condition B rubric was more definitive.

In regards to the qualitative feedback mentioned in the second research question, survey results regarding the directors and raters' opinions on the items showed that both the directors and raters appreciated the fleshed out nature of the condition B rubric items and felt that it provided adequate feedback. In response to the third research question that addresses usability, there were some instances when more detail or potential rewording was requested, particularly in instances when absolute words such as "never" or "perfectly" seemed to limit the choices that raters could select. The raters mentioned that more favorable words might include "consistently" or "somewhat". It was also mentioned that a few aspects of string-specific items were missing, such as bow placement, bow usage, and specific items that address the releases of notes (not just the attacks). Comments were also made in regards to a lack of items that directly address correct rhythms. Replication of this study could be used to research, develop, and validate new item stems while also refining anchors and descriptors for each item in order to provide more comprehensive feedback to directors. These changes could help to improve usability for the raters. In short, the issues with the condition B rubric item stems can be adjusted and reworked in order to refine the evaluation system that might improve validity, reliability, and become even more usable for future use in performance evaluation systems.

In terms of raters, both sets of raters were somewhat comparable within their group. When considering the first research question, the numerical results show that the condition B raters were the more effective group whereas the condition A raters had one rater that was misfit and one that was almost misfit. Fleece's Kappa showed that there was relatively little agreement amongst the condition A raters. The advantage with the condition B method is that it allows for

future compatibility research, whereas the condition A method does not. This capability allows raters to be trained and/or certified to be a rater before they evaluate for their first time. This also allows for continuous monitoring to ensure proper practice is maintained over time. For this reason, the condition B rubric is more favorable because of the higher level of agreement that is largely due to the anchors of the rubric system. The anchors provided the raters with strong statements that they could use to guide their responses, thus preventing the presence of possible arbitrary ratings. This information can be used to hire (and train) raters to better understand how to adequately rate performances in the future.

Survey results also showed that both the directors and raters would like to have the opportunity for raters to leave a personal final statement describing the overall performance. Adding a final overall written comment would help to advance the qualitative results. These changes could be made in favor or eliminating the moment-by-moment commentary that is currently given during performances at large ensemble performance evaluation, which would contribute to the overall usability of the condition B rubric . Though there may be some merit in hearing the raters speak during the performance, survey results showed that the raters were able to better focus on thoroughly rating the performance by listening critically because they did not have to speak into a recorder in real time. This directly accounts for the usability factor of the condition B system in that raters must be able to accomplish everything during the real-time performance. On that same topic, one director mentioned on the survey that he/she felt the raters were able to give their undivided attention to truly be able to listen and evaluate the performance.

The final aspect of investigation, the performances, provides the opportunity for the highest amount of comparison. The real advantage of the condition B rubric becomes evident in how it accurately reconciles the issue of true versus estimated performance. The condition B

rubric provided a very distinct and specific measurement for each performance with only a few ties. In contrast, the condition A rating scale allowed for many more ties and duplicate rankings for certain performances. The reliability of separation for the condition B rubric was higher. In addition, the Z-Scores were more evenly distributed as is visually evident in Figure 4.4. Since performance achievement is the latent construct being measured, it should be the facet that carries the most weight in terms of deciding which method to use for performance evaluation. The condition B method provides a higher level of confidence that the actual level of true performance was reported.

One disadvantage that was mentioned several times in relation to the qualitative feedback gleaned from the condition B results was the lack of specificity for particular pieces, or measure numbers. Because only one rubric was filled out for each performance, the clarity in terms of what was done well in each piece versus what could have been done better was perhaps unclear. Raters being able to leave a final overall written comment by satisfy this concern so that specific overall ideas and issues can be addressed in relation to each piece, or perhaps even some particularly problematic sections within each of the pieces. With that being said, a lack of moment-by-moment commentary on the raters part might also create the opportunity for directors and music educators to listen more critically to their recordings in order to glean how the rater reached their decisions and how the results accurately reflect the true performance of their ensemble. If raters are unable to comment about specific measures and pieces, directors (and students) may learn how to actively and critically listen in order to be able to understand how the raters reached their conclusions.

Transition to the condition B rubric for large ensemble performance evaluation would require additional work, specifically in consideration of how it accounts for items and raters.

New items (and possibly new categories) would need to be investigated and refined in order to increase qualitative feedback and usability. A formal process for the training of raters would also need to be developed, and specific details as to how to complete the rubrics would need some attention. Switching to the condition B rubric would also require standard setting to be investigated. The condition B rubric would allow for the establishment of categories such as superior, excellent, good, fair, and poor in order to maintain the common language already used within the condition A rating scale. In the survey results, many directors commented that they would appreciate if an overall rating could be added to the final report of the performance achievement in order to improve the qualitative nature of the condition B rubric. The qualitative comments are of course very helpful, but the assignment of a final overall standard of the actual performance and how it measured up against other performances is important to directors.

Regardless of some of the shortcomings mentioned, survey results showed that some directors and raters were excited about the possibility of a new evaluation system. Though there were some that are completely satisfied with the current system, there is certainly an interest in at least considering possible changes to our condition A rating scale. Perhaps future adjustments could include some of the current methods, such as a final written overall comment, and an overall standard that is tied to and empirically based on the results from the condition B rubric. As mentioned earlier, these are all adjustments that can be done with future study and replication.

Though there would be considerable adjustments ahead in order to transition to a new system, perhaps it is time that we as music educators consider a different approach to performance evaluation. The development and validation process would help us to align our beliefs and understandings of string performance concepts in order to develop a truly accurate

representation of what constitutes an exemplary string ensemble performance. Being able to entertain such discussions and produce a more valid measurement system would only help to improve teaching and learning, not only in our classrooms but perhaps in our pre-service teacher setting as well. In addition to aiding the conversation to improve practices in our field, a valid and reliable performance achievement measure would satisfy policy makers and administrators by allowing us to properly communicate evidence of student achievement through the use of common language. Further research, additional considerations, training, collaboration, and study are needed to replicate and finalize the findings that have been mentioned here, but such interventions and actions can only help to improve the work being done in our discipline. The results from such an overhaul can only help to improve our students' learning experience, community members' perceptions and appreciation, policy makers' understanding of what it is that we do, and administrators' approval of student achievement levels. Perhaps Bond and Bond (2010) emphasize the importance considering the use of Rasch in music education programs more clearly by stating, "it forms an important part of reflective, empirically-informed pedagogical practice" (p. 6).

References

Andrich, D. (1996). Measurement criteria for choosing among models with graded responses. In Eye, A. V., & Clogg, C. C. (Eds.), *Categorical Variables in Developmental Research* (pp. 3–35). San Diego: Academic Press.

Asmus, E. P. (1999). Music assessment concepts. *Music Educators Journal*, *86*(2), 19–24.

Bond, T., & Bond, M. (2010). Measure for measure: Curriculum requirements and children's achievement in music education. *Journal Of Applied Measurement*, 11(4), 368–383.

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

Brewer, C., Knoeppel, R. C., & Lindle, J. C. (2014). Consequential validity of accountability policy: Public understanding of assessments. *Educational Policy*, *29*(5), 711–745.

Brookhart, S. M. (2013). The public understanding of assessment in educational reform in the United States. *Oxford Review of Education*, *39*(1), 52–71.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston, 1986.

Engelhard, G. J. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: development, implementation, and analysis* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*(4), 585–602.

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

Ferrero, A., Petri, D., Carbone, P., & Catelani, M. (Eds.). (2015). *Modern Measurements: Fundamentals and Applications*. John Wiley & Sons.

Fisher, R. (2008). Debating Assessment in Music Education. *Research And Issues In Music Education*, 6(1),

Gruijter, D. N., and van der Kamp, L. J. (1984). *Statistical models in psychological and educational testing*. n.p.: Lisse : Swets & Zeitlinger, 1984.

Hambleton, R. K., Rogers, H. J., & Swaminathan, H. (1991). *Fundamentals of item response theory*. Newbury Park, Calif. : Sage Publications, 1991.

Hash, P. M. (2013). Large-group contest ratings and music teacher evaluation: Issues and recommendations. *Arts Education Policy Review*, 114(4), 163–169.

Hope, S., & Wait, M. (2013). Assessment on our own terms. *Arts Education Policy Review*, *114*(1), 2–12.

Johnson, R. L., Gordon, B., & Penny, J. A. (2009). *Assessing performance: designing, scoring, and validating performance tasks*. New York: The Guilford Press, c2009.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science*. San Francisco: Chandler, 1964.

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

Linacre, J. M. (1994). *Many-facets Rasch measurement, 2nd ed.* Chicago: MESA Press.

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

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

Messick, S. (1989). Validity. In R.L. Linn (Eds.), *Educational Measurement*, 3rd ed. (pp. 13–103). New York, NY, England: Macmillan Publishing Co, Inc.

Parkes, K. (2010). Performance assessment: lessons from performers. *International Journal of Teaching and Learning in Higher Education*, *22*(1), 98–106.

Pellegrino, K., Conway, C. M., & Russell, J. A. (2015). Assessment in performance-based secondary music classes. *Music Educators Journal*, *102*(1), 48–55.

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009). The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades. *Washington, DC: Center for Educator Compensation Reform. Retrieved February, 18, 2011*.

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational And Psychological Measurement, 76*(2), 325–338.

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management*, 54(2), 189–203.

Scott, S. J. (2012). Rethinking the roles of assessment in music education. *Music Educators Journal*, *98*(3), 31–35.

Springer, M. G. (2009). Rethinking teacher compensation policies: Why now, why again. *Performance incentives: Their growing impact on American K–12 education*, 1–21.

State Music Educators Association. (2017). *Large group performance evaluation*. Retrieved from http://opus.?mea.org/Pages/Forum/ViewForum.aspx?Forum=5.

Swan, G., & Mazur, J. (2011). Examining data driven decision-making via formative assessment: A confluence of technology, data interpretation heuristics and curricular policy. *Contemporary Issues In Technology And Teacher Education (CITE Journal),* 11(2), 205–222.

Wesolowski, B. C., Wind, S. A., & Engelhard, J. G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147–170.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016a). Rater analyses in music performance assessment: Application of the many facet Rasch model. In *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education* (pp. 335–356).

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

Whitcomb, R. (1999). Writing rubrics for the music classroom. *Music Educators Journal, 85*(6), 26.

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

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

Zaleski, D. J. (2014). An introduction to classroom assessment for today's music educator. *Illinois Music Educator*, *75*(1), 58.

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, *50*(3), 245.

Tables

Table 4.1 – Condition A Results

Summary Statistics from the PC-MFR Model

	Facets		
	Performance (θ)	**Rater (λ)**	**Item (δ)**
Measure (Logits)			
Mean	2.41	0.00	0.00
SD	1.03	0.26	0.65
N	34	3	21
Infit *MSE*			
Mean	0.97	1.00	1.00
SD	0.22	0.18	0.12
Std. Infit *MSE*			
Mean	-0.10	-0.10	0.10
SD	1.10	2.80	0.70
Outfit *MSE*			
Mean	0.96	0.96	0.96
SD	0.31	0.18	0.22
Std. Outfit *MSE*			
Mean	0.00	-0.60	0.00
SD	1.20	1.80	1.00
Separation Statistics			
Reliability of Separation	0.91	0.89	0.87
Chi-Square	501.6*	30.2*	201.8*
Degrees of Freedom	33	2	20

* $p < 0.01$

Table 4.2 – Condition B Results

Summary Statistics from the PC-MFR Model

	Facets		
	Performance (θ)	**Rater (λ)**	**Item (δ)**
Measure (Logits)			
Mean	1.60	0.00	0.00
SD	1.28	0.62	1.00
N	34	3	26
Infit *MSE*			
Mean	1.01	0.99	.99
SD	0.18	0.07	0.19
Std. Infit *MSE*			
Mean	-0.10	-0.30	-0.10
SD	1.20	1.70	1.60
Outfit *MSE*			
Mean	1.09	1.09	1.08
SD	0.43	.29	.64
Std. Outfit *MSE*			
Mean	.00	0.20	.30
SD	1.00	2.00	2.10
Separation Statistics			
Reliability of Separation	0.94	0.98	0.94
Chi-Square	411.9*	173.5*	437.6*
Degrees of Freedom	33	2	25

* $p < 0.01$

Table 4.3 – Condition A Results

Calibration of the Item Facet

Item Number	**Observed Average**	**Measure**	**Standard Error**	**Infit *MSE***	**Std. Infit**	**Outfit *MSE***	**Std. Outfit**
Intonation 3	4.13	2.00	0.18	0.94	-0.40	0.94	-0.40
Technique 2	4.39	0.94	0.19	0.96	-0.20	0.96	-0.20
Technique 1	4.36	0.75	0.21	0.91	-0.70	0.90	-0.80
Intonation 1	4.25	0.48	0.17	1.04	0.30	0.99	0.00
Intonation 2	4.25	0.48	0.18	0.95	-0.30	0.94	-0.30
Technique 3	4.44	0.38	0.21	1.03	0.20	1.05	0.40
Interpretation 2	4.59	0.10	0.21	1.00	0.00	0.97	-0.10
Balance 2	4.76	-0.03	0.24	1.04	0.20	1.00	0.00
Tone 2	4.68	-0.07	0.22	0.85	-1.00	0.78	-1.10
Interpretation 1	4.52	-0.12	0.21	1.06	0.60	1.08	0.60
Tone 3	4.81	-0.14	0.26	0.76	-1.00	0.75	-0.50
Balance 3	4.72	-0.14	0.23	1.05	0.30	1.24	1.00
Interpretation 3	4.60	-0.27	0.22	1.28	2.30	1.55	3.40
Musical Effect 2	4.79	-0.31	0.25	0.96	-0.10	0.76	-0.70
Musical Effect 3	4.81	-0.36	0.26	1.12	0.60	1.08	0.30
Tone 1	4.82	-0.38	0.27	0.91	-0.30	0.79	-0.50
General Effect 2	4.84	-0.43	0.28	0.84	-0.50	0.50	-1.30
General Effect 3	4.80	-0.45	0.24	1.20	0.80	1.17	0.50
Balance 1	4.82	-0.74	0.27	0.99	0.00	1.24	0.70
Musical Effect 1	4.84	-0.79	0.29	1.10	0.40	0.85	-0.30
General Effect 1	4.88	-0.90	0.32	0.95	-0.10	0.66	-0.60
Mean	4.62	0.00	0.23	1.00	-0.10	0.96	0.00
SD	0.23	0.65	0.04	0.12	0.70	0.22	1.00

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

Table 4.4 – Condition B Results

Calibration of the Item Facet

Item Number	Observed Average	Measure	Std. Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
21	2.40	2.05	0.24	0.93	-0.60	0.91	-0.30
12	2.51	1.45	0.23	1.23	2.30	2.77	6.00
2	2.56	1.19	0.23	1.14	1.50	1.06	0.30
17	2.08	1.08	0.28	0.98	0.00	0.93	-0.10
11	2.58	1.05	0.23	0.89	-1.10	0.80	-0.70
22	2.60	0.98	0.23	0.75	-2.80	0.63	-1.60
24	2.61	0.93	0.23	1.23	2.20	1.26	1.00
8	2.62	0.87	0.23	0.92	-0.80	0.89	-0.30
16	2.20	0.84	0.19	0.98	-0.10	0.93	-0.30
7	2.64	0.76	0.23	1.14	1.40	2.23	3.40
1	3.73	0.25	0.25	0.86	-1.20	0.69	-0.60
9	2.78	-0.11	0.27	0.90	-0.70	0.69	-0.60
19	2.43	-0.17	0.21	0.78	-1.80	0.74	-1.30
3	2.79	-0.22	0.27	0.98	-0.10	0.75	-0.40
18	2.46	-0.24	0.20	0.80	-1.50	0.75	-1.20
23	2.45	-0.34	0.21	1.07	0.50	1.00	0.00
15	2.37	-0.35	0.22	0.74	-2.10	0.70	-1.70
14	2.63	-0.69	0.21	0.88	-0.80	0.76	-0.70
26	3.44	-0.85	0.22	0.87	-1.10	0.80	-1.00
6	1.88	-0.93	0.33	1.01	0.10	0.77	-0.10
10	2.48	-0.94	0.22	1.62	4.60	3.15	7.00
20	3.89	-1.10	0.34	1.02	0.10	1.07	0.30
27	3.64	-1.25	0.23	1.02	0.20	1.04	0.20
28	2.72	-1.39	0.24	1.29	1.80	1.54	1.30
13	2.72	-1.40	0.24	0.76	-1.70	0.60	-1.00
25	1.92	-1.46	0.39	1.01	0.10	0.70	0.00
Mean	2.66	0.00	0.24	0.99	-0.10	1.08	0.30
SD	0.49	1.02	0.04	0.20	1.70	0.65	2.10

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

Table 4.5 – Condition A Results

Item Behavior of Categorical Usage

Item	Category Usage (%)				Average Observed Measure (Average Expected Measure)				Outfit *MSE*			
	2	3	4	5	2	3	4	5	2	3	4	5
Tone 1	-	2(2)	14(14)	86(84)	-	0.03(0.66)	1.96(1.93)	2.99(2.98)	-	.30	.80	1.00
Intonation 1	1(1)	11(11)	52(51)	38(37)	-1.18(-0.37)	0.93(0.76)	1.78(1.79)	2.49(2.50)	.50	1.00	.90	1.10
Technique 1	-	3(3)	59(58)	40(39)	-	-0.23(0.09)	1.34(1.38)	2.28(2.20)	-	.80	.90	.90
Balance 1	-	1(1)	16(16)	85(85)	-	0.04(0.98)	2.53(2.27)	3.31(3.34)	-	.30	1.40	1.00
Interpretation 1	-	1(1)	47(46)	54(53)	-	0.13(0.75)	2.18(2.08)	2.88(2.96)	-	1.00	.90	1.30
Musical Effect 1	-	1(1)	14(14)	87(85)	-	0.80(1.00)	2.39(2.27)	3.36(3.38)	-	.50	.80	1.10
General Effect 1	-	1(1)	10(10)	91(89)	-	0.90(1.03)	2.07(2.28)	3.47(3.44)	-	.50	.60	1.00
Tone 2	-	2(2)	29(28)	71(70)	-	-0.28(0.56)	1.77(1.86)	2.84(2.78)	-	.50	.80	.90
Intonation 2	1(1)	9(9)	56(55)	36(35)	-1.18(-0.40)	0.54(0.71)	1.85(1.79)	2.50(2.52)	.40	.80	1.00	1.00
Technique 2	-	5(5)	52(51)	45(44)	-	-0.30(-0.06)	1.19(1.18)	1.98(1.97)	-	.90	1.00	1.00
Balance 2	-	3(3)	18(18)	81(79)	-	0.13(0.44)	1.90(1.71)	2.64(2.67)	-	.50	1.10	1.10
Interpretation 2	-	2(2)	38(37)	62(61)	-	-0.45(0.49)	1.89(1.80)	2.65(2.68)	-	.60	1.00	1.10
Musical Effect 2	-	2(2)	17(17)	83(81)	-	-0.04(0.64)	1.95(1.92)	2.94(2.93)	-	.30	.80	1.10
General Effect 2	-	2(2)	12(12)	88(86)	-	0.08(0.67)	1.67(1.93)	3.06(3.01)	-	.30	.50	1.00
Tone 3	-	3(3)	13(13)	86(84)	-	0.49(0.47)	1.29(1.73)	2.81(2.74)	-	.90	.70	.80
Intonation 3	-	15(15)	59(58)	28(27)	-	-0.83(-0.68)	0.40(0.37)	1.10(1.08)	-	.80	1.10	1.00
Technique 3	-	2(2)	53(52)	47(46)	-	-0.73(0.35)	1.77(1.67)	2.44(2.51)	-	.70	1.00	1.10
Balance 3	-	2(2)	25(25)	75(74)	-	-0.20(0.59)	2.18(1.88)	2.75(2.83)	-	.40	1.40	1.10
Interpretation 3	-	1(1)	39(38)	62(61)	-	0.28(0.82)	2.56(2.15)	2.80(3.05)	-	.90	1.80	1.30
Musical Effect 3	-	2(2)	15(15)	85(83)	-	0.28(0.65)	2.25(1.93)	2.92(2.96)	-	.40	1.20	1.20
General Effect 3	1(1)	1(1)	15(15)	85(83)	1.98(0.21)	0.30*(0.80)	1.89(2.05)	3.08(3.06)	5.90	.20	.90	1.00

Note. Categories 1-5 vary in usage depending on the item, however, 1 always equals lesser performance ability and 5 always equals higher performance ability.

*Violation of monotonicity

Table 4.6 - Condition B Results

Item Behavior of Category Usage

Item	Category Usage (%)				Average Observed Measure (Average Expected Measure)				Outfit *MSE*			
	1	2	3	4	1	2	3	4	1	2	3	4
1	-	-	28(27)	74(73)	-	-	.19(.41)	1.79(1.71)	-	-	.60	.90
2	-	45(44)	57(56)	-	-	-.23(-.36)	.91(1.01)	-	-	.90	1.200	-
3	-	21(21)	81(79)	-	-	.71(.80)	2.10(2.08)	-	-	.70	1.00	-
4	-	102(100)	-	-	-	-	-	-	-	-	-	-
5	-	102(100)	-	-	-	-	-	-	-	-	-	-
6	12(12)	89(88)	-	-	1.44(1.49)	2.71(2.70)	-	-	.70	1.00	-	-
7	-	37(36)	65(64)	-	-	.23(-.01)	1.18(1.32)	-	-	2.90	1.10	-
8	-	39(38)	63(62)	-	-	-.18(-.10)	1.29(1.24)	-	-	.80	1.10	-
9	-	22(22)	79(78)	-	-	.52(.72)	2.06(2.01)	-	-	.60	.90	-
10	1(1)	51(50)	50(49)	-	2.23(1.04)	2.46(1.86)	2.62(3.26)	-	1.00	5.10	1.60	-
11	-	42(42)	59(58)	-	-	-.37(-.25)	1.2(1.12)	-	-	.70	.90	-
12	-	50(49)	52(51)	-	-	-.28(-.57)	.55(.83)	-	-	4.30	1.20	-
13	1(1)	26(26)	74(73)	-	1.42(1.33)	1.69(2.06)	3.49(3.36)	-	.90	.50	.80	-
14	3(3)	32(32)	67(66)	-	.64(.73)	1.32(1.47)	2.83(2.76)	-	.80	.70	.90	-
15	2(2)	60(59)	40(39)	-	.03(.53)	1.22(1.39)	3.15(2.88)	-	1.00	.50	.70	-
16	12(12)	58(57)	32(31)	-	-.17(-.47)	.24(.39)	2.06(1.90)	-	1.20	.70	.90	-
17	5(5)	84(82)	13(13)	-	-1.27(-.74)	.36(.32)	2.27(2.30)	-	.70	1.00	1.10	-
18	4(4)	47(46)	51(50)	-	.00(.39)	1.05(1.18)	2.70(2.56)	-	.80	.70	.80	-
19	4(4)	50(49)	48(47)	-	.06(.35)	.99(1.15)	2.74(2.55)	-	.90	.70	.70	-
20	-	-	11(11)	91(89)	-	-	1.65(1.56)	2.83(2.84)	-	-	1.10	1.00
21	-	61(60)	41(40)	-	-	-1.10(-1.05)	.52(.44)	-	-	1.20	.70	-
22	-	41(40)	61(60)	-	-	-.45(-.19)	1.34(1.16)	-	-	.60	.70	-
23	3(3)	50(49)	48(48)	-	1.16(.48)	1.27(1.29)	2.68(2.70)	-	1.20	.90	1.00	-
24	-	40(39)	62(61)	-	-	.10(-.14)	1.04(1.20)	-	-	1.20	1.30	-
25	8(8)	92(92)	-	-	1.76(1.88)	3.19(3.18)	-	-	.70	1.00	-	-
26	-	1(1)	55(54)	46(45)	-	1.12(.98)	1.69(1.81)	3.37(3.25)	-	1.00	.70	.90
27	-	1(1)	35(34)	66(65)	-	.93(1.25)	2.06(2.01)	3.29(3.32)	-	.80	1.10	1.00
28	1(1)	27(26)	74(73)	-	1.07(1.33)	2.58(2.06)	3.17(3.35)	-	.80	1.70	1.20	-

Note. Categories 1-4 vary in usage depending on the item, however, 1 always equals lesser performance ability and 4 always equals higher performance ability.

*Violation of monotonicity

Table 4.7 - Condition A Results

Calibration of the Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
36	4.90	4.24	0.45	1.01	0.10	0.72	-0.02
38	4.90	4.24	0.45	0.85	-0.30	0.52	-0.60
15	4.86	3.72	0.38	0.82	-0.60	0.87	0.00
37	4.86	3.72	0.38	0.82	-0.60	0.62	-0.60
20	4.84	3.58	0.37	0.99	0.00	1.06	0.20
29	4.83	3.45	0.36	0.93	-0.20	0.82	-0.20
30	4.81	3.33	0.35	0.94	-0.10	0.72	-0.50
1	4.79	3.21	0.34	1.08	0.40	1.46	1.10
27	4.76	2.99	0.32	1.04	0.20	1.11	0.40
3	4.75	2.89	0.31	0.91	-0.30	0.64	-1.00
9	4.75	2.89	0.31	0.61	-2.10	0.46	-1.80
26	4.75	2.89	0.31	0.90	-0.40	0.96	0.00
6	4.71	2.70	0.30	0.93	-0.20	1.06	0.20
7	4.71	2.70	0.30	0.72	-1.40	0.81	-0.50
14	4.70	2.61	0.30	1.22	1.00	1.09	0.30
31	4.68	2.52	0.29	0.92	-0.30	0.82	-0.50
10	4.67	2.44	0.29	0.69	-1.60	0.64	-1.30
13	4.67	2.44	0.29	0.98	0.00	1.18	0.70
28	4.65	2.35	0.29	0.98	0.00	1.19	0.70
21	4.63	2.27	0.28	0.95	-0.10	0.88	-0.30
22	4.62	2.19	0.28	0.84	-0.70	0.79	-0.80
34	4.60	2.12	0.27	0.86	-0.60	0.83	-0.60
4	4.59	2.04	0.27	0.94	-0.20	0.97	0.00
19	4.59	2.04	0.27	0.93	-0.30	0.82	-0.70
32	4.59	2.04	0.27	0.91	-0.30	0.86	-0.50
17	4.56	1.90	0.27	1.78	3.20	1.85	3.10
11	4.54	1.83	0.26	1.02	0.10	0.99	0.00
33	4.52	1.76	0.26	1.02	0.10	0.99	0.00
16	4.46	1.50	0.25	1.10	0.50	1.11	0.60
18	4.46	1.50	0.25	0.86	-0.60	0.88	-0.50
12	4.44	1.44	0.25	0.73	-1.40	0.71	-1.50
23	4.27	0.83	0.23	0.84	-0.80	0.83	-0.80
2	3.92	-0.17	0.20	1.59	3.00	1.86	3.90
24	3.86	-0.33	0.20	1.08	0.50	1.16	0.90
Mean	4.62	2.41	0.30	0.97	-0.10	0.96	0.00
SD	0.23	1.03	0.06	0.22	1.10	0.32	1.20

Note. Presented in measure order from highest achievement to lowest achievement.

Table 4.8 - Condition B Results

Calibration of the Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
36	3.05	5.42	0.76	1.26	0.50	2.76	1.20
38	3.01	4.32	0.50	1.26	0.70	2.04	1.10
37	3.00	4.08	0.46	1.30	0.80	1.86	1.10
15	2.95	3.39	0.37	0.95	-0.10	1.18	0.40
3	2.88	2.80	0.32	0.77	-1.10	0.54	-1.20
1	2.87	2.70	0.31	0.99	0.00	1.13	0.40
27	2.86	2.60	0.31	1.32	1.60	1.62	1.50
33	2.82	2.34	0.29	1.35	1.90	1.48	1.30
26	2.81	2.26	0.29	1.13	0.80	1.04	0.20
6	2.79	2.17	0.28	0.93	-0.40	0.79	-0.60
14	2.79	2.12	0.28	0.75	-1.70	0.65	-1.20
29	2.76	1.95	0.27	0.98	-0.10	0.81	-0.60
13	2.73	1.80	0.27	0.88	-0.80	0.95	-0.10
4	2.67	1.46	0.26	0.95	-0.30	0.97	0.00
20	2.64	1.33	0.25	1.14	1.00	1.08	0.40
30	2.64	1.33	0.25	0.85	-1.20	0.86	-0.60
9	2.62	1.21	0.25	0.71	-2.40	0.63	-2.10
7	2.60	1.14	0.25	1.09	0.70	1.02	0.10
18	2.60	1.14	0.25	0.80	-1.60	0.79	-1.10
34	2.60	1.14	0.25	0.98	-0.10	0.90	-0.50
22	2.59	1.08	0.25	0.92	-0.60	0.94	-0.20
10	2.54	0.84	0.25	0.81	-1.50	0.77	-1.40
23	2.54	0.84	0.25	1.07	0.50	1.11	0.60
11	2.53	0.77	0.25	1.19	1.40	1.11	0.60
31	2.53	0.77	0.25	0.75	-2.00	0.71	-1.80
19	2.52	0.75	0.25	1.16	1.10	1.17	0.90
28	2.53	0.74	0.25	1.15	1.10	1.13	0.90
32	2.50	0.65	0.25	0.99	0.00	0.98	0.00
12	2.49	0.59	0.25	1.10	0.70	1.06	0.40
17	2.47	0.53	0.25	0.72	-2.20	0.64	-2.40
16	2.46	0.47	0.24	0.89	-0.70	0.85	-0.80
21	2.45	0.41	0.25	1.07	0.50	1.09	0.50
2	2.29	-0.31	0.25	1.13	0.80	1.05	0.30
24	2.28	-0.48	0.25	1.02	0.10	1.05	0.30
Mean	2.66	1.60	0.29	1.01	-0.10	1.09	0.00
SD	0.19	1.28	0.10	0.18	1.20	0.43	1.00

Note. Presented in measure order from highest achievement to lowest achievement.

Table 4.9 - Condition A Results

Calibration of the Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
2	4.55	0.37	0.08	0.94	-0.90	0.85	-1.80
3	4.66	-0.17	0.09	1.25	3.60	1.22	1.90
1	4.66	-0.20	0.09	0.81	-3.10	0.82	-1.70
Mean	4.62	0.00	0.09	1.00	-0.10	0.96	-0.60
SD	0.05	0.26	0.00	0.18	2.80	0.18	1.80

Note. Presented in measure order from most severe to least severe.

Table 4.10 - Condition B Results

Calibration of the Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit *MSE*	Std. Infit	Outfit *MSE*	Std. Outfit
1	2.52	0.76	0.08	1.02	0.40	1.06	0.60
2	2.66	0.00	0.08	0.89	-2.60	0.75	-2.40
3	2.79	-0.76	0.09	1.07	1.40	1.46	2.40
Mean	2.66	0.00	.08	0.99	-0.30	1.09	0.20
SD	0.11	0.62	.00	0.07	1.70	0.29	2.00

Note. Presented in measure order from most severe to least severe.

Table 4.11 - Condition A Results

CTT Item Analysis of Condition A Results

	Difficulty	Discrimination	Correlation	Kappa	Standard Error
Tone 1	0.965	0.094	0.631	0.092	0.089
Intonation 1	0.849	0.184	0.520	0.036	0.077
Technique 1	0.873	0.148	0.489	0.255	0.093
Balance 1	0.965	0.049	0.468	0.232	0.094
Interpretation 1	0.904	0.100	0.404	0.092	0.096
Musical Effect 1	0.969	0.065	0.521	0.149	0.094
General Effect 1	0.976	0.067	0.638	0.252	0.093
Tone 2	0.935	0.139	0.660	0.290	0.093
Intonation 2	0.849	0.161	0.575	-0.039	0.080
Technique 2	0.878	0.155	0.556	0.079	0.087
Balance 2	0.953	0.088	0.592	0.242	0.088
Interpretation 2	0.918	0.124	0.578	0.162	0.093
Musical Effect 2	0.959	0.109	0.671	0.114	0.090
General Effect 2	0.969	0.109	0.764	0.387	0.089
Tone 3	0.963	0.109	0.684	0.491	0.089
Intonation 3	0.825	0.165	0.547	-0.156	0.075
Technique 3	0.888	0.100	0.515	0.127	0.091
Balance 3	0.943	0.064	0.527	0.041	0.092
Interpretation 3	0.920	0.051	0.351	0.129	0.096
Musical Effect 3	0.963	0.081	0.520	0.066	0.090
General Effect 3	0.961	0.103	0.468	0.241	0.090
Average	0.925	0.108	0.556	0.156	0.089

Table 4.12

Performance Comparison Chart

Condition A Raw Score			Condition A Rasch Score			Condition B Rasch Score		
Performance	Raw Score	Z-Score	Performance	Logit Score	Z-Score	Performance	Logit Score	Z-Score
36	309	1.21	36	4.24	1.78	36	5.42	2.99
38	309	1.21	38	4.24	1.78	38	4.32	2.13
15	306	1.00	15	3.72	1.27	37	4.08	1.94
37	306	1.00	37	3.72	1.27	15	3.39	1.40
20	305	0.93	20	3.58	1.14	3	2.80	0.94
29	304	0.87	29	3.45	1.01	1	2.70	0.86
30	303	0.80	30	3.33	0.89	27	2.60	0.78
1	302	0.73	1	3.21	0.78	33	2.34	0.58
27	300	0.59	27	2.99	0.56	26	2.26	0.52
3	299	0.52	3	2.89	0.47	6	2.17	0.45
9	299	0.52	9	2.89	0.47	14	2.12	0.41
26	299	0.52	26	2.89	0.47	29	1.95	0.27
6	297	0.39	6	2.70	0.28	13	1.80	0.16
7	297	0.39	7	2.70	0.28	4	1.46	-0.11
14	296	0.32	14	2.61	0.20	20	1.33	-0.21
31	295	0.25	31	2.52	0.11	30	1.33	-0.21
10	294	0.18	10	2.44	0.03	9	1.21	-0.30
13	294	0.18	13	2.44	0.03	7	1.14	-0.36
28	293	0.11	28	2.35	-0.06	18	1.14	-0.36
21	292	0.04	21	2.27	-0.13	34	1.14	-0.36
22	291	-0.02	22	2.19	-0.21	22	1.08	-0.41
34	290	-0.09	34	2.12	-0.28	10	0.84	-0.59
4	289	-0.16	4	2.04	-0.36	23	0.84	-0.59
19	289	-0.16	19	2.04	-0.36	11	0.77	-0.65
32	289	-0.16	32	2.04	-0.36	31	0.77	-0.65
17	287	-0.30	17	1.90	-0.49	19	0.75	-0.66
11	286	-0.37	11	1.83	-0.56	28	0.74	-0.67
33	285	-0.43	33	1.76	-0.63	32	0.65	-0.74
16	281	-0.71	16	1.50	-0.88	12	0.59	-0.79
18	281	-0.71	18	1.50	-0.88	17	0.53	-0.83
12	280	-0.78	12	1.44	-0.94	16	0.47	-0.88
23	269	-1.53	23	0.83	-1.53	21	0.41	-0.93
2	247	-3.03	2	-0.17	-2.50	2	-0.31	-1.49
24	243	-3.31	24	-0.33	-2.65	24	-0.48	-1.62

Figures

Figure 4.1 - Condition A Rating Scale (State Music Educators Association, 2017)

Criteria	Sel. 1	Sel. 2	Sel. 3	General Comments (May be continued on back)
Tone Quality Beauty, Blend Control				
Intonation Chords, Melodic Line, Tutti				
Technique Bowing, Facility, Precision, Rhythm				
Balance Ensemble, Sectional				
Interpretation Style, Phrasing, Tempo/Dynamics, Articulation				
Musical Effect Artistry, Fluency				
General Effect Choice of Music, Discipline, Instrumentation, Appearance				

Add Columns: _____+_____+_____ = _____ Total

Figure 4.2 - Condition B Rubric for String Performance Evaluation

Tone Production				
1. Tone quality in varying registers	Tone quality is poor	Tone quality is fair	Tone quality is good	Tone quality is very good
2. Consistency of attacks	Unclear attacks always detract from performance	Unclear attacks sometimes detract from the performance	Unclear attacks never detract from the performance	
3. Tone while executing expressive gestures	The execution of expressive gestures has a major negative effect on tone quality	The execution of expressive gestures has a moderate negative effect on tone quality	The execution of expressive gestures does not have a negative effect on tone quality	
4. Consistency of tone across sections	Tone quality across sections detracts very much from the performance	Tone quality across sections detracts very little from the performance		
Rhythm and Pulse Accuracy				
5. Expressive pulse and tempo fluctuations	Expressive changes in tempo and pulse are inappropriate for the style	Expressive changes in tempo and pulse are appropriate for the style		
6. Sustained notes	Notes are not consistently held for full value	Notes are consistently held for full value		
7. Precision of attacks	Attacks are rarely executed with precision across the ensemble	Attacks are sometimes executed with precision across the ensemble	Attacks are consistently executed with precision across the ensemble	
8. Consistency of articulation	Rhythmic articulations are often inconsistent with the style of music and consistently lack ensemble uniformity	Rhythmic articulations are occasionally inconsistent with the style of music and sometimes lack ensemble uniformity	Rhythmic articulations are consistent with style of music and maintain ensemble uniformity	
9. Consistency of rhythmic stress	Rhythmic stress does not effectively communicate proper musical style	Rhythmic stress somewhat effectively communicates proper musical style	Rhythmic stress effectively communicates proper musical style	
10. Steadiness of pulse	A lack of steady pulse detracts much from the continuous flow of the music	Wavering steady pulse sometimes detracts from the continuous flow of the music	Control of steady pulse does not detract from the continuous flow of the music	
11. Appropriateness of tempo in technical passages	Tempo fluctuations during technical passages are a serious problem	Tempo fluctuations during technical passages are a moderate problem	Tempo fluctuations during technical passages are not at all a problem	

12. Subdivision of the rhythm	Inaccurate performance of subdivisions frequently detracts from solidly communicated tempo and meter	Inaccurate performance of subdivisions occasionally detracts from solidly communicated tempo and meter		Accurate performance of subdivisions contribute to solidly communicated tempo and meter
Intonation Accuracy				
13. Intonation of cadential points	Cadential points are rarely in tune	Cadential points are occasionally in tune		Cadential points are consistently in tune
14. Centered pitch	The pitch is rarely centered	The pitch is occasionally centered		The pitch is centered a great deal of the time
15. Overall intonation accuracy	Maintaining consistently good intonation in all registers is a serious problem during performance	Maintaining consistently good intonation in all registers is a moderate problem during performance		Maintaining consistently good intonation in all registers is not a problem during performance
16. Pitch adjustments	It is rarely evident that players are able to accurately and quickly adjust pitch when necessary	It is sometimes evident that players are able to accurately and quickly adjust pitch when necessary		It is frequently evident that players are able to accurately and quickly adjust pitch when necessary
17. Half step intonation	Half step intonation is unacceptable	Half step intonation is slightly unacceptable		Half step intonation is perfectly acceptable
18. Chromatic alterations intonation	Chromatic alterations are rarely in tune	Chromatic alterations are sometimes in tune		Chromatic alterations are consistently in tune
19. Presence of wrong notes	Wrong notes detract from the performance a great deal	Wrong notes occasionally detract from the performance		Wrong notes do not detract from the performance
20. Open string intonation	Out of tune open strings is a serious problem	Out of tune open strings is a moderate problem	Out of tune open strings is a minor problem	Out of tune open strings is not at all a problem
21. Intonation in technical passages	Intonation fluctuations during technical passages are a serious problem	Intonation fluctuations during technical passages are a moderate problem		Intonation fluctuations during technical passages are not at all a problem
Expressive Qualities/Stylistic Interpretation				
22. Presence of crescendo and diminuendo	Crescendo and diminuendo are not at all influential on effective expression	Crescendo and diminuendo are somewhat influential on effective expression		Crescendo and diminuendo are extremely influential on effective expression
23. Balance between melody and accompaniment	Balance between melody and accompaniment is undesirable	Balance between melody and accompaniment is desirable		Balance between melody and accompaniment is very desirable

24. Stylistically appropriate articulations	Stylistically appropriate articulations are never evident	Stylistically appropriate articulations are sometimes evident	Stylistically appropriate articulations are always evident	
25. Connection of phrases	Ensemble does not meaningfully connect phrases	Ensemble meaningfully connects phrases		
26. Articulation	Articulations are inconsistent in passages with notes of a similar style, resulting in a very dissatisfactory performance	Articulations are often inconsistent in passages with notes of a similar style, resulting in a dissatisfactory performance	Articulations are sometimes inconsistent in passages with notes of a similar style, resulting in a satisfactory performance	Articulations are consistent in passages with notes of a similar style, resulting in a highly satisfactory performance
27. Contrast in dynamics	Dynamic contrasts are never evident	Dynamic contrasts are almost never evident	Dynamic contrasts are sometimes evident	Dynamic contrasts are frequently evident
28. Expressive modifications (>, sfz., rit., ten., cantabile)	Stylistic or expressive modifications are rarely appropriate or present in performance	Stylistic or expressive modifications are typically appropriate and somewhat present in performance	Stylistic or expressive modifications are appropriate and consistently present in performance.	

Figure 4.3 - Variable Map

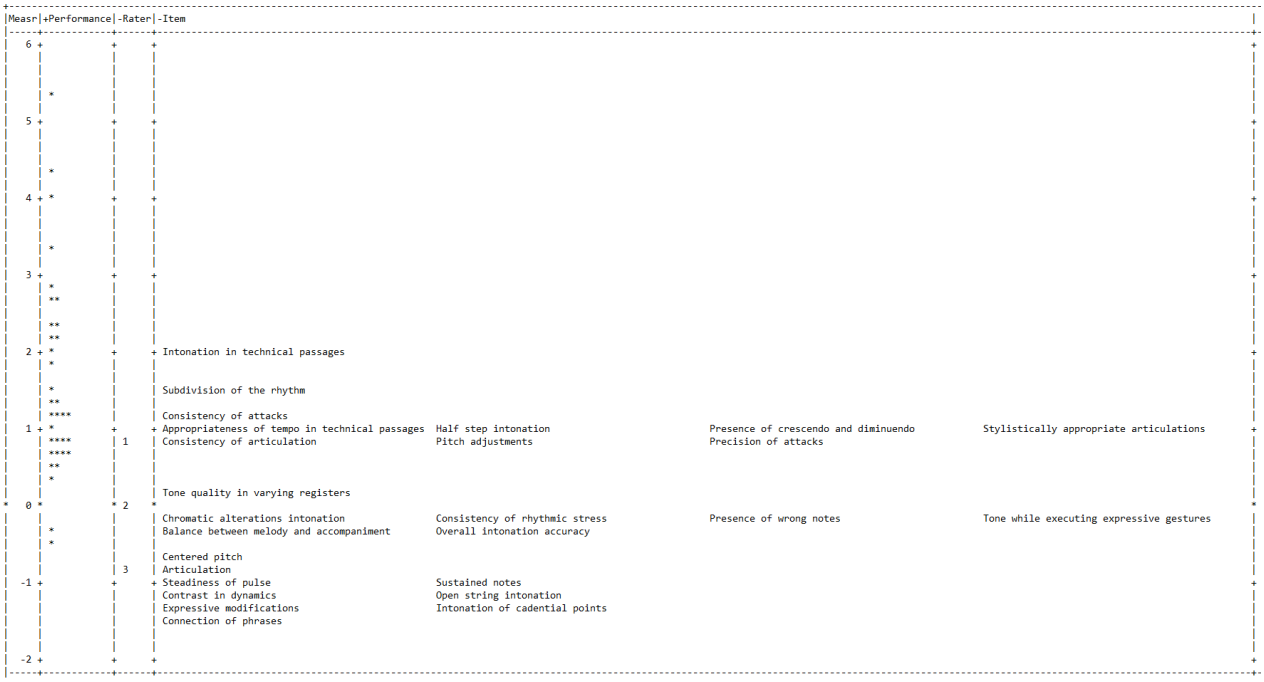


Figure 4.4

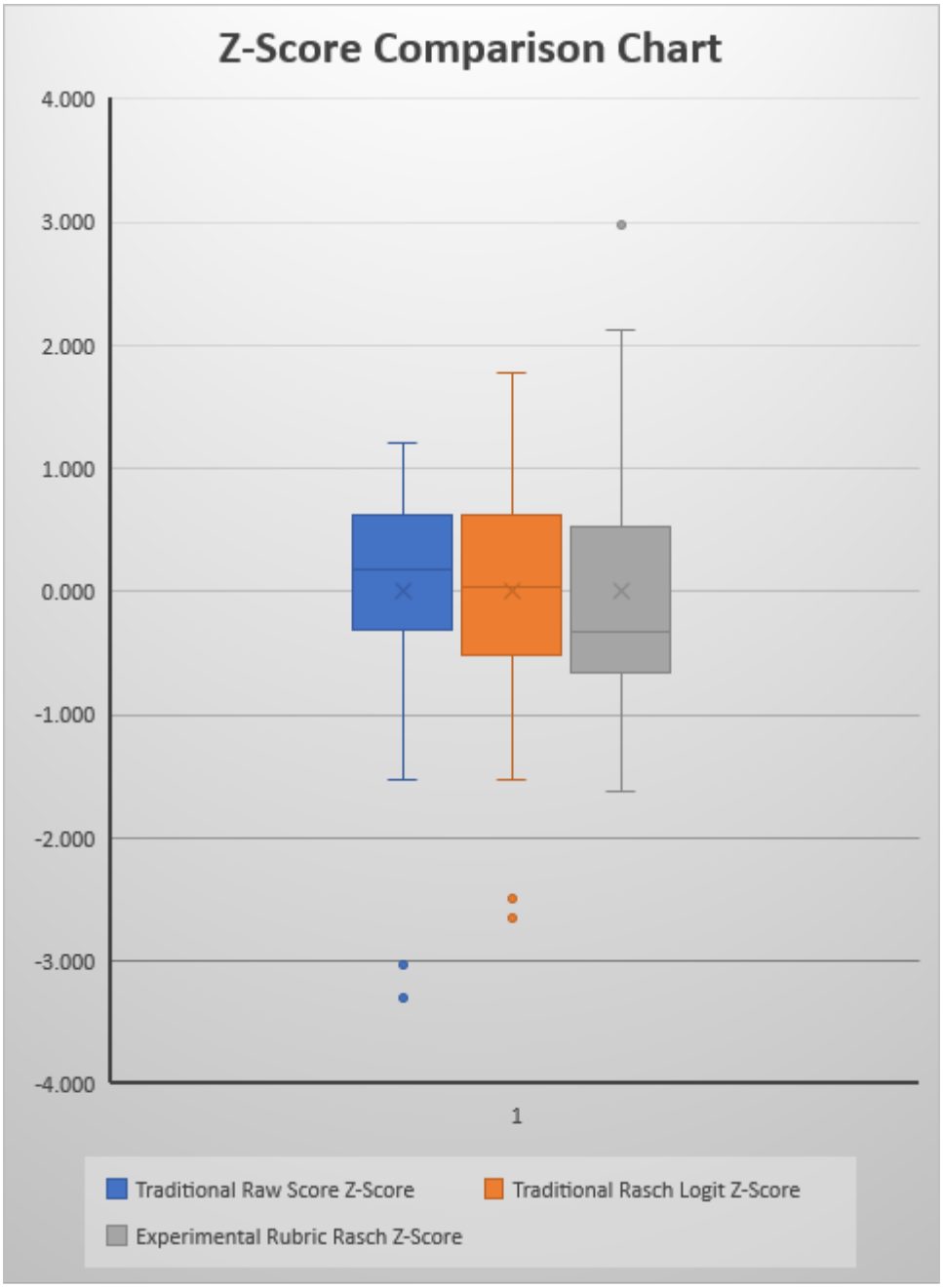
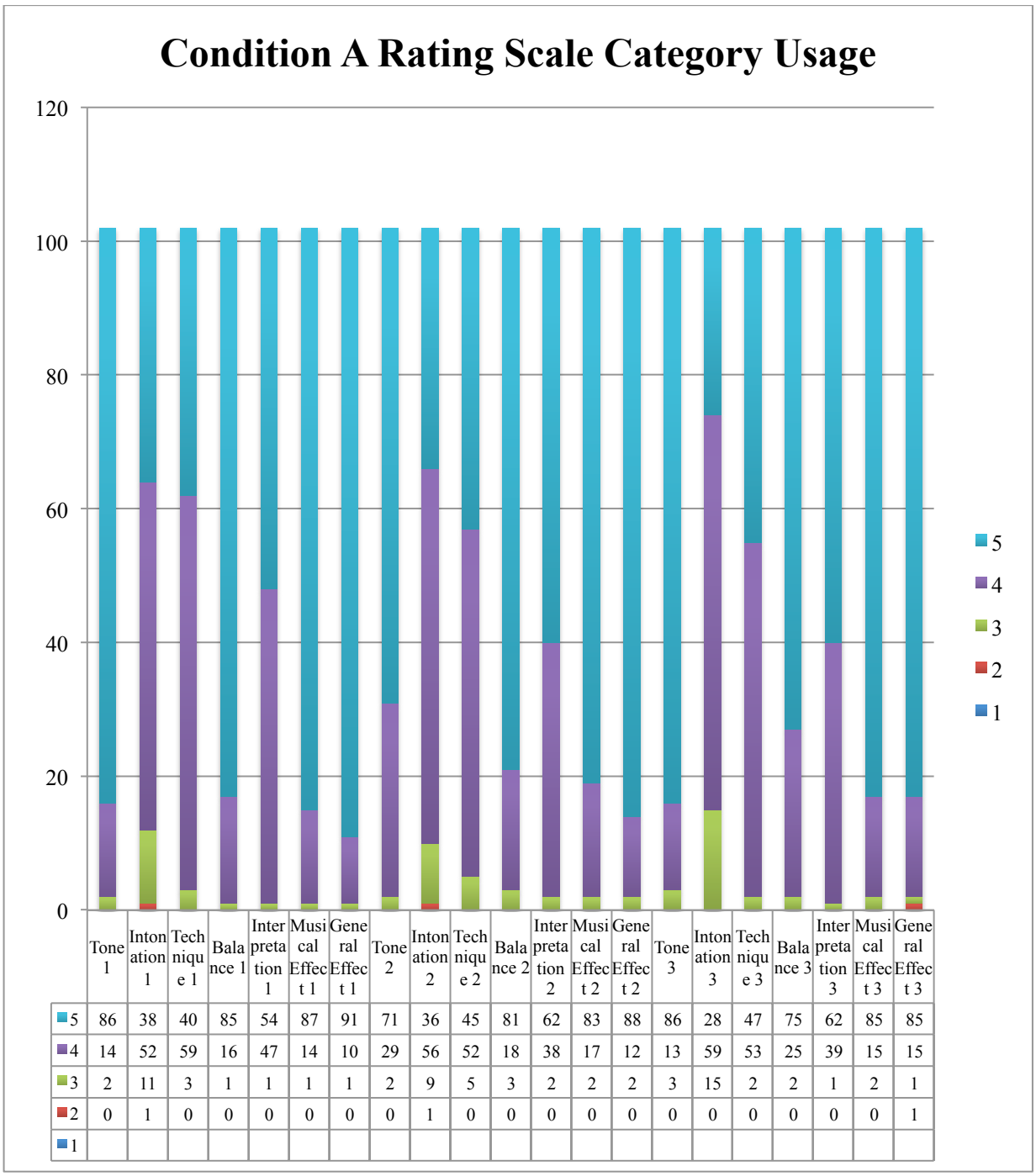
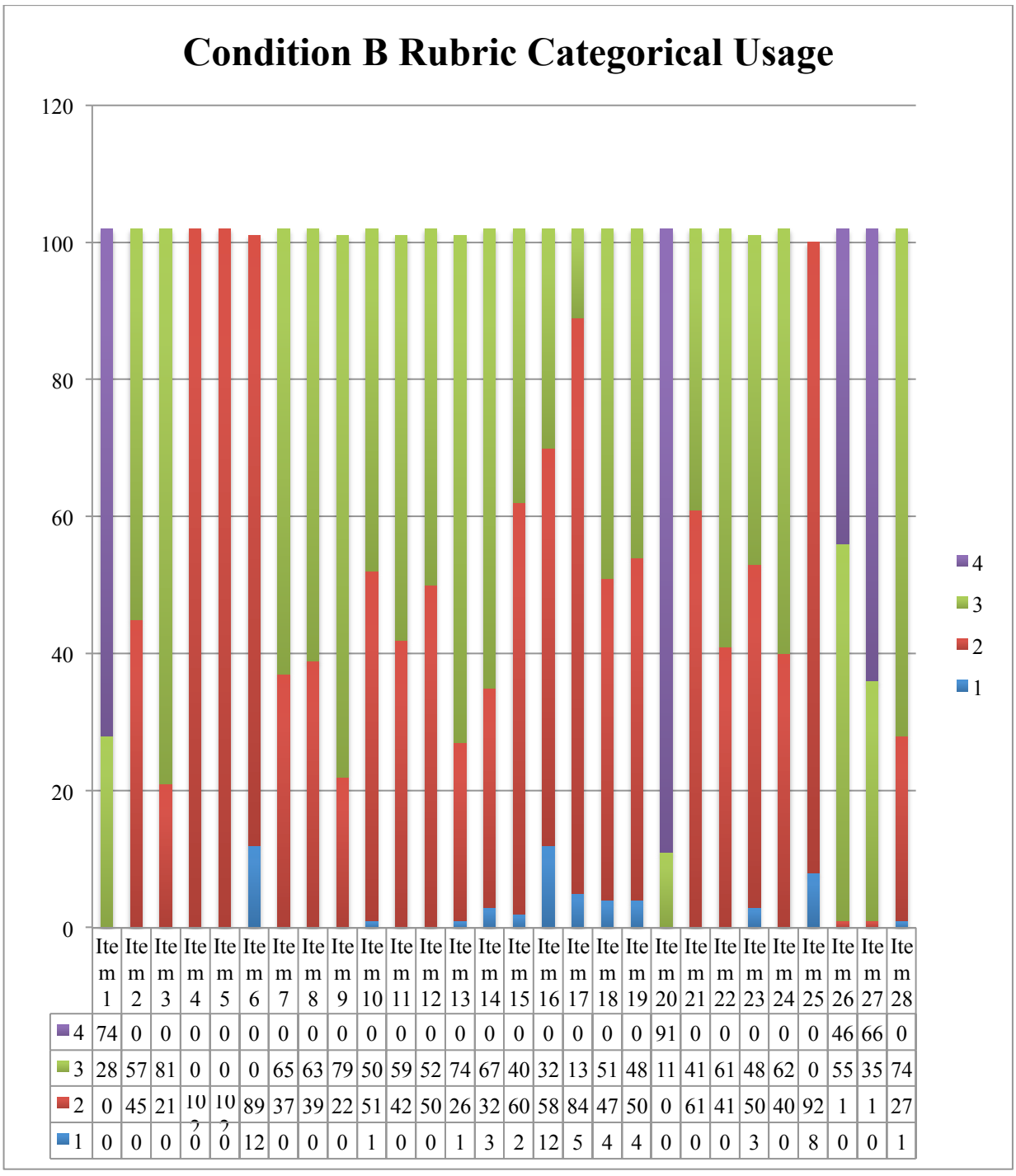

Z-Score Comparison Chart
4.000
3.000
2.000
1.000
0.000
-1.000
-2.000
-3.000
-4.000
1
Traditional Raw Score Z-Score
Traditional Rasch Logit Z-Score
Experimental Rubric Rasch Z-Score

Figure 4.5



	Tone 1	Intonation 1	Technique 1	Balance 1	Interpretation 1	Musical Effect 1	General Effect 1	Tone 2	Intonation 2	Technique 2	Balance 2	Interpretation 2	Musical Effect 2	General Effect 2	Tone 3	Intonation 3	Technique 3	Balance 3	Interpretation 3	Musical Effect 3	General Effect 3
5	86	38	40	85	54	87	91	71	36	45	81	62	83	88	86	28	47	75	62	85	85
4	14	52	59	16	47	14	10	29	56	52	18	38	17	12	13	59	53	25	39	15	15
3	2	11	3	1	1	1	1	2	9	5	3	2	2	2	3	15	2	2	1	2	1
2	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
1																					

Figure 4.6



	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19	Item 20	Item 21	Item 22	Item 23	Item 24	Item 25	Item 26	Item 27	Item 28
4	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91	0	0	0	0	0	46	66	0
3	28	57	81	0	0	0	65	63	79	50	59	52	74	67	40	32	13	51	48	11	41	61	48	62	0	55	35	74
2	0	45	21	102	102	89	37	39	22	51	42	50	26	32	60	58	84	47	50	0	61	41	50	40	92	1	1	27
1	0	0	0	0	0	12	0	0	0	1	0	0	1	3	2	12	5	4	4	0	0	0	3	0	8	0	0	1

CHAPTER 5

CONCLUSION

These papers discuss the nature of assessment in the music performance setting. The second chapter posits that assessment in the music classroom and large ensemble performance setting should not be used to advocate, but should instead be primarily used for the purpose of providing empirical evidence of student growth. The third chapter explains the process involved in the development of a new rubric. The intention was to use the new method for performance evaluation in order to measure student performance achievement in a valid and reliable manner. The fourth chapter then shows how the new (condition B) rubric operates in comparison to the current (condition A) system.

The second chapter presents the idea of silent advocacy for performing arts programs. This notion challenges music educators to let their programs speak for themselves in terms of how well the students can perform and how merely being a member of the program can benefit their preparation for being citizens in society. A limitation of the practitioner manuscript in the second chapter is that no real solution for a music performance assessment is actually recommended. A set of beliefs is only presented, coupled with a challenge to advocate silently without using assessment results as the only reason for keeping music programs.

The limitations for the second chapter are resolved in the third chapter as the rubric is developed based on the beliefs presented in the first article. The rubric is an authentic assessment that can be used to show student growth in a way that will ultimately helping to improve teaching and learning. The performance-based rubric allows for music educators to continue to

teach and assess in a manner that fits the nature of the music classroom. Instruction can be properly aligned with the nature of the assessment, therefore not compromising the nature of the discipline. Clear indicators are utilized in a way that makes it possible for both those in the music education profession and others outside of the profession to understand how we are being effective in the music classroom. The rubric provides the means for important communication and feedback between teacher and students, parents, administrators, policy makers, and other stakeholders.

The third chapter provides details on how the rubric has been validated. However, there are two primary limitations that further research could help to resolve. Some of the items from the rubric contain gaps, particularly in how exemplary tone production for string ensemble performances is described. Improvements would require additional writing of stems and evaluation of performances in order to validate an updated version of the rubric. A second limitation is that the content experts that were used to help develop the rubric were recruited using convenience sampling. The content experts were music educators in the area that were willing to listen to and evaluate four recordings in order to validate the items. Further research might allow for more strict guidelines in terms of how the content experts are chosen and recruited to listen to and rate the performances.

Despite the limitations, results from the fourth chapter demonstrate that the rubric can be used in the current large ensemble performance evaluation setting with relatively few changes in terms of how the feedback will be shared with directors and students. Based on the metrics gleaned form the evaluation, the rubric seems to be proven valid, reliable, and usable. Furthermore, because the metrics fall within the high stakes standards for the Rasch Measurement Model, invariant measurement can be assumed. This allows the results to be

considered sample independent. Thus, the data that is gleaned from the use of this rubric can be used to inform future evaluations (Engelhard, 2013). Further research and development would allow for potential standard setting, meaning the specific measures provided for each performance could qualify it for one of the traditional performance labels such as Superior, Excellent, Good, Fair or Poor.

The development of the condition B rubric and the comparison made in relation to the condition A rating scale serves as a catalyst for future research and development that can be used to further refine the performance evaluation process. As with the third chapter, there is a need for replication of this study in order to help minimize some of the above-mentioned limitations in addition to the limitations that surface in the fourth chapter. The side-by-side comparison only included one large ensemble performance evaluation site, in an area that has a relatively consistent socioeconomic status throughout. Furthermore, the particular area where the rubric was piloted contains many very strong string ensembles. Further replication would allow for the opportunity to see how the condition B rubric would operate in an area with more variation in terms of both socioeconomic status and the level to which the string large ensemble perform. Lastly, there was a small pool of raters that used the condition B rubric in the actual large ensemble performance setting. A further complication lies in the fact that the there were two sets of raters and each set used a different evaluation system: either the condition A or the condition B rubric. Therefore, comparison of the evaluation results is slightly tainted because one set of raters did not use both systems when rating the performances in order to provide for a more sound opportunity for true comparison. Further replication would allow for more data to be explored in terms of how effective the condition B rubric is for music performance assessment, when used by a consistent group of raters.

The primary goal of the three papers is to advocate for an authentic music performance assessment that can be used to accurately rate large ensemble string performances, while also having the potential to be used to inform instructional decisions made in the music classroom. Even though the rubric was initially developed using large ensemble recordings and is primarily intended for large ensemble performance evaluation, it can also be used in the classroom setting to guide curricular settings. Because it was developed under the strict conditions of large ensemble performance evaluation, it can be easily used in the less restrictive assessment environment in the classroom. Further research and development would allow for this assessment to be used in conjunction with a written classroom assessment that would help to confirm that students also understand the music concepts behind the performance skills that they are demonstrating.

Continued study must involve experts that are willing to collaborate in order to develop a valid, reliable, and useable rubric for music performance assessment. This will require the involvement of musicians, music educators, statisticians, administrators, and policy makers. The development of an assessment for the music performance setting should be an ongoing process with continued research, development, implementation, trial and error, adjustments, and reflection. We should aim to develop and use assessments for the purpose of improving student achievement, while being sure to not impede or compromise the aesthetic nature and other immeasurable benefits of music education as described by Oxley (1996), "all music students show us that music is a force for healing, building community and transcending obstacles. Music is a tool in learning living and finding life for both student *and* teacher" (p. 21).

REFERENCES

Asmus, E. P. (1999). Music assessment concepts. *Music Educators Journal*, *86*(2), 19–24.

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

Brewer, C., Knoeppel, R. C., & Lindle, J. C. (2014). Consequential validity of accountability policy: public understanding of assessments. *Educational Policy*, *29*(5), 711–745.

Brookhart, S. M. (2013). The public understanding of assessment in educational reform in the United States. *Oxford Review of Education*, *39*(1), 52–71.

Colwell, R. (1970). *The evaluation of music teaching and learning.* Engelwood Cliffs, NJ: Prentice-Hall.

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

Fisher, R. (2008). Debating assessment in music education. *Research And Issues In Music Education*, 6(1),

Hope, S., & Wait, M. (2013). Assessment on our own terms. *Arts Education Policy Review*, *114*(1), 2–12.

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

Mark, M., & Gary, C. (2007). *A history of American music education.* Lanham: Rowman & Littlefield.

Messick, S. (1989). Validity. In R.L. Linn (Eds.), *Educational Measurement*, 3rd ed. (pp. 13–103). New York, NY, England: Macmillan Publishing Co, Inc.

Morrison, S. J. (1994). Music students and academic growth: Steven J. Morrison finds that music students generally do well in the areas of academics and student leadership. *Music Educators Journal*, *81*(2), 33–36.

Oxley, M. (1996). Students and stories. *American Music Teacher*, *45*(6), 18–21.

Pellegrino, K., Conway, C. M., & Russell, J. A. (2015). Assessment in performance-based secondary music classes. *Music Educators Journal*, *102*(1), 48–55.

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

Wesolowski, B. C. (2012). Understanding and developing rubrics for music performance assessment. *Music Educators Journal*, *98*(3), 36–42.

Wesolowski, B. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal*, *101*(1), 77. doi:10.1177/0027432114540475

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016a). Rater analyses in music performance assessment : Application of the many facet Rasch model. In *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education* (pp. 335–356).

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

Zaleski, D. J. (2014). An introduction to classroom assessment for today's music educator. *Illinois Music Educator*, *75*(1), 58.

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, *50*(3), 245.

Appendix A – CRME Consent to Publish

UNIVERSITY OF ILLINOIS PRESS

FROM:
Bulletin of the Council for Research in Music Education
c/o University of Illinois Press
1325 South Oak Street
Champaign, Illinois 61820-6903

TO:

The University of Illinois Press is pleased to have the privilege of publishing your contribution entitled:

In *Bulletin of the Council for Research in Music Education* issue

So that you as Author and we as Publishers of the Journal may be protected from the consequences of unauthorized use of its contents, we consider it essential to secure a copyright. To this end, we ask you to grant us rights to the article as described below.

1. CONSENT TO PUBLISH: Whereas the University of Illinois Press is publisher of the Journal above named and the undersigned is Author of one or more parts, the Author grants and assigns exclusively to the Board of Trustees of the University of Illinois for its use any and all rights of whatsoever kind or nature now or hereafter protected by the Copyright Laws of the United States and all foreign countries in all languages in and to the above article, including all subsidiary rights. The Press, in turn, as part of the consideration of this agreement, grants to the Author the right of republication in any book of which he/she is the author, subject only to his/her giving proper credit in the book to the original publication of the article by the University of Illinois Press.

2. PREVIOUS PUBLICATION: The Author guarantees that the article furnished for the Journal has not been published previously elsewhere. If any part of the contribution (including illustrations, examples, and other supplementary materials) has been previously published elsewhere, permission has been obtained in writing for publication in the Journal and the Author will submit copy for credit lines with his/her manuscript. The Author holds the Editor, the Journal, and the University of Illinois Press harmless against all copyright claims.

3. PROOFREADING: The Author will be given an opportunity to read and correct manuscript and/or page proof, but if he/she fails to return corrections by the deadline(s), production and publication may proceed without the Author's approval.

4. SUBSIDIARY RIGHTS AND COMPENSATION: The Author will receive no payment from the University of Illinois for the use of his/her materials, but should the Press receive any requests to reproduce or translate all or any portion of the Author's article, the Press will attempt to obtain the Author's approval for the requested use. Should any fee be charged for this use, monies collected over $50.00 (US) will be divided between the Author (40%) and the Press (60%). Fees collected by the Press will be paid to the Author within three months after the end of the Press's fiscal year.

FOR THE UNIVERSITY OF ILLINOIS PRESS:

Laurie C. Matheson, Director

ACCEPTED AND APPROVED:

Kinsey E. Edwards
Author (signature must be by hand, not electronic)

Date: 5/20/17

Permanent Address:
5741 Crest Hill Drive
Buford, GA 30518

E-mail Address: kinsed@uga.edu

Office #:

Home/mobile #: 770-722-3482

RETURN THIS FORM TO:

Janet Barrett. Editor
Bulletin of the Council for Research in Education
School of Music
University of Illinois at Urbana-Champaign
1114 W. Nevada Street
Urbana, IL 61801

Alternately you may send a scan of a printed form with your original (**not electronic**) signature to crme@illinois.edu.