THE DEVELOPMENT AND VALIDATION OF INSTRUMENT-SPECIFIC RATING SCALES FOR SECONDARY-LEVEL INSTRUMENTAL MUSIC PERFORMANCE ASSESSMENT

By

WESLEY R YORK

(Under the Direction of Brian C. Wesolowski)

ABSTRACT

The purpose of this study is to develop and validate instrument-specific rating scales to evaluate the classroom performances of middle and high school students. The study is guided by the following research questions:

1. What does Rasch measurement analysis reveal about the psychometric properties (i.e., validity and reliability) of items, raters, and performers in the context of solo music performance assessment?
2. How do items vary in difficulty, raters in severity, and performers in achievement?
3. How does the rating scale structure vary across raters and performers?

A total of 160 secondary level student performances will be recorded and evaluated by music content experts (*N* = 40). Data will be analyzed using the Many-Facet Rasch Partial Credit model to determine if the requirements of invariant measurement are met, resulting in a set of music performance rating scales that may improve teaching and learning in the instrumental music classroom.

INDEX WORDS: music education, assessment, performance evaluation, invariant measurement, Rasch, rating scale

THE DEVELOPMENT AND VALIDATION OF INSTRUMENT-SPECIFIC RATING SCALES FOR SECONDARY-LEVEL INSTRUMENTAL MUSIC PERFORMANCE ASSESSMENT

By

WESLEY RALPH YORK
B.Mus., The University of Georgia, 2003
M.M.E., The University of Georgia, 2008

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

2022

THE DEVELOPMENT AND VALIDATION OF INSTRUMENT-SPECIFIC RATING SCALES FOR SECONDARY-LEVEL INSTRUMENTAL MUSIC PERFORMANCE ASSESSMENT

By

WESLEY RALPH YORK

Major Professor:	Brian C. Wesolowski
Committee:	Jaclyn Hartenberger Michael Robinson

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2022

DEDICATION

I dedicate this study to all my students – past, present, and future – including my own sons: Ashton, Collin, and Embry. Having you in my classroom has been far and away the greatest joy of my professional career. Seeing you walk into the band room each day inspires me to give my best.

ACKNOWLEDGEMENTS

I would like to thank Dr. Brian Wesolowski for encouraging me to pursue this study and for his constant example of setting high professional standards. I am grateful to Dr. Jaclyn Hartenberger for her example of exceptional musicianship as well as her incredible joy and passion for teaching. As a graduate student in music education, I am so encouraged to see Dr. Michael Robinson teaching and guiding future music educators. I am thankful for Dr. Myriam Athanas and Ms. Jing Li who helped me with data analysis. I would also like to thank my former band directors Richard Brasco, Reid Hall, Elisa Hunt, Kevin Baggott, Katie Bennett, and Steve Pruitt for inspiring me to teach.

TABLE OF CONTENTS

Page

CHAPTER

APPENDICES

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

Public schools in the United States operate under systems of accountability based largely upon data from standardized test scores (Davison & Fischer, 2019). Standardized test data is used as a primary factor in any number of decisions, affecting everything from state and federal funding to yearly teacher performance evaluations (Brewer et al., 2014). As the demand for data has increased, so has the scrutiny placed on assessment measures used to gather the data (Hopkins et al., 1985). Developers of high stakes tests must regularly produce data to support the reliability and validity of their assessment tools (Hughes et al., 2019). Assessments in non-tested or performance-based subjects however, such as music, are largely lacking in assessment measures that meet the same psychometric standards as those used in core subjects such as mathematics (Kelly et al., 2019).

Stakeholders, such as parents and administrators, often make value judgements on the quality of school music programs based on public performances (Asmus, 1999). While students appear to be learning, as evidenced by high quality performing ensembles in schools, there are potentially missed opportunities for improving individual student learning and understanding (Tindal & Marston, 1990). What is missing in many of these programs is a systematic focus on individual student accountability and achievement (Russell, 2010).

The lack of meaningful assessment may have a negative effect on the student, teacher, and ultimately the instrumental music program as a whole (Asmus, 1999). Assessment is a primary mechanism for teachers to communicate achievement levels and expectations to students and parents (Andrade, 2013). Often, a student's band grade in secondary schools throughout the United States is only a reflection of their daily participation in class or their attendance at after school activities such as concerts (Goolsby, 1999). A student's grade in band, just like any other subject, should communicate their progress toward real learning goals (Goolsby). The unique challenges that are presented in a typical beginning band classroom (i.e., student to teacher ratio, heterogeneous instrument groupings, administrative tasks) can make grading a difficult task. In many cases, students are in a class of heterogeneous instrument groupings with one teacher (Kimpton 2019). Students exiting music education programs are well trained in leading large ensemble classes, but perhaps still lacking in instrument-specific knowledge. The teacher is also expected to perform the administrative tasks required to run a band program, when they often teach more students than any other teachers at their school (Cooper, 2004). Kimpton refers to this condition as the "performance treadmill."

> Music classes often revolve around performance activities, having to "cover" pieces to fill a concert; we call this the performance treadmill where music teachers haphazardly select music, superficially fix mistakes, play a concert, and then race to prepare for the next performance. Unfortunately, this way of teaching leaves little room for bigger ideas to guide teaching and no plan to measure and assess in an effort to ensure students are learning. It is our belief that music educators must jump off the treadmill and commit to purposeful instructional design, and that begins with developing quality assessments. (p. 325)

As a result of these and other factors, individual performance assessment can fall lower on a teacher's list of priorities (Pellegrino et al., 2015). Band directors often assign grades based on participation (Lehman, 1998). It is much easier to quickly assess whether or not a student is prepared with their instrument, music, and other necessary materials, or to gauge their level of participation, than it is to determine and document proficiency. A grading system built primarily on class participation fails to communicate to students and parents levels of student achievement and understanding (Shepard, 2006). Parents appreciate honest communication about their child's level of achievement, and student success is hindered when this communication is absent (Cooper, 2004). Sadler (2015) states:

> The pure concept of achievement has become corrupted by the practice of adding into a total mark, score or grade a variety of elements that are not part of achievement itself. This includes credits for participation in various learning activities (such as class discussions, contributions to online forums, and journals or logs of activities engaged in). These may all assist learning to occur, but they themselves are not the learning. (p. 11)

Grades lose meaning without accurate measures of student performance (Andrade, 2013). Shepard suggests that the inclusion of non-achievement factors in a student's grade not only undermines the validity of the grade, but it also causes inequities among students' grades because factors such as effort and attitude are not accurately measured. This practice harms the credibility of the teacher and the band program (Asmus, 1999).

In recent years, experts in the field of educational measurement have recognized the need for further investigation into classroom assessment, not only in the music classroom, but throughout all subject areas (Shepard, 2006). For much of the twentieth century the primary focus of research on assessment was placed on standardized tests, particularly as large-scale testing and accountability became the norm in the 1980's and 1990's. The broad implementation of high stakes testing required new methods of validating individual test scores as well as the inferences drawn about testing populations (Koretz & Hamilton, 2006). This also increased demand for studies on the impact of standardized testing on teacher preparation and instructional strategies, as well as their competence in assessment (Cohen, 1995; Haertel, 1999).

Another reason for a lack of interest in CA is that the education community has historically believed assessment development is better left to experts, and teachers should either not be burdened with test creation or they were underprepared to do so (Cook, 1951; McMilan, 2013; Stiggins, 1991). Much of the impetus for change has derived from studies performed by subject-matter experts searching for alternatives to standardized tests (Shepard, 2000). Research of this type reaffirms the crucial role of the classroom teacher and the importance of daily formative-type observations taken throughout the course of routine classroom activities (Goodman, 1985). The sum of these observations and interactions often proves more valuable to teachers in assessing student ability than the results of a single test. In the case of reading ability, for example, such classroom observations are found to be more effective than standardized placement tests in determining students' developmental level (Clay, 1985).

Another factor of evolving classroom assessment is a wide acceptance of constructivist learning theory during the last half of the 20th century (Shepard, 2006). Constructivism was developed by Vygotsky in the early 20th century and suggests that learners construct knowledge by combining what they already know with new learning experiences in the classroom (Bruning et al., 2004). Skinner's behaviorism was the prevailing theory of learning throughout the mid-20th century and promoted a reliance on conditioning students' behavior through a system of stimulus and response (Woolfolk, 2001). Constructivism differs from the widely held behaviorist views of the time in that emphasis is placed on ongoing interactions and experiences in the classroom as a way to gain and demonstrate knowledge, as opposed to students simply achieving at a certain level on an assignment or test (Bruning et al). The constructivist approach to learning also places greater emphasis on authenticity in assessment, meaning the assessment should mirror a real-life situation as closely as possible (Wiggins, 1993). Researchers agree that performance assessments are far preferable to selected-response type assessments when it comes to maximizing authenticity (Johnson et al., 2009). This renewed interest in performance assessment also satisfies the need for more direct assessments of students' high-level reasoning skills (Lane & Stone, 2006). For these reasons, the concept of performance assessment in general has played a large role in education reform beginning in the late 20th century (Lane, 2013).

Assessment in the music classroom takes the form of performance assessment (Asmus, 1999). Performance assessments in music, as opposed to selected-response items or writing formats, are valued for their authenticity because they require students to demonstrate evidence of learning through action. (Johnson et al., 2009). Performance assessments can "assess important learning outcomes that cannot be assessed by selected-response item formats and...may require students to carry out a complex, extended process such as play(ing) a musical instrument. (Lane & Stone, 2006, p. 387)." Performance assessments are more closely related to the construct being assessed and therefore provide a more accurate measure of student achievement than traditional tests, such as…(Lane & Stone). Motivation to learn also increases when students view their work as meaningful (Woolfolk). Use of performance assessments promotes a deeper student understanding of content (Lane, 2013).

When determining the level of authenticity of a particular performance assessment, researchers often speak in terms of meaningfulness (Johnson et al., 2009). Kimpton (2019) states that meaningful assessments "(1) emerge purposeful instruction, (2) engage students in authentic musical work, and (3) demand high-level musicianship (p. 331)." Performance assessment is also meaningful when the task requires complex cognitive processes, which can increase the level of motivation for the learner (Johnson et al). Performance assessment is now recognized as a factor in promoting complex and advanced student learning (McMillan, 2013).

A lack of systematic performance assessment coupled with an absence of standardized testing in the arts has proven to be detrimental to music educators when it comes to professional evaluations. Legislation mandating standardized tests focuses solely on high-stakes subjects such as math and science (Mark, 1996). This puts pressure on those teachers to ensure individuals are progressing in their understanding of concepts and will be prepared to perform well on standardized tests. Most states do not utilize standardized tests in music, and the few that do only offer them as an optional assessment (McCaffrey & Lovins, 2019). Band directors, and music teachers in general, are not held to the level of accountability as teachers of high stakes subjects when it comes to documenting individual student learning (Russell, 2010). Some music educators see this as a benefit because they have more latitude in determining curricular goals for their students without the pressure of teaching toward a standardized test (Pellegrino et al., 2015). However, they are not meeting their professional obligations if they fail to measure student learning.

Moreover, student performance on standardized tests is included as a component of teacher evaluations (Shuler, 2012). States have begun to include student achievement data as criteria for teacher evaluations in order to receive funding from the American Recovery and Reinvestment Act (Davison & Fisher, 2019). This is done to motivate teachers to ensure students demonstrate adequate growth from year to year, which is one of the four core reform areas of the ARRA legislation. Teachers' yearly performance reviews are impacted by their students' standardized test scores in their subject. With no standardized music test scores to consider, music teacher evaluations are often impacted by the average of all standardized test scores by all students in all subjects in their school, whether or not they have ever taught the students (Shuler).

Solo Music Performance Assessment

Studies of solo music performance assessment (MPA) date back to 1942 with the Waktins cornet performance scale, and the resulting Watkins-Farnum Performance Scale. This scale was widely used throughout the latter half of the 20th century, and is considered to be the first systematic research effort in solo music performance assessment (Zdzinski, 1991). Another landmark study is Abeles' (1971) research on clarinet performance assessment, as he introduced the facet-factorial method into MPA research. Numerous other researchers followed this method, including Bergee (1987) and Zdzinski (2002). The facet-factorial approach of scale item construction became more commonly used as a way to identify both observable and unobservable constructs used to evaluate music performance. The benefit, as evidenced in the Abeles study, is improved content validity and interjudge reliability (Zdzinski, 1991).

Despite researchers' best efforts to construct valid and reliable measurement scales, there have been limitations due to the unique nature of music performance and the difficulty in measuring a latent, or unobservable, variable (Wesolowski et al., 2016). Significant developments in the behavioral sciences, specifically the use of the Multifaceted Rasch Partial Credit Model (MFR-PC), have allowed researchers to more precisely measure latent variables (Masters & Keeves, 1999). Through the MFR-PC method of data analysis, researchers can account for multiple variables independently of one another, including the performer, item difficulty, and rater severity.

Research Problem

This study will address two issues in music performance assessment. The first is the lack of research-based performance assessments developed specifically for use in the instrumental music classroom. The second is the lack of specificity in student performance assessment and feedback. There are many successful studies that have resulted in research-based assessment tools (Abeles, 1973; Bergee, 1987; Russell, 2010; Zdzinski & Barnes, 2002). The scope of those studies has been limited to one specific instrument or instrument group. The goal of this current study is to expand upon those previous studies and apply the principles to include several of the wind instruments common to public school band classes.

The purpose of this study is to develop a suite of instrument-specific performance rating scales that could be used for performance assessment in the classroom, as well as improve the audition experience and results for competitive region and state level band events such as solo and ensemble festivals, region honor bands, and all-state bands. Band directors and students throughout the United States prioritize competitive events such as these, but current scoring methods and student feedback are often lacking in specificity (Asums, 1999; Keene, 2009). An instrument-specific performance rating scale could not only result in more accurate placement results in competitive events, but also communicate achievement level more clearly to students and directors (Shepard, 2006; DeLuca & Bolden, 2014).

The questions that guided this study include:

1. What does Rasch measurement analysis reveal about the psychometric properties (i.e., validity and reliability) of items, raters, and performers in the context of solo music performance assessment?

2. How do items vary in difficulty, raters in severity, and performers in achievement?
3. How does the rating scale structure vary across raters and performers?

Instrument-specific performance rating scales will be developed for flute, oboe, clarinet, saxophone, trumpet, french horn, trombone, euphonium, and tuba. Rating scales will be constructed using a facet-factorial approach and will include items sourced from a variety of pedagogical literature as well as existing score sheets from around the United States. Each scale will then be evaluated for content validity by an accomplished performer and teacher who specializes in that specific instrument. Video performances will be collected from middle and high school students performing short excerpts to be assessed using the rating scales. Content experts will be recruited to view the videos and assess each student's performance using the rating scales. Data will be analyzed using the Multifaceted Rasch Partial Credit Model, and response functions for items, persons, and raters were analyzed for data-model fit. If the requirements of invariance are met, this will ultimately yield valid, research-based performance rating scales for each wind instrument typically found in public secondary school band programs (Bond & Fox, 2020).

CHAPTER 2

Classroom assessment permeates all aspects of teaching and learning. In undertaking a study of classroom assessment in the instrumental music classroom, this literature review will cover a broad range of topics including trends in classroom assessment, federal education policy as it relates to assessment, national and state music assessment tools, current CA practice in music, music performance assessment studies of the past, and finally how advances in the field of measurement can inform future MPA studies. The literature reviewed here will prove CA as a worthy research topic in music, but will also highlight the disconnect between policy and practice when it comes to CA in the music classroom. It will also explore the potential benefits of applying the concepts of item response theory and invariant measurement to current and future MPA studies.

This study is undertaken as a form of action research, with the researcher having identified a problem in his own classroom, and in the classrooms of colleagues, that needs to be addressed. The term "action research" was first used in the 1940's by Kurt Lewin (Willis & Edwards, 2014). Lewin defined action research as any study that did not separate the investigation from the action needed to solve the problem (McFarland & Stansell, 1993). Action research is based on the following four assertions by Watts (1985):

1. Teachers and principals work best on problems they have identified for themselves.
2. Teachers and principals become more effective when encouraged to examine and assess their own work and then consider ways of working differently.

3. Teachers and principals help each other by working collaboratively.
4. Working with colleagues helps teachers and principals in their professional development.

Action research can occur on the classroom, school, or district level. Implications can be limited to just the classroom in question or can be extended to impact colleagues and students in other classrooms or schools. Benefits of action research for the researcher include a focus on (a) school issue, problem, or area of interest; (b) form of teacher professional development; (c) collegial interactions; (d) potential to impact school change; I reflect on own practice; and (f) improved communications (Brown University, 2000).

Surveys of music educators in the early 2000's revealed a consensus that research in music education is desirable, while at the same time raising concerns that most research was not readily applicable in the classroom (Laprise, 2017). Action research offers a solution to bridge the gap. This goes beyond a casual reflection of teaching and requires the educator to collect and analyze data. The result is an improvement in teaching and learning in the classroom that is immediate, research-based, and shareable with colleagues (Laprise).

Classroom Assessment

Tindal and Marston (1990) define assessment as the "systematic process used to gather data that allow educators to instruct students more effectively (p. 9)". Research on assessment has largely focused on high-stakes standardized tests. Often overlooked until recent years, however, is the importance of classroom assessment (McMillan, 2013). McMillan identifies CA as the most critical type of measurement for influencing student learning and achievement.

CA deals with any type of assessment, formal or informal, that is administered within the classroom by the teacher. This may include anything from anecdotal evidence collected by the teacher to a midterm exam. Educators categorize CA into two types: formative and summative

(Bonner, 2013). Formative assessments include graded homework, quizzes, or assignments which carry less weight in a student’s final grade but are used to guide learning and instructional practice. Summative assessments occur less often and include significant tests or projects with a grade that may be reported to stakeholders. Greater consequences for the student, teacher, and perhaps school are attached to performance on summative assessments.

CA is particularly useful because of its impact on student motivation to learn and the information it provides teachers (Andrade, 2013). Andrade states that “effective CA articulates the learning targets, provides feedback to teachers and students about where they are in relation to those targets, and prompts adjustments to instruction by teachers as well as changes to learning processes and revision of work products by students. (p. 21)” In order for CA to be effective, teachers must set appropriately challenging learning goals and clearly define them for students (Hattie, 2009). Teachers can also maximize the self-regulation benefit for students when they vary certain aspects of CA such as frequency, type of task, and feedback (Brookhart, 1997).

There is a growing research interest in CA (McMillan 2013). After years of focusing only on standardized tests, focus is shifting to the importance of CA and the relationship between the two. It is becoming clear that greater correlations are needed between high-stakes assessments and common classroom assessments (Schneider et al., 2013). Grades earned on daily or interim assessments do not always accurately predict a student’s performance on a state test. Implementing a backward design curriculum has been shown to improve alignment of classroom assessments with state test objectives (Wiggins & McTighe, 2005). This is not effective, however, if teachers and district administrators do not share the same interpretation of state standards as that of test developers or raters. Considering CA in light of year-end tests also yields unintended consequences such as teachers constructing assessments to mimic not only the content of a year-end test, but also the item format. With multiple-choice being the most common standardized test format, this means teachers are increasing the frequency with which they administer multiple choice tests to their students while foregoing other proven testing methods, namely high-quality performance assessments (Roediger & Karpicke, 2006).

Lane (2013) defines a performance assessment as a "demonstration of mastery that emulates the context or conditions in which the intended knowledge or skills are actually applied (p. 313)." Performance assessments may take the form of a demonstration, presentation, or performance and lend themselves well to individual work as well as collaborative work. Performance assessments resemble a real-life experience more so than written assessments. They can be employed seamlessly at any point of instruction and work well as formative or summative assessments. Through performance assessments students can demonstrate understanding and mastery of specific objectives that are otherwise difficult to assess (Lane). Perhaps most importantly, performance assessments are valuable because of their authenticity, especially in the music classroom (Asmus, 1999). A student is required to demonstrate mastery as opposed to simply demonstrating conceptual knowledge through a written assessment. While this type of assessment is already inherent to the music classroom, it is essential that performance assessment data becomes more widely used as a component of teacher evaluations (Davison & Fisher, 2019).

Assessment and Education Policy

Standards and assessment in public school education have been a topic of importance in politics since the 1960's(Burrack & Parkes, 2019). President Lyndon B. Johnson included education reform as an essential pillar of his "War on Poverty" platform. Citing a national poverty rate approaching 19%, Johnson believed education reform was the most effective solution to poverty. Beginning with the Elementary and Secondary Education Act of 1965, the public has been concerned with holding teachers and schools accountable for improving student learning outcomes (Ornstein & Levine, 2000). The Elementary and Secondary Education Act (ESEA) secured federal funding for schools designated as Title I Schools, or schools which have a high percentage of students living in poverty. Funds were tied to a stipulation that students would meet specific academic goals as demonstrated by standardized test scores. Title I funds are intended to be used by schools to create programs to help low-income students succeed academically, and ultimately to close the achievement gap between students attending rural or urban schools and students who attend suburban schools. This is the first instance of government funding for education tied to assessment results and is considered one of the most impactful pieces of legislation affecting public schools in the United States (Ornstein & Levine).

In the 1980's, the political focus on education had less to do with achievement gaps among students within the United States and more to do with a perceived disparity between the United States and other nations (Mark, 1996). Comparisons were drawn since the late 1950's when Russia surpassed the United States in the Space Race with the launch and successful three-month orbit of Sputnik I. Politicians claimed this signaled the future decline of America's workforce (Labuta & Smith, 1997). President Ronald Reagan's National Commission on Excellence in Education published a report titled *A Nation at Risk: The Imperative for Educational Reform*, in which they called for greater rigor in the classroom and higher standards of achievement to be documented through standardized test scores (Ornstein & Levine, 2000). The commission cited falling SAT scores and also listed numerous tests in which American students placed last among other advanced nations. The report demanded that teachers be held accountable for student achievement and listed 38 recommendations in the areas of a) content, b) standards and expectations, c) time, d) teaching, and e) leadership and fiscal support. As it relates to educational standards, the report recommended increasing the number and consequence of standardized tests as students progress through levels of secondary schooling as well as raising entrance requirements for colleges and universities. The commission identified high school grade inflation and low college admission standards as causes of declining rigor in education (Ornstein & Levine).

The Clinton era Goals 2000 reform bill expanded standards-based education and applied the requirements to multiple disciplines including English, mathematics, science, foreign language, civics and government, economics, history, geography, and the arts. This was the first time standards in the arts were included in legislation (Mark, 1996). Four divisions of arts standards were identified as a) creating and performing, b) perception and analysis, c) cultural and historical context, and d) the nature and value of the arts. There are nine general music content standards and accompanying achievement standards. These do not comprise a curriculum on their own, but were intended to be considered when developing music curricula at the state and local level. No accompanying assessments were included (Burrack & Parkes, 2019).

In 2001 the ESEA was reauthorized in the form of the No Child Left Behind act (NCLB). With the passage of NCLB, Title I funding was maintained and standards-based education remained in the forefront of education reform. NCLB increased the scope and consequence of standardized testing, requiring both state and national tests to be administered each year (Kelly et al., 2019). Individual states were charged with choosing and developing their own standardized assessments and reporting data. Under NCLB, student progress in reading and math must be measured and reported annually. Federal funding was attached to compliance with the testing mandates, and schools which failed to meet the standard of Adequate Yearly Progress were designated for improvement plans. A school that fails to demonstrate Adequate Yearly Progress for a number of consecutive years is subject to a complete restructuring including termination of faculty and administration.

The American Recovery and Reinvestment Act of 2009 (ARRA) brought CA to the forefront of education reform (Kelly et al., 2019). Among several other objectives designed to stimulate the economy, the ARRA offered grant funding to states for innovation in education through a program titled Race to the Top. Focus areas for receiving grant money included a) adopting assessments to prepare students for college, b) improving student growth measures, c) recruiting and retaining effective teachers, and d) improving student achievement in lower performing schools. One of the major platforms within ARRA addresses standards and assessments. The objectives were to develop and adopt common standards within each state and to develop and implement high-quality assessments. States that could demonstrate the most improvement toward these goals and the greatest degree of success in closing achievement gaps were awarded funding ranging from tens of millions to hundreds of millions of dollars. Each of these goals depends upon measuring student growth. The U.S. Department of Education offered three suggested means of measuring growth: Student Learning Objectives, standardized assessments, and school-wide collective performance measures.

These measurements are all problematic for music educators due to the lack of a national curriculum and corresponding assessments. Without a curriculum and without standard practice and tools in CA, teachers are left to devise their own assessment methods (Kelly et al., 2019). In the case of Student Learning Objectives (SLOs), the burden of creating the measurement tool and documenting growth falls on the classroom teacher. It is then up to the local school administrator to approve the measure and confirm documentation. Unless the administrator is a content expert in music they may not be equipped to make judgment calls on the validity of a music assessment tool (Davidson and Fisher 2019). When a music supervisor is available the task is easier, but most school districts in the United States do not employ a music supervisor. The advantage of using an SLO is that because it is teacher created, it will align with the curriculum and practice already in place in that teacher's classroom (Wesolowski, 2014).

A lack of standardized music assessments compels states to adopt school-based performance measures to document SLO achievement for music students. In this case the average standardized test scores of all students in a given school population are factored into teacher evaluation scores for those who teach non-tested subjects. As a result, music educators are evaluated based on school-wide student performance on standardized assessments in math and reading. This leads to decreased morale and a feeling of loss of professional control for music teachers. This also becomes a deterrent to working in lower achieving schools for fear of receiving lower evaluation scores due to collective performance scoring (Davison & Fisher, 2019).

State and National Music Assessment Tools

Music assessments at the state and federal levels have largely followed the trends of standardized testing in that they are disconnected from any practical application of CA. The inclusion of music in the National Assessment of Educational Progress in the 1970's was a first step in national assessment but was met with challenges (Mark, 1996). First, the music portion of the assessment was omitted throughout the 1980's and most of the 1990's due to lack of funding. There were also limitations of testing time constraints, unreliable reporting, and the omission of performing tasks in the second administration of the test that caused concern for music educators (Colwell, 1999). Connecting the assessment to some form of state accountability could have resulted in improved student achievement (Hamilton, 2008). The NAEP is designed to serve as a national report card and does not provide reports on individual schools or students, so usefulness at the state or local level is limited.

The Educational Testing Service of Princeton, New Jersey developed the first music objectives for the 1971 NAEP (Mark, 1996). Categories included (a) perform a piece of music; (b) read standard musical notation; (c) listen to music with understanding; (d) be knowledgeable about some musical instruments, some of the terminology of music, methods of performance, some of the standard literature of music, and some aspects of the history of music; I know about the musical resources of the community and seek musical experiences by performing music; (f) make judgments about music, and (g) value the personal worth of music. A panel of music educators appointed by the MENC board met with the NAEP staff to offer commentary on the test results. Three general concerns were expressed. First, the random sample of citizens taking the test may or may not have had music instruction in school. Second, some of the knowledge and skills measured are more likely to be influenced by the culture than by music instruction. Third, the test attempted to cover too many broad objectives with too few exercises devoted to each, which harms the validity of the test (Mark).

The first administration of the NAEP was of little value on its own, and it was not until data became available from the second administration in 1979 that any conclusions could be drawn about student growth. The information proved valuable, but not as much from a student achievement perspective as much as a commentary on the state of music programs in schools. Students scored lower in almost every category, but the number of music educators also declined in the time period between the two tests (Mark, 1996). A decrease in funding prevented the inclusion of performing and creating tasks on the second assessment, which are the tasks deemed most valuable by music teachers.

The Goals 2000 legislation included the first national music standards in 1994. They identified nine skills of a) singing, b) performing on instruments, c) improvising, d) composing, e) reading and notating, f) analyzing, g) evaluating, h) understanding interdisciplinary connections, and i) understanding historical and cultural connections in music. Minimum competencies were included for each grade level. Music educators lacked assessment tools to measure student learning in relation to these standards because the standards did not match common instructional processes in the classroom (Brophy, 1997). Without a national curriculum and without assessment measures to accompany the standards, individual states began to address these issues by allocating funding to create common assessments. A common thread among many of the state assessments discussed below is the inclusion of and priority placed upon authentic performance tasks in addition to multiple choice or short response items.

The Connecticut Department of Education provided a grant for the development of school improvement plans based on the 1994 National Arts Standards, resulting in the Connecticut Common Arts Assessment. The CCMA utilized key musical tasks as outlined in Goals 2000: (a) solo performance, (b) ensemble performance, (c) ensemble critique, (d) sight-reading, and (e) arranging. Performance tasks were weighted more heavily than written assessments in order to emphasize what is most valued by music educators (Wiggins & McTighe, 2005). For this reason, the artistic processes of creating and/or performing were involved in each assessment task. Special consideration was given to test reliability and validity.

Upon the introduction of the 2003 South Carolina Visual and Performing Arts Curriculum Standards, the South Carolina Department of Education began a collaborative effort with the University of South Carolina and the Office of Program Evaluation to create the South Carolina Arts Assessment Program. This is a statewide assessment administered at any elementary school receiving funds from the Distinguished Arts Program grant from the SCDE. The test is taken by fourth grade students and includes multiple choice and performance tasks. The assessment meets high standards of reliability, validity, item fit, and test bias. The developers have used data from Rasch item analysis to continually refine the assessment. The test has not been consistently administered due to varied funding levels and waning school participation. It is intended for use as a program evaluation tool and not to track individual student achievement (Lewis et al., 2019).

President Obama’s Race to the Top initiative included a provision for special funding designated for “hard to measure” subjects such as the fine arts. A handful of states used these funds to develop statewide music assessments. A team of Florida music educators developed the Florida Performing Fine Arts Assessment. They followed the artistic process model of creating, performing, and responding. The portion of the assessment accounted for by each of these elements was dictated by the results of a teacher survey in which Florida music educators indicated their class time consisted of 70% performance, 25% knowledge and skills, and 5% creating. The assessment included both prepared and at-sight test items. Students were given two weeks for the prepared assignments. The result of the project was a test item bank made available to music teachers to use as appropriate for their classroom (Kelly et al., 2019).

The revised 2014 National Music Standards include general music standards by grade level through 8th grade and group ensemble standards by level of experience, taking into account the fact that music is not required for middle and high school students and they may enroll in music classes at different grade levels. They also now include sample performance assessments known as Model Cornerstone Assessments (MCAs). MCAs are modeled after the Connecticut assessments (Shuler, 2016). MCAs are tied to the standards and are scored using a rubric. They are designed to be easily implemented with any lesson material the teacher is already. The language in the rubrics is broad and is compatible with any curricular content as well as any type of curricular ensemble such as voice, strings, winds, or percussion. Along with the MCAs, NafME provides exemplars of student work on their website. As of 2019, MCAs are still being tested for reliability and validity (Burrack & Parkes 2019).

Current Classroom Assessment Trends in Music Education

Current classroom assessment trends in music education show a disconnect between local and state policy and curriculum/assessment. In light of the evidence that the quality and type of CA profoundly affects student learning in all subjects, music educators are missing an opportunity to leverage CA to the benefit of their own students. Current CA practices in music education do not promote rigor or individual accountability. The teacher, student, and the music program as a whole are negatively impacted by poor CA practices. Burrack and Parkes (2019) suggest that "the key to effective local policy is designing a system of music learning assessments through sequential, standards-based music instruction and alignment of internal (district level) and external (state and national level) accountability measures so that teachers are able to explain their student learning as part of their own professional development plans (p. 654)." The alignment of classroom assessments with state level achievement standards is the basis of standards-based learning and promotes a culture of accountability within individual schools (Stites & Malin, 2008). When these conditions are met, classroom assessment accurately provides evidence of student achievement and growth.

Grading procedures in music classrooms across the United States often fail to accurately measure individual student achievement. Pellegrino, Conway, and Russell (2015) suggest that student grades are often based on non-achievement considerations such as participation and attendance, for example. Such nonmusical skills such as attendance or participation are not included in the national or state standards and should not be included as a component of a student's grade (Lehman, 1998). When grading policies include elements such as these that are outside of the state standards, teachers may actually be leaving themselves vulnerable to litigation by parents and guardians of the impacted students (Russell, 2011). Teachers must ensure that their classroom grading policies are aligned with county and state policies.

The time and effort required to implement meaningful CA contributes to the lack of rigor and individual accountability in many instrumental music classes (Goolsby, 1999). With large numbers of students to teach, instrumental music directors rely on error detection techniques to gauge group progress toward ensemble performance goals. Goolsby states that without individual formative assessments, the student ensemble members are not held accountable for learning and instead are simply being conditioned to perform to the teacher's tolerance level for mistakes. The attempt by ensemble directors to save rehearsal time by foregoing more rigorous CA is ill-advised, as concepts may require re-teaching due to the lack of individual accountability.

Music educators are not only responsible for presenting their students with a rigorous curriculum, but also documenting student growth in a way that can be understood by parents and administrators. Documenting achievement demonstrates that learning is taking place. Classroom assessment data is essential for teachers, parents, administrators, and the community for determining the effectiveness of instruction (Asmus, 1999). Music teachers are often unable to explain student learning because they lack a system of music learning assessments (Asmus). Assessments must align with sequential, standards-based instruction and local policy.

Studies in Solo Music Performance Assessment

Considering the lack of rigor in music classrooms and the disconnect between standards and assessment, there is a need for further exploration and implementation of individual music performance assessment (Bergee, 2003). Studies in solo music performance assessment (MPA) have shown a gradual increase in reliability and validity, but as of yet they have not yielded classroom-friendly tools for educators and therefore are not widely used in the classroom. An examination of past studies in MPA reveal inherent difficulties in measuring solo music performance, namely rater bias. Rater mediated assessment is by nature subjective and will always reflect the conscious or subconscious biases held by individual raters (Boyle, 1992). Beginning with the Watkins-Farnum Performance Scale for Band Instruments, it is helpful to examine several studies that have attempted to solve problems in MPA.

The Watkins-Farnum Performance Scale for Band instruments in 1942 is considered the first systematic effort in solo music performance assessment (Zdzinski, 1991). This study began as a scale for rating solo cornet performance and was later transposed for other wind instruments. It consists of a set of exercises that progressively increase in difficulty. Scoring was determined by the number of measures played correctly. Both prepared and at-sight performances were scored, and two equivalent forms were used. The scale is shown to have a high degree of reliability (Zdzinski).

Kidd's 1975 study of trombone performance rating is modeled after the Watkins-Farnum study (Abeles, 1981). One significant difference is an improvement in validity by using a content analysis. Kidd performed a content analysis of more than forty trombone solos to identify salient characteristics of trombone performance. Fifty characteristics were chosen to evaluate range, slide technique, and articulation. Three judges rated performances using the scale and items were discarded that had poor interjudge agreement. Two equivalent forms were created. Utilizing specific evaluative criteria as opposed to simply scoring a measure as correct or incorrect resulted in increased validity and a more accurate measure of achievement.

Abeles' 1971 study is significant because of the use of the facet-factorial approach of item construction (Zdzinski, 1991). Facet-factorial design systematically determines common evaluative criteria within a broad discipline such as music performance. Abeles' aim was to decrease subjectivity in MPA. In developing his Clarinet Performance Rating Scale (CPRS), Abeles performed a content analysis of clarinet pedagogical literature, essays, articles, as well as other related literature to identify performance criteria. These items were grouped into seven categories, which were ultimately reduced to six after initial trials and a factor analysis: (a) interpretation, (b) intonation, (c) rhythmic continuity, (d) tempo, I articulation, and (f) tone. A total of thirty performance descriptive statements were paired with a five-point Likert scale. The CPRS demonstrated high interjudge reliability and content validity. Through the CPRS, Abeles successfully assessed multiple facets of a complex behavior using facet-factorial design (Boyle, 1992). Abeles' work also inspired subsequent research applying the facet-factorial method to ensemble performance evaluation (Cooksey, 1977; DeCamp, 1980).

Bergee (1987) developed a Euphonium and Tuba Performance Rating Scale (ETPRS) using the facet-factorial approach to item construction. Evaluative criteria were selected following a content analysis of tuba and euphonium pedagogical literature. One difference between Bergee's study and Abeles' is that Bergee used existing adjudication sheets in addition to pedagogical literature and articles. Five Likert-type response options were used initially, and then reduced to four after responses were factor analyzed. Like Abeles' CPRS, the ETPRS exhibited high validity.

Saunders and Holahan (1997) conducted a study using the facet-factorial method of rating scale design, but sought to improve the quality of instructional feedback to students by developing a criteria-specific rating scale. They argue that a typical Likert response scale rating does not provide the performer with adequate information from the rater with which to improve their performance. The criteria-specific scale was intended to improve diagnostic feedback from rater to performer. Criteria-specific scales differ from Likert scales in that they include descriptions of specific levels of achievement from the performer. An example for scoring tone quality may be "characteristic tone quality in most ranges but distorts occasionally in some passages, "or "tone quality is not characteristic of the instrument." Their study compared the criteria-specific method to other rating scales in use at the time. A solo evaluation form was used to score 926 woodwind and brass players auditioning for the Connecticut All-State Band in 1994. The Woodwind and Brass Solo Evaluation Form was found to be substantially reliable and held high diagnostic validity.

Considering the number of individual skills involved in the complex task of instrumental music performance, Dressman (1990) developed and validated rating scales which divided performance tasks into two broad categories: executive skills and performance skills (Boyle, 1992). He defined executive skills as (a) embouchure, (b) posture and playing position, (c) hand/finger position and technique, (d) breathing, anIe) tongue movement. Performance skills included (a) tone quality, (b) performance of dynamic markings, (c) phrasing, (d) articulatioI(e) common interpretive notation, and (f) performance of rhythms in common meters. Dressman's scales demonstrated high nter- and intrajudge reliability.

Progress Toward Invariant Measurement

The studies discussed above demonstrate significant improvements over time in the quality of test construction, particularly in the areas of item construction and inter- and intra-rater reliability. However, there has been progress in the field of measurement over the past fifty years that has not yet been utilized in MPA research (Wesolowski et al., 2016). Masters and Keeves (1999) state "ThI is...a strong body of well-established theory that awaits widespread application in programs of assessment, evaluation and research where more accurate measurement is required to provide new understandings of educational achievement and educational, psychological and societal processes. Unfortunately, information on the wide variety of powerful procedures which now exist to advance measurement in these areas would appear to be hidden away in journals and reference works (p. 1)." Researchers in the field of music performance have long accepted that rater severity will always be an uncontrolled variable when it comes to assessment (Engelhard, 2013). While subjectivity may always play a role in assessment, accuracy and precision of performance evaluation are positively impacted and objectivity is increased by the implementation of invariant measurement techniques (Boyle, 1992). Application of the Rasch Measurement Theory offers researchers the possibility of developing models of invariant measurement in music assessment that can account for multiple factors, including rater subjectivity (Wesolowski et al., 2016).

Much of the work in the field of measurement research $^{\text{in}}$ the 20th century was dominated by the test-score tradition, also known as Classical Test Theory (CTT) (Engelhard, 2013). The basic premise of CTT is that errors will always occur in measurement, and the researcher must have a way to account for the error. Spearman is attributed with developing the most widely used formula to account for error as a random variable (Wilson, 2005). It is important to note that measurement in the CTT tradition is always sample dependent. This is not desirable if the goal of assessment research is to yield a measurement tool that is consistently useful across multiple test items, persons, and raters.

An outgrowth of CTT that is particularly useful in performance-based assessments is Generalizability Theory (G Theory). G Theory is still sample dependent, but it seeks to make broader generalizations based on an analysis of multiple potential sources of error including rater error, for example (Engelhard, 2013). Performance-based assessments involve judgment calls by raters, so in this case G Theory does offer a step towards accounting for rater error in music performance (Bergee, 2007).

Wilson (2005) defines the central purpose of measurement as "providing a reasonable and consistent way to summarize the responses that people make to express their achievements, attitudes, or personal points of view through instruments such as attitude scales, achievement tests, questionnaires, surveys, and psychological scales (p. 5)." As it pertains here, music educators are concerned with measuring the construct of musical achievement.

One of the difficulties in measuring music performance is that music performance, or musical ability is not an observable trait (Boyle, 1992). We know individuals possess musical ability but the ability itself cannot be directly measured. This type of trait is referred to as a latent trait, or latent construct. A construct is a psychological quality that we assume exists in order to explain some aspect of human behavior (Gronlunch & Linn, 1990). We can listen to a musical performance and rate the performance, but we are only rating the quality of that particular performance at that moment in time. Music educators, however, are primarily concerned with assessing a student's musical ability and documenting their growth over time. In this case, music educators are attempting to measure the latent trait of musical performance ability.

In the field of measurement, latent variables are often expressed in the form of variable maps (Wilson, 2005). Engelhard states that “The goal of measurement is to develop variable maps that can be used to represent the locations of both persons and items on a latent variable of substantive and theoretical interest (Engelhard, 2013, p.13).” The scaling tradition which emerged in the second half of the 20th century provides the tools necessary to represent these variable maps for latent traits, and therefore could prove to be the most useful type of measurement theory for music educators. The scaling tradition offers a means of calibrating items and responses, which is a step toward invariance. Once the items, persons, and raters are calibrated they can be represented on a variable map (Engelhard). Rasch is unique from other theorists in that his model calibrates both items and persons on the same map.

Among the different measurement theories within the scaling tradition, item response theory (IRT) is the dominant model today (Engelhard, 2013). IRT is also the primary theory addressing student achievement in the field of education. The Rasch IRT model is predicated on his belief that measurement should be independent of particular items and particular respondents (Bond & Fox, 2020). This stands in direct contrast to CTT models which are sample dependent. Rasch states “The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli with the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared on the same or on some other occasion (Rasch, 1961).” The concept of sample independence is a necessary step toward invariant measurement, which is identified by Engelhard (2013) as the foundation for modern rules of measurement in the 21st century.

The goal of invariant measurement is to create a measurement tool that is accurate and useful regardless of which persons or items are used and is consistent with the latent construct (Bond & Fox, 2020). In order to achieve invariant measurement, five conditions must be met: (a) item-invariant measurement of persons (i.e., the measurement of persons must be independent of the particular items that happen to be used for the measuring), (b) non-crossing person response functions (i.e., a more able person must always have a better chance of success on an item than a less able person), (c) person-invariant calibration of test items (i.e., the calibration of the items must be independent of the particular persons used for calibration), (d) non-crossing item response functions (i.e., any person must have a better chance of success on an easy item than on a more difficult item), and (e) variable map (i.e., items and person must be simultaneously located on a single underlying latent variable) (Engelhard).

In order to determine whether the five conditions of invariant measurement are met, item and person response functions must be examined (Bond & Fox 2020). Item response functions chart the probability of success on an item or items as a function of person ability. If the condition of non-crossing item response functions is met, this confirms that success on an easy item is more likely for all persons than success on a difficult item. Person response functions chart the probability of success by persons as a function of item difficulty. If the condition of non-crossing person response functions is met, this confirms that a more able person will always have a better chance of success than a less able person.

The Rasch method differs from other scaling theories in the way the data are compared to the measurement model. The Rasch method is considered an ideal-type model, meaning that the model perfectly meets the requirements of invariance (Engelhard, 2013). For this reason, data obtained in a given study must approach the ideal model to be considered valid. Rasch's idealization of the data as a straight line on the variable map is unachievable, but provides a tool with which to measure real data (Bond & Fox, 2020).

As discussed earlier, use of Likert-type responses in performance assessment is a common practice. This points to another useful characteristic of Rasch measurement, which is the possibility of assigning partial credit for success on individual items using the Partial Credit Rasch Model (Bond & Fox, 2020). Likert items may have varied numbers of response options according to what makes the most sense for the particular item. The partial credit model allows for differing numbers of Likert response options on the same rating scale.

Invariant measurement in rater-mediated assessments such as music performance evaluations can only be reached if rater severity is accounted for. Another aspect of Rasch measurement is the ability to measure two or more variables along the variable map using the Many-Facets Rasch Model (Bond & Fox, 2020). With the many-facets model the researcher is no longer limited to person and item data, but can also plot rater data on the map.

Music Performance Assessment Studies using the Multifaceted Rasch Partial Credit Model

A number of recent studies from 2016 to present have utilized the Multifaceted Rasch Partial Credit Model for validating music performance assessment measures (Ooi and Engelhard, 2019; Wesolowski et al., 2016; Wesolowski et al., 2017; Wesolowski, 2019). The MFR-PC model can "more clearly detect variability in rater judgment and improve model-data fit, thereby enhancing the objectivity, fairness, and precision of rating quality in the music assessment process (Wesolowski et al., 2016, p. 662)." This was found to be the case in each of the studies listed above, and yielded successful results for both individual and ensemble performance assessments.

Wesolowski (2017) led a cohort of thirteen graduate students in music education in to create the Music Performance Rubric for Secondary-Level Wind Instrument Solos. The rubric was designed to be applicable to the most common beginning band instruments: flute, clarinet, saxophone, trumpet, and trombone. The researchers formed an initial pool of 47 items related to technique, tone, articulation, visual, melody, and time/rhythm. A four-point Likert scale was then applied to each item. The researchers collected 75 videos of secondary level performances, which were subsequently rated by each researcher in order to create a complete assessment network. Through Rasch analysis, the rubric demonstrated high levels of reliability, precision, and validity. Seventeen of the 47 items failed to meet the requirements of acceptable data-model fit and were removed. The partial credit analysis revealed to the researchers which items needed fewer than four response options.

The Jazz Big Band Performance Rating Scale (JBBPRS) is an example of an ensemble performance measurement tool developed using factor analysis and validated using Rasch analysis (Wesolowski et al., 2016). Twenty-three raters scored recordings of middle school, high school, and college/professional big bands using the JBBPRS, which combined twenty-two performance statements with a four-point Likert scale. Rasch data analysis revealed differences in rater severity. Additionally, the use of the partial credit model showed significant differences in how each rater interacted with each item and how they perceived item structure. Discrepancies in raters' interpretation of a rating scale can negatively impact fairness of the assessment (Wesolowski).

These studies further highlight the shortcomings of CTT as a tool for the development of rater-mediated performance assessments, and point to the need for further investigation into MPA scale construction using Rasch measurement, and specifically the MFR-PC model. With MPA depending entirely upon rater mediation, the necessity of addressing rater bias and providing fairness in assessment is paramount (Ooi & Engelhard, 2019).

CHAPTER 3

While making judgments is an inevitable aspect of daily living, the majority of human judgment is made purely on intuition and instinct (Hogarth, 1980). This is sufficient in many cases, but not if judgments carry greater consequences. In the era of high-stakes testing in schools, it is more important than ever to consider just how judgments are made by those evaluating student work and how assessment procedures can be improved based on our understanding of human judgement.

Music performance assessment relies strictly on rater-mediated assessments. This holds true in the case of individual classroom assessment, auditions for region honor bands or all-state bands, solo and ensemble evaluations, and large group performance evaluations. Rater-mediated assessments offer rich benefits for teaching and learning, but also come with potential problems that often go unaddressed (Engelhard, 2013). This has become an issue not only in the case of music assessment, but also in other subjects as constructed-response type questions have become more common in the 1980s and 1990s.

Brunswik's (1952) lens model offers a starting point for identifying the concerns of rater subjectivity in rater-mediated assessments. Five key points surrounding rater judgments in Brunswick's Lens Model include interrelatedness, correspondence, accuracy, uncertainty, and individual-task interactions. Rater judgments cannot be separated from the environment in which they are made (interrelatedness). Environmental factors include not only the performance to be rated and the physical circumstances surrounding the performance, but also the rater's own schema (i.e., their own mental representation of the world around them) and how the interaction between their schema and reality impact the quality of assessment made by the rater (Engelhard, 2013). Another factor to be considered when accounting for a rater's particular "lens" is the idea that raters subconsciously take cues from the structure of individual test items and the design of the assessment as a whole (Hogarth, 1987).

Hogarth (1980) frames the concept of the lens model in terms of points of reference or cues. Brunswik framed his conception of judgment using the relationship between real cues and events in the environment and the cues and predictions in the mind of the individual making the judgment. These imagined cues can even include memories that the individual deems relevant to the decision-making task at hand. Hogarth states the "accuracy of (judgment) clearly depends on the extent to which the 'model of the environment' is matched by the 'model of the person', i.e. in terms of cues, relationships between cues, and between cues and the target event, as well as the relative importance of the cues (p. 7).

Engelhard (2013) outlines how the lens model can be applied in rater-mediated assessments such as music performance assessment. When measuring the latent variable of music performance ability, several variables contribute to the environment of cues that will contextualize the task for the rater. These can include a) the construction of the rating scale itself, b) the particular items used, c) the response categories, d) the individual rater's degree of severity in scoring, and e) the rater's past experiences in music teaching and learning. It follows that all aspects of a rating scale must be carefully examined including item functions, rater use, and rater precision (Wesolowski et al., 2016).

In order to validate the measurement tool developed in this study, rater-invariant measurement must be achieved. Rasch analysis offers the best path to meet the requirements of invariant measurement for a rater-mediated study (Engelhard, 2013). In the context of rater-mediated assessment, the five requirements of invariant measurement are:

1. Rater-invariant measurement of persons (i.e., the measurement of persons must be independent of the particular raters that happen to be used for the measuring).
2. Non-crossing person response functions (i.e., a more able person must always have a better chance of obtaining higher ratings from raters than a less able person)
3. Person-invariant calibration of raters (i.e., the calibration of the raters must be independent of the particular persons used for calibration)
4. Non-crossing rater response functions (i.e., any person must have a better chance of obtaining a higher rating from lenient raters than from more severe raters)
5. Variable map (i.e., persons and raters must be simultaneously located on a single underlying latent variable)

The Rasch method is considered an ideal-type model, meaning that the model perfectly meets the requirements of invariance (Bond & Fox, 2020). For this reason, data obtained in a given study must approach the ideal model to be considered valid. Rasch's idealization of the data as a straight line on the variable map is unachievable, but provides a tool with which to measure real data. The Rasch model is different from other measurement models in that the data must closely fit the ideal model, instead of the model being constructed to fit the data (Engelhard, 2013).

In order to determine whether or not the five conditions of invariant measurement are met, item and person response functions must be examined (Bond & Fox 2020). Item response functions chart the probability of success on an item or items as a function of person ability. If the condition of non-crossing item response functions is met, this confirms that success on an easy item is more likely for all persons than success on a difficult item. Person response functions chart the probability of success by persons as a function of item difficulty. If the condition of non-crossing person response functions is met, this confirms that a more able person will always have a better chance of success than a less able person.

A Rasch analysis yields item difficulty and person ability estimates expressed on an interval called a logit scale (Bond & Fox, 2020). A logit (or log odds) scale is constructed in such a way that the distance between each unit of measurement is uniform. Item and person fit are determined by examining the distance of each one from the expected location based on the ideal model. Acceptable data-model fit is achieved when items and persons fall within +/-2 from the model (Bond & Fox).

Considering the multiple variables involved in MPA (performer, item, and rater) and the type of performance being rated, the Partial Credit Many-Facets Rasch analysis provides the best avenue to achieve invariant measurement. The many-facets Rasch model allows separate parameterization of each of the three variables along the same latent variable (Bond & Fox, 2020). The partial credit model is a variation of the many-facets Rasch model which addresses the true difference in ability required to achieve various marks on the Likert scale response for each item. The partial credit model is the preferred model whenever there are more than two ordered response categories (Masters & Keeves, 1999). The research questions to be answered by Rasch analysis are:

1. What does Rasch measurement analysis reveal about the psychometric properties (i.e., validity and reliability) of items, raters, and performers in the context of solo music performance assessment?
2. How do items vary in difficulty, raters in severity, and performers in achievement?
3. How does the rating scale structure vary across raters and performers?

Item Pool Development

A rating scale format was chosen for its usefulness in collecting judgments in rater-mediated assessment (Engelhard, 2013). Initial item pools for each instrument were compiled from a thorough content analysis of pedagogical literature, existing rating scales, and consultation with content experts (Abeles, 1971). Pedagogical literature included instrument specific method books and journal articles. Solo and Ensemble or All-State audition score sheets from 26 states were examined for woodwind, brass, and percussion instruments and analyzed for common items used in evaluating and rating student performance. Content experts included school band directors and professional players who also teach privately (Wesolowski et al., 2015). A comprehensive list of performance traits was compiled for each woodwind and brass instrument commonly found in school band programs (appendix…) Redundant items were removed as well as items that could not be measured or scored (e.g., an assessment of the shape of the inside of a subject's oral cavity). Content experts were then consulted again to evaluate the rating scale drafts and offered edits to the remaining items. Likert-type response options were chosen as they are the most appropriate choice for a rating scale (Masters & Keeves, 1999). Items were paired with four-point Likert-type scales which were developed using response anchors chosen from Vagias' (2006) *Likert-type Scale Response Anchors*.

Subjects

The subjects are 160 middle and high school band students recruited from a band program in a suburban school district in northeast Georgia. Instruments represented within this group of subjects include flute (*n* = 28), oboe (*n* = 3), clarinet (*n* = 34), bass clarinet (*n* = 1), alto saxophone (*n* = 17), tenor saxophone (*n* = 2), trumpet (*n* = 38), french horn (*n* = 9), trombone (*n* = 18), euphonium (*n* = 2), and tuba (*n* = 8). It is acknowledged that sample sizes may be too small in some cases (e.g., oboe) to meet the minimum requirements of validity (Wright & Stone, 1979). This study is grounded in the principles of action research and will serve as a starting point for improving classroom instruction and assessment, and may or may not offer broad generalizations applicable beyond the particular band program involved (Laprise, 2017). Students represent a wide range of ability, and there was no stipulation of achievement level in order to take part in the study. All students were invited to participate. Each is given a short exercise appropriate for their developmental level, as determined by their school band director, two weeks in advance of recording. Subjects will not receive any instruction on how to play the music and will not be directed to practice a certain amount of time. Subjects will then be video recorded performing the exercise. Videos are securely stored and labeled using a numbering system to protect students' personal information.

Performance Stimuli

The exercises performed by the subjects are taken from the Essential Elements for Band series published by Hal Leonard. The primary considerations when choosing the exercises were availability of music to the students, appropriateness of technical and musical demand, and the inclusion of multiple musical elements to be demonstrated by the students. In order to secure a large number of participants, especially middle school students, it was advantageous to choose music from a method book that they already were responsible for keeping up with on a daily basis for band class and home practice. This was done to reduce the number of students who would potentially misplace their music had they been given a special handout with music specifically for the research study. Technical demands needed to be developmentally appropriate for all participating, so three exercises were chosen to represent beginning, intermediate, and advanced performance ability. Each exercise was already transposed for each instrument since it appeared in a method book for use in a heterogeneous band class. Each exercise was evaluated to ensure multiple musical demands were required, including varied note values, articulations, dynamics, etc. in order for raters to utilize all items on the rating scale. The length of each exercise ranged from 20 to 80 seconds which meets the minimum length requirements established by Geringer and Johnson (2007).

All performances took place at one of two school buildings. Middle school students played in a meeting room near their media center and high school students played in a practice room attached to the band room. Performances were videoed using an iPad with a PadCaster external microphone. Videos were controlled for consistent angle and distance from each participant. Videos were framed to allow raters to fully evaluate all visual aspects of playing (i.e., posture, hand position).

Rater Pool and Network

Raters (*N* = 40) include middle and high school band directors from the state of Georgia. Expert teachers were selected in order to provide the best chance of fair and equitable assessment (Wesolowski, 2015). Criteria for selection of raters includes education, teacher certification, and experience teaching middle and high school band. All have at minimum a bachelor's degree in music education and hold a current Georgia teaching certificate in music Pre-K through 12th grade. Years of teaching experience range from one to 38 (*M* = 12.6). Each rater will view and score eight videos in an incomplete overlapping rater network (Engelhard, 1997), with four overlapping performances per rater (e.g., rater 1 scores videos 1-8 and rater 2 scores videos 5-12) Raters are instructed to score each item as best they can for each performer, even if certain items are difficult to score. They are asked not to make assumptions about the age of the performer, how long the performer may have studied his/her instrument, or whether or not the performer received instruction on the particular excerpt, but simply to score based on what they hear and see in the video performance. Raters are told they can view each video multiple times if needed in order to address each item on the rating scale. Rater data will be compiled and analyzed with the MFR-PC Rasch model using FACETS software (Linacre, 1989). The three facets (items, persons, and raters) will be represented as linear measures. The partial credit model will describe how each rater viewed the available response anchors for each item on the scale. Response functions for items, persons, and raters will be analyzed for data-model fit to determine if the requirements of invariance have been met (Bond & Fox, 2020).

CHAPTER 4

RESULTS

The purpose of this study was to develop and validate a set of instrument-specific rating scales for classroom use at the secondary level. A Multifaceted Rasch Partial Credit Model (MFR-PC) analysis was performed to examine the psychometric qualities of each scale. The study was guided by the following research questions:

1. What does Rasch measurement analysis reveal about the psychometric properties (i.e., validity and reliability) of items, raters, and performers in the context of solo music performance assessment?
2. How do items vary in difficulty, raters in severity, and performers in achievement?
3. How does the rating scale structure vary across raters and performers?

All scoring data was compiled for each of the rating scales and a MFR-PC analysis was performed for each data set using *FACETS* software (Linacre, 2009). The MFR-PC analysis translates observed scores into linear measures to show the achievement level of the performers, the severity of the raters, and the difficulty level of the items. These results were then used to determine the validity and reliability of each individual rating scale. Due to the nature of classroom research, it is acknowledged that some of the sample sizes are small. In particular, oboe and euphonium rating scale data are invalid due to the number of participants falling below the suggested minimum of 10 (Linacre, 2002).

Summary statistics are presented below for each rating scale along with calibration tables for items and performers. Rater calibration tables are included in Appendix (A?). Salient statistics discussed in this chapter include reliability of separation for each facet (performer, rater, item) and Mean Square Error values (MSE). High reliability of separation indicates adequate spread of elements within each facet along a single latent variable. This allows for an objective ordering of elements for comparison. In the case of this study, the latent variable investigated is the measure of music performance ability. Infit (information sensitive) and Outfit (outlier sensitive) MSE values are used to describe how closely the data for each facet follow predictable patterns expected by the Rasch model. Rasch is an ideal-type model, meaning that the validity of a set of data is determined by how closely it fits the ideal model. Mean Square Error values (MSE) close to 1.00 indicate good data to model fit. In the case of rater mediated assessments using a rating scale, MSE values between 0.6 and 1.4 are considered valid (Bond & Fox, 2020). Variable maps provide a graphic representation of the locations of each individual performer, rater, and item along the latent variable. The facets are ordered along a logit (log odds unit) scale which is an equal interval scale created by the Rasch model for the purpose of objective comparisons of elements within each facet.

Rater Characteristics

Raters included 41 practicing and retired public school band directors in the state of Georgia. Table 4.1 describes rater characteristics including primary instrument and number of years taught. The largest instrument groups represented by the raters were clarinet and trumpet with eight players each (19.51%). None of the raters played oboe. Years of teaching experience ranged from 1 to 42, with 18 (43.90%) of the raters having taught 21 years or more. While the raters represent a broad distribution of primary instruments, they were not specifically chosen to score performances by players of their instrument. Any scoring trends related to primary instrument will be discussed in chapter 5.

Table 4.1

Rater Characteristics by Primary Instrument and Years Taught

	Number of Raters (*N* = 41)	Percentage of Raters
Instrument		
Flute	3	7.32
Oboe	0	0.00
Bassoon	1	2.44
Clarinet	8	19.51
Saxophone	4	9.76
Trumpet	8	19.51
Horn	2	4.88
Trombone	5	12.20
Euphonium	1	2.44
Tuba	3	7.32
Percussion	6	14.63
Years Taught		
1-10	13	31.71
11-20	10	24.39
21+	18	43.90

Flute Performance Rating Scale Data

Multifaceted Rasch Partial Credit Model Results

Table 4.2 presents summary statistics from the analysis of performers (*n* = 28), raters (*n* = 9), and items (*n* = 69) on the Flute Performance Rating Scale (FPRS). Overall significant differences in chi-square are indicated for performers (2511.30), raters (1316.40), and items (1098.10). The probability for each facet was less than 0.01 and reliability of separation for each was high with performers and raters both at 0.99 and items at 0.94. This indicates adequate spread of elements within each facet along a single measure of music performance ability. Infit MSE values on the FPRS are well targeted with 1.03 for performers, 1.01 for raters, and 0.98 for items. Outfit MSE values were higher for raters which was 1.48, meaning there was up to 48% more variation than expected in rater behaviors. High outfit values can signal random response patterns that will be investigated in the discussion section.

Table 4.2

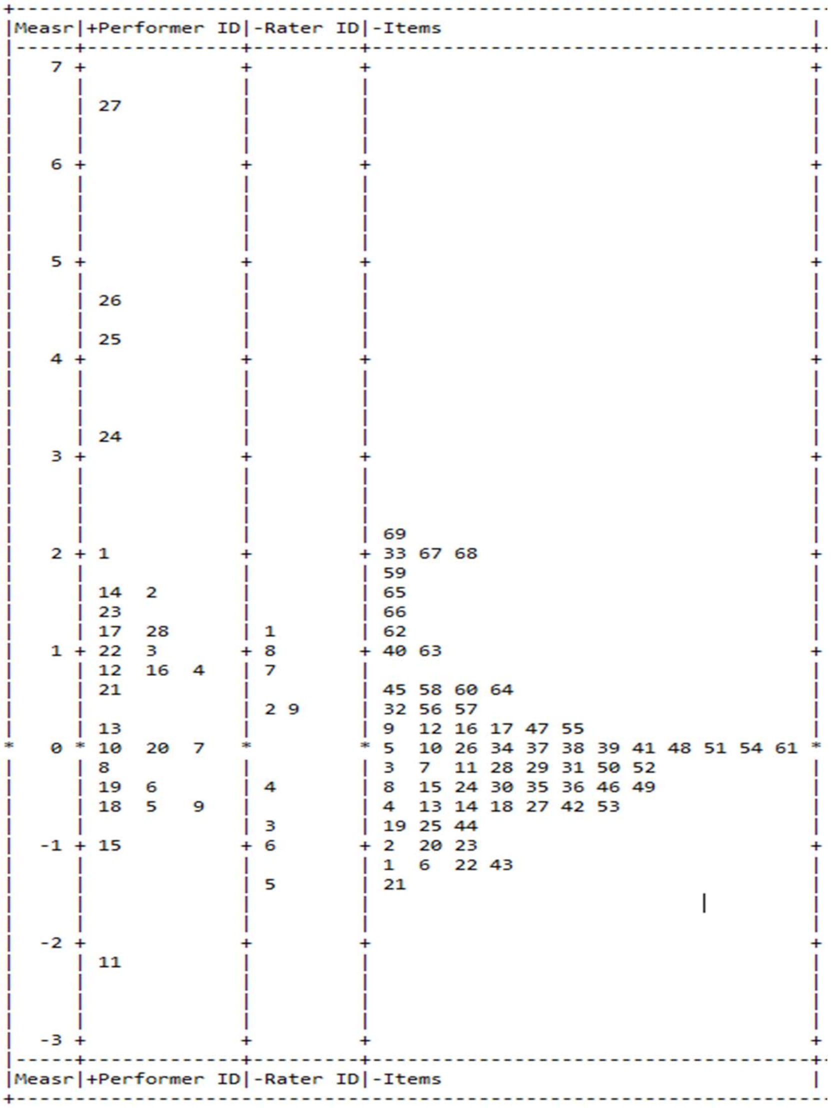
FPRS Summary Statistics from the PC-MFR Model

	Facets		
	Performance (θ)	**Rater** (λ)	**Item** (δ)
Measure (Logits)			
Mean	0.95	0.00	0.00
SD	1.78	0.91	0.81
N	28	9	69
Infit *MSE*			
Mean	1.03	1.01	0.98
SD	0.31	0.17	0.34
Std. Infit *MSE*			
Mean	0.00	-0.10	-0.30
SD	2.00	1.60	1.60
Outfit *MSE*			
Mean	1.27	1.48	1.25
SD	0.94	1.04	1.56
Std. Outfit *MSE*			
Mean	0.10	0.90	-0.10
SD	2.20	2.10	1.50
Separation Statistics			
Reliability of Separation	0.99	0.99	0.94
Chi-Square	2511.30	1316.40	1098.10
Degrees of Freedom	27	8	68

*$p < 0.01$

Figure 4.1

Variable Map, Flute Performance Rating Scale



```
+---------------------------------------------------------------------
|Measr|+Performer ID|-Rater ID|-Items                               |
|-----+-------------+---------+-------------------------------------+
|   7 +             +         +                                     +
|     |             |         |                                     |
|     | 27          |         |                                     |
|     |             |         |                                     |
|     |             |         |                                     |
|   6 +             +         +                                     +
|     |             |         |                                     |
|     |             |         |                                     |
|     |             |         |                                     |
|     |             |         |                                     |
|   5 +             +         +                                     +
|     |             |         |                                     |
|     | 26          |         |                                     |
|     |             |         |                                     |
|     | 25          |         |                                     |
|   4 +             +         +                                     +
|     |             |         |                                     |
|     |             |         |                                     |
|     |             |         |                                     |
|     | 24          |         |                                     |
|   3 +             +         +                                     +
|     |             |         |                                     |
|     |             |         |                                     |
|     |             |         |                                     |
|     |             |         | 69                                  |
|   2 + 1           +         + 33 67 68                            +
|     |             |         | 59                                  |
|     | 14   2      |         | 65                                  |
|     | 23          |         | 66                                  |
|     | 17   28     | 1       | 62                                  |
|   1 + 22   3      + 8       + 40 63                               +
|     | 12   16   4 | 7       |                                     |
|     | 21          |         | 45 58 60 64                         |
|     |             | 2 9     | 32 56 57                            |
|     | 13          |         | 9  12 16 17 47 55                   |
*   0 * 10   20   7 *         * 5  10 26 34 37 38 39 41 48 51 54 61 *
|     | 8           |         | 3  7  11 28 29 31 50 52             |
|     | 19   6      | 4       | 8  15 24 30 35 36 46 49             |
|     | 18   5    9 |         | 4  13 14 18 27 42 53                |
|     |             | 3       | 19 25 44                            |
|  -1 + 15          + 6       + 2  20 23                            +
|     |             |         | 1  6  22 43                         |
|     |             | 5       | 21                                  |
|     |             |         |                                     |
|     |             |         |                                     |
|  -2 +             +         +                                     +
|     | 11          |         |                                     |
|     |             |         |                                     |
|     |             |         |                                     |
|     |             |         |                                     |
|  -3 +             +         +                                     +
|-----+-------------+---------+-------------------------------------+
|Measr|+Performer ID|-Rater ID|-Items                               |
+---------------------------------------------------------------------
```

Variable Map

Figure 4.1 is a variable map representing flute performance ability as a latent variable. Included on the map are the calibrations of each facet examined in the study: performers (column 2), raters (column 3), and items (column 4). Facets are ordered from top to bottom according to high to low ability of performers, severity to leniency of raters, and highest to lowest difficulty of items. The logit scale provides consistent value to the locations of elements for objective comparisons. (Bond & Fox, 2020).

Column 2 on the map shows the location of each performer along the latent variable, with numbers representing the performer number. Student achievement ranged from 6.52 to -2.16 logits (*M* = 0.95, *SD* = 1.78, *n* = 28). Misfit performers, or performers which exist outside of the linear measure of music performance ability, are identified by infit and outfit MSE values lower than 0.60 or higher than 1.40 (Bond & Fox, 2020). Underfitting performers included numbers 14, 17, and 24. Underfit indicates unpredictability or randomness in achievement level (e.g., a higher achieving performer scoring poorly on an easier item or a less able performer scoring well on a difficult item). There were no overfitting performances. Table 4.3 shows the complete calibration and statistics of performers.

Column 3 shows the calibration of raters along the latent construct, with numbers representing each rater number. All raters fell within the acceptable range for productive parameter-level MSE. Rater 5 was the most severe (observed average = 3.30, logit measure -.149) and Rater 1 was the most lenient (observed average = 2.90, logit measure 1.20).

Column 4 shows the calibration of items, with numbers representing each item number. The most difficult item was Item 69 (choice of vibrato depth and speed, observed average = 1.79, logit measure = 2.14). The easiest item was Item 21 (mouthplate placement on chin, observed average = 3.75, logit measure = -1.35). Items demonstrating underfit include 2, 3, 12, 15, 46, 52, 60, and 61. Underfitting items display less predictability than expected and could indicate problems such as varied interpretations by raters. Items demonstrating overfit include 29, 34, 41, 55, and 57. Overfitting items display more predictability than is desirable for the Rasch model. Table 4.4 shows the complete calibration and statistics for items.

Table 4.3

Calibration of Flute Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
27	3.99	6.52	0.69	1.00	0.20	5.18	2.10
26	3.88	4.55	0.26	0.93	-0.10	0.73	-0.50
25	3.83	4.10	0.21	1.03	0.20	1.38	1.10
24	3.61	3.11	0.16	1.56	3.30	1.58	2.20
1	3.20	1.94	0.13	0.89	-0.80	0.96	-0.10
14	3.67	1.54	0.17	2.03	4.80	2.72	4.40
2	3.00	1.51	0.12	0.95	-0.30	0.96	-0.20
23	2.92	1.42	0.11	0.62	-3.40	0.67	-2.60
28	2.93	1.23	0.11	0.92	-0.60	1.11	0.80
17	3.64	1.14	0.16	1.62	3.40	3.04	5.50
22	3.14	1.09	0.12	0.79	-1.60	0.78	-1.30
3	3.20	0.91	0.13	1.28	1.90	1.55	2.80
4	3.18	0.88	0.13	1.31	2.10	1.40	2.10
16	3.54	0.83	0.15	0.93	-0.40	0.82	-0.80
12	3.43	0.75	0.14	0.66	-2.60	0.69	-1.60
21	2.88	0.55	0.12	0.98	0.00	0.79	-1.40
13	3.21	0.22	0.13	1.26	1.70	1.65	3.20
20	2.62	0.09	0.11	0.88	-1.00	0.75	-1.80
7	3.01	0.08	0.12	0.91	-0.60	0.85	-1.00
10	2.93	-0.08	0.12	0.88	-0.90	0.86	-0.90
8	2.85	-0.22	0.11	0.78	-1.80	0.75	-1.90
19	2.38	-0.32	0.11	1.26	2.00	0.57	1.10
6	2.47	-0.40	0.11	0.66	-3.30	1.15	-3.80
5	2.41	-0.51	0.11	1.03	0.20	0.96	-0.20
18	2.94	-0.59	0.12	0.91	-0.60	0.87	-0.80
9	2.59	-0.65	0.11	0.81	-1.70	0.77	-1.90
15	2.72	-0.98	0.11	0.79	-1.80	0.70	-2.50
11	1.83	-2.16	0.11	1.30	2.30	1.22	1.30
Mean	3.07	0.95	0.15	1.03	0.00	1.27	0.10
SD	0.50	1.78	0.11	0.31	2.00	0.94	2.20

Presented in measure order from highest to lowest achievement.

Table 4.4

Calibration of Flute Item Facet

Item Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
69	1.79	2.14	0.19	1.06	0.30	0.78	-0.30
33	1.87	2.01	0.19	1.23	0.90	1.10	0.30
67	2.00	2.01	0.20	1.19	0.90	1.03	0.20
68	2.07	1.91	0.19	0.93	-0.30	0.87	-0.50
59	2.09	1.80	0.20	1.03	0.20	0.95	-0.10
65	2.23	1.65	0.18	1.03	0.20	1.01	0.10
66	2.38	1.34	0.18	1.24	1.30	1.27	1.20
62	2.44	1.12	0.20	0.67	-1.90	0.66	-1.60
63	2.52	1.03	0.19	0.66	-2.10	0.63	-1.80
40	2.57	0.95	0.19	0.84	-0.80	0.80	-0.90
58	2.78	0.65	0.18	0.76	-1.30	0.74	-0.80
45	2.63	0.65	0.21	0.92	-0.30	0.92	-0.20
60	2.77	0.59	0.17	1.43	2.10	2.12	2.50
64	2.75	0.58	0.19	0.81	-1.00	0.82	-0.70
57	3.00	0.36	0.18	0.57	-2.30	0.59	-1.50
32	2.95	0.35	0.19	0.76	-1.20	0.71	-1.10
56	2.96	0.33	0.18	0.69	-1.70	0.66	-1.20
16	2.91	0.26	0.20	0.64	-2.10	0.59	-1.80
47	2.95	0.15	0.21	0.63	-2.20	0.61	-1.80
55	3.07	0.12	0.20	0.58	-2.30	0.55	-1.70
9	3.04	0.12	0.19	1.09	0.50	1.05	0.20
17	3.00	0.12	0.18	0.93	-0.30	1.08	0.30
12	3.13	0.11	0.16	1.55	2.50	1.91	1.40
5	3.41	0.08	0.21	1.39	2.00	1.87	1.30
48	3.07	0.08	0.21	1.03	0.20	1.05	0.20
61	3.18	0.08	0.16	2.01	3.70	7.33	4.20
26	3.00	0.07	0.19	0.72	-1.60	0.61	-1.10
10	3.04	0.03	0.19	1.21	1.10	1.16	0.50
54	3.13	0.00	0.19	0.60	-2.20	0.51	-1.50
51	3.11	-0.02	0.20	0.88	-0.50	0.98	0.00
38	3.02	-0.03	0.19	0.64	-2.20	0.57	-1.40
37	3.02	-0.04	0.19	0.72	-1.70	0.68	-0.90
39	2.95	-0.06	0.20	0.60	-2.50	0.59	-1.60
34	3.07	-0.06	0.20	0.51	-3.10	0.46	-2.20
41	3.09	-0.07	0.20	0.58	-2.40	0.54	-1.80
52	3.27	-0.13	0.18	1.41	1.60	1.20	0.50
50	3.11	-0.18	0.23	0.85	-0.70	0.81	-0.60
7	3.18	-0.20	0.20	1.24	1.10	1.38	1.10
31	3.20	-0.20	0.22	0.61	-1.90	0.58	-1.50
11	3.23	-0.20	0.19	1.19	0.90	1.30	0.70
29	3.11	-0.22	0.21	0.58	-2.40	0.57	-1.70
3	3.30	-0.22	0.18	2.07	3.80	8.26	5.60
28	3.14	-0.25	0.22	0.94	-0.20	0.96	0.00
24	3.23	-0.40	0.21	0.81	-0.90	0.78	-0.60
8	3.34	-0.42	0.20	0.88	-0.50	0.78	-0.40
30	3.21	-0.42	0.19	0.62	-2.40	0.48	-1.20
36	3.23	-0.43	0.20	0.74	-1.40	0.76	-0.50
46	3.44	-0.44	0.20	1.50	1.80	1.40	0.70
49	3.27	-0.45	0.21	0.82	-0.80	0.84	-0.30
35	3.29	-0.47	0.22	0.80	-0.90	0.81	-0.40

15	3.54	-0.48	0.20	1.48	1.40	3.45	1.90
14	3.50	-0.52	0.20	1.17	0.60	2.22	1.50
27	3.14	-0.53	0.21	0.61	-2.40	0.56	-1.50
18	3.11	-0.54	0.19	0.76	-1.40	0.62	-0.80
42	3.16	-0.55	0.21	1.00	0.00	0.89	-0.20
53	3.45	-0.55	0.21	1.22	0.90	1.12	0.30
4	3.45	-0.55	0.20	1.38	1.40	9.00	5.90
13	3.50	-0.62	0.20	1.21	0.80	0.96	0.20
25	3.30	-0.73	0.23	0.72	-1.50	0.63	-1.10
44	3.52	-0.78	0.23	1.23	0.90	0.94	0.00
19	3.54	-0.80	0.22	1.01	0.10	0.69	-0.30
2	3.64	-0.94	0.24	1.57	1.70	1.23	0.50
23	3.50	-1.00	0.24	0.78	-0.90	0.67	-0.60
20	3.75	-1.09	0.26	1.00	0.10	0.45	-0.20
43	3.58	-1.11	0.24	1.12	0.50	0.95	0.10
22	3.63	-1.17	0.25	1.03	0.10	0.81	-0.10
6	3.64	-1.17	0.23	0.72	-1.00	0.66	0.00
1	3.70	-1.27	0.26	1.39	1.20	1.17	0.40
21	3.75	-1.35	0.28	0.97	0.00	0.55	-0.20
Mean	3.07	0.00	0.20	0.98	-0.30	1.25	-0.10
SD	0.44	0.81	0.02	0.34	1.60	1.56	1.50

Presented in measure order from highest to lowest difficulty.

Response Category Diagnostics

The MFR-PC analysis provides detailed response category statistics shown in table 4.5 below. The table shows the usage of each response category (number of instances and percentage of uses), average observed logit measure compared to the average expected logit measure, and the outfit mean squared error (MSE) value for each item in the rating scale. This data can be used to optimize response categories by eliminating underused categories and/or combining categories with too small of a threshold between them. This will be discussed in chapter 5. A review of response category diagnostics for the FPRS shows many opportunities for optimization. For instance, in the case of item 1 (Upper Body Position), response category 1 (Unacceptable) was only used once out of a total of 56 scoring opportunities. Based on Linacre's (2002) recommendations, this response category should be eliminated when revising the FPRS. Other items would require elimination of categories based an outfit MSE greater than or equal to 2.00. This applies to several of the categories in the FPRS, most of which will already be eliminated based on insufficient usage. All four response categories for Item 3, “Placement of Feet,” show a value greater than 2.00, but this item also would be eliminated based on item fit statistics.

Response categories would also be evaluated for proper step ordering (Linacre, 2002). In other words, the level of difficulty of achieving each category should increase across responses 1 through 4. As an example, category 2 on item 10 shows a logit measure of 0.12 which is less than category 1 with a measure of 0.26. Therefore, categories 1 and 2 would be collapsed into one response category.

Table 4.5. FPRS Category Diagnostics for Fit Items: Category Usage, Average Observed and Expected Logit Measure, Outfit Mean Squared Error (*MSE*).

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
1	1(2)	2(4)	10(18)	43(77)	0.08(0.27)	1.40(0.82)	1.91(1.56)	2.66(2.76)	0.50	1.50	1.10	1.40
2	2(4)	2(4)	10(18)	42(75)	0.37(0.05)	0.76(0.61)	1.74(1.31)	2.35(2.47)	1.00	0.90	1.30	1.70
3	5(9)	6(11)	12(21)	33(59)	0.56(-0.31)	0.95(0.28)	1.63(0.88)	1.50(2.02)	6.10	7.00	9.90	2.90
4	3(5)	5(9)	12(21)	36(64)	-0.19(-0.15)	1.09(0.45)	1.64(1.10)	1.99(2.25)	0.60	3.10	9.90	1.70
5	0(0)	11(20)	11(20)	34(61)		0.24(-0.20)	0.52(0.50)	1.54(1.69)		2.70	0.50	1.90
6	1(2)	6(11)	5(9)	44(79)	0.50(0.30)	0.52(0.87)	1.61(1.55)	2.71(2.68)	0.70	0.60	0.70	0.70
7	3(5)	7(13)	23(41)	23(41)	0.25(-0.40)	0.57(0.26)	0.79(0.97)	2.33(2.31)	2.90	1.60	0.60	1.00
8	3(5)	5(9)	18(32)	30(54)	-0.76(-0.25)	0.49(0.37)	1.11(1.05)	2.28(2.28)	0.40	0.90	0.70	1.00
9	5(9)	10(18)	19(34)	22(39)	-0.58(-0.54)	0.24(0.10)	0.77(0.76)	2.03(2.08)	1.00	1.50	0.70	1.10
10	4(7)	13(23)	16(29)	23(41)	0.26(-0.46)	0.12(0.19)	0.73(0.86)	2.16(2.14)	2.30	1.00	0.80	0.90
11	4(7)	8(14)	15(27)	29(52)	0.02(-0.33)	0.19(0.29)	1.06(0.93)	2.05(2.13)	2.90	0.60	0.90	1.20
12	8(14)	8(14)	9(16)	31(55)	-0.28(-0.44)	0.77(0.13)	0.44(0.67)	1.66(1.79)	2.50	2.90	0.60	1.70
13	3(5)	6(11)	7(13)	40(71)	0.39(-0.08)	0.52(0.51)	0.83(1.13)	2.25(2.23)	2.20	0.50	0.70	1.00
14	4(7)	3(5)	10(18)	39(70)	0.24(-0.18)	0.50(0.40)	0.79(1.02)	2.15(2.14)	7.70	0.70	0.20	1.10
15	5(9)	2(4)	7(13)	42(75)	0.43(-0.19)	0.30(0.36)	0.79(0.96)	2.00(2.04)	9.90	0.50	0.10	1.30
16	4(7)	13(23)	23(41)	16(29)	-1.24(-0.68)	0.01(0.01)	0.63(0.76)	2.59(2.23)	0.50	0.70	0.50	0.70
17	5(9)	14(25)	13(23)	24(43)	-0.83(-0.49)	0.29(0.15)	0.82(0.78)	2.02(2.03)	0.70	1.20	1.50	0.90
18	1(2)	17(30)	13(23)	25(45)	-0.65(-0.02)	0.51(0.67)	1.47(1.37)	2.73(2.63)	0.90	0.60	0.30	0.80
19	2(4)	4(7)	12(21)	38(68)	-0.18(-0.01)	0.47(0.59)	1.32(1.27)	2.45(2.44)	0.50	0.70	0.50	1.10
20	2(4)	1(2)	6(11)	47(84)	0.11(0.12)	0.78(0.64)	0.97(1.34)	2.55(2.50)	0.40	0.50	0.40	0.90
21	1(2)	1(2)	9(16)	45(80)	0.16(0.29)	0.57(0.82)	1.46(1.57)	2.82(2.79)	0.50	0.30	0.50	1.00
22	1(2)	2(4)	14(25)	39(70)	-0.01(0.20)	0.96(0.78)	1.49(1.53)	2.77(2.76)	0.50	1.00	0.70	1.00
23	1(2)	3(5)	19(34)	33(59)	-0.18(0.10)	0.29(0.73)	1.42(1.49)	2.84(2.75)	0.60	0.60	0.60	0.90
24	2(4)	7(13)	23(41)	24(43)	-0.58(-0.27)	0.16(0.40)	1.18(1.13)	2.53(2.47)	0.60	0.70	0.80	0.90
25	1(2)	6(11)	24(43)	25(45)	-0.46(-0.05)	0.17(0.63)	1.38(1.40)	2.90(2.75)	0.60	0.50	0.40	0.90
26	4(7)	14(25)	16(29)	22(39)	-1.02(-0.49)	0.17(0.17)	0.77(0.85)	2.31(2.15)	0.60	0.60	0.50	0.80

(continued)

Table 4.5. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
27	1(2)	12(21)	21(38)	22(39)	-1.14(-0.11)	0.29(0.60)	1.37(1.35)	2.90(2.71)	0.50	0.50	0.40	0.70
28	2(4)	7(13)	28(50)	19(34)	-0.38(-0.41)	-0.02(0.29)	1.20(1.07)	2.46(2.53)	1.20	0.60	1.00	1.00
29	2(4)	9(16)	26(46)	19(34)	-1.12(-0.40)	0.26(0.30)	0.86(1.07)	2.90(2.51)	0.40	0.90	0.30	0.60
30	2(4)	12(21)	14(25)	28(50)	-0.49(-0.16)	0.25(0.49)	1.05(1.16)	2.58(2.39)	0.70	0.50	0.10	0.70
31	3(5)	4(7)	27(49)	21(38)	-0.99(-0.46)	0.04(0.20)	0.85(0.95)	2.63(2.38)	0.30	0.70	0.40	0.80
32	6(11)	8(14)	25(45)	17(30)	-0.88(-0.74)	-0.23(-0.09)	0.56(0.63)	2.30(2.07)	0.90	0.50	0.60	0.80
33	31(57)	9(17)	4(7)	10(19)	-1.52(-1.59)	-0.84(-0.82)	-0.68(-0.02)	1.36(1.25)	1.90	0.40	2.10	0.50
34	3(5)	10(18)	23(41)	20(36)	-1.02(-0.48)	0.09(0.20)	0.66(0.93)	2.78(2.32)	0.50	0.60	0.20	0.50
35	2(4)	5(9)	24(43)	25(45)	-0.16(-0.25)	0.25(0.41)	0.98(1.15)	2.69(2.49)	1.30	0.70	0.70	0.80
36	2(4)	10(18)	17(30)	27(48)	-0.02(-0.19)	0.09(0.47)	1.23(1.16)	2.51(2.42)	1.10	0.50	0.80	0.80
37	3(5)	15(27)	16(29)	22(39)	-0.46(-0.41)	0.11(0.26)	0.83(0.95)	2.47(2.26)	0.90	0.90	0.20	0.70
38	3(5)	14(25)	18(32)	21(38)	-0.48(-0.44)	0.00(0.23)	0.81(0.94)	2.56(2.28)	0.90	0.60	0.30	0.60
39	2(4)	17(30)	19(34)	18(32)	-1.37(-0.42)	0.14(0.29)	1.10(1.04)	2.67(2.46)	0.60	0.60	0.50	0.70
40	9(16)	17(30)	19(34)	11(20)	-1.31(-1.10)	-0.46(-0.42)	0.55(0.37)	1.91(1.96)	0.80	0.60	0.80	1.00
41	3(5)	9(16)	24(43)	20(36)	-1.10(-0.48)	0.17(0.20)	0.72(0.93)	2.69(2.33)	0.40	0.70	0.30	0.60
42	1(2)	11(20)	22(39)	22(39)	-1.12(-0.11)	0.66(0.59)	1.38(1.35)	2.71(2.72)	0.50	0.80	0.50	1.30
43	1(2)	3(5)	14(25)	37(67)	-0.56(0.18)	0.99(0.77)	1.72(1.51)	2.68(2.75)	0.30	1.00	1.00	1.20
44	2(4)	2(4)	17(30)	35(63)	-0.13(-0.07)	0.53(0.54)	1.46(1.26)	2.39(2.48)	1.00	0.70	0.80	1.20
45	4(7)	24(43)	17(30)	11(20)	-1.12(-0.92)	-0.19(-0.18)	0.76(0.68)	2.30(2.28)	0.90	1.20	0.70	1.00
46	4(7)	4(7)	11(20)	36(65)	0.13(-0.22)	0.76(0.36)	0.98(1.00)	2.07(2.15)	1.90	2.20	0.60	1.50
47	3(5)	12(21)	26(46)	15(27)	-0.77(-0.64)	-0.27(0.06)	0.83(0.85)	2.73(2.38)	0.90	0.50	0.40	0.60
48	4(7)	6(11)	28(50)	18(32)	-0.74(-0.62)	0.21(0.05)	0.85(0.80)	2.17(2.26)	0.90	1.10	1.10	1.00
49	2(4)	6(11)	23(41)	25(45)	-0.98(-0.25)	0.38(0.41)	1.23(1.15)	2.47(2.48)	0.30	0.70	1.00	1.00
50	2(4)	7(13)	30(54)	17(30)	-0.73(-0.47)	0.15(0.23)	1.00(1.04)	2.71(2.56)	0.80	0.90	0.70	0.90
51	4(7)	8(14)	22(39)	22(39)	-1.05(-0.50)	0.21(0.16)	1.10(0.85)	2.04(2.19)	0.40	0.80	1.50	1.10
52	6(11)	5(9)	12(22)	32(58)	-0.27(-0.36)	0.23(0.22)	1.26(0.79)	1.76(1.95)	1.00	0.60	1.10	2.00
53	3(5)	4(7)	14(25)	35(63)	-0.47(-0.17)	0.55(0.44)	1.62(1.10)	2.08(2.27)	0.40	0.80	1.40	1.50

(continued)

Table 4.5. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
54	5(9)	7(13)	19(35)	24(44)	-0.70(-0.49)	-0.28(0.14)	0.77(0.78)	2.27(2.08)	0.60	0.30	0.30	0.80
55	5(9)	6(11)	25(45)	20(36)	-0.82(-0.61)	-0.25(0.04)	0.59(0.74)	2.47(2.13)	0.60	0.90	0.20	0.70
56	7(13)	8(14)	21(38)	20(36)	-0.94(-0.67)	0.08(-0.05)	0.40(0.61)	2.23(1.96)	0.60	1.40	0.20	0.70
57	8(15)	4(7)	23(42)	20(36)	-0.95(-0.72)	-0.17(-0.10)	0.39(0.57)	2.24(1.91)	0.90	0.40	0.30	0.70
58	10(19)	9(17)	18(33)	17(31)	-1.10(-0.85)	-0.20(-0.24)	0.35(0.43)	2.01(1.83)	0.90	0.70	0.60	0.70
59	18(32)	22(39)	9(16)	7(13)	-1.81(-1.64)	-0.55(-0.88)	-0.47(0.07)	1.90(1.71)	0.90	1.10	1.20	0.60
60	10(18)	13(23)	13(23)	20(36)	-0.49(-0.75)	-0.11(-0.16)	0.79(0.48)	1.42(1.77)	4.30	1.00	1.50	1.40
61	9(16)	5(9)	8(15)	33(60)	0.39(-0.42)	0.71(0.12)	0.70(0.64)	1.42(1.74)	9.90	2.20	3.40	1.80
62	8(15)	24(44)	14(25)	9(16)	-1.32(-1.24)	-0.67(-0.51)	0.60(0.39)	2.31(2.06)	0.90	0.50	0.60	0.60
63	10(18)	18(32)	17(30)	11(20)	-1.44(-1.13)	-0.45(-0.46)	0.38(0.34)	2.15(1.90)	0.70	0.60	0.60	0.60
64	6(11)	16(29)	20(36)	14(25)	-1.05(-0.86)	-0.16(-0.18)	0.52(0.58)	2.26(2.08)	0.80	1.00	0.90	0.70
65	18(32)	14(25)	17(30)	7(13)	-1.48(-1.54)	-1.02(-0.87)	0.12(0.01)	1.70(1.75)	1.70	0.30	0.50	1.20
66	14(25)	16(29)	17(30)	9(16)	-0.98(-1.33)	-0.85(-0.66)	0.07(0.17)	1.82(1.80)	2.10	0.90	0.90	0.80
67	21(38)	20(36)	9(16)	6(11)	-1.75(-1.78)	-0.87(-1.00)	-0.38(-0.03)	1.80(1.68)	1.30	0.90	1.00	0.80
68	20(36)	18(32)	12(21)	6(11)	-1.76(-1.72)	-0.98(-0.98)	-0.03(-0.03)	1.89(1.72)	1.30	0.30	0.80	0.80
69	33(59)	11(20)	3(5)	9(16)	-1.63(-1.69)	-1.17(-0.90)	-0.05(-0.06)	1.38(1.26)	1.50	0.80	0.20	0.50

Clarinet Performance Rating Scale Data

Multifaceted Rasch Partial Credit Model Results

Table 4.6 presents summary statistics from the analysis of performers (*n* = 35), raters (*n* = 11), and items (*n* = 64) on the Clarinet Performance Rating Scale (CPRS). Overall significant differences in chi-square are indicated for performers (2192.30), raters (720.10), and items (1375.90). The probability for each facet was less than 0.01 and reliability of separation for each was high with performers at 0.99, raters at 0.96, and items at 0.95. This indicates adequate spread of elements within each facet along a single measure of music performance ability. All MSE measures were well targeted with values close to 1.00. Infit MSE values on the CPRS are 1.02 for performers, 1.02 for raters, and 1.01 for items. Outfit MSE values were also well targeted with values of 1.12 for performers, 1.14 for raters, and 1.13 for items.

Table 4.6

CPRS Summary Statistics from the PC-MFR Model

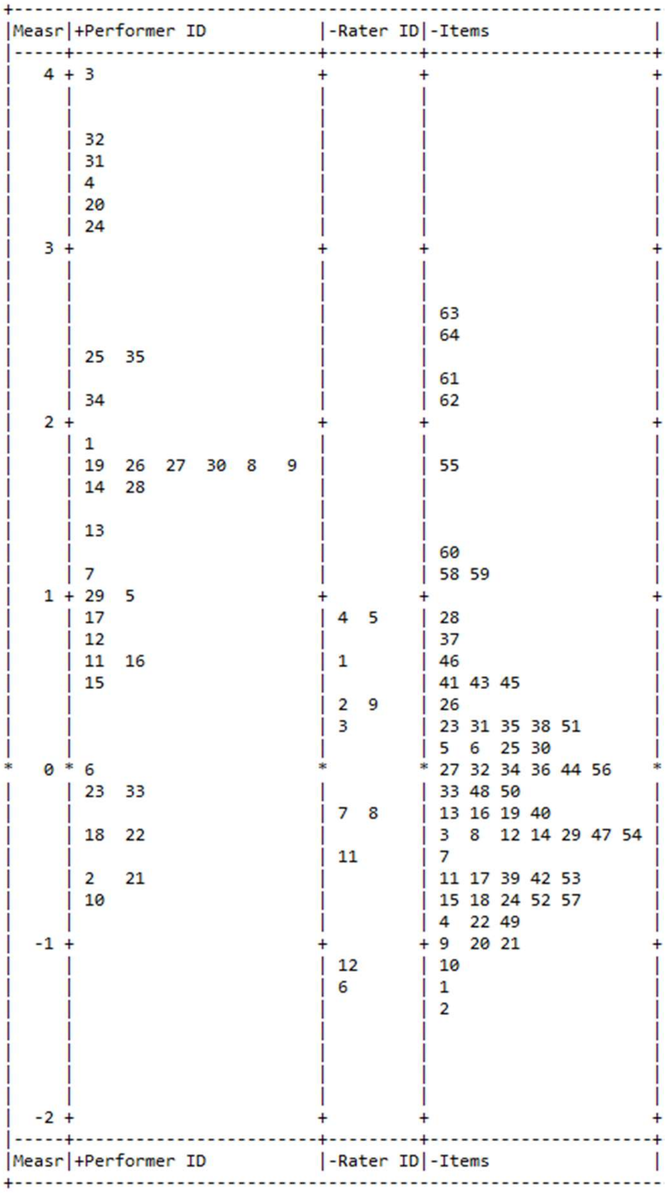
			Facets	
		Performance (θ)	**Rater** (λ)	**Item** (δ)
Measure (Logits)				
	Mean	1.38	0.00	0.00
	SD	1.29	0.72	0.87
	N	35	11	64
Infit *MSE*				
	Mean	1.02	1.02	1.01
	SD	0.29	0.15	0.28
Std. Infit *MSE*				
	Mean	-0.10	0.20	-0.10
	SD	2.30	2.40	1.40
Outfit *MSE*				
	Mean	1.12	1.14	1.13
	SD	0.42	0.28	0.73

Std. Outfit *MSE*			
Mean	0.20	0.50	-0.10
SD	2.20	2.70	1.60
Separation Statistics			
Reliability of Separation	0.99	0.96	0.95
Chi-Square	2192.30	720.10	1375.90
Degrees of Freedom	34	10	63

**p* < 0.01

Figure 4.2

Variable Map, Clarinet Performance Rating Scale



```
+-----------------------------------------------------------------
|Measr|+Performer ID           |-Rater ID|-Items                  |
|-----+------------------------+---------+------------------------+
|   4 + 3                      +         +                        +
|     |                        |         |                        |
|     |                        |         |                        |
|     | 32                     |         |                        |
|     | 31                     |         |                        |
|     | 4                      |         |                        |
|     | 20                     |         |                        |
|     | 24                     |         |                        |
|   3 +                        +         +                        +
|     |                        |         |                        |
|     |                        |         |                        |
|     |                        |         | 63                     |
|     |                        |         | 64                     |
|     | 25  35                 |         |                        |
|     |                        |         | 61                     |
|     | 34                     |         | 62                     |
|   2 +                        +         +                        +
|     | 1                      |         |                        |
|     | 19  26  27  30  8   9  |         | 55                     |
|     | 14  28                 |         |                        |
|     |                        |         |                        |
|     | 13                     |         |                        |
|     |                        |         | 60                     |
|     | 7                      |         | 58 59                  |
|   1 + 29  5                  +         +                        +
|     | 17                     | 4   5   | 28                     |
|     | 12                     |         | 37                     |
|     | 11  16                 | 1       | 46                     |
|     | 15                     |         | 41 43 45               |
|     |                        | 2   9   | 26                     |
|     |                        | 3       | 23 31 35 38 51         |
|     |                        |         | 5  6  25 30            |
*   0 * 6                      *         * 27 32 34 36 44 56      *
|     | 23  33                 |         | 33 48 50               |
|     |                        | 7   8   | 13 16 19 40            |
|     | 18  22                 |         | 3  8  12 14 29 47 54   |
|     |                        | 11      | 7                      |
|     | 2   21                 |         | 11 17 39 42 53         |
|     | 10                     |         | 15 18 24 52 57         |
|     |                        |         | 4  22 49               |
|  -1 +                        +         + 9  20 21               +
|     |                        | 12      | 10                     |
|     |                        | 6       | 1                      |
|     |                        |         | 2                      |
|     |                        |         |                        |
|     |                        |         |                        |
|     |                        |         |                        |
|     |                        |         |                        |
|  -2 +                        +         +                        +
|-----+------------------------+---------+------------------------+
|Measr|+Performer ID           |-Rater ID|-Items                  |
+-----------------------------------------------------------------
```

Variable Map

Figure 4.2 is a variable map representing clarinet performance ability as a latent variable. Included on the map are the calibrations of performers (column 2), raters (column 3), and items (column 4). Facets are ordered from top to bottom according to high to low ability of performers, severity to leniency of raters, and highest to lowest difficulty of items.

Column 2 on the map shows the location of each performer along the latent variable, with numbers representing each performer number. Student achievement ranged from 3.97 to -0.74 logits (*M* = 1.38, *SD* = 1.29, *n* = 35). Underfitting performers included 5, 19, 23, and 29. Overfitting performers included 8 and 12. Table 4.7 shows the complete calibration and statistics for performers.

Column 3 shows the calibration of raters along the latent construct, with numbers representing the rater number. All raters fell within the acceptable range for productive parameter-level MSE. Rater 4 was the most severe (observed average = 2.65, logit measure 0.89) and Rater 6 was the most lenient (observed average = 3.40, logit measure -1.30).

Column 4 shows the calibration of items, with numbers representing the item number. The most difficult item was Item 63 (dynamics used to create tension and release, observed average = 1.83, logit measure = 2.67). The easiest item was Item 2 (head position, observed average = 3.66, logit measure = -1.33). Items demonstrating underfit include 3, 4, 11, 12, and 48. Underfitting items display less predictability than expected and could indicate problems such as varied interpretations by raters. Items demonstrating overfit include 50 and 52. Overfitting items display more predictability than is desirable for the Rasch model. Table 4.8 shows the complete calibration and statistics for items.

Table 4.7

Calibration of Clarinet Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
3	3.85	3.97	0.25	1.35	1.40	2.90	22.30
32	3.77	3.62	0.21	0.92	-0.30	0.98	0.00
41	3.73	3.46	0.20	0.91	-0.40	1.41	1.40
4	3.75	3.32	0.20	1.14	0.80	1.69	2.10
20	3.88	3.24	0.27	1.18	0.70	2.12	1.90
24	3.80	3.10	0.22	1.11	0.60	1.30	0.90
25	3.66	2.44	0.18	0.89	-0.60	1.14	0.60
35	3.79	2.42	0.22	0.96	-0.10	1.22	0.70
34	3.36	2.14	0.14	1.04	0.30	1.18	1.10
1	3.27	1.90	0.14	0.88	-0.90	1.00	0.00
8	3.19	1.80	0.14	0.58	-3.90	0.56	-3.70
30	3.37	1.79	0.15	0.66	-2.80	0.61	-2.80
27	3.37	1.79	0.15	0.72	-2.20	0.87	-0.70
19	3.62	1.78	0.17	1.44	2.40	1.98	3.30
9	3.17	1.76	0.14	1.14	1.10	1.12	0.80
26	3.46	1.72	0.15	1.22	1.50	1.25	1.40
28	3.33	1.68	0.14	0.77	-1.80	0.73	-1.80
14	2.94	1.57	0.13	1.07	0.60	1.14	1.10
13	2.85	1.38	0.12	0.73	-2.50	0.70	-2.80
7	2.86	1.09	0.13	0.82	-1.60	0.79	-1.80
5	2.96	1.01	0.13	1.81	5.60	1.82	5.50
29	3.01	0.94	0.13	1.47	3.50	1.39	2.80
17	3.08	0.89	0.14	0.73	-2.30	0.74	-1.80
12	2.54	0.80	0.12	0.51	-5.30	0.54	-4.80
16	2.99	0.69	0.13	0.68	-2.80	0.62	-2.90
11	2.42	0.56	0.12	1.14	1.20	1.11	0.90
15	2.91	0.49	0.13	0.69	-2.70	0.67	-2.50
6	2.43	0.01	0.12	0.96	-0.30	0.90	-0.80
33	2.27	-0.08	0.12	0.97	-0.20	0.92	-0.60
23	2.62	-0.13	0.12	1.47	3.60	1.14	3.20
18	2.53	-0.33	0.13	0.81	-1.60	0.75	-2.00
22	2.76	-0.40	0.12	1.29	2.30	1.20	1.50
2	2.00	-0.64	0.13	1.33	2.40	1.16	1.10
21	2.62	-0.67	0.12	1.01	0.10	1.07	0.60
10	1.94	-0.74	0.14	1.24	1.70	1.08	0.60
Mean	3.09	1.38	0.15	1.02	-0.10	1.12	0.20
SD	0.54	1.29	0.04	0.29	2.30	0.42	2.20

Presented in measure order from highest to lowest achievement.

Table 4.8

Calibration of Clarinet Item Facet

Item Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
63	1.83	2.67	0.19	0.92	-0.40	0.86	-0.60
67	1.89	2.51	0.18	1.08	0.50	1.02	0.10
61	2.01	2.21	0.18	1.08	0.50	1.02	0.10
62	2.06	2.10	0.18	0.97	-0.10	0.91	-0.40
55	2.20	1.78	0.18	0.91	-0.50	0.90	-0.50
60	2.50	1.20	0.16	0.76	-1.40	0.69	-1.70
59	2.51	1.19	0.17	0.79	-1.30	0.87	-0.70
58	2.54	1.11	0.17	0.81	-1.10	0.78	-1.30
28	2.69	0.85	0.18	0.77	-1.50	0.75	-1.60
37	2.73	0.76	0.18	0.80	-1.30	0.81	-1.20
46	3.17	0.67	0.22	0.78	-1.40	0.75	-1.50
41	2.83	0.52	0.18	0.91	-0.50	0.90	-0.50
43	2.83	0.46	0.19	0.92	-0.40	0.88	-0.60
45	3.23	0.45	0.22	0.82	-1.10	0.77	-1.40
26	2.93	0.39	0.18	0.84	-0.90	0.85	-0.80
51	2.97	0.29	0.17	0.89	-0.60	0.97	0.00
38	2.96	0.26	0.18	1.03	0.20	0.94	-0.20
23	3.36	0.23	0.20	0.78	-1.30	0.59	-1.40
35	2.91	0.21	0.18	0.78	-1.30	0.78	-1.10
31	2.91	0.19	0.19	0.81	-1.10	0.83	-1.00
25	2.99	0.18	0.18	0.85	-0.90	0.85	-0.70
5	3.37	0.13	0.21	0.94	-0.20	1.17	0.70
6	3.39	0.12	0.21	0.91	-0.40	0.92	-0.10
30	3.03	0.07	0.19	0.76	-1.50	0.78	-1.30
32	3.07	0.05	0.18	0.81	-1.10	0.73	-1.40
56	3.01	0.02	0.18	1.28	1.50	1.37	1.40
34	3.04	0.02	0.18	0.75	-1.50	0.72	-1.40
44	3.03	0.02	0.18	1.00	0.00	0.85	-0.60
36	3.03	0.00	0.19	0.97	0.00	0.96	-0.10
27	3.00	-0.04	0.21	0.75	-1.60	0.74	-1.60
50	3.41	-0.14	0.23	0.55	-3.20	0.48	-2.80
48	3.56	-0.16	0.23	1.54	2.20	2.43	1.80
33	3.19	-0.18	0.19	0.69	-1.90	0.58	-2.20
19	3.44	-0.20	0.23	0.88	-0.60	0.75	-1.00
40	3.17	-0.21	0.18	1.23	1.20	1.18	0.60
16	3.51	-0.22	0.21	0.90	-0.40	0.60	-1.00
13	3.14	-0.30	0.20	1.09	0.50	1.12	0.60
29	3.14	-0.32	0.20	0.71	-1.80	0.71	-1.60
12	3.14	-0.33	0.19	1.46	2.40	1.70	3.00
47	3.13	-0.33	0.19	1.07	0.40	1.02	0.10
3	3.29	-0.36	0.18	2.13	4.60	4.20	5.90
54	3.17	-0.38	0.19	0.75	-1.50	0.77	-1.10
14	3.17	-0.39	0.19	1.15	0.80	1.08	0.40
8	3.01	-0.40	0.19	1.25	1.40	1.81	2.90
7	3.10	-0.52	0.19	1.20	1.10	1.68	2.60
42	3.44	-0.57	0.19	1.28	1.30	0.83	-0.30
39	3.40	-0.59	0.19	1.31	1.40	1.22	0.80
53	3.19	-0.62	0.20	0.73	-1.70	0.86	-0.60
17	3.33	-0.63	0.19	1.21	1.10	1.22	0.70
11	3.36	-0.67	0.19	1.58	2.70	1.37	1.10

24	3.21	-0.69	0.19	1.01	0.10	1.18	0.70
57	3.52	-0.74	0.22	1.40	1.60	1.32	0.80
18	3.41	-0.75	0.20	1.01	0.10	0.86	-0.30
52	3.30	-0.79	0.21	0.57	-2.90	0.64	-1.80
15	3.29	-0.79	0.20	1.13	0.70	1.05	0.20
22	3.49	-0.86	0.20	1.31	1.40	0.94	0.00
49	3.33	-0.86	0.19	1.24	1.30	0.96	0.00
4	3.61	-0.90	0.20	1.56	1.90	2.79	2.60
9	3.69	-0.96	0.25	0.78	-1.00	0.54	-1.00
21	3.44	-1.01	0.21	0.95	-0.10	0.82	-0.60
20	3.44	-1.03	0.20	1.06	0.30	0.91	-0.10
10	3.50	-1.10	0.21	1.28	1.30	1.63	1.60
1	3.60	-1.25	0.21	1.26	1.10	4.65	4.10
2	3.66	-1.33	0.23	1.16	0.70	2.18	2.20
Mean	3.09	0.00	0.19	1.01	-0.10	1.13	-0.10
SD	0.42	0.87	0.02	0.28	1.40	0.73	1.60

Presented in measure order from highest to lowest difficulty.

Response Category Diagnostics

Table 4.9 shows response category data for the CPRS including the usage of each response category, average observed and average expected logit measures, and the outfit mean squared error (MSE) value for each item in the rating scale. This data can be used to optimize response categories by eliminating underused categories and/or combining categories with too small of a threshold between them. Linacre (2002) also suggests eliminating response categories with a MSE value greater than 2.00. Based on this criteria, response categories for item 1 (Upper Body Position) could be optimized to include only two options: Unacceptable and Acceptable. Fifty-one items would also require collapsing of response categories due to insufficient usage.

Examination of Differential Item Functioning

An analysis of differential item functioning (DIF) was performed to determine if there were statistically significant differences in scores between soprano clarinet and bass clarinet performances which were judged using the CPRS. Specifically, an item by instrument-type DIF analysis was performed to examine a total of 129 interactions. Results showed no statistically significant effect of DIF (χ^2 = 67.2, df = 128, *p* = 1) which is likely due to the small sample size of bass clarinet performances (*n* = 1). Appendix (B?) shows DIF analysis statistics.

Table 4.9. CPRS Category Diagnostics for Fit Items: Category Usage, Average Observed and Expected Logit Measure, Outfit Mean Squared Error (*MSE*).

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
1	1(1)	7(10)	11(16)	51(73)	-0.01(0.36)	1.22(0.81)	2.28(1.57)	2.76(2.96)	0.60	9.90	2.70	1.30
2	1(1)	3(4)	15(21)	51(73)	0.07(0.37)	0.82(0.80)	2.11(1.56)	2.85(3.01)	0.50	0.70	3.20	1.20
3	4(6)	11(16)	16(23)	39(56)	0.45(-0.30)	1.16(0.24)	1.45(1.08)	1.90(2.39)	9.90	5.20	1.80	2.20
4	3(4)	3(4)	12(17)	52(74)	0.44(0.04)	1.43(0.47)	1.53(1.19)	2.42(2.58)	1.60	3.30	3.50	1.20
5		11(16)	22(31)	37(53)		-0.66(-0.39)	0.80(0.52)	1.83(1.92)		0.70	1.70	1.00
6		12(17)	19(27)	39(56)		-0.42(-0.39)	0.42(0.50)	1.93(1.88)		0.80	1.10	0.80
7	1(1)	20(29)	20(29)	29(41)	0.59(-0.11)	0.57(0.53)	1.74(1.54)	2.64(2.83)	1.00	2.80	0.80	1.50
8	1(1)	24(34)	18(26)	27(39)	-0.55(-0.18)	0.68(0.51)	1.63(1.54)	2.59(2.79)	1.00	3.20	0.90	1.40
9		4(6)	14(20)	52(74)		0.10(0.39)	0.95(1.16)	2.70(2.62)		0.50	0.40	0.80
10	1(1)	7(10)	18(26)	44(63)	1.16(0.24)	0.95(0.73)	1.66(1.55)	2.86(2.96)	1.80	1.50	2.00	1.20
11	2(3)	11(16)	17(24)	40(57)	0.88(-0.06)	0.88(0.47)	1.16(1.32)	2.57(2.66)	1.90	1.80	0.80	1.60
12	2(3)	14(20)	26(37)	28(40)	0.37(-0.34)	0.68(0.28)	1,27(1.28)	2.38(2.63)	2.30	2.60	1.10	1.30
13	2(3)	11(16)	32(46)	25(36)	-0.48(-0.40)	0.36(0.21)	1.28(1.25)	2.57(2.65)	1.00	1.10	1.20	1.10
14	2(3)	15(22)	21(30)	31(45)	-0.39(-0.27)	0.59(0.33)	1.12(1.30)	2.63(2.63)	0.90	1.30	1.00	1.00
15	1(1)	14(20)	19(27)	36(51)	1.42(0.07)	0.64(0.64)	1.48(1.55)	2.89(2.89)	1.80	0.90	1.10	1.10
16		10(14)	14(20)	46(66)		-.08(-0.15)	0.25(0.66)	2.16(2.05)		0.90	0.20	0.80
17	2(3)	12(17)	17(24)	39(56)	0.47(-0.08)	0.61(0.45)	1.22(1.32)	2.62(2.65)	1.60	0.80	1.60	1.00
18	2(3)	9(13)	17(24)	42(60)	0.17(-0.02)	0.49(0.49)	1.19(1.32)	2.73(2.68)	1.00	0.80	0.80	1.00
19		7(10)	25(36)	38(54)		-0.01(-0.16)	0.47(0.75)	2.35(2.20)		1.20	0.40	0.80
20	1(1)	11(16)	14(20)	44(36)	1.65(0.24)	0.54(0.74)	1.57(1.56)	2.93(2.92)	2.60	0.50	0.80	1.20
21	1(1)	7(10)	22(31)	40(57)	0.80(0.17)	0.47(0.68)	1.47(1.54)	3.02(2.96)	1.40	0.70	0.70	1.00
22	2(3)	8(11)	14(20)	46(66)	1.57(0.06)	0.38(0.54)	1.08(1.32)	2.73(2.69)	4.00	0.60	0.20	1.20
23		14(20)	17(24)	39(56)		-0.65(-0.46)	0.32(0.43)	1.90(1.78)		0.70	0.20	0.90
24	1(1)	16(23)	20(29)	33(47)	-0.26(0.00)	0.57(0.59)	1.73(1.54)	2.78(2.87)	0.90	0.80	1.70	1.10
25	5(7)	15(22)	25(36)	24(35)	-0.95(-0.73)	-0.24(-0.08)	1.12(0.95)	2.21(2.25)	0.70	0.60	1.00	1.00
26	7(10)	13(19)	28(40)	22(31)	-1.21(-0.88)	-0.22(-0.22)	0.93(0.81)	2.06(2.09)	0.60	0.80	0.90	1.00

(continued)

Table 4.9. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
27	2(3)	14(20)	36(51)	18(26)	-0.54(-0.61)	-0.25(0.06)	1.19(1.21)	2.85(2.59)	1.00	0.70	0.60	0.80
28	8(11)	20(29)	28(40)	14(20)	-1.39(-1.23)	-0.70(-0.45)	0.95(0.71)	1.87(1.90)	0.70	0.60	0.60	1.10
29	2(3)	13(19)	28(40)	27(39)	-0.41(-0.36)	-0.12(0.26)	1.32(1.27)	2.77(2.64)	0.90	0.50	0.70	0.80
30	4(6)	14(20)	28(40)	24(34)	-0.61(-0.66)	-0.32(-0.01)	1.05(1.02)	2.48(2.34)	0.90	0.60	0.80	0.80
31	3(4)	18(26)	31(44)	18(26)	-0.85(-0.76)	-0.30(-0.06)	1.22(1.08)	2.42(2.39)	0.90	0.70	0.90	0.90
32	5(7)	13(19)	24(34)	28(40)	-0.58(-0.62)	-0.24(0.00)	0.92(0.97)	2.41(2.26)	0.90	0.70	0.50	0.90
33	4(6)	10(14)	25(36)	31(44)	-0.68(-0.47)	-0.13(0.11)	0.92(1.05)	2.60(2.39)	0.70	0.50	0.40	0.80
34	4(6)	16(23)	23(33)	27(39)	-0.88(-0.59)	-0.06(0.05)	1.02(1.05)	2.47(2.33)	0.70	0.60	0.70	0.80
35	4(6)	21(30)	22(31)	23(33)	-1.07(-0.72)	-0.01(-0.02)	0.90(1.04)	2.46(2.28)	0.80	0.80	0.90	0.70
36	3(4)	16(23)	26(38)	24(35)	-0.45(-0.60)	0.14(0.08)	0.98(1.14)	2.54(2.43)	1.00	1.00	1.10	0.08
37	8(11)	19(27)	27(39)	16(23)	-1.22(-1.15)	-0.51(-0.40)	0.72(0.73)	2.13(1.93)	0.80	0.90	0.80	0.70
38	6(9)	17(24)	21(30)	26(37)	-0.33(-0.75)	-0.21(-0.08)	0.75(0.91)	2.26(2.14)	1.40	1.10	0.60	0.80
39	3(4)	6(9)	21(30)	40(57)	0.08(-0.18)	0.95(0.33)	0.99(1.17)	2.53(2.55)	1.20	1.80	1.00	1.20
40	4(6)	16(23)	14(20)	36(51)	-0.08(-0.38)	0.25(0.20)	1.20(1.08)	2.25(2.34)	1.00	1.00	1.10	1.60
41	7(10)	19(27)	23(33)	21(30)	-1.07(-0.95)	-0.27(-0.23)	0.88(0.83)	2.06(2.03)	0.90	0.70	1.00	1.00
42	4(6)	6(9)	15(21)	45(64)	0.05(-0.17)	0.40(0.31)	0.94(1.08)	2,45(2,43)	1.10	0.60	0.40	1.60
43	5(7)	20(29)	26(38)	18(26)	-0.90(-0.94)	-0.42(-0.21)	1.11(0.92)	2.12(2.17)	1.10	0.70	0.70	1.10
44	4(6)	19(27)	18(26)	29(41)	-0.13(-0.55)	0.02(0.09)	0.89(1.07)	2.41(2.31)	1.40	0.80	0.70	0.80
45		10(14)	34(49)	26(37)		-0.66(-0.65)	0.21(0.40)	2.11(1.86)		1.00	0.50	0.70
46		13(19)	32(46)	25(36)		-0.92(-0.81)	0.12(0.26)	1.92(1.68)		0.90	0.60	0.70
47	2(3)	17(24)	21(30)	30(43)	0.33(-0.30)	0.27(0.32)	1.31(1.30)	2.58(2.60)	1.30	0.80	1.20	1.00
48		10(16)	7(11)	45(73)		0.37(-0.16)	0.72(0.64)	1.93(2.06)		3.10	2.00	1.80
49	1(1)	14(20)	16(23)	39(56)	1.83(0.13)	0.74(0.68)	1.36(1.55)	2.90(2.89)	2.70	1.00	0.40	1.20
50		7(10)	27(39)	36(51)		-0.67(-0.21)	0.45(0.72)	2.47(2.18)		0.60	0.30	0.60
51	7(10)	14(20)	23(33)	26(37)	-0.67(-0.78)	-0.49(-0.13)	1.14(0.84)	2.00(2.10)	1.00	0.50	1.40	0.90
52	1(1)	10(14)	26(37)	33(47)	-0.46(0.02)	0.10(0.58)	1.50(1.53)	3.11(2.93)	0.70	0.40	0.80	0.70
53	1(1)	14(20)	26(37)	29(41)	-0.63(-0.09)	0.22(0.52)	1.66(1.52)	2.92(2.89)	0.70	0.60	1.30	0.80

(continued)

Table 4.9. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
54	2(3)	14(20)	24(34)	30(43)	-0.21(-0.29)	0.07(0.32)	1.28(1.29)	2.75(2.63)	1.00	0.60	0.90	0.70
55	16(23)	33(47)	12(17)	9(13)	-2.04(-1.83)	-0.68(-0.78)	0.41(0.45)	1.36(1.31)	0.80	0.70	1.10	1.10
56	4(6)	20(29)	16(23)	29(42)	-0.88(-0.55)	0.09(0.09)	1.89(1.07)	1.90(2.31)	0.80	0.70	2.00	1.80
57	2(3)	5(8)	14(23)	41(66)	0.14(0.06)	1.41(0.45)	1.03(1.28)	2.69(2.72)	0.90	2.90	0.90	1.20
58	12(17)	23(33)	20(29)	15(21)	-1.57(-1.35)	-0.51(-0.52)	0.65(0.60)	1.77(1.67)	0.80	0.80	0.60	0.90
59	16(23)	19(27)	18(26)	17(24)	-1.56(-1.35)	-0.48(-0.54)	0.65(0.51)	1.55(1.55)	0.70	1.30	0.60	0.90
60	17(24)	19(27)	16(23)	18(26)	-1.59(-1.34)	-0.42(-0.52)	0.56(0.50)	1.61(1.52)	0.70	0.50	0.70	0.80
61	26(37)	24(34)	13(19)	7(10)	-2.06(-2.06)	-0.91(-0.97)	0.10(0.18)	0.91(0.95)	1.00	0.80	1.30	1.00
62	25(36)	24(34)	13(19)	8(11)	-1.96(-1.98)	-0.99(-0.92)	0.37(0.23)	0.99(1.02)	1.10	0.70	1.00	1.00
63	31(44)	24(34)	11(16)	4(6)	-2.42(-2.39)	-1.18(-1.18)	0.04(-0.04)	0.72(0.64)	1.00	0.70	1.00	0.70
64	29(41)	25(36)	11(16)	5(7)	-2.18(-2.27)	-1.24(-1.09)	0.22(0.06)	0.60(76)	1.10	1.00	0.80	1.10

Saxophone Performance Rating Scale Data

Multifaceted Rasch Partial Credit Model Results

Table 4.10 presents summary statistics from the analysis of performers (*n* = 19), raters (*n* = 6), and items (*n* = 67) on the Saxophone Performance Rating Scale (SPRS). Overall significant differences in chi-square are indicated for performers (832.90), raters (495.70), and items (650.20). The probability for each facet was less than 0.01 and reliability of separation for each was high with performers at 0.98, raters at 0.99, and items at 0.91. This indicates adequate spread of elements within each facet along a single measure of music performance ability. Infit MSE values on the SPRS are 1.00 for performers, 0.99 for raters, and 0.99 for items. Outfit MSE values were also well targeted with values of 1.06 for performers, 1.03 for raters, and 1.06 for items.

Table 4.10

SPRS Summary Statistics from the PC-MFR Model

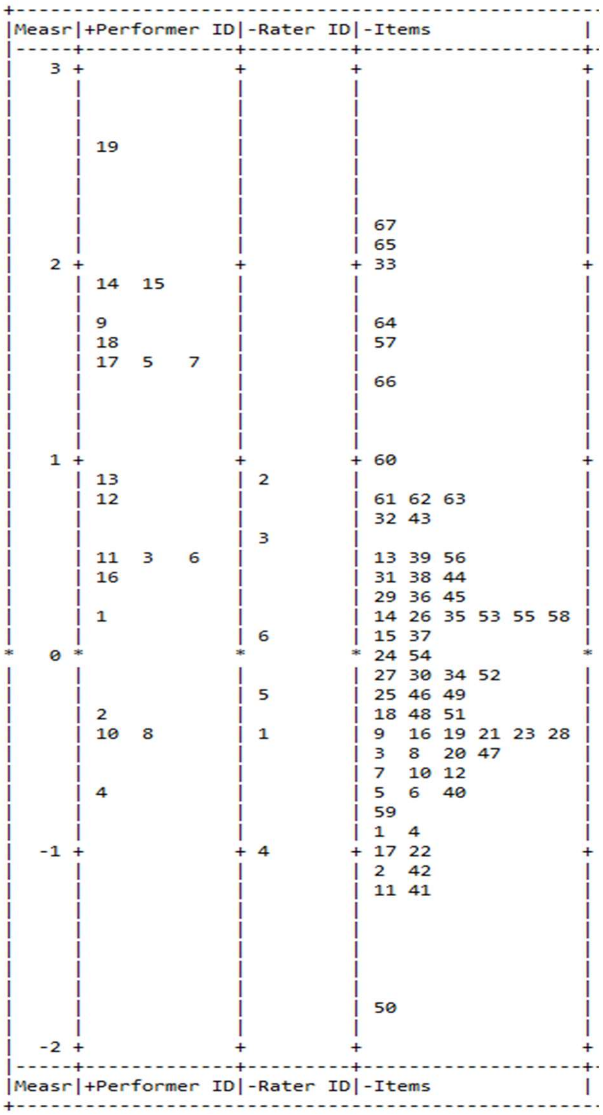
			Facets	
		Performance (θ)	**Rater** (λ)	**Item** (δ)
Measure (Logits)				
	Mean	0.85	0.00	0.00
	SD	0.91	0.62	0.82
	N	19	6	67
Infit *MSE*				
	Mean	1.00	0.99	0.99
	SD	0.22	0.11	0.27
Std. Infit *MSE*				
	Mean	-0.20	-0.20	-0.10
	SD	1.90	1.40	1.10
Outfit *MSE*				
	Mean	1.06	1.03	1.06
	SD	0.27	0.17	0.38
Std. Outfit *MSE*				

Mean	0.20	0.30	0.10
SD	1.80	1.70	1.20
Separation Statistics			
Reliability of Separation	0.98	0.99	0.91
Chi-Square	832.90	495.70	650.20
Degrees of Freedom	18	5	66

**p* < 0.01

Figure 4.3

Variable Map, Saxophone Performance Rating Scale



```
+---------------------------------------------------+
|Measr|+Performer ID|-Rater ID|-Items             |
|-----+-------------+---------+-------------------+
|   3 +             +         +                   +
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     | 19          |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         | 67                |
|     |             |         | 65                |
|   2 +             +         + 33                +
|     | 14  15      |         |                   |
|     |             |         |                   |
|     | 9           |         | 64                |
|     | 18          |         | 57                |
|     | 17  5   7   |         |                   |
|     |             |         | 66                |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|   1 +             +         + 60                +
|     | 13          | 2       |                   |
|     | 12          |         | 61 62 63          |
|     |             |         | 32 43             |
|     |             | 3       |                   |
|     | 11  3   6   |         | 13 39 56          |
|     | 16          |         | 31 38 44          |
|     |             |         | 29 36 45          |
|     | 1           |         | 14 26 35 53 55 58 |
|     |             | 6       | 15 37             |
*   0 *             *         * 24 54             *
|     |             |         | 27 30 34 52       |
|     |             | 5       | 25 46 49          |
|     | 2           |         | 18 48 51          |
|     | 10  8       | 1       | 9  16 19 21 23 28 |
|     |             |         | 3  8  20 47       |
|     |             |         | 7  10 12          |
|     | 4           |         | 5  6  40          |
|     |             |         | 59                |
|     |             |         | 1  4              |
|  -1 +             + 4       + 17 22             +
|     |             |         | 2  42             |
|     |             |         | 11 41             |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         | 50                |
|     |             |         |                   |
|  -2 +             +         +                   +
|-----+-------------+---------+-------------------+
|Measr|+Performer ID|-Rater ID|-Items             |
+---------------------------------------------------+
```

Variable Map

Figure 4.3 is a variable map representing saxophone performance ability as a latent variable. Included on the map are the calibrations of performers (column 2), raters (column 3), and items (column 4). Facets are ordered from top to bottom according to high to low ability of performers, severity to leniency of raters, and highest to lowest difficulty of items.

Column 2 on the map shows the location of each performer along the latent variable, with numbers representing the performer number. Student achievement ranged from 2.61 to -0.74 logits (*M* = 0.85, *SD* = 0.91, *n* = 19). Performer 7 was the only underfitting performer with a MSE value of 1.53. Performer 6 was the only overfitting performer at 0.52. Table 4.11 shows the complete calibration and statistics for performers.

Column 3 shows the calibration of raters along the latent construct, with numbers representing the rater number. All raters fell within the acceptable range for productive parameter-level MSE. Rater 2 was the most severe (observed average = 2.25, logit measure 0.88) and Rater 4 was the most lenient (observed average = 3.40, logit measure -0.96).

Column 4 shows the calibration of items, with numbers representing the item number. The most difficult item was Item 67 (choice of vibrato depth and speed, observed average = 1.45, logit measure = 2.15). The easiest item was Item 50 (accuracy of tied notes, observed average = 3.76, logit measure = -1.77). Items demonstrating underfit include 7, 10, 44, 58, and 59. Items demonstrating overfit include 21, 39, 45, and 52. Table 4.12 shows the complete calibration and statistics for items.

Table 4.11

Calibration of Saxophone Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
19	3.63	2.61	0.18	1.13	0.80	0.96	0.00
15	3.57	1.86	0.17	1.26	1.60	1.52	2.10
14	3.57	1.86	0.17	0.81	-1.30	0.80	-0.90
9	3.39	1.73	0.15	0.75	-1.80	1.12	0.60
18	3.34	1.58	0.15	1.17	1.20	1.39	2.20
7	2.99	1.53	0.13	1.53	3.50	1.56	3.50
17	3.31	1.50	0.14	1.09	0.60	1.15	0.90
5	2.96	1.47	0.12	0.89	-0.80	0.86	-1.00
13	3.28	0.90	0.14	1.07	0.50	1.34	2.00
12	3.04	0.83	0.13	1.24	1.80	1.22	1.40
11	2.91	0.54	0.12	0.98	-0.10	1.13	0.90
6	2.43	0.51	0.11	0.52	-5.30	0.50	-4.60
3	2.69	0.49	0.12	1.01	0.00	1.07	0.50
16	3.07	0.39	0.13	1.20	1.50	1.19	1.30
1	2.51	0.15	0.12	0.86	-1.20	0.85	-1.00
2	2.23	-0.34	0.12	0.95	-0.40	0.87	-0.80
10	2.45	-0.36	0.12	0.84	-1.30	0.86	-1.00
8	1.89	-0.41	0.12	0.91	-0.70	1.01	0.00
4	2.01	-0.74	0.12	0.77	-2.10	0.71	-1.90
Mean	2.91	0.85	0.13	1.00	-0.20	1.06	0.20
SD	0.52	0.91	0.02	0.22	1.90	0.27	1.80

Presented in measure order from highest to lowest achievement.

Table 4.12

Calibration of Saxophone Item Facet

Item Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
67	1.45	2.15	0.27	1.01	0.10	0.65	-0.40
65	1.89	2.15	0.28	1.05	0.20	1.05	0.20
33	1.47	2.05	0.26	0.96	0.00	0.62	-0.40
64	2.11	1.72	0.24	1.25	1.10	1.33	1.40
57	2.03	1.64	0.25	1.28	1.10	1.47	1.80
66	2.24	1.41	0.24	0.78	-1.00	0.78	-1.00
60	2.47	1.05	0.25	0.82	-0.70	0.83	-0.70
63	2.00	0.82	0.27	1.23	1.10	1.28	1.20
62	2.55	0.82	0.25	0.89	-0.40	0.90	-0.30
61	2.55	0.76	0.23	0.77	-1.00	0.75	-1.20
43	2.61	0.71	0.23	1.03	0.20	0.97	0.00
32	2.61	0.71	0.23	0.74	-1.20	0.77	-1.00
39	2.63	0.55	0.24	0.58	-2.20	0.59	-2.10
13	2.61	0.51	0.22	1.08	0.40	1.07	0.30
56	2.71	0.47	0.25	0.86	-0.60	0.86	-0.50
38	2.71	0.43	0.23	0.79	-0.90	0.78	-1.00
31	2.66	0.43	0.25	0.85	-0.60	0.85	-0.60
44	2.84	0.35	0.21	1.41	1.50	1.54	1.70
36	2.76	0.33	0.24	0.91	-0.30	1.02	0.10
29	2.71	0.29	0.24	0.99	0.00	0.96	-0.10

45	2.76	0.28	0.25	0.56	-2.30	0.56	-2.30
55	2.82	0.25	0.24	0.77	-1.00	0.77	-1.00
14	2.74	0.24	0.24	0.91	-0.30	0.93	-0.20
26	2.84	0.22	0.24	1.18	0.80	1.19	0.80
35	2.79	0.21	0.25	0.97	0.00	0.95	-0.10
58	2.82	0.20	0.23	1.51	2.10	1.59	2.30
53	2.82	0.18	0.25	0.63	-1.80	0.65	-1.70
15	2.89	0.11	0.23	0.98	0.00	0.95	-0.10
37	2.84	0.06	0.26	0.91	-0.30	0.91	-0.30
54	2.89	0.02	0.24	0.70	-1.40	0.75	-1.10
24	2.87	-0.03	0.25	1.24	1.10	1.25	1.10
27	2.95	-0.07	0.23	0.89	-0.40	0.85	-0.60
30	2.95	-0.10	0.23	0.77	-1.00	0.75	-0.90
52	3.00	-0.13	0.24	0.48	-2.80	0.48	-2.70
34	3.00	-0.13	0.24	0.84	-0.60	0.83	-0.60
25	2.76	-0.15	0.25	0.67	-1.60	0.66	-1.50
49	3.03	-0.20	0.23	0.88	-0.40	0.87	-0.30
46	3.16	-0.23	0.22	0.92	-0.20	0.82	-0.40
48	3.08	-0.26	0.23	0.86	-0.50	0.87	-0.40
51	3.16	-0.30	0.22	1.35	1.30	1.99	1.90
18	3.13	-0.33	0.23	0.97	0.00	1.01	0.10
23	3.05	-0.35	0.24	0.77	-1.00	0.82	-0.60
16	3.08	-0.35	0.25	1.00	0.00	1.10	0.50
9	3.08	-0.35	0.25	1.09	0.40	1.19	0.80
28	2.95	-0.40	0.25	0.77	-1.10	0.74	-1.10
21	3.26	-0.44	0.23	0.47	-2.30	0.52	-1.30
19	3.21	-0.44	0.24	0.89	-0.30	1.11	0.40
20	3.21	-0.45	0.23	0.98	0.00	1.32	0.90
47	3.13	-0.45	0.24	0.68	-1.40	0.72	-1.10
3	3.24	-0.47	0.24	1.31	1.20	1.41	1.40
8	3.18	-0.54	0.24	1.19	0.80	1.28	0.90
12	3.32	-0.59	0.24	1.29	1.10	1.36	1.00
7	3.24	-0.59	0.25	1.42	1.60	1.61	2.10
10	3.34	-0.63	0.24	1.57	1.80	2.00	2.00
40	3.13	-0.67	0.25	1.04	0.20	0.93	-0.20
6	3.29	-0.67	0.25	0.99	0.00	1.08	0.30
5	3.39	-0.70	0.24	1.20	0.70	1.64	1.50
59	3.55	-0.84	0.31	1.97	3.10	2.08	2.40
1	3.47	-0.93	0.26	1.01	0.10	1.16	0.50
4	3.47	-0.93	0.26	0.87	-0.30	1.48	1.00
22	3.37	-1.01	0.26	0.77	-0.80	0.73	-0.60
17	3.37	-1.01	0.26	0.75	-0.90	0.97	0.00
42	3.45	-1.11	0.27	1.26	0.90	1.05	0.20
2	3.45	-1.12	0.26	1.15	0.60	1.92	1.80
11	3.47	-1.15	0.27	0.97	0.00	0.83	-0.30
41	3.50	-1.19	0.28	1.20	0.70	1.05	0.20
50	3.76	-1.77	0.39	1.38	1.10	2.03	1.60
Mean	2.91	0.00	0.25	0.99	-0.10	1.06	0.10
SD	0.46	0.82	0.02	0.27	1.10	0.38	1.20

Presented in measure order from highest to lowest difficulty.

Response Category Diagnostics

Table 4.13 shows response category data for the SPRS including the usage of each response category, average observed and average expected logit measures, and the outfit mean squared error (MSE) value for each item in the rating scale. This data can be used to optimize response categories by eliminating underused categories, combining categories with too small of a threshold between them, or eliminating categories with high outfit MSE values. Linacre (2002) also recommends collapsing adjacent categories when they violate monotonicity, meaning the response categories do not follow the expected ordering of lower to higher difficulty. In the case of item 37, for example, category 2 shows a lower logit measure (-0.56) than category 1 (-0.46). Based on Linacre's recommendations, these two categories would be collapsed. Forty-three items would require a collapsing of response categories due to insufficient usage.

Examination of Differential Item Functioning

An analysis of differential item functioning (DIF) was performed to determine if there were statistically significant differences in scores between alto saxophone and tenor saxophone performances which were judged using the SPRS. Specifically, an item by instrument-type DIF analysis was performed to examine a total of 135 interactions. Results showed no statistically significant effect of DIF ($\chi^2 = 59$, df = 134, $p = 1$) which could be due to the small sample size of tenor saxophone performances ($n = 2$). Appendix (B?) shows DIF analysis statistics.

Table 4.13. SPRS Category Diagnostics for Fit Items: Category Usage, Average Observed and Expected Logit Measure, Outfit Mean Squared Error (*MSE*).

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
1	2(5)	2(5)	10(26)	24(63)	-0.53(-0.12)	0.24(0.32)	1.66(1.16)	2.09(2.25)	0.40	0.40	1.60	1.20
2	1(3)	5(13)	8(21)	24(63)	-0.17(0.09)	0.68(0.55)	1.91(1.41)	2.28(2.47)	0.60	2.50	2.20	1.30
3	3(8)	3(8)	14(37)	18(47)	-0.75(-0.46)	0.59(0.08)	1.50(0.99)	1.57(1.99)	0.50	1.20	2.00	1.40
4	2(5)	3(8)	8(21)	25(66)	-0.53(-0.09)	0.23(0.34)	1.73(1.17)	2.11(2.24)	0.50	0.30	2.60	1.10
5	3(8)	2(5)	10(26)	23(61)	-0.61(-0.29)	1.22(0.18)	1.41(1.03)	1.85(2.07)	0.40	3.10	2.00	1.40
6	2(5)	4(11)	13(34)	19(50)	-0.79(-0.30)	0.94(0.22)	0.97(1.14)	2.19(2.16)	0.40	2.30	0.80	0.90
7	2(5)	4(11)	15(39)	17(45)	-0.87(-0.37)	0.86(0.18)	1.60(1.13)	1.62(2.14)	0.40	2.10	2.00	1.60
8	2(5)	8(21)	9(24)	19(50)	-0.92(-0.34)	0.56(0.25)	1.39(1.16)	1.91(2.09)	0.50	1.30	1.80	1.20
9	2(5)	6(16)	17(45)	13(34)	-1.10(-0.55)	0.67(0.08)	1.01(1.08)	1.94(2.04)	0.50	1.90	1.20	1.00
10	3(8)	4(11)	8(21)	23(61)	-0.68(-0.31)	1.30(0.19)	1.48(1.04)	1.73(2.03)	0.40	3.30	2.20	1.70
11	1(3)	4(11)	9(24)	24(63)	-0.14(0.10)	0.29(0.54)	1.66(1.40)	2.44(2.49)	0.60	0.40	0.90	1.20
12	3(8)	3(8)	11(29)	21(55)	0.63(-0.36)	-0.50(0.15)	0.85(1.02)	2.07(2.03)	3.80	0.00	1.00	0.90
13	5(13)	15(39)	8(21)	10(26)	-0.87(-.1.06)	-0.27(-0.20)	0.81(0.71)	1.35(1.43)	1.10	1.30	0.90	1.00
14	3(8)	13(34)	13(34)	9(24)	-0.95(-0.93)	-0.16(-0.13)	0.83(0.87)	1.78(1.67)	1.00	1.20	0.70	0.80
15	4(11)	7(18)	16(42)	11(29)	-0.49(-0.87)	-0.55(-0.17)	0.79(0.80)	1.79(1.67)	1.70	0.40	0.70	0.90
16	2(5)	6(16)	17(45)	13(34)	-1.10(-0.55)	0.52(0.08)	1.03(1.08)	1.99(2.04)	0.50	1.80	1.00	1.00
17	1(3)	6(16)	9(24)	22(58)	-0.28(0.02)	0.55(0.52)	1.13(1.40)	2.55(2.43)	0.70	2.00	0.30	0.70
18	3(8)	6(16)	12(32)	17(45)	-0.98(-0.53)	0.48(0.08)	0.92(1.00)	1.92(1.92)	0.50	1.80	0.80	1.00
19	3(8)	5(13)	11(29)	19(50)	-0.87(-0.45)	0.49(0.12)	1.00(1.01)	1.94(1.96)	0.40	1.50	1.40	0.90
20	3(8)	6(16)	9(24)	20(53)	-0.73(-0.42)	0.41(0.15)	1.14(1.02)	1.87(1.96)	0.50	2.00	1.60	0.90
21	4(11)	4(11)	8(21)	22(58)	-0.85(-0.44)	0.07(0.10)	0.82(0.93)	2.00(1.88)	0.30	0.50	0.60	0.60
22	1(3)	6(16)	9(24)	22(58)	-0.28(0.02)	0.36(0.52)	1.39(1.40)	2.49(2.43)	0.70	0.60	0.80	0.80
23	2(5)	9(24)	12(32)	15(39)	-1.07(-0.49)	0.13(0.16)	1.17(1.12)	2.08(2.02)	0.60	1.00	0.70	0.80
24	2(5)	10(26)	17(45)	9(24)	-1.39(-0.77)	0.23(-0.04)	1.20(1.00)	1.36(1.88)	0.60	1.50	1.10	1.50
25	1(3)	16(42)	12(32)	9(24)	-1.08(-0.60)	-0.06(0.22)	1.63(1.26)	2.11(2.06)	0.90	0.50	0.40	0.90

(continued)

Table 4.13. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
26	4(11)	7(18)	18(47)	9(24)	-0.48(-0.97)	-0.26(-0.25)	0.56(0.75)	1.80(1.62)	1.80	1.10	1.10	0.80
27	3(8)	9(24)	13(34)	13(34)	-0.35(-0.71)	-0.21(-0.01)	0.80(0.95)	2.01(1.81)	1.50	0.60	0.80	0.70
28	1(3)	11(29)	15(39)	11(29)	-0.57(-0.45)	0.02(0.26)	1.39(1.29)	2.30(2.19)	0.90	0.60	0.70	0.80
29	3(8)	13(34)	14(37)	8(21)	-0.72(-0.99)	-0.22(-0.17)	0.74(0.84)	1.79(1.64)	1.20	0.90	1.10	0.80
30	3(8)	11(29)	9(24)	15(39)	-0.32(-0.64)	-0.19(0.06)	0.90(0.98)	1.99(1.81)	1.40	0.40	0.90	0.60
31	3(8)	13(34)	16(42)	6(16)	-0.85(-1.12)	-0.52(-0.28)	0.82(0.76)	1.78(1.56)	1.30	0.80	0.60	0.80
32	6(16)	9(24)	17(45)	6(16)	-1.53(-1.30)	-0.39(-0.48)	0.34(0.50)	1.80(1.27)	0.90	1.00	0.60	0.60
33	27(71)	4(11)	7(18)		-1.68(-1.67)	-0.80(-0.61)	0.22(0.07)		1.20	0.40	0.50	
34	3(8)	7(18)	15(39)	13(34)	-0.54(-0.69)	-0.46(-0.03)	1.10(0.94)	1.87(1.84)	1.00	0.50	0.90	0.90
35	3(8)	10(26)	17(45)	8(21)	-0.89(-0.96)	-0.40(-0.19)	1.01(0.83)	1.52(1.68)	1.00	0.90	0.80	1.10
36	4(11)	9(24)	17(45)	8(21)	-1.25(-1.05)	-0.09(-0.28)	0.63(0.73)	1.67(1.56)	0.70	2.00	0.50	0.90
37	2(5)	9(24)	20(53)	7(18)	-0.46(-0.88)	-0.56(-0.14)	1.12(0.94)	1.75(1.84)	1.30	0.60	0.70	1.10
38	5(13)	10(26)	14(37)	9(24)	-1.36(-1.07)	-0.37(-0.28)	0.87(0.67)	1.42(1.46)	0.60	0.70	0.70	1.00
39	4(11)	12(32)	16(42)	6(16)	-1.77(-1.20)	-0.45(-0.36)	0.73(0.65)	1.80(1.44)	0.50	0.60	0.50	0.70
40	1(3)	8(21)	14(37)	15(39)	-0.30(-0.25)	0.50(0.36)	1.20(1.35)	2.38(2.31)	0.90	0.90	0.90	0.90
41	1(3)	3(8)	10(26)	24(63)	0.22(0.11)	1.55(0.54)	0.87(1.40)	2.59(2.51)	0.90	2.60	0.50	0.90
42	1(3)	4(11)	10(26)	23(61)	0.15(0.07)	1.44(0.52)	0.82(1.40)	2.57(2.48)	0.90	2.30	0.40	0.90
43	7(18)	8(21)	16(42)	7(18)	-1.25(-1.27)	-0.51(-0.46)	0.49(0.48)	1.23(1.24)	1.00	0.70	1.10	1.00
44	8(21)	3(8)	14(37)	13(34)	-0.73(-0.97)	-0.26(-0.28)	0.58(0.58)	1.22(1.38)	2.80	0.60	0.90	1.30
45	3(8)	10(26)	18(47)	7(18)	-1.55(-1.02)	-0.39(-0.25)	0.76(0.80)	2.18(1.64)	0.50	0.60	0.50	0.60
46	5(13)	4(11)	9(24)	20(53)	-0.57(-0.58)	-0.02(0.00)	0.68(0.84)	1.82(1.74)	1.20	0.70	0.60	0.80
47	2(5)	7(18)	13(34)	16(42)	-0.97(-0.44)	0.05(0.17)	1.14(1.13)	2.18(2.07)	0.50	0.60	0.90	0.80
48	3(8)	7(18)	12(32)	16(42)	-0.93(-0.57)	0.30(0.06)	0.78(0.99)	2.01(1.89)	0.60	1.20	0.90	0.80
49	3(8)	9(24)	10(26)	16(42)	-0.90(-0.59)	0.11(0.07)	1.03(0.99)	1.87(1.86)	0.70	0.90	1.00	0.90
50		1(3)	7(18)	30(79)		0.80(0.86)	2.40(1.61)	2.68(2.86)		0.70	2.60	1.30
51	4(11)	7(18)	6(16)	21(55)	0.25(-0.50)	-0.05(0.10)	0.73(0.94)	1.78(1.81)	6.10	0.60	1.80	0.80

(continued)

Table 4.13. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
52	3(8)	7(18)	15(39)	13(34)	-1.14(-0.69)	-0.34(-0.03)	0.90(0.94)	2.16(1.84)	0.50	0.30	0.30	0.70
53	3(8)	9(24)	18(47)	8(21)	-1.46(-0.95)	-0.38(-0.20)	0.90(0.83)	1.92(1.70)	0.50	0.60	0.70	0.80
54	3(8)	9(24)	15(39)	11(29)	-1.29(-0.80)	-0.32(-0.08)	1.21(0.91)	1.70(1.77)	0.50	0.40	0.90	1.00
55	4(11)	8(21)	17(45)	9(24)	-1.30(-0.98)	-0.23(-0.24)	0.75(0.75)	1.76(1.60)	0.50	0.90	0.70	0.90
56	4(11)	9(24)	19(50)	6(16)	-1.60(-1.18)	-0.23(-0.39)	0.66(0.65)	1.51(1.48)	0.50	1.10	0.80	1.00
57	10(26)	20(53)	5(13)	3(8)	-1.62(-1.85)	-0.72(-0.78)	-0.46(0.12)	0.46(0.67)	1.20	0.90	2.60	1.60
58	4(11)	10(26)	13(34)	11(29)	0.02(-0.90)	0.03(-0.14)	0.33(0.81)	1.70(1.62)	3.40	1.30	1.10	0.90
59		3(8)	11(29)	24(63)		1.29(0.12)	1.59(0.99)	1.73(2.15)		3.40	1.40	2.00
60	6(16)	11(29)	18(47)	3(8)	-1.72(-1.60)	-0.72(-0.72)	0.22(0.30)	1.77(1.04)	0.90	1.00	0.80	0.60
61	6(16)	11(29)	15(39)	6(16)	-1.74(-1.32)	-0.42(-0.46)	0.67(0.50)	1.16(1.25)	0.50	0.70	0.60	1.10
62	5(13)	11(29)	18(47)	4(11)	-1.55(-1.43)	-0.59(-0.57)	0.46(0.45)	1.38(1.23)	1.00	1.00	0.70	0.90
63	11(29)	16(42)	11(29)		-0.75(-1.03)	-0.12(0.03)	0.89(0.95)		1.40	0.90	1.50	
64	11(29)	14(37)	11(29)	2(5)	-1.65(-1.96)	-1.04(-0.96)	-0.37(-0.06)	1.10(0.53)	1.40	1.10	1.90	0.60
65	11(29)	21(55)	5(13)	1(3)	-2.08(-2.30)	-1.31(-1.17)	-0.25(-0.25)	0.68(0.25)	1.20	1.10	1.00	0.60
66	9(24)	14(37)	12(32)	3(8)	-1.82(-1.75)	-0.92(-0.78)	0.29(0.14)	1.04(0.78)	0.90	0.90	0.50	0.90
67	27(71)	5(13)	6(16)		-1.78(-1.76)	-0.62(-0.68)	0.04(0.00)		1.10	0.40	0.60	

Trumpet Performance Rating Scale Data

Multifaceted Rasch Partial Credit Model Results

Table 4.14 presents summary statistics from the analysis of performers (*n* = 38), raters (*n* = 10), and items (*n* = 65) on the Trumpet Performance Rating Scale (TPRS). Overall significant differences in chi-square are indicated for performers (1620.50), raters (453.10), and items (1726.30). The probability for each facet was less than 0.01 and reliability of separation for each was high with performers at 0.98, raters at 0.98, and items at 0.96. This indicates adequate spread of elements within each facet along a single measure of music performance ability. Infit MSE values on the TPRS are 1.02 for performers, 1.06 for raters, and 1.01 for items. Outfit MSE values were also well targeted with values of 1.06 for performers, 1.21 for raters, and 1.07 for items.

Table 4.14

TPRS Summary Statistics from the PC-MFR Model

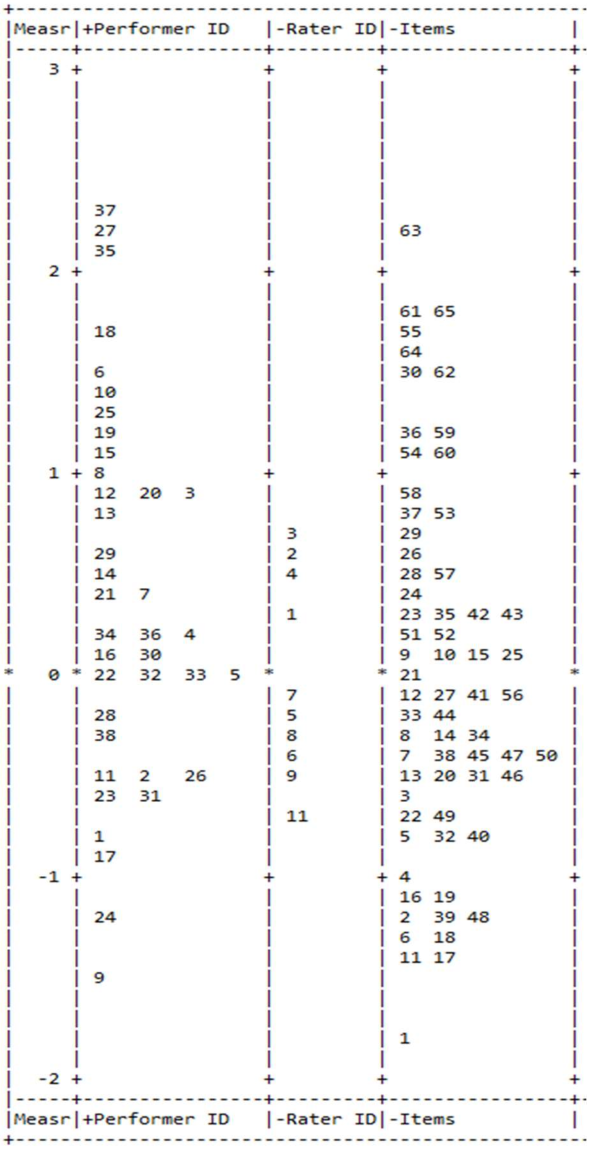
		Facets		
		Performance (θ)	**Rater** (λ)	**Item** (δ)
Measure (Logits)				
	Mean	0.38	0.00	0.00
	SD	0.92	0.47	0.92
	N	38	10	65
Infit *MSE*				
	Mean	1.02	1.06	1.01
	SD	0.28	0.31	0.24
Std. Infit *MSE*				
	Mean	-0.10	0.00	0.00
	SD	2.10	3.60	1.50
Outfit *MSE*				
	Mean	1.06	1.21	1.07
	SD	0.46	0.65	0.52

Std. Outfit *MSE*			
Mean	0.00	0.60	0.00
SD	2.20	3.20	1.80
Separation Statistics			
Reliability of Separation	0.98	0.98	0.96
Chi-Square	1620.50	453.10	1726.30
Degrees of Freedom	37	9	64

*$p < 0.01$

Figure 4.4

Variable Map, Trumpet Performance Rating Scale



```
+--------------------------------------------------
|Measr|+Performer ID   |-Rater ID|-Items          |
|-----+----------------+---------+----------------+
|   3 +                +         +                +
|     |                |         |                |
|     |                |         |                |
|     |                |         |                |
|     |                |         |                |
|     |                |         |                |
|     |                |         |                |
|     | 37             |         |                |
|     | 27             |         | 63             |
|     | 35             |         |                |
|   2 +                +         +                +
|     |                |         |                |
|     |                |         | 61 65          |
|     | 18             |         | 55             |
|     |                |         | 64             |
|     | 6              |         | 30 62          |
|     | 10             |         |                |
|     | 25             |         |                |
|     | 19             |         | 36 59          |
|     | 15             |         | 54 60          |
|   1 + 8              +         +                +
|     | 12  20  3      |         | 58             |
|     | 13             |         | 37 53          |
|     |                | 3       | 29             |
|     | 29             | 2       | 26             |
|     | 14             | 4       | 28 57          |
|     | 21  7          |         | 24             |
|     |                | 1       | 23 35 42 43    |
|     | 34  36  4      |         | 51 52          |
|     | 16  30         |         | 9  10 15 25    |
*   0 * 22  32  33  5  *         * 21             *
|     |                | 7       | 12 27 41 56    |
|     | 28             | 5       | 33 44          |
|     | 38             | 8       | 8  14 34       |
|     |                | 6       | 7  38 45 47 50 |
|     | 11  2   26     | 9       | 13 20 31 46    |
|     | 23  31         |         | 3              |
|     |                | 11      | 22 49          |
|     | 1              |         | 5  32 40       |
|     | 17             |         |                |
|  -1 +                +         + 4              +
|     |                |         | 16 19          |
|     | 24             |         | 2  39 48       |
|     |                |         | 6  18          |
|     |                |         | 11 17          |
|     | 9              |         |                |
|     |                |         |                |
|     |                |         |                |
|     |                |         | 1              |
|     |                |         |                |
|  -2 +                +         +                +
|-----+----------------+---------+----------------+
|Measr|+Performer ID   |-Rater ID|-Items          |
+--------------------------------------------------
```

Variable Map

Figure 4.4 is a variable map representing trumpet performance ability as a latent variable. Included on the map are the calibrations of performers (column 2), raters (column 3), and items (column 4). Facets are ordered from top to bottom according to high to low ability of performers, severity to leniency of raters, and highest to lowest difficulty of items.

Column 2 on the map shows the location of each performer along the latent variable, with numbers representing the performer number. Student achievement ranged from 2.30 to -1.49 logits (*M* = 0.38, *SD* = 0.92, *n* = 38). Underfitting performers included 3, 9, 33, and 37. Performer 19 was the only overfitting performer with an infit MSE of 0.53. Table 4.15 shows the complete calibration and statistics for performers.

Column 3 shows the calibration of raters along the latent construct, with numbers representing the rater number. Rater 3 was the most severe (observed average = 2.40, logit measure 0.73) and Rater 11 was the most lenient (observed average = 3.24, logit measure -0.66) and was also the only misfit rater with an infit MSE of 1.79.

Column 4 shows the calibration of items, with numbers representing the item number. The most difficult item was Item 63 (dynamics used to create tension and release, observed average = 1.64, logit measure = 2.19). The easiest item was Item 1 (upper body position, observed average = 3.6, logit measure = -1.77). Items demonstrating underfit included 1, 5, 9, and 10. There were no overfitting items. Table 4.16 shows the complete calibration and statistics for items.

Table 4.15

Calibration of Trumpet Performance Facet

Performance Number	**Observed Average**	**Measure**	**Standard Error**	**Infit MSE**	**Std. Infit**	**Outfit MSE**	**Std. Outfit**
37	3.61	2.30	0.17	1.56	3.00	3.27	5.90
27	3.60	2.24	0.17	0.68	-2.20	0.72	-1.20
35	3.55	2.13	0.16	1.32	1.90	1.27	1.20
18	3.47	1.66	0.16	0.62	-2.90	0.57	-2.40
6	3.02	1.47	0.13	1.18	1.30	1.34	2.30
10	2.98	1.40	0.13	0.78	-1.80	0.76	-1.90
25	3.28	1.27	0.14	1.05	0.40	1.08	0.50
19	3.30	1.16	0.15	0.53	-3.90	0.51	-3.50
15	3.06	1.12	0.13	1.17	1.30	1.09	0.60
8	2.77	0.99	0.12	0.92	-0.60	0.88	-1.00
12	2.73	0.92	0.12	0.67	-3.20	0.65	-3.30
3	2.85	0.91	0.12	1.55	4.00	1.42	3.00
20	3.19	0.89	0.14	1.37	2.40	1.82	4.30
13	2.90	0.77	0.13	0.81	-1.60	0.78	-1.70
29	3.09	0.60	0.14	1.18	1.30	1.21	1.30
14	2.77	0.52	0.12	0.89	-0.90	0.85	-1.20
21	2.94	0.43	0.12	0.98	-0.10	1.02	0.20
7	2.43	0.39	0.12	0.99	0.00	0.97	-0.20
4	2.45	0.19	0.12	1.01	0.10	0.96	-0.30
34	2.73	0.19	0.12	0.91	-0.70	0.90	-0.80
36	2.71	0.16	0.12	0.94	-0.50	0.94	-0.40
16	2.54	0.11	0.12	0.86	-1.20	0.87	-1.10
30	2.83	0.09	0.12	0.73	-2.40	0.75	-2.10
32	2.79	0.02	0.12	1.13	1.00	1.17	1.30
5	2.21	-0.01	0.12	0.92	-0.60	0.84	-1.30
22	2.70	-0.02	0.12	1.04	0.30	1.08	0.60
33	2.64	-0.04	0.12	1.48	3.60	1.50	3.60
28	2.56	-0.17	0.12	0.66	-3.40	0.67	-3.20
38	2.50	-0.30	0.12	1.11	0.90	1.13	1.00
11	1.98	-0.46	0.13	0.89	-0.90	0.85	-1.10
26	2.40	-0.47	0.12	0.69	-3.00	0.72	-2.70
2	2.05	-0.53	0.12	1.15	1.20	1.07	0.60
23	2.38	-0.60	0.12	0.81	-1.70	0.90	-0.80
31	2.43	-0.62	0.12	1.33	2.60	1.34	2.70
1	1.95	-0.75	0.13	1.38	2.80	1.28	1.90
17	2.25	-0.93	0.12	0.75	-2.30	0.76	-2.20
24	2.03	-1.24	0.12	1.08	0.70	0.94	-0.40
9	1.57	-1.49	0.16	1.53	3.00	1.55	2.30
Mean	2.72	0.38	0.13	1.02	-0.10	1.06	0.00
SD	0.48	0.92	0.02	0.28	2.10	0.46	2.20

Presented in measure order from highest to lowest achievement.

Table 4.16

Calibration of Trumpet Item Facet

Item Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
63	1.64	2.19	0.19	0.83	-1.00	0.80	-1.10
65	1.59	1.80	0.22	0.94	-0.20	0.91	-0.20
61	1.81	1.75	0.18	0.75	-1.60	0.72	-1.70
55	1.71	1.73	0.18	0.89	-0.50	0.83	-0.90
64	1.72	1.62	0.17	0.83	-0.90	0.75	-1.20
62	1.88	1.54	0.17	0.78	-1.50	0.74	-1.50
30	1.74	1.45	0.20	1.21	1.00	1.22	0.80
59	2.00	1.22	0.16	0.91	-0.50	0.87	-0.70
36	2.07	1.16	0.15	0.94	-0.30	0.90	-0.60
54	2.24	1.12	0.16	0.90	-0.60	0.85	-0.90
60	2.15	1.11	0.16	1.05	0.30	1.00	0.00
58	2.18	0.91	0.14	0.79	-1.50	0.72	-1.70
37	2.24	0.85	0.16	0.90	-0.60	0.91	-0.50
53	2.32	0.78	0.15	0.93	-0.40	0.86	-0.80
29	2.32	0.72	0.15	0.99	0.00	0.91	-0.50
26	2.42	0.62	0.17	0.86	-0.90	0.84	-1.00
57	2.34	0.54	0.15	1.16	0.90	1.11	0.50
28	2.42	0.51	0.15	0.90	-0.60	0.86	-0.90
24	2.47	0.38	0.17	0.75	-1.70	0.74	-1.80
35	2.53	0.35	0.17	0.80	-1.30	0.81	-1.30
42	2.47	0.35	0.14	1.29	1.80	2.12	4.70
23	2.46	0.34	0.16	0.82	-1.20	0.81	-1.20
43	2.42	0.31	0.16	0.63	-2.70	0.62	-2.60
52	2.61	0.21	0.17	0.63	-2.70	0.64	-2.60
51	2.54	0.17	0.19	0.74	-1.70	0.75	-1.70
15	2.61	0.11	0.15	0.92	-0.40	0.91	-0.50
10	2.59	0.08	0.15	1.53	2.80	2.15	4.10
9	2.72	0.06	0.13	1.65	3.50	4.31	7.90
25	2.68	0.05	0.18	0.70	-2.00	0.69	-2.10
21	3.12	0.03	0.17	1.36	2.50	1.39	2.10
56	2.76	-0.07	0.14	0.95	-0.30	0.91	-0.50
41	2.66	-0.09	0.16	0.67	-2.50	0.64	-2.60
27	2.79	-0.11	0.18	0.87	-0.70	0.89	-0.50
12	2.71	-0.12	0.16	0.82	-1.20	0.80	-1.30
44	2.80	-0.16	0.14	0.74	-1.90	0.68	-2.10
33	2.75	-0.25	0.17	1.10	0.60	1.10	0.60
8	3.26	-0.26	0.17	1.24	1.60	1.22	1.00
14	2.87	-0.27	0.16	0.86	-0.90	0.86	-0.80
34	2.72	-0.30	0.18	1.15	1.00	1.13	0.80
38	2.87	-0.36	0.15	0.92	-0.50	0.87	-0.70
47	2.89	-0.36	0.16	0.79	-1.40	0.76	-1.50
7	3.33	-0.40	0.17	1.34	2.20	1.67	2.50
45	2.89	-0.41	0.17	0.99	0.00	0.94	-0.20
50	2.70	-0.42	0.18	0.70	-2.10	0.69	-2.10
31	2.79	-0.45	0.19	1.00	0.00	1.00	0.00
46	2.93	-0.52	0.19	0.82	-1.00	0.80	-1.20
20	3.32	-0.54	0.19	1.18	1.20	1.19	1.10
13	3.01	-0.54	0.15	0.86	-0.90	0.84	-0.80
3	3.42	-0.61	0.18	1.33	2.00	1.38	1.40
49	3.11	-0.73	0.18	1.00	0.00	1.02	0.10

22	2.92	-0.74	0.17	1.19	1.20	1.17	1.00
5	3.49	-0.78	0.18	1.45	2.50	2.08	3.10
32	2.84	-0.81	0.19	1.05	0.30	1.04	0.30
40	3.16	-0.82	0.16	1.02	0.10	0.93	-0.20
4	3.28	-1.04	0.16	1.04	0.20	1.24	1.10
19	3.17	-1.06	0.17	1.27	1.60	1.34	1.80
16	3.24	-1.11	0.18	1.11	0.60	1.09	0.50
39	3.28	-1.16	0.17	0.98	0.00	0.88	-0.50
2	3.61	-1.22	0.21	1.24	1.20	1.23	0.80
48	3.31	-1.24	0.24	0.98	0.00	0.94	-0.10
6	3.58	-1.27	0.21	1.17	1.00	1.07	0.30
18	3.20	-1.32	0.17	1.15	1.00	1.24	1.30
17	3.24	-1.36	0.18	1.32	1.90	1.39	2.00
11	3.47	-1.39	0.19	1.36	1.70	1.42	1.60
1	3.61	-1.77	0.19	1.56	2.30	1.29	0.80
Mean	2.71	0.00	0.17	1.01	0.00	1.07	0.00
SD	0.52	0.92	0.02	0.24	1.50	0.52	1.80

Presented in measure order from highest to lowest difficulty.

Response Category Diagnostics

Table 4.17 shows response category data for the TPRS including the usage of each response category, average observed and average expected logit measures, and the outfit mean squared error (MSE) value for each item in the rating scale. This data can be used to optimize response categories by eliminating underused categories, combining categories with too small of a threshold between them, or eliminating categories with high outfit MSE values. Based on response category diagnostic data, many of the items in the TPRS would be revised. Thirty-five of the items would require collapsing response categories 1 and 2 based on usage alone. Items requiring revision due to disordered logit measures include items 5, 6, 8, 9, 10, 11, 17, 18, 34, 40, and 42.

Table 4.17. TPRS Category Diagnostics for Fit Items: Category Usage, Average Observed and Expected Logit Measure, Outfit Mean Squared Error (*MSE*).

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
1	1(1)	8(11)	11(14)	56(74)	1.05(0.42)	1.58(1.03)	1.59(1.63)	2.28(2.36)	1.20	1.50	0.70	1.70
2		6(8)	18(24)	52(68)		0.88(0.39)	1.05(1.06)	1.78(1.84)		1.70	0.80	1.30
3		13(17)	18(24)	45(59)		0.25(-0.02)	0.75(0.60)	1.20(1.34)		1.30	1.30	1.60
4	3(4)	11(14)	24(32)	38(50)	-0.43(-0.07)	0.65(0.54)	1.30(1.12)	1.73(1.85)	0.60	1.00	2.20	1.00
5		11(14)	17(22)	48(63)		0.79(0.09)	0.64(0.72)	1.32(1.46)		3.60	0.60	1.30
6		5(7)	22(29)	49(64)		1.15(0.43)	0.97(1.13)	1.91(1.91)		1.50	0.80	1.00
7		16(21)	19(25)	41(54)		0.20(-0.16)	0.48(0.45)	1.02(1.18)		1.30	2.50	1.70
8		18(24)	20(26)	38(50)		0.11(-0.26)	0.06(0.35)	1.06(1.08)		1.30	1.20	1.20
9	20(27)	10(14)	15(20)	29(39)	-0.05(-0.64)	-0.18(-0.14)	-0.08(0.35)	0.83(1.00)	8.60	1.30	4.40	0.90
10	16(28)	8(14)	18(31)	16(28)	-0.22(-0.75)	-0.61(-0.26)	0.01(0.21)	0.78(0.92)	2.90	1.50	2.80	1.00
11	2(3)	5(7)	24(32)	45(59)	1.29(0.09)	1.00(0.71)	1.26(1.33)	2.04(2.09)	3.90	1.20	0.90	1.20
12	7(9)	25(33)	27(36)	17(22)	-1.17(-0.64)	-0.05(-0.03)	0.77(0.58)	1.29(1.35)	0.60	0.70	0.80	1.10
13	6(8)	16(21)	25(33)	29(38)	-0.72(-0.35)	0.23(0.23)	0.86(0.79)	1.53(1.51)	0.60	0.70	1.00	1.00
14	7(9)	16(21)	33(43)	20(26)	-1.02(-0.56)	0.07(0.03)	0.75(0.62)	1.32(1.40)	0.60	0.80	1.00	1.10
15	10(13)	25(33)	26(34)	15(20)	-1.21(-0.79)	-0.10(-0.19)	0.62(0.41)	0.96(1.19)	0.60	0.90	0.80	1.20
16	2(3)	10(13)	32(42)	32(42)	-1.02(-0.05)	1.13(0.60)	1.19(1.22)	1.92(1.99)	0.30	1.50	0.90	1.10
17	1(1)	13(17)	29(38)	33(43)	0.26(0.18)	1.39(0.84)	1.36(1.47)	2.11(2.23)	0.90	1.80	1.20	1.20
18	1(1)	16(21)	26(34)	33(43)	1.22(0.19)	0.89(0.85)	1.53(1.46)	2.10(2.21)	1.50	1.30	1.50	1.00
19	2(3)	15(20)	27(36)	32(42)	-1.07(-0.03)	1.16(0.61)	1.16(1.21)	1.81(1.96)	0.40	2.00	0.80	1.30
20		10(13)	31(41)	34(45)		0.10(-0.09)	0.67(0.59)	1.26(1.38)		1.30	1.10	1.20
21		20(26)	27(36)	29(38)		-0.24(-0.47)	0.28(0.16)	0.65(0.93)		1.20	1.50	1.50
22	2(3)	22(29)	32(42)	20(26)	-0.66(-0.24)	0.50(0.43)	1.27(1.07)	1.52(1.87)	0.80	0.90	1.00	1.50
23	11(14)	30(39)	24(32)	11(14)	-1.45(-0.96)	-0.20(-0.34)	0.32(0.29)	1.15(1.09)	0.60	0.90	0.90	0.90
24	9(12)	30(39)	29(38)	8(11)	-1.31(-1.04)	-0.50(-0.41)	0.38(0.27)	1.38(1.11)	0.80	0.60	0.70	0.80
25	6(8)	21(28)	39(52)	9(12)	-1.28(-0.87)	-0.42(-0.23)	0.56(0.43)	1.50(1.31)	0.70	0.50	0.60	0.90
26	10(13)	29(38)	32(42)	5(7)	-1.45(-1.26)	-0.71(-0.62)	0.28(0.09)	0.65(0.96)	0.90	0.70	0.70	1.20

(continued)

Table 4.17. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
27	5(7)	16(21)	45(59)	10(13)	-0.89(-0.79)	-0.40(-0.14)	0.68(0.54)	1.27(1.42)	0.90	0.60	0.90	1.10
28	17(22)	21(28)	27(36)	11(14)	-1.15(-1.05)	-0.53(-0.48)	0.25(0.13)	0.87(0.91)	0.90	0.40	0.90	1.00
29	19(25)	22(29)	27(36)	8(11)	-1.19(-1.22)	-0.81(-0.64)	0.19(0.00)	0.56(0.80)	1.10	0.50	0.70	1.20
30	28(53)	14(26)	8(15)	3(6)	-1.71(-1.81)	-1.32(-1.17)	-0.30(-0.39)	0.06(0.48)	1.10	1.80	0.90	1.20
31	2(3)	23(30)	40(53)	11(14)	-0.80(-0.51)	0.18(0.20)	0.96(0.89)	1.61(1.77)	0.80	1.10	0.70	1.10
32	1(1)	24(32)	37(49)	14(18)	-0.51(-0.19)	0.50(0.52)	1.29(1.21)	1.88(2.05)	0.90	1.10	0.70	1.10
33	4(5)	24(32)	35(46)	13(17)	-0.34(-0.62)	0.04(0.04)	0.72(0.70)	1.41(1.53)	1.30	1.00	1.00	1.10
34	3(4)	27(36)	34(45)	12(16)	0.58(-0.58)	-0.07(0.10)	0.91(0.78)	1.35(1.62)	1.80	0.90	0.80	1.30
35	9(12)	26(34)	33(43)	8(11)	-1.50(-1.03)	-0.31(-0.40)	0.28(0.26)	1.27(1.11)	0.50	1.10	0.80	0.90
36	26(34)	24(32)	21(28)	5(7)	-1.54(-1.52)	-0.91(-0.92)	-0.31(-0.24)	0.92(0.56)	1.00	0.90	0.90	0.70
37	17(22)	30(39)	23(30)	6(8)	-1.46(-1.34)	-0.64(-0.72)	-0.13(-0.04)	1.10(0.79)	0.80	1.10	1.00	0.70
38	6(8)	20(26)	28(37)	22(29)	-0.44(-0.48)	0.05(0.12)	0.65(0.71)	1.59(1.46)	1.20	0.70	0.80	0.80
39	2(3)	10(13)	29(38)	35(46)	-0.97(-0.01)	0.94(0.63)	1.17(1.25)	2.04(2.00)	0.30	1.20	0.60	1.00
40	4(5)	14(19)	22(30)	34(46)	-0.26(-0.19)	0.72(0.41)	0.56(0.99)	1.88(1.72)	1.30	1.10	0.60	0.80
41	7(9)	29(38)	23(30)	17(22)	-0.92(-0.65)	-0.15(-0.03)	0.58(0.58)	1.65(1.35)	0.90	0.50	0.50	0.60
42	21(28)	15(20)	20(27)	18(24)	-0.60(-0.87)	-0.21(-0.35)	-0.22(0.19)	0.94(0.92)	3.60	1.80	1.50	1.00
43	11(14)	35(46)	17(22)	13(17)	-1.25(-0.90)	-0.32(-0.28)	0.38(0.35)	1.47(1.12)	0.80	0.70	0.40	0.50
44	10(13)	18(24)	25(33)	23(30)	-0.71(-0.58)	-0.08(-0.02)	0.35(0.53)	1.57(1.26)	0.90	0.60	0.50	0.70
45	4(5)	16(21)	40(53)	16(21)	-0.66(-0.53)	0.19(0.12)	0.71(0.76)	1.68(1.59)	0.70	1.10	1.00	1.00
46	3(4)	14(18)	44(58)	15(20)	-0.57(-0.49)	-0.08(0.18)	0.87(0.85)	1.88(1.70)	0.80	0.70	0.70	0.90
47	6(8)	17(22)	32(42)	21(28)	-0.96(-0.51)	0.14(0.10)	0.66(0.69)	1.62(1.46)	0.60	0.80	0.70	0.90
48	1(3)	4(10)	16(41)	18(46)	-0.08(0.45)	0.56(0.71)	1.29(1.11)	1.73(1.83)	0.50	0.70	1.20	1.10
49	6(11)	4(8)	21(40)	22(42)	-0.53(-0.36)	0.31(0.16)	0.81(0.67)	1.22(1.34)	0.80	1.50	0.90	1.10
50	2(3)	31(41)	31(41)	12(16)	-1.71(-0.46)	0.15(0.24)	0.96(0.92)	2.09(1.76)	0.60	0.80	0.60	0.70
51	4(5)	33(43)	33(43)	6(8)	-1.47(-0.95)	-0.36(-0.26)	0.57(0.47)	1.73(1.36)	0.80	0.70	0.90	0.70
52	8(11)	23(30	36(47)	9(12)	-1.29(-0.95)	-0.49(-0.32)	0.39(0.34)	1.73(1.19)	0.70	0.50	0.60	0.70

(continued)

Table 4.17. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
53	20(27)	18(24)	30(40)	7(9)	-1.33(-1.28)	-0.80(-0.70)	0.05(-0.06)	0.67(0.75)	0.90	0.50	0.80	1.10
54	19(25)	23(30)	31(41)	3(4)	-1.55(-1.61)	-1.21(-1.00)	-0.24(-0.30)	1.24(0.56)	1.10	0.60	0.80	0.70
55	36(47)	29(38)	8(11)	3(4)	-2.01(-1.91)	-1.18(-1.23)	-0.06(-0.45)	-0.05(0.28)	0.90	0.90	0.40	1.30
56	11(14)	17(22)	27(36)	21(28)	-0.78(-0.65)	0.04(-0.09)	0.34(0.47)	1.34(1.21)	0.90	1.00	0.60	1.00
57	19(33)	13(22)	13(22)	13(22)	-0.80(-0.84)	-0.44(-0.44)	0.17(0.08)	0.74(0.88)	1.10	1.50	0.50	1.50
58	27(36)	16(21)	25(33)	8(11)	-1.38(-1.30)	-0.86(-0.74)	-0.03(-0.13)	0.83(0.64)	0.80	0.70	0.60	0.80
59	27(36)	27(36)	17(22)	5(7)	-1.64(-1.55)	-0.92(-0.93)	-0.10(-0.24)	0.44(0.54)	0.90	0.70	0.70	1.30
60	20(27)	28(37)	23(31)	4(5)	-1.50(-1.55)	-0.95(-0.92)	-0.24(-0.21)	0.72(0.62)	1.10	0.80	1.00	1.00
61	31(42)	28(38)	13(18)	2(3)	-2.10(-2.00)	-1.36(-1.32)	-0.29(-0.54)	0.72(0.23)	0.90	0.60	0.70	0.60
62	30(41)	26(35)	15(20)	3(4)	-1.91(-1.81)	-1.15(-1.16)	-0.39(-0.41)	1.10(0.36)	0.90	0.60	0.80	0.50
63	38(51)	27(36)	9(12)	1(1)	-2.41(-2.35)	-1.66(-1.64)	-0.56(-0.80)	0.42(-0.06)	0.90	0.80	0.70	0.60
64	38(50)	25(33)	9(12)	4(5)	-1.83(-1.79)	-1.23(-1.13)	0.00(-0.39)	0.38(0.33)	0.90	0.80	0.40	0.90
65	32(59)	14(26)	6(11)	2(4)	-2.07(-2.10)	-1.56(-1.42)	-0.56(-0.48)	0.99(0.35)	1.10	0.70	1.10	0.40

French Horn Performance Rating Scale Data

Multifaceted Rasch Partial Credit Model Results

Table 4.18 presents summary statistics from the analysis of performers (*n* = 9), raters (*n* = 4), and items (*n* = 63) on the French Horn Performance Rating Scale (HPRS). Overall significant differences in chi-square are indicated for performers (205.70), raters (297.50), and items (239.50). The probability for each facet was less than 0.01 and reliability of separation for each was moderate to high with performers at 0.96, raters at 0.98, and items at 0.77. This indicates adequate spread of elements within each facet along a single measure of music performance ability. Infit MSE values on the HPRS are 0.97 for performers, 1.09 for raters, and 0.98 for items. Outfit MSE values were also well targeted with values of 1.06 for performers, 1.26 for raters, and 1.06 for items.

Table 4.18

HPRS Summary Statistics from the PC-MFR Model

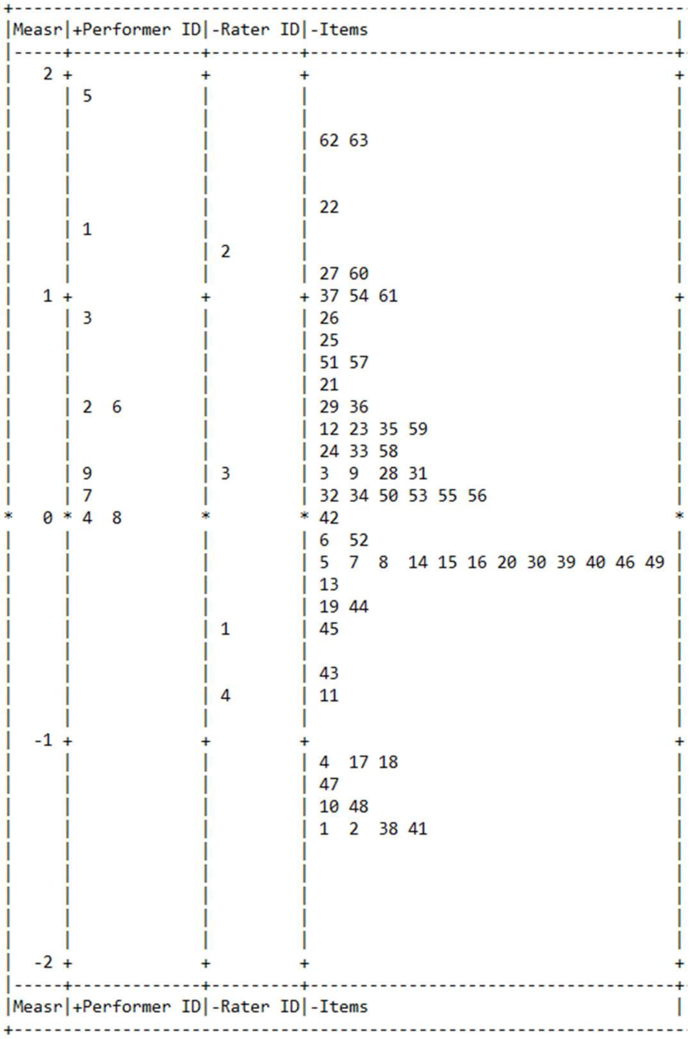
		Facets	
	Performance (θ)	**Rater** (λ)	**Item** (δ)
Measure (Logits)			
Mean	0.60	0.00	0.00
SD	0.62	0.77	0.76
N	9	4	63
Infit *MSE*			
Mean	0.97	1.09	0.98
SD	0.16	0.28.	0.34
Std. Infit *MSE*			
Mean	-0.30	-0.10	0.00
SD	1.30	4.00	1.10
Outfit *MSE*			
Mean	1.06	1.26	1.06
SD	0.22	0.39	0.49

Std. Outfit *MSE*			
Mean	0.20	0.60	0.20
SD	1.40	3.40	1.20
Separation Statistics			
Reliability of Separation	0.96	0.98	0.77
Chi-Square	205.70	297.50	239.50
Degrees of Freedom	8	3	62

**p* < 0.01

Figure 4.5

Variable Map, French Horn Performance Rating Scale



```
+-------------------------------------------------------------------
|Measr|+Performer ID|-Rater ID|-Items                               |
|-----+-------------+---------+-------------------------------------+
|   2 +             +         +                                     +
|     | 5           |         |                                     |
|     |             |         |                                     |
|     |             |         | 62 63                               |
|     |             |         |                                     |
|     |             |         |                                     |
|     |             |         | 22                                  |
|     | 1           |         |                                     |
|     |             | 2       |                                     |
|     |             |         | 27 60                               |
|   1 +             +         + 37 54 61                            +
|     | 3           |         | 26                                  |
|     |             |         | 25                                  |
|     |             |         | 51 57                               |
|     |             |         | 21                                  |
|     | 2  6        |         | 29 36                               |
|     |             |         | 12 23 35 59                         |
|     |             |         | 24 33 58                            |
|     | 9           | 3       | 3  9  28 31                         |
|     | 7           |         | 32 34 50 53 55 56                   |
*   0 * 4  8        *         * 42                                  *
|     |             |         | 6  52                               |
|     |             |         | 5  7  8  14 15 16 20 30 39 40 46 49 |
|     |             |         | 13                                  |
|     |             |         | 19 44                               |
|     |             | 1       | 45                                  |
|     |             |         |                                     |
|     |             |         | 43                                  |
|     |             | 4       | 11                                  |
|     |             |         |                                     |
|  -1 +             +         +                                     +
|     |             |         | 4  17 18                            |
|     |             |         | 47                                  |
|     |             |         | 10 48                               |
|     |             |         | 1  2  38 41                         |
|     |             |         |                                     |
|     |             |         |                                     |
|     |             |         |                                     |
|     |             |         |                                     |
|  -2 +             +         +                                     +
|-----+-------------+---------+-------------------------------------+
|Measr|+Performer ID|-Rater ID|-Items                               |
+-------------------------------------------------------------------
```

Variable Map

Figure 4.5 is a variable map representing horn performance ability as a latent variable. Included on the map are the calibrations of performers (column 2), raters (column 3), and items (column 4). Facets are ordered from top to bottom according to high to low ability of performers, severity to leniency of raters, and highest to lowest difficulty of items.

Column 2 on the map shows the location of each performer along the latent variable, with numbers representing the performer number. All performers fell within the acceptable range for productive parameter-level MSE. Student achievement ranged from 1.93 to -0.02 logits (*M* = 0.60, *SD* = 0.62, *n* = 9). Table 4.19 shows the complete calibration and statistics for performers.

Column 3 shows the calibration of raters along the latent construct, with numbers representing the rater number. Rater 2 was the most severe (observed average = 2.52, logit measure 1.17) and Rater 4 was the most lenient (observed average = 3.23, logit measure -0.84) and was also the only misfit rater with an infit MSE of 1.42.

Column 4 shows the calibration of items, with numbers representing the item number. The most difficult item was Item 62 (dynamics used to create tension and release, observed average = 1.50, logit measure = 1,71). The easiest item was Item 1 (upper body position, observed average = 3.83, logit measure = -1.39). Items demonstrating underfit included 4, 5, 6, 9, 14, 20, and 56. Overfitting items included 13, 52, 57, 58, and 60. Table 4.20 shows the complete calibration and statistics for items.

Table 4.19

Calibration of F Horn Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
5	3.31	1.93	0.13	1.20	1.40	1.09	0.50
1	3.14	1.31	0.13	1.17	1.30	1.43	2.10
3	2.75	0.85	0.12	0.81	-1.60	0.85	-1.10
6	3.10	0.49	0.12	1.15	1.20	1.36	2.10
2	2.55	0.47	0.12	0.84	-1.40	0.84	-1.20
9	2.95	0.23	0.12	0.97	-0.20	1.24	1.60
7	2.90	0.13	0.12	0.91	-0.70	0.89	-0.70
8	2.86	0.04	0.12	0.80	-1.70	0.84	-1.20
4	2.30	-0.02	0.13	0.84	-1.30	0.96	-0.20
Mean	2.87	0.60	0.13	0.97	-0.30	1.06	0.20
SD	0.29	0.62	0.00	0.16	1.30	0.22	1.40

Presented in measure order from highest to lowest achievement.

Table 4.20

Calibration of F Horn Item Facet

Item Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
62	1.50	1.71	0.33	0.83	-0.20	0.73	-0.20
63	1.56	1.66	0.33	0.82	-0.30	0.71	-0.40
22	2.56	1.45	0.39	1.21	0.70	1.41	1.10
27	2.06	1.09	0.34	0.91	-0.10	0.85	-0.30
60	2.06	1.07	0.27	0.54	-1.70	0.54	-1.30
37	2.72	1.05	0.36	0.87	-0.30	0.83	-0.40
54	2.11	0.97	0.36	0.84	-0.30	0.84	-0.30
61	2.00	0.96	0.27	0.65	-1.10	0.57	-1.10
26	2.33	0.86	0.33	0.97	0.00	1.00	0.00
25	2.22	0.85	0.29	1.03	0.10	1.28	0.90
57	2.33	0.72	0.30	0.57	-1.60	0.58	-1.50
51	2.22	0.69	0.32	0.99	0.00	0.96	0.00
21	2.89	0.64	0.34	1.38	1.30	1.49	1.50
36	2.44	0.53	0.28	0.77	-0.70	0.83	-0.40
29	2.33	0.50	0.30	0.91	-0.20	0.85	-0.30
59	2.44	0.44	0.27	0.74	-0.80	0.81	-0.50
12	2.50	0.42	0.33	1.18	0.60	1.17	0.60
23	2.56	0.39	0.33	1.06	0.20	1.05	0.20
35	2.56	0.38	0.24	0.62	-1.40	0.62	-0.90
58	2.56	0.32	0.26	0.43	-2.40	0.46	-1.90
24	2.33	0.27	0.36	0.91	-0.10	0.80	-0.30
33	2.61	0.27	0.28	0.63	-1.30	0.64	-1.20
9	3.11	0.19	0.34	1.45	1.60	1.52	1.60
28	2.56	0.16	0.29	0.83	-0.40	0.86	-0.30
3	3.56	0.16	0.51	1.34	1.80	1.57	2.30
31	3.11	0.16	0.36	0.67	-1.20	0.69	-1.10
55	2.72	0.15	0.27	0.67	-1.20	0.67	-1.10
50	2.61	0.14	0.30	0.95	0.00	1.01	0.10
32	2.61	0.14	0.30	0.94	-0.10	1.00	0.00
34	2.61	0.13	0.35	0.79	-0.60	0.79	-0.60

53	2.61	0.06	0.28	0.89	-0.20	0.96	0.00
56	2.78	0.06	0.27	2.42	3.60	3.09	4.20
42	2.56	0.03	0.33	1.05	0.20	1.05	0.20
52	2.83	-0.10	0.28	0.59	-1.50	0.61	-1.30
6	3.22	-0.12	0.37	1.62	1.90	1.77	2.20
15	2.89	-0.17	0.27	0.91	-0.20	0.98	0.00
49	2.78	-0.17	0.32	0.92	-0.10	1.00	0.00
40	2.78	-0.17	0.32	1.07	0.30	1.08	0.30
20	2.83	-0.20	0.33	1.50	1.50	1.60	1.70
39	3.33	-0.21	0.33	1.33	1.10	1.46	1.10
7	3.33	-0.21	0.33	1.36	1.20	1.70	1.50
30	3.28	-0.22	0.36	0.70	-1.10	0.74	-0.80
16	3.28	-0.22	0.36	1.08	0.30	1.32	1.00
8	3.28	-0.22	0.36	1.15	0.60	1.16	0.50
5	3.28	-0.22	0.36	1.48	1.50	1.50	1.50
46	3.00	-0.22	0.28	0.94	0.00	1.17	0.50
14	2.89	-0.23	0.33	1.78	2.00	1.93	2.40
13	3.06	-0.29	0.28	0.56	-1.50	0.58	-1.10
19	3.00	-0.41	0.31	1.33	1.00	1.53	1.50
44	3.06	-0.44	0.32	0.81	-0.40	0.81	-0.50
45	3.17	-0.54	0.33	0.76	-0.60	0.78	-0.50
43	3.28	-0.70	0.30	0.81	-0.40	0.71	-0.50
11	3.56	-0.76	0.39	0.98	0.00	1.27	0.60
18	3.56	-1.07	0.43	0.78	-0.60	0.73	-0.60
17	3.61	-1.15	0.44	0.72	-0.70	0.59	-0.90
4	3.56	-1.15	0.44	1.45	1.20	1.71	1.40
47	3.67	-1.22	0.45	0.81	-0.30	0.76	-0.20
10	3.72	-1.28	0.46	0.69	-0.60	1.31	0.60
48	3.78	-1.35	0.48	0.99	0.10	0.54	-0.30
41	3.83	-1.39	0.66	0.70	-0.60	0.47	-0.90
38	3.83	-1.39	0.66	1.07	0.20	1.15	0.40
2	3.83	-1.39	0.66	0.88	-0.10	0.71	-0.30
1	3.83	-1.39	0.66	1.38	0.90	2.71	2.10
Mean	2.87	0.00	0.35	0.98	0.00	1.06	0.20
SD	0.56	0.76	0.10	0.34	1.10	0.49	1.20

Presented in measure order from highest to lowest difficulty.

Response Category Diagnostics

Table 4.21 shows response category data for the HPRS including the usage of each response category, average observed and average expected logit measures, and the outfit mean squared error (MSE) value for each item in the rating scale. This data can be used to optimize response categories by eliminating underused categories, combining categories with too small of a threshold between them, or eliminating categories with high outfit MSE values. Based on response category diagnostic data, many of the items in the HPRS would be revised. Thirty-five of the items would require response categories to be collapsed, with 11 of them collapsing into dichotomous response options. Additional items requiring revision due to high MSE values include items 15, 25, 46, and 56.

Table 4.21. HPRS Category Diagnostics for Fit Items: Category Usage, Average Observed and Expected Logit Measure, Outfit Mean Squared Error (*MSE*).

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
1			3(17)	15(83)			2.53(1.33)	1.66(1.90)			3.00	1.20
2			3(17)	15(83)			1.04(1.33)	1.96(1.90)			0.70	0.90
3			8(44)	10(56)			0.36(-0.05)	0.17(0.50)			1.70	1.40
4	1(6)		5(28)	12(67)	2.21(0.76)		1.25(1.23)	1.64(1.77)	3.50		0.60	1.30
5		3(17)	7(39)	8(44)		0.63(0.01)	0.51(0.48)	0.74(1.00)		1.60	1.50	1.30
6		3(17)	8(44)	7(39)		0.66(-0.06)	0.57(0.42)	0.45(0.93)		1.90	1.70	1.70
7		4(22)	4(22)	10(56)		0.44(0.00)	0.43(0.45)	0.78(0.94)		2.00	2.40	1.00
8		3(17)	7(39)	8(44)		0.54(0.01)	0.20(0.48)	1.04(1.00)		1.40	1.10	0.90
9		5(28)	6(33)	7(39)		0.06(-0.30)	0.17(0.16)	0.39(0.66)		1.90	1.20	1.20
10		1(6)	3(17)	14(78)		0.09(0.85)	1.54(1.31)	1.85(1.85)		0.20	2.50	0.80
11		2(11)	4(22)	12(67)		0.61(0.43)	0.81(0.88)	1.40(1.40)		1.80	0.80	0.90
12	2(11)	7(39)	7(39)	2(11)	-1.17(-0.69)	0.22(-0.25)	-0.06(0.24)	0.64(0.70)	0.60	1.80	1.40	1.10
13	2(11)	3(17)	5(28)	8(44)	-0.66(-0.13)	0.48(0.24)	0.30(0.66)	1.39(1.12)	0.40	1.10	0.10	0.60
14	1(6)	4(22)	9(50)	4(22)	1.30(-0.21)	0.94(0.21)	0.30(0.69)	0.96(1.18)	2.90	2.30	1.80	1.20
15	2(11)	5(28)	4(22)	7(39)	-0.34(-0.19)	0.38(0.20)	0.32(0.62)	1.16(1.07)	0.70	0.90	3.00	0.70
16		3(17)	7(39)	8(44)		0.62(0.01)	0.08(0.48)	1.12(1.00)		2.40	0.10	0.80
17		1(6)	5(28)	12(67)		0.96(0.76)	0.67(1.23)	1.99(1.77)		1.00	0.20	0.70
18		1(6)	6(33)	11(61)		0.75(0.72)	0.88(1.19)	1.89(1.72)		0.90	0.60	0.80
19	1(6)	4(22)	7(39)	6(33)	1.29(-0.04)	0.54(0.36)	0.45(0.81)	1.38(1.30)	2.90	2.00	0.40	0.90
20	1(6)	5(28)	8(44)	4(22)	1.08(-0.21)	0.55(0.21)	0.36(0.69)	1.12(1.18)	2.40	1.70	2.00	1.00
21		7(39)	6(33)	5(28)		-0.11(-0.65)	-0.83(-0.17)	0.36(0.32)		1.90	2.30	0.80
22		10(56)	6(33)	2(11)		-1.15(-1.32)	-1.02(-0.80)	-0.48(-0.32)		1.30	1.70	1.30
23	2(11)	6(33)	8(44)	2(11)	-0.90(-0.69)	-0.03(-0.25)	0.14(0.24)	0.66(0.71)	0.70	1.60	1.00	1.00
24	1(6)	12(67)	3(17)	2(11)	-0.59(-0.51)	-0.06(-0.02)	-0.57(0.50)	1.12(0.93)	1.10	1.10	0.50	0.70
25	5(28)	6(33)	5(28)	2(11)	-1.06(-0.97)	-0.36(-0.53)	-0.07(-0.07)	0.02(0.34)	0.80	1.10	2.40	1.10

(continued)

Table 4.21. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
26	3(17)	7(39)	7(39)	1(6)	-1.45(-1.07)	-0.33(-0.61)	-0.20(-0.12)	0.11(0.32)	0.70	1.80	0.90	1.10
27	4(22)	10(56)	3(17)	1(6)	-1.45(-1.19)	-0.61(-0.70)	-0.16(-0.20)	0.24(0.20)	0.80	1.20	0.70	0.80
28	2(11)	8(44)	4(22)	4(22)	-1.11(-0.43)	0.12(0.00)	0.60(0.46)	0.87(0.90)	0.60	1.40	0.80	0.80
29	3(17)	9(50)	3(17)	3(17)	-1.14(-0.68)	-0.12(-0.23)	0.51(0.24)	0.52(0.66)	0.70	1.10	0.70	0.90
30		3(17)	7(39)	8(44)		-0.24(0.01)	0.37(0.48)	1.19(1.00)		0.70	1.00	0.70
31		4(22)	8(44)	6(33)		-0.55(-0.29)	0.13(0.20)	0.98(0.71)		0.70	0.90	0.60
32	2(11)	7(39)	5(28)	4(22)	-0.65(-0.43)	0.08(0.00)	0.45(0.46)	0.89(0.91)	0.80	1.60	1.00	0.80
33	3(17)	5(28)	6(33)	4(22)	-1.01(-0.54)	-0.09(-0.13)	0.43(0.31)	0.88(0.76)	0.50	0.50	0.80	0.70
34	1(6)	7(39)	8(44)	2(11)	-0.84(-0.48)	-0.16(-0.02)	0.63(0.48)	1.01(0.96)	0.80	0.80	0.70	0.90
35	6(33)	2(11)	4(22)	6(33)	-0.75(-0.57)	-0.06(-0.20)	0.24(0.19)	0.70(0.60)	0.50	1.10	0.90	0.60
36	4(22)	5(28)	6(33)	3(17)	-0.93(-0.74)	-0.33(-0.32)	0.26(0.13)	0.56(0.56)	0.70	1.70	0.40	0.80
37		8(44)	7(39)	3(17)		-1.18(-0.99)	-0.23(-0.49)	-0.10(0.00)		0.80	0.60	1.00
38			3(17)	15(83)			1.50(1.33)	1.87(1.90)			1.20	1.10
39		4(22)	4(22)	10(56)		0.18(0.00)	0.79(0.45)	0.74(0.94)		1.30	1.50	1.70
40	1(6)	6(33)	7(39)	4(22)	-1.02(-0.21)	0.41(0.22)	0.78(0.69)	0.92(1.17)	0.50	1.40	0.80	1.20
41			3(17)	15(83)			0.65(1.33)	2.04(1.90)			0.40	0.80
42	1(6)	9(50)	5(28)	3(17)	-1.22(-0.33)	0.26(0.12)	0.63(0.61)	0.92(1.07)	0.70	1.50	0.70	1.10
43	1(6)	3(17)	4(22)	10(56)	-0.01(0.17)	0.31(0.54)	1.13(0.97)	1.46(1.44)	0.60	0.50	1.00	0.90
44	1(6)	3(17)	8(44)	6(33)	0.57(-0.05)	-0.21(0.35)	0.79(0.82)	1.53(1.31)	1.60	0.30	0.60	0.80
45	1(6)	2(11)	8(44)	7(39)	0.68(0.02)	-0.41(0.41)	0.73(0.86)	1.65(1.36)	1.80	0.10	0.40	0.70
46	2(11)	3(17)	6(33)	7(39)	-0.72(-0.19)	0.96(0.19)	0.41(0.62)	1.09(1.09)	0.40	3.30	0.30	0.90
47		1(6)	4(22)	13(72)		0.51(0.81)	1.14(1.27)	1.87(1.81)		0.50	0.90	0.80
48		1(6)	2(11)	15(83)		0.64(0.88)	1.26(1.34)	1.91(1.88)		0.40	0.50	1.10
49	1(6)	6(33)	7(39)	4(22)	-0.15(-0.21)	0.15(0.22)	0.72(0.69)	1.20(1.17)	1.00	1.40	0.70	0.80
50	2(11)	7(39)	5(28)	4(22)	-1.08(-0.43)	0.23(0.00)	0.49(0.46)	0.78(0.91)	0.60	2.00	0.70	0.90
51	3(17)	10(56)	3(17)	2(11)	-1.43(-0.84)	-0.14(-0.37)	0.05(0.12)	0.36(0.53)	0.70	1.10	1.20	1.00
52	2(11)	5(28)	5(28)	6(33)	-0.42(-0.25)	-0.12(0.14)	0.66(0.58)	1.23(1.03)	0.70	0.40	0.90	0.60

(continued)

Table 4.21. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
53	2(11)	8(44)	3(17)	5(28)	-0.58(-0.34)	0.20(0.08)	0.11(0.53)	1.12(0.97)	0.80	1.80	0.90	0.60
54	3(17)	11(61)	3(17)	1(6)	-1.38(-1.11)	-0.54(-0.61)	-0.26(-0.10)	0.84(0.31)	0.80	1.10	1.10	0.50
55	3(17)	4(22)	6(33)	5(28)	-0.40(-0.45)	-0.52(-0.05)	0.50(0.38)	1.02(0.83)	0.80	0.30	0.70	0.70
56	3(17)	4(22)	5(28)	6(33)	1.06(-0.37)	-0.18(0.01)	0.56(0.43)	0.19(0.88)	5.10	0.10	0.90	3.60
57	4(22)	6(33)	6(33)	2(11)	-1.24(-0.90)	-0.50(-0.47)	0.13(0.00)	0.85(0.43)	0.60	0.30	0.60	0.60
58	4(22)	5(28)	4(22)	5(28)	-0.92(-0.54)	-0.25(-0.14)	0.56(0.28)	0.89(0.71)	0.50	0.10	0.50	0.60
59	4(22)	6(33)	4(22)	4(22)	-1.04(-0.63)	-0.03(-0.22)	0.36(0.22)	0.62(0.64)	0.60	1.50	0.80	0.70
60	8(44)	3(17)	5(28)	2(11)	-1.34(-1.10)	-0.34(-0.68)	-0.20(-0.24)	0.50(0.15)	0.50	1.40	0.30	0.50
61	8(44)	5(28)	2(11)	3(17)	-1.24(-0.97)	-0.20(-0.54)	-0.09(-0.10)	0.40(0.27)	0.60	0.80	0.30	0.50
62	12(67)	4(22)	1(6)	1(6)	-1.59(-1.53)	-0.83(-1.02)	-1.03(-0.58)	0.10(-0.28)	0.80	0.70	1.30	0.30
63	11(61)	5(28)	1(6)	1(6)	-1.59(-1.51)	-0.80(-1.00)	-0.98(-0.55)	0.15(-0.23)	0.90	0.60	1.30	0.40

Trombone Performance Rating Scale Data

Multifaceted Rasch Partial Credit Model Results

Table 4.22 presents summary statistics from the analysis of performers (*n* = 18), raters (*n* = 6), and items (*n* = 65) on the Trombone Performance Rating Scale (TBPRS). Overall significant differences in chi-square are indicated for performers (1044.90), raters (551.40), and items (744.70). The probability for each facet was less than 0.01 and reliability of separation for each was high with performers at 0.99, raters at 0.99, and items at 0.92. This indicates adequate spread of elements within each facet along a single measure of music performance ability. Infit MSE values on the TBPRS are 1.01 for performers, 1.08 for raters, and 1.00 for items. Outfit MSE values were also well targeted with values of 1.03 for performers, 1.09 for raters, and 1.02 for items.

Table 4.22

TBPRS Summary Statistics from the PC-MFR Model

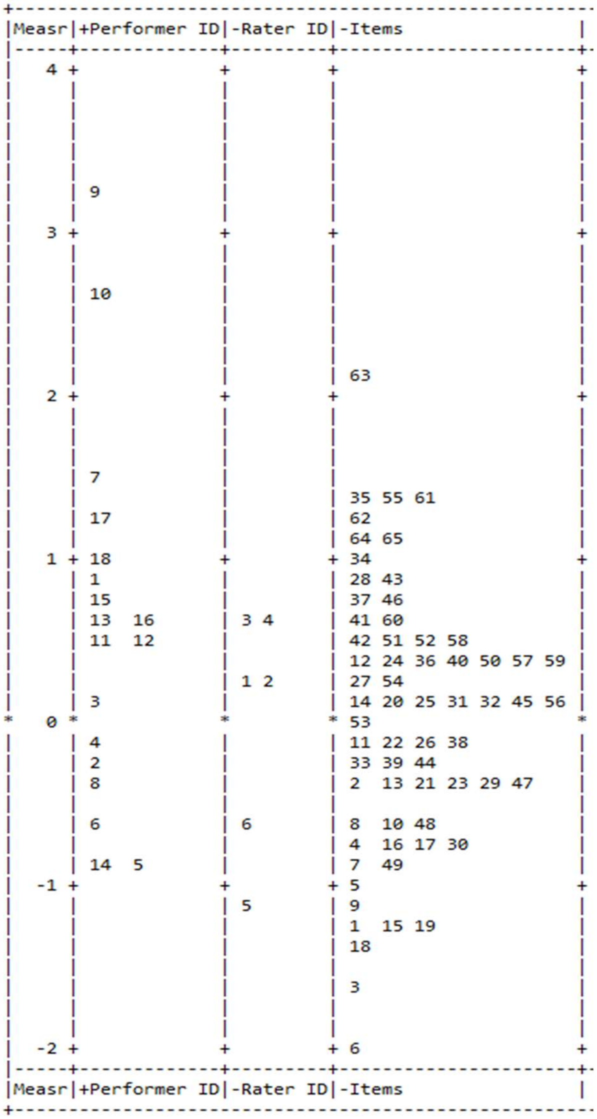
		Facets		
		Performance (θ)	**Rater** (λ)	**Item** (δ)
Measure (Logits)				
	Mean	0.57	0.00	0.00
	SD	1.08	0.66	0.80
	N	18	6	65
Infit *MSE*				
	Mean	1.01	1.08	1.00
	SD	0.14	0.25	2.40
Std. Infit *MSE*				
	Mean	0.00	0.40	0.00
	SD	1.20	3.30	1.10
Outfit *MSE*				
	Mean	1.03	1.09	1.02
	SD	0.15	0.27	0.34

Std. Outfit *MSE*			
Mean	0.10	0.60	0.00
SD	1.30	2.60	1.10
Separation Statistics			
Reliability of Separation	0.99	0.99	0.92
Chi-Square	1044.90	551.40	744.70
Degrees of Freedom	17	5	64

*$p < 0.01$

Figure 4.6

Variable Map, Trombone Performance Rating Scale



```
+-----------------------------------------------------
|Measr|+Performer ID|-Rater ID|-Items                |
|-----+-------------+---------+----------------------+-
|   4 +             +         +                      +
|     |             |         |                      |
|     |             |         |                      |
|     |             |         |                      |
|     |             |         |                      |
|     |             |         |                      |
|     | 9           |         |                      |
|     |             |         |                      |
|   3 +             +         +                      +
|     |             |         |                      |
|     |             |         |                      |
|     | 10          |         |                      |
|     |             |         |                      |
|     |             |         |                      |
|     |             |         |                      |
|     |             |         | 63                   |
|   2 +             +         +                      +
|     |             |         |                      |
|     |             |         |                      |
|     |             |         |                      |
|     | 7           |         |                      |
|     |             |         | 35 55 61             |
|     | 17          |         | 62                   |
|     |             |         | 64 65                |
|   1 + 18          +         + 34                   +
|     | 1           |         | 28 43                |
|     | 15          |         | 37 46                |
|     | 13  16      | 3 4     | 41 60                |
|     | 11  12      |         | 42 51 52 58          |
|     |             |         | 12 24 36 40 50 57 59 |
|     |             | 1 2     | 27 54                |
|     | 3           |         | 14 20 25 31 32 45 56 |
*   0 *             *         * 53                   *
|     | 4           |         | 11 22 26 38          |
|     | 2           |         | 33 39 44             |
|     | 8           |         | 2  13 21 23 29 47    |
|     |             |         |                      |
|     | 6           | 6       | 8  10 48             |
|     |             |         | 4  16 17 30          |
|     | 14  5       |         | 7  49                |
|  -1 +             +         + 5                    +
|     |             | 5       | 9                    |
|     |             |         | 1  15 19             |
|     |             |         | 18                   |
|     |             |         |                      |
|     |             |         | 3                    |
|     |             |         |                      |
|     |             |         |                      |
|  -2 +             +         + 6                    +
|-----+-------------+---------+----------------------+-
|Measr|+Performer ID|-Rater ID|-Items                |
+-----------------------------------------------------
```

Variable Map

Figure 4.6 is a variable map representing trombone performance ability as a latent variable. Included on the map are the calibrations of performers (column 2), raters (column 3), and items (column 4). Facets are ordered from top to bottom according to high to low ability of performers, severity to leniency of raters, and highest to lowest difficulty of items.

Column 2 on the map shows the location of each performer along the latent variable, with numbers representing the performer number. Student achievement ranged from 3.3 to -0.9 logits (*M* = 0.57, *SD* = 1.08, *n* = 18). Underfitting performers included 3, 9, 33, and 37. Performer 19 was the only overfitting performer with an infit MSE of 0.53. Table 4.23 shows the complete calibration and statistics for performers.

Column 3 shows the calibration of raters along the latent construct, with numbers representing the rater number. Rater 3 was the most severe (observed average = 2.50, logit measure 0.68) and Rater 5 was the most lenient (observed average = 3.3, logit measure -1.16). Rater 6 was the only misfit rater with an infit MSE value of 1.52.

Column 4 shows the calibration of items, with numbers representing the item number. The most difficult item was Item 63 (dynamics used to create tension and release, observed average = 1.72, logit measure = 2.11). The easiest item was Item 16 (chin position, observed average = 3.87, logit measure = -1.96). Underfitting items included items 4 and 57. There were no overfitting items. Table 4.24 shows the complete calibration and statistics for items.

Table 4.23

Calibration of Trombone Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
9	3.7	3.3	0.2	1.0	0.1	1.1	0.6
10	3.5	2.6	0.2	1.1	1.0	1.2	0.9
7	3.0	1.5	0.1	0.9	-0.6	1.1	0.5
17	3.3	1.2	0.1	1.0	0.1	1.1	0.3
18	3.2	1.0	0.1	0.9	-1.4	0.9	-1.0
1	2.8	0.9	0.1	0.9	-1.5	0.8	-1.7
15	3.1	0.7	0.1	1.0	0.0	1.0	0.1
13	2.9	0.6	0.1	1.1	1.3	1.1	0.4
16	3.0	0.6	0.1	1.3	2.6	1.4	3.3
12	2.8	0.5	0.1	0.9	-1.1	0.9	-0.9
11	2.8	0.5	0.1	0.9	-0.7	1.2	2.0
3	2.4	0.1	0.1	0.9	-0.8	0.9	-1.0
4	2.3	-0.1	0.1	1.1	0.6	1.1	0.4
2	2.2	-0.3	0.1	0.9	-1.1	0.9	-1.6
8	2.1	-0.3	0.1	0.9	-1.3	0.9	-1.1
6	2.0	-0.6	0.1	1.2	1.3	1.1	0.5
5	1.9	-0.9	0.1	1.3	1.9	1.1	0.8
14	2.1	-0.9	0.1	1.0	-0.2	0.9	-0.6
Mean	2.72	0.57	0.12	1.01	0.00	1.03	0.10
SD	0.53	1.08	0.02	0.14	1.20	0.15	1.30

Presented in measure order from highest to lowest achievement.

Table 4.24

Calibration of Trombone Item Facet

Item Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
63	1.72	2.11	0.25	1.09	0.40	1.10	0.50
61	2.09	1.40	0.22	0.80	-1.00	0.79	-1.00
55	2.02	1.36	0.22	1.33	1.50	1.25	1.20
35	2.13	1.33	0.22	0.96	-0.10	0.94	-0.20
62	2.15	1.20	0.21	0.84	-0.80	0.82	-0.80
65	1.82	1.17	0.23	0.99	0.00	0.88	-0.30
64	1.76	1.07	0.26	1.03	0.20	0.98	0.00
34	2.30	0.96	0.22	1.06	0.30	1.08	0.40
28	2.03	0.90	0.22	1.12	0.50	1.06	0.30
43	2.30	0.84	0.22	0.72	-1.50	0.73	-1.50
37	2.37	0.71	0.20	0.67	-1.80	0.70	-1.50
46	2.89	0.70	0.25	0.73	-1.60	0.72	-1.60
41	2.39	0.65	0.22	0.83	-0.80	0.84	-0.80
60	2.46	0.58	0.20	0.92	-0.30	0.91	-0.30
51	2.48	0.51	0.20	0.78	-1.20	0.79	-1.00
58	2.42	0.49	0.23	0.97	0.00	0.96	-0.10
52	2.46	0.49	0.20	0.92	-0.30	0.94	-0.20
42	2.43	0.44	0.19	0.75	-1.20	0.73	-1.20
50	2.50	0.43	0.19	0.70	-1.70	0.69	-1.60
36	2.41	0.43	0.24	0.88	-0.50	0.85	-0.60
59	2.48	0.42	0.21	1.04	0.20	1.04	0.20

24	2.52	0.38	0.21	1.08	0.40	1.10	0.50
40	2.50	0.37	0.19	0.93	-0.20	1.03	0.20
57	2.53	0.33	0.25	1.83	3.10	1.81	3.00
12	3.02	0.32	0.24	1.18	0.90	1.24	1.20
54	2.57	0.23	0.20	0.71	-1.50	0.76	-1.10
27	2.63	0.19	0.20	1.15	0.80	1.22	1.10
25	2.61	0.18	0.20	0.95	-0.20	0.92	-0.30
45	3.07	0.18	0.27	0.74	-1.40	0.74	-1.50
14	3.09	0.14	0.24	1.30	1.50	1.31	1.30
56	2.61	0.13	0.21	1.05	0.30	1.05	0.30
31	2.61	0.12	0.21	0.70	-1.60	0.68	-1.70
20	2.63	0.11	0.20	1.08	0.40	1.10	0.50
32	2.61	0.08	0.21	0.84	-0.80	0.83	-0.70
53	2.76	-0.05	0.20	0.71	-1.60	0.67	-1.70
22	2.70	-0.11	0.20	1.16	0.80	1.14	0.60
11	3.20	-0.13	0.24	1.09	0.50	1.28	1.10
38	2.85	-0.17	0.19	0.87	-0.60	0.99	0.00
26	2.83	-0.18	0.20	1.01	0.10	0.90	-0.40
44	2.76	-0.21	0.21	1.06	0.30	1.03	0.20
33	2.80	-0.25	0.21	1.25	1.20	1.29	1.30
39	2.87	-0.27	0.20	0.71	-1.50	0.70	-1.50
29	2.85	-0.32	0.21	1.05	0.30	1.10	0.50
47	2.87	-0.34	0.21	0.71	-1.50	0.72	-1.40
21	2.76	-0.38	0.23	0.96	-0.10	0.99	0.00
13	2.80	-0.38	0.23	0.87	-0.50	0.89	-0.40
23	2.89	-0.39	0.24	0.98	0.00	0.94	-0.20
2	3.30	-0.41	0.24	1.20	1.00	1.29	1.00
10	2.98	-0.57	0.24	0.84	-0.70	0.88	-0.50
48	3.08	-0.67	0.24	0.67	-1.60	0.64	-1.70
8	3.15	-0.67	0.21	1.03	0.20	1.07	0.30
30	3.00	-0.71	0.22	1.01	0.10	0.94	-0.20
4	3.22	-0.75	0.19	1.89	3.20	2.82	3.70
17	3.13	-0.75	0.21	1.33	1.40	1.24	1.00
16	3.11	-0.75	0.21	1.26	1.20	1.20	0.80
7	3.11	-0.82	0.22	1.12	0.50	1.07	0.30
49	3.08	-0.93	0.23	0.78	-1.00	0.67	-1.40
5	3.17	-0.95	0.21	1.28	1.30	1.29	1.00
9	3.57	-1.19	0.27	1.35	1.50	2.08	2.30
15	3.22	-1.19	0.23	0.99	0.00	1.02	0.10
1	3.41	-1.21	0.23	0.95	-0.10	0.90	-0.20
19	3.28	-1.29	0.23	1.10	0.50	0.93	-0.10
18	3.37	-1.40	0.23	1.16	0.70	1.02	0.10
3	3.54	-1.57	0.26	1.20	0.70	1.34	0.90
6	3.87	-1.96	0.46	0.92	-0.20	0.94	0.10
Mean	2.74	0.00	0.22	1.00	0.00	1.02	0.00
SD	0.45	0.80	0.04	0.24	1.10	0.34	1.10

Presented in measure order from highest to lowest difficulty.

Response Category Diagnostics

Table 4.25 shows response category data for the TBPRS including the usage of each response category, average observed and average expected logit measures, and the outfit mean squared error (MSE) value for each item in the rating scale. This data can be used to optimize response categories by eliminating underused categories, combining categories with too small of a threshold between them, or eliminating categories with high outfit MSE values. Based on response category diagnostic data, many of the items in the TBPRS would be revised. Thirty-nine items would require collapsing response categories due to insufficient usage. Additional items that would require revision due to high MSE values include items 4 and 57. Items requiring revision due to disordered logit measures include items 3, 18, 19, 22, 28, 34, 38, 57, and 65.

Table 4.25. TBPRS Category Diagnostics for Fit Items: Category Usage, Average Observed and Expected Logit Measure, Outfit Mean Squared Error (*MSE*).

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
1	2(4)	3(7)	15(33)	26(57)	0.23(0.12)	0.26(0.44)	1.00(1.06)	2.21(2.16)	1.00	0.50	1.00	0.90
2		9(20)	14(30)	23(50)		-0.22(-0.28)	0.60(0.39)	1.33(1.49		1.50	1.10	1.20
3	1(2)	2(4)	14(30)	29(63)	0.10(0.42)	1.96(0.72)	1.32(1.31)	2.34(2.42)	0.60	3.60	1.00	1.10
4	5(11)	5(11)	11(24)	25(54)	0.42(-0.23)	0.63(0.14)	0.95(0.76)	1.47(1.79)	5.70	2.30	1.30	2.30
5	2(4)	9(20)	14(30)	21(46)	-0.57(-0.05)	0.79(0.34)	1.06(1.04)	1.95(2.11)	0.60	2.00	0.90	1.40
6			6(13)	40(87)			1.28(1.41)	2.52(2.50)			0.90	0.90
7	2(4)	8(17)	19(41)	17(37)	0.05(-0.17)	0.31(0.24)	0.97(1.00)	2.06(2.09)	1.30	1.00	1.00	1.10
8	4(9)	4(9)	19(41)	19(41)	-0.33(-0.32)	0.05(0.07)	0.85(0.78)	1.80(1.87)	1.00	0.80	1.30	1.00
9		4(9)	12(26)	30(65)		0.71(0.33)	1.35(0.91)	1.79(2.02)		2.30	2.50	1.30
10	2(4)	8(17)	25(54)	11(24)	-0.83(-0.41)	0.03(0.04)	0.90(0.91)	2.11(2.01)	0.60	1.20	0.80	0.90
11		11(24)	15(33)	20(43)		-0.27(-0.50)	0.02(0.23)	1.35(1.32)		2.20	0.30	0.90
12		13(28)	19(41)	14(30)		-0.53(-0.85)	-0.23(-0.01)	1.08(1.08)		1.60	1.30	0.80
13	2(4)	15(33)	19(41)	10(22)	-0.81(-0.51)	0.04(0.00)	0.82(0.93)	2.14(1.94)	0.80	1.10	1.00	0.70
14		13(28)	16(35)	17(37)		-0.55(-0.71)	0.20(0.08)	0.92(1.16)		1.10	1.40	1.50
15	1(2)	8(17)	17(37)	20(43)	-0.13(0.16)	0.74(0.54)	1.11(1.26)	2.42(2.36)	0.80	1.00	1.40	0.80
16	3(7)	9(20)	14(30)	20(43)	-0.36(-0.21)	0.36(0.19)	1.10(0.90)	1.76(1.95)	0.80	0.90	1.50	1.40
17	3(7)	7(15)	17(37)	19(41)	0.00(-0.23)	0.39(0.16)	0.91(0.88)	1.81(1.96)	1.30	1.50	1.00	1.30
18	1(2)	7(15)	12(26)	26(57)	-0.16(0.33)	1.14(0.67)	1.03(1.30)	2.40(2.38)	0.60	1.80	0.40	1.00
19	1(2)	8(17)	14(30)	23(50)	1.25(0.25)	0.54(0.61)	1.18(1.28)	2.41(2.37)	1.90	1.00	0.40	1.10
20	8(17)	14(30)	11(24)	13(28)	-0.59(-0.85)	-0.55(-0.32)	0.74(0.54)	1.35(1.44)	1.20	1.10	1.20	1.00
21	2(4)	18(39)	15(33)	11(24)	-0.82(-0.49)	0.14(0.04)	0.80(0.97)	2.06(1.94)	0.90	1.30	1.10	0.70
22	5(11)	18(39)	9(20)	14(30)	-0.19(-0.68)	-0.37(-0.15)	1.29(0.72)	1.39(1.64)	1.20	0.80	1.40	1.20
23	2(4)	9(20)	27(59)	8(17)	-1.04(-0.57)	-0.31(-0.10)	1.11(0.84)	1.38(1.93)	0.50	0.70	0.90	1.40
24	7(15)	15(33)	17(37)	7(15)	-1.15(-1.12)	-0.38(-0.54)	0.32(0.45)	1.32(1.33)	0.90	1.60	1.10	0.90
25	8(17)	13(28)	14(30)	11(24)	-0.93(-0.93)	-0.46(-0.39)	0.54(0.50)	1.42(1.40)	0.90	0.70	1.10	0.90
26	6(13)	10(22)	16(35)	14(30)	-0.44(-0.67)	-0.33(-0.20)	0.46(0.61)	1.78(1.61)	1.20	0.70	0.90	0.80

(continued)

Table 4.25. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
27	7(15)	12(26)	18(39)	9(20)	-0.65(-0.97)	-0.44(-0.44)	0.31(0.49)	1.52(1.42)	1.20	1.90	1.20	0.80
28	15(39)	11(29)	8(21)	4(11)	-1.25(-1.49)	-1.40(-0.91)	0.31(-0.06)	0.55(0.83)	1.30	1.30	0.80	0.90
29	4(9)	13(28)	15(33)	14(30)	-0.39(-0.55)	0.00(-0.07)	0.58(0.76)	1.84(1.76)	1.10	1.80	0.70	0.90
30	2(4)	12(26)	16(35)	16(35)	-0.09(-0.23)	0.14(0.21)	1.05(1.00)	2.05(2.06)	1.10	0.70	0.90	1.10
31	5(11)	17(37)	15(33)	9(20)	-0.98(-0.90)	-0.54(-0.33)	0.72(0.62)	1.80(1.53)	0.90	0.40	0.70	0.60
32	5(11)	18(39)	13(28)	10(22)	-1.07(-0.86)	-0.25(-0.29)	0.50(0.66)	1.78(1.55)	0.80	1.10	0.80	0.60
33	4(9)	14(30)	15(33)	13(28)	-0.71(-0.60)	0.23(-0.11)	0.52(0.75)	1.67(1.74)	0.90	2.20	0.70	1.40
34	10(22)	15(33)	18(39)	3(7)	-1.24(-1.62)	-1.33(-0.94)	0.19(0.11)	1.05(0.89)	1.60	0.90	0.70	0.90
35	13(28)	16(35)	15(33)	2(4)	-1.97(-1.90)	-1.09(-1.14)	-0.08(-0.09)	0.56(0.59)	0.80	1.10	1.00	1.00
36	4(9)	24(52)	13(28)	5(11)	-1.14(-1.17)	-0.60(-0.49)	0.79(0.60)	1.44(1.40)	1.00	0.80	0.70	0.90
37	12(26)	11(24)	17(37)	6(13)	-1.68(-1.36)	-0.60(-0.75)	0.35(0.22)	1.06(1.04)	0.50	1.00	0.50	0.90
38	7(15)	8(17)	16(35)	15(33)	-0.50(-0.67)	-0.60(-0.21)	0.61(0.57)	1.65(1.57)	1.80	0.60	0.60	0.80
39	5(11)	10(22)	17(37)	14(30)	-0.66(-0.61)	-0.43(-0.14)	0.69(0.67)	1.89(1.69)	0.80	0.60	0.60	0.80
40	11(24)	12(26)	12(26)	11(24)	-1.12(-1.05)	-0.31(-0.49)	0.19(0.40)	1.35(1.25)	0.80	1.20	1.60	0.70
41	8(17)	17(37)	16(35)	5(11)	-1.41(-1.35)	-0.76(-0.70)	0.32(0.33)	1.49(1.15)	0.90	0.90	0.80	0.70
42	11(24)	15(33)	9(20)	11(24)	-1.15(-1.09)	-0.69(-0.49)	0.85(0.42)	1.23(1.23)	0.90	0.50	0.70	0.80
43	9(20)	18(39)	15(33)	4(9)	-1.70(-1.50)	-0.86(-0.81)	0.31(0.24)	1.44(1.01)	0.80	0.90	0.60	0.70
44	4(9)	16(35)	13(28)	13(28)	-0.58(-0.62)	-0.21(-0.11)	1.08(0.76)	1.52(1.73)	1.00	1.10	0.90	1.10
45		9(20)	25(54)	12(26)		-0.98(-0.77)	0.01(0.09)	1.55(1.23)		0.80	0.80	0.70
46		15(33)	21(46)	10(22)		-1.38(-1.16)	-0.19(-0.23)	1.09(0.83)		0.70	0.80	0.70
47	4(9)	12(26)	16(35)	14(30)	-0.69(-0.54)	-0.32(-0.07)	0.88(0.75)	1.87(1.77)	0.80	0.40	0.90	0.80
48	2(5)	8(20)	15(38)	15(38)	-0.56(-0.33)	-0.01(0.15)	0.88(1.03)	2.35(2.08)	0.70	0.50	0.70	0.60
49	2(5)	8(21)	14(36)	15(38)	-0.37(-0.07)	0.26(0.29)	0.70(0.89)	2.11(1.89)	0.70	0.70	0.40	0.80
50	11(24)	10(22)	16(35)	9(20)	-1.36(-1.13)	-0.62(-0.57)	0.53(0.34)	1.21(1.21)	0.60	0.40	0.80	0.90
51	10(22)	11(24)	18(39)	7(15)	-1.48(-1.22)	-0.36(-0.64)	0.13(0.31)	1.60(1.18)	0.60	1.30	0.80	0.70
52	10(22)	13(28)	15(33)	8(17)	-1.24(-1.18)	-0.53(-0.59)	0.25(0.36)	1.39(1.21)	1.10	0.70	1.10	0.70
53	7(15)	10(22)	16(35)	13(28)	-0.79(-0.77)	-0.41(-0.28)	0.30(0.55)	1.94(1.53)	0.90	0.70	0.60	0.50

(continued)

Table 4.25. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
54	8(17)	15(33)	12(26)	11(24)	-1.08(-0.95)	-0.56(-0.39)	0.77(0.51)	1.43(1.39)	0.80	0.40	1.20	0.70
55	15(33)	18(39)	10(22)	3(7)	-1.53(-1.85)	-1.18(-1.02)	-0.21(0.01)	0.69(0.61)	1.60	0.90	1.40	0.90
56	6(13)	16(35)	14(30)	10(22)	-0.93(-0.91)	-0.20(-0.35)	0.34(0.58)	1.60(1.49)	0.90	1.70	0.90	0.80
57	4(11)	14(39)	13(36)	5(14)	0.08(-1.15)	-0.37(-0.50)	0.24(0.57)	0.98(1.48)	2.60	1.50	1.40	1.70
58	5(11)	21(47)	14(31)	5(11)	-1.14(-1.24)	-0.60(-0.57)	0.48(0.53)	1.50(1.33)	1.10	0.90	1.00	0.80
59	7(15)	17(37)	15(33)	7(15)	-1.39(-1.15)	-0.17(-0.53)	0.00(0.46)	1.66(1.31)	0.80	1.60	1.30	0.70
60	10(22)	11(24)	19(41)	6(13)	-1.40(-1.29)	-0.60(-0.70)	0.22(0.27)	1.30(1.14)	0.80	0.80	1.10	0.80
61	14(30)	16(35)	14(30)	2(4)	-2.13(-1.94)	-1.04(-1.15)	-0.10(-0.11)	0.91(0.54)	0.70	1.00	0.80	0.80
62	14(30)	14(30)	15(33)	3(7)	-1.90(-1.76)	-0.93(-1.02)	0.02(0.00)	0.88(0.70)	0.70	1.00	0.80	0.90
63	20(43)	20(43)	5(11)	1(2)	-2.32(-2.43)	-1.47(-1.39)	-0.67(-0.43)	0.52(0.02)	1.20	1.10	1.40	0.60
64	19(41)	19(41)	8(17)		-1.32(-1.43)	-0.64(-0.43)	0.79(0.56)		1.20	1.00	0.70	
65	18(47)	13(34)	3(8)	4(11)	-1.63(-1.65)	-1.18(-0.97)	1.27(-0.06)	0.28(0.72)	1.00	1.00	0.30	1.00

Tuba Performance Rating Scale Data

Multifaceted Rasch Partial Credit Model Results

Table 4.26 presents summary statistics from the analysis of performers (*n* = 8), raters (*n* = 5), and items (*n* = 63) on the Tuba Performance Rating Scale (TUPRS). Overall significant differences in chi-square are indicated for performers (387.0), raters (110.7), and items (172.8). The probability for each facet was less than 0.01 and reliability of separation for each was high with performers (0.98) and raters (0.97). Reliability of separation was lower for items (0.65). Mean Square Error values (MSE) close to 1.00 indicate good data to model fit (Bond & Fox, 2020). Infit MSE values on the TUPRS are 1.09 for performers, 1.27 for raters, and 1.01 for items. Outfit MSE values were also well targeted with values of 1.08 for performers, 1.31 for raters, and 1.07 for items

Table 4.26

TUPRS Summary Statistics from the PC-MFR Model

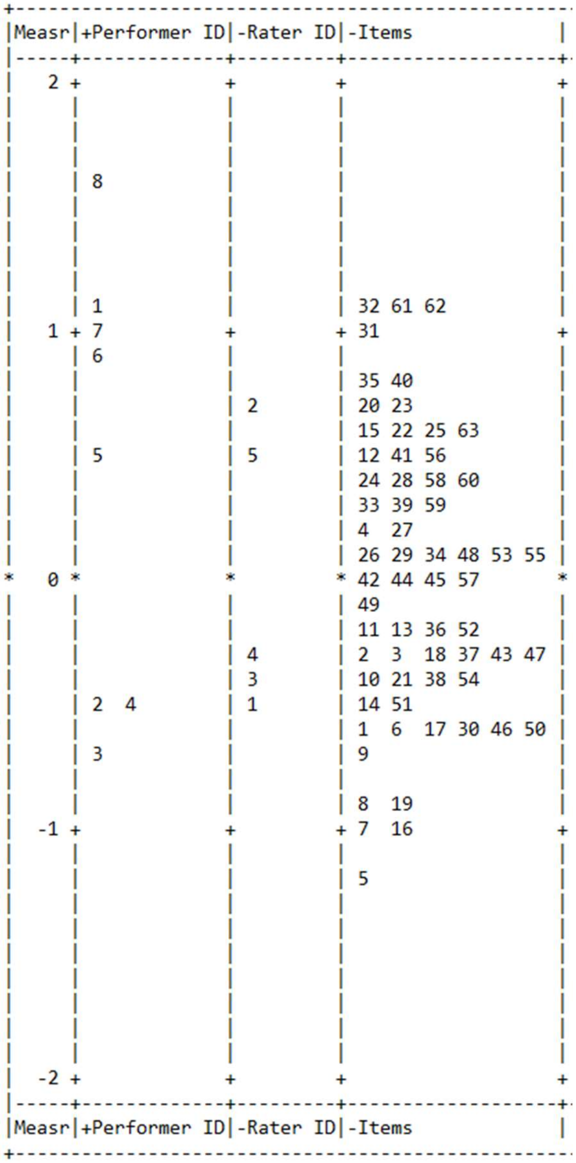
			Facets	
		Performance (θ)	**Rater** (λ)	**Item** (δ)
Measure (Logits)				
	Mean	0.43	0.00	0.00
	SD	0.81	0.52	0.56
	N	8	5	63
Infit *MSE*				
	Mean	1.09	1.27	1.01
	SD	0.39	0.63	0.37
Std. Infit *MSE*				
	Mean	0.30	0.90	0.00
	SD	3.60	4.70	1.10
Outfit *MSE*				
	Mean	1.08	1.31	1.07
	SD	0.39	0.69	0.58

Std. Outfit *MSE*			
Mean	0.10	1.10	0.10
SD	3.60	4.70	1.20
Separation Statistics			
Reliability of Separation	0.98	0.97	0.65
Chi-Square	387.00	110.70	172.80
Degrees of Freedom	7	4	62

**p* < 0.01

Figure 4.7

Variable Map, Tuba Performance Rating Scale



```
+--------------------------------------------------
|Measr|+Performer ID|-Rater ID|-Items             |
|-----+-------------+---------+-------------------+
|   2 +             +         +                   +
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     | 8           |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     | 1           |         | 32 61 62          |
|   1 + 7           +         + 31                +
|     | 6           |         |                   |
|     |             |         | 35 40             |
|     |             | 2       | 20 23             |
|     |             |         | 15 22 25 63       |
|     | 5           | 5       | 12 41 56          |
|     |             |         | 24 28 58 60       |
|     |             |         | 33 39 59          |
|     |             |         | 4  27             |
|     |             |         | 26 29 34 48 53 55 |
*   0 *             *         * 42 44 45 57       *
|     |             |         | 49                |
|     |             |         | 11 13 36 52       |
|     |             | 4       | 2  3  18 37 43 47 |
|     |             | 3       | 10 21 38 54       |
|     | 2  4        | 1       | 14 51             |
|     |             |         | 1  6  17 30 46 50 |
|     | 3           |         | 9                 |
|     |             |         |                   |
|     |             |         | 8  19             |
|  -1 +             +         + 7  16             +
|     |             |         |                   |
|     |             |         | 5                 |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|     |             |         |                   |
|  -2 +             +         +                   +
|-----+-------------+---------+-------------------+
|Measr|+Performer ID|-Rater ID|-Items             |
+--------------------------------------------------
```

Variable Map

Figure 4.7 is a variable map representing tuba performance ability as a latent variable. Included on the map are the calibrations of performers (column 2), raters (column 3), and items (column 4). Facets are ordered from top to bottom according to high to low ability of performers, severity to leniency of raters, and highest to lowest difficulty of items.

Column 2 on the map shows the location of each performer along the latent variable, with numbers representing the performer number. Student achievement ranged from 1.58 to -0.66 logits (*M* = 0.43, *SD* = 0.81, *n* = 8). Underfitting performers included numbers 1 and 2, and there were no overfitting performers. Table 4.27 shows the complete calibration and statistics for performers.

Column 3 shows the calibration of raters along the latent construct, with numbers representing the rater number. Rater 2 was the most severe (observed average = 2.55, logit measure 0.74) and Rater 1 was the most lenient (observed average = 3.08, logit measure -0.54). Rater 2 was the only misfit rater with an infit MSE value of 2.42.

Column 4 shows the calibration of items, with numbers representing the item number. The most difficult item was Item 62 (multiple stylistic elements utilized to create character, observed average = 2,14, logit measure = 1.11). The easiest item was Item 5 (instrument angle, observed average = 3.68, logit measure = -1.24). Items demonstrating underfit included 4, 6, 17, 28, 46, and 63. There were no overfitting items. Table 4.28 shows the complete calibration and statistics for items.

Table 4.27

Calibration of Tuba Performance Facet

Performance Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
8	3.43	1.58	0.15	1.09	0.70	1.12	0.80
1	3.29	1.09	0.12	1.80	5.90	1.65	4.70
7	3.21	1.00	0.14	1.36	2.50	1.27	1.90
6	3.39	0.92	0.12	1.09	0.70	1.00	0.00
5	3.22	0.49	0.11	0.69	-3.30	0.68	-3.20
2	2.56	-0.47	0.10	1.42	3.80	1.63	5.30
4	2.73	-0.50	0.10	0.61	-4.50	0.64	-4.40
3	2.64	-0.66	0.10	0.69	-3.60	0.65	-4.20
Mean	3.06	0.43	0.12	1.09	0.30	1.08	0.10
SD	0.33	0.81	0.02	0.39	3.60	0.39	3.60

Presented in measure order from highest to lowest achievement.

Table 4.28

Calibration of Tuba Item Facet

Item Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
62	2.14	1.11	0.34	0.83	-0.40	0.79	-0.50
61	2.14	1.11	0.34	0.81	-0.40	0.77	-0.60
32	2.77	1.06	0.31	1.00	0.00	1.11	0.40
31	2.82	0.98	0.31	1.20	0.80	1.23	0.80
35	2.91	0.83	0.35	0.70	-1.10	0.71	-1.10
40	2.36	0.77	0.24	1.27	1.00	1.25	0.80
23	2.95	0.72	0.37	0.73	-1.00	0.73	-1.00
20	2.95	0.72	0.37	1.16	0.60	1.26	0.90
15	3.50	0.59	0.46	1.03	0.20	1.12	0.70
25	3.00	0.59	0.33	0.84	-0.60	0.84	-0.50
22	3.00	0.59	0.33	0.86	-0.50	0.86	-0.50
63	2.45	0.58	0.22	2.72	4.00	4.42	5.20
41	2.64	0.51	0.28	0.93	-0.10	1.07	0.30
12	3.05	0.47	0.34	1.13	0.50	1.23	0.90
56	2.73	0.46	0.32	0.68	-0.90	0.69	-0.80
60	2.50	0.45	0.29	0.75	-0.80	0.73	-0.90
28	2.64	0.45	0.26	1.63	2.00	1.70	2.10
24	3.09	0.37	0.33	0.82	-0.70	0.80	-0.70
58	2.68	0.35	0.30	0.65	-1.30	0.65	-1.20
33	2.73	0.33	0.30	0.79	-0.60	0.81	-0.50
59	2.55	0.32	0.33	0.76	-0.80	0.75	-0.80
39	2.77	0.28	0.27	0.96	0.00	0.92	-0.10
27	3.14	0.23	0.34	0.79	-0.80	0.78	-0.80
4	2.86	0.22	0.25	1.96	2.90	2.40	3.10
53	2.18	0.14	0.34	0.95	-0.10	0.93	-0.10
34	2.82	0.12	0.28	0.81	-0.60	0.80	-0.60
48	2.95	0.12	0.32	0.69	-0.80	0.69	-0.70
29	3.18	0.06	0.36	0.97	0.00	0.97	0.00
26	3.18	0.06	0.36	0.79	-0.70	0.78	-0.80
55	3.23	0.06	0.33	1.22	0.80	1.25	0.90
42	2.86	0.04	0.28	0.86	-0.40	0.88	-0.30

44	2.77	-0.02	0.31	0.78	-0.70	0.80	-0.60
45	3.00	-0.03	0.29	0.87	-0.30	0.94	0.00
57	2.82	-0.04	0.31	0.69	-1.10	0.70	-1.00
49	2.95	-0.05	0.27	1.06	0.30	1.14	0.50
36	3.09	-0.16	0.28	0.70	-0.90	0.70	-0.90
52	3.23	-0.17	0.39	0.77	-0.80	0.77	-0.80
11	3.14	-0.19	0.29	0.82	-0.40	0.88	-0.20
13	2.91	-0.21	0.29	0.65	-1.40	0.64	-1.40
37	3.36	-0.26	0.34	0.72	-1.00	0.69	-1.00
3	3.36	-0.26	0.34	1.40	1.40	1.44	1.30
2	3.45	-0.28	0.32	1.11	0.40	1.03	0.20
18	3.68	-0.29	0.49	1.19	0.90	1.28	1.00
43	3.05	-0.29	0.32	0.80	-0.50	0.87	-0.30
47	3.09	-0.32	0.33	0.68	-0.90	0.68	-0.90
38	3.27	-0.35	0.28	0.70	-0.80	0.68	-0.70
10	3.05	-0.37	0.29	1.01	0.10	0.92	-0.10
54	3.09	-0.40	0.29	0.60	-1.50	0.60	-1.50
21	3.14	-0.42	0.30	0.73	-0.80	0.72	-0.90
14	3.41	-0.49	0.36	1.25	0.90	1.40	1.30
51	3.27	-0.54	0.43	0.86	-0.40	0.87	-0.30
17	3.27	-0.59	0.29	1.42	1.20	1.57	1.40
50	3.32	-0.64	0.42	0.80	-0.70	0.80	-0.70
30	3.32	-0.64	0.42	1.12	0.50	1.12	0.50
6	3.32	-0.64	0.29	1.96	2.40	2.99	3.50
1	3.32	-0.64	0.29	1.30	0.90	1.12	0.40
46	3.36	-0.64	0.32	1.56	1.30	1.48	1.20
9	3.55	-0.69	0.36	1.35	1.10	1.34	0.80
8	3.45	-0.87	0.40	1.05	0.20	1.07	0.30
19	3.50	-0.94	0.40	1.13	0.50	1.10	0.40
16	3.55	-1.00	0.40	1.23	0.80	1.18	0.60
7	3.55	-1.00	0.40	1.01	0.10	0.96	0.00
5	3.68	-1.24	0.43	1.09	0.30	0.95	0.00
Mean	3.03	0.00	0.33	1.01	0.00	1.07	0.10
SD	0.36	0.56	0.05	0.37	1.10	0.58	1.20

Presented in measure order from highest to lowest difficulty.

Response Category Diagnostics

Table 4.29 shows response category data for the TUPRS including the usage of each response category, average observed and average expected logit measures, and the outfit mean squared error (MSE) value for each item in the rating scale. This data can be used to optimize response categories by eliminating underused categories, combining categories with too small of a threshold between them, or eliminating categories with high outfit MSE values. Based on response category diagnostic data, many of the items in the TUPRS would be revised. Fifty-nine out of 63 items would require collapsing adjacent categories due to insufficient usage. Of the remaining four items, only items 39 () and would not require any revision based on response category diagnostics.

Table 4.29. TUPRS Category Diagnostics for Fit Items: Category Usage, Average Observed and Expected Logit Measure, Outfit Mean Squared Error (*MSE*).

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
1	1(5)	3(14)	6(27)	12(55)	0.35(0.19)	0.67(0.58)	1.22(1.04)	1.43(1.55)	0.90	0.70	1.30	1.60
2		4(18)	4(18)	14(64)		0.16(0.14)	0.82(0.62)	1.08(1.14)		0.60	1.40	1.50
3		3(14)	8(36)	11(50)		0.24(0.14)	1.09(0.65)	0.82(1.18)		0.90	2.10	1.60
4	4(18)	2(9)	9(41)	7(32)	0.13(-0.45)	0.42(-0.06)	0.52(0.41)	0.25(0.86)	2.60	2.10	1.00	3.00
5	1(5)		4(18)	17(77)	0.96(0.90)		1.43(1.40)	1.97(1.98)	0.80		1.00	1.10
6	1(5)	3(14)	6(27)	12(55)	1.69(0.19)	1.30(0.58)	1.23(1.04)	1.15(1.55)	4.90	4.10	1.50	1.80
7		1(5)	8(36)	13(59)		1.35(0.75)	1.11(1.27)	1.89(1.84)		1.40	0.80	0.90
8		1(5)	10(45)	11(50)		1.22(0.66)	1.14(1.19)	1.75(1.75)		1.30	1.10	1.00
9		2(9)	6(27)	14(64)		1.15(0.47)	0.93(0.96)	1.43(1.51)		1.80	0.80	1.40
10	1(5)	5(23)	8(36)	8(36)	0.72(0.01)	0.03(0.43)	1.16(0.94)	1.35(1.41)	1.60	0.30	0.80	1.20
11	2(9)	1(5)	11(50)	8(36)	-0.24(-0.20)	0.09(0.21)	0.56(0.69)	1.39(1.18)	1.20	0.50	0.70	0.80
12		5(23)	11(50)	6(27)		-0.60(-0.44)	0.38(0.10)	0.20(0.59)		0.70	0.90	1.80
13	1(5)	6(27)	9(41)	6(27)	-1.00(-0.11)	0.17(0.33)	0.96(0.85)	1.47(1.31)	0.50	0.50	0.80	0.80
14		2(9)	9(41)	11(50)		1.40(0.33)	0.63(0.85)	1.37(1.39)		2.40	0.60	1.20
15			11(50)	11(50)			-0.25(-0.29)	0.23(0.28)			1.00	1.20
16		1(5)	8(36)	13(59)		0.72(0.75)	1.54(1.27)	1.67(1.84)		0.80	1.30	1.30
17	1(5)	3(14)	7(32)	11(50)	0.94(0.15)	0.88(0.55)	0.91(1.02)	1.43(1.53)	2.00	2.10	0.80	1.40
18			7(32)	15(68)			0.75(0.47)	0.92(1.05)			1.30	1.20
19		1(5)	9(41)	12(55)		0.84(0.71)	1.34(1.24)	1.71(1.80)		1.00	1.10	1.20
20		5(23)	13(59)	4(18)		-0.60(-0.66)	0.00(-0.10)	-0.01(0.39)		1.00	0.60	1.70
21	1(5)	3(14)	10(45)	8(36)	-0.79(0.01)	0.23(0.43)	1.07(0.93)	1.44(1.43)	0.30	0.50	1.00	1.00
22		6(27)	10(45)	6(27)		-0.63(-0.53)	-0.02(0.01)	0.65(0.49)		1.00	0.70	0.80
23		5(23)	13(59)	4(18)		-1.06(-0.66)	-0.01(-0.10)	0.61(0.39)		0.70	0.70	0.80
24		5(23)	10(45)	7(32)		-0.72(-0.36)	0.34(0.18)	0.69(0.67)		0.60	0.80	1.00
25		6(27)	10(45)	6(27)		-0.60(-0.53)	-0.05(0.01)	0.67(0.49)		0.70	1.40	0.80
26		3(14)	12(55)	7(32)		-0.60(-0.12)	0.48(0.43)	1.07(0.95)		0.60	0.90	0.90

(continued)

Table 4.29. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
27		4(18)	11(50)	7(32)		-0.63(-0.25)	0.36(0.29)	0.89(0.79)		0.60	0.90	0.90
28	4(18)	5(23)	8(36)	5(23)	0.09(-0.60)	-0.29(-0.17)	0.23(0.33)	0.45(0.72)	2.50	1.10	1.10	1.50
29		3(14)	12(55)	7(32)		-0.10(-0.12)	0.39(0.43)	1.01(0.95)		1.10	0.90	0.90
30		1(5)	13(59)	8(36)		0.51(0.48)	1.14(1.04)	1.42(1.59)		1.00	1.10	1.20
31		9(41)	8(36)	5(23)		-0.50(-0.82)	-0.71(-0.28)	0.29(0.17)		1.40	2.10	0.80
32		10(45)	7(32)	5(23)		-0.86(-0.87)	-0.31(-0.33)	0.07(0.10)		1.00	1.70	1.00
33	2(9)	5(23)	12(55)	3(14)	-0.76(-0.59)	-0.21(-0.13)	0.36(0.40)	1.28(0.83)	0.90	0.70	0.90	0.70
34	2(9)	5(23)	10(45)	5(23)	-0.55(-0.39)	-0.07(0.05)	0.55(0.56)	1.19(1.00)	0.90	0.60	0.90	0.80
35		6(27)	12(55)	4(18)		-1.15(-0.75)	-0.05(-0.19)	0.49(0.28)		0.60	0.60	0.80
36	2(9)	2(9)	10(45)	8(36)	-0.49(-0.19)	0.26(0.21)	0.54(0.69)	1.43(1.18)	0.60	0.80	0.60	0.80
37		3(14)	8(36)	11(50)		-0.23(0.14)	0.58(0.65)	1.32(1.18)		0.60	0.80	0.80
38	2(9)	1(5)	8(36)	11(50)	-0.30(-0.06)	0.20(0.32)	0.67(0.78)	1.42(1.28)	0.60	0.40	0.70	0.80
39	3(14)	4(18)	10(45)	5(23)	-0.56(-0.51)	-0.08(-0.08)	0.40(0.42)	0.93(0.85)	1.00	0.80	1.00	0.90
40	7(32)	5(23)	5(23)	5(23)	-0.70(-0.78)	0.02(-0.33)	-0.44(0.14)	0.58(0.47)	1.00	1.20	2.70	0.80
41	3(14)	5(23)	11(50)	3(14)	-0.31(-0.71)	-0.64(-0.25)	0.22(0.26)	1.09(0.67)	2.00	0.10	0.80	0.70
42	2(9)	5(23)	9(41)	6(27)	-0.70(-0.32)	0.27(0.11)	0.54(0.62)	1.17(1.06)	0.60	1.20	0.90	0.90
43	1(5)	3(14)	12(55)	6(27)	-0.92(-0.11)	0.65(0.33)	0.76(0.85)	1.51(1.34)	0.30	1.50	0.70	0.90
44	1(5)	7(32)	10(45)	4(18)	-1.20(-0.27)	0.33(0.20)	0.59(0.73)	1.54(1.17)	0.50	1.10	1.00	0.70
45	2(9)	2(9)	12(55)	6(27)	-0.34(-0.31)	0.00(0.11)	0.55(0.61)	1.25(1.09)	1.20	0.50	0.90	0.90
46	1(5)	1(5)	9(41)	11(50)	1.12(0.13)	0.49(0.54)	1.13(1.03)	1.38(1.55)	2.60	0.60	1.10	1.40
47	1(5)	2(9)	13(59)	6(27)	-0.90(-0.12)	0.19(0.32)	0.85(0.85)	1.53(1.35)	0.30	0.60	0.80	0.90
48	2(9)	1(5)	15(68)	4(18)	-0.77(-0.47)	-0.27(-0.04)	0.42(0.96)	1.41(0.96)	0.60	0.40	0.70	0.80
49	2(9)	4(18)	9(41)	7(32)	-0.63(-0.25)	0.91(0.16)	0.32(0.66)	1.24(1.12)	0.60	2.60	0.60	0.90
50		1(5)	13(59)	8(36)		0.54(0.48)	0.87(1.04)	1.86(1.59)		1.00	0.70	0.80
51		1(5)	14(64)	7(32)		0.45(0.41)	0.87(0.98)	1.73(1.52)		1.00	0.80	0.80
52		2(9)	13(59)	7(32)		-0.49(0.07)	0.61(0.63)	1.36(1.16)		0.70	0.70	0.80
53	4(18)	10(45)	8(36)		0.02(-0.18)	0.11(0.35)	1.06(0.86)		1.20	0.50	0.80	

(continued)

Table 4.29. (continued)

	Category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit SE			
Item	1	2	3	4	1	2	3	4	1	2	3	4
54	1(5)	4(18)	9(41)	8(36)	-0.82(0.01)	0.28(0.43)	0.90(0.93)	1.64(1.42)	0.40	0.50	0.70	0.70
55		4(18)	9(41)	9(41)		0.08(-0.11)	0.47(0.41)	0.77(0.92)		1.10	0.80	1.60
56	2(9)	4(18)	14(64)	2(9)	-0.89(-0.73)	-0.67(-0.27)	0.32(0.27)	1.32(0.71)	1.00	0.20	0.60	0.80
57	1(5)	6(27)	11(50)	4(18)	-1.17(-0.27)	0.19(0.20)	0.67(0.73)	1.57(1.18)	0.40	0.80	0.90	0.70
58	2(9)	6(27)	11(50)	3(14)	-0.78(-0.59)	-0.44(-0.12)	0.49(0.40)	1.25(0.83)	0.90	0.30	0.60	0.70
59	1(5)	10(45)	9(41)	2(9)	-1.53(-0.54)	-0.04(-0.03)	0.58(0.52)	1.23(0.91)	0.60	0.90	0.80	0.70
60	2(9)	10(45)	7(32)	3(14)	-0.88(-0.61)	-0.21(-0.11)	0.53(0.41)	1.02(0.79)	0.80	0.70	0.70	0.70
61	3(14)	14(64)	4(18)	1(5)	-1.24(-1.14)	-0.66(-0.59)	0.20(-0.08)	0.51(0.23)	1.00	0.80	0.60	0.70
62	3(14)	14(64)	4(18)	1(5)	-1.24(-1.14)	-0.65(-0.59)	0.16(-0.08)	0.51(0.23)	1.00	0.80	0.60	0.70
63	7(32)	6(27)	1(5)	8(36)	0.06(-0.60)	0.15(-0.17)	-0.68(0.29)	-0.08(0.62)	2.40	2.30	3.00	7.20

Data from the PC-MFR analyses of each of the rating scales can be used to determine how closely they meet the standards of invariant measurement, which specifically addresses research question 1. The reliability of separation of each facet and the MSE values support the validity of each of the scales examined (flute, clarinet, saxophone, trumpet, horn, trombone, and tuba). The oboe and euphonium scales could not be validated due to an insufficient number of participants.

Research questions 2 and 3 can be addressed by examining the variable maps for each of the rating scales. The elements of each facet could be reasonably separated along the latent construct, and they demonstrate adequate unidimensionality along the measure. All performers, raters, and items were objectively ordered.

Chapter 5

Discussion and Conclusion

The purpose of this study was to develop and validate a set of instrument-specific performance rating scales for classroom use at the secondary level. Once all scoring data was compiled a MFR-PC analysis was conducted for each rating scale. The MFR-PC was chosen to account for the inherent challenges of rater mediated music performance assessment (e.g., measuring a latent trait, rater bias) as well as to determine how well the data meet the requirements of invariant measurement. Data from the MFR-PC analysis was used to determine the validity and reliability of each scale. It also shows the achievement level of student performers, the level of severity of the raters, and the level of difficulty of the items on each scale. Finally, an analysis of response category usage across multiple raters offers insight into how to best optimize each scale for future classroom use.

Overall, the rating scales were shown to be reliable based on an examination of person, rater, and item response functions as well as probability, reliability of separation, and infit/outfit mean statistics. A comprehensive analysis of results is provided in chapter 4. Based on the research problem identified earlier in this study as well as the examination of related literature in the fields of educational measurement and music performance assessment in chapter 2, these results may be useful for music educators desiring to accurately assess individual student growth and achievement using research-based assessment methods. This chapter will include a summary of the findings, a discussion of the research questions, and offer suggestions for future research.

Research Questions

Research Question 1. What does Rasch measurement analysis reveal about the psychometric properties (i.e., validity and reliability) of items, raters, and performers in the context of solo music performance assessment? Rasch is an ideal-type model, which means the model perfectly meets the requirements of invariant measurement (Bond & Fox, 2020). Comparing item, rater, and performer data from this study to the ideal Rasch model reveals the degree to which these rating scales could be useful to accurately assess other performances regardless of the students performing, the raters judging, or the items being used for scoring. Five conditions must be met in order to achieve invariance: (a) item-invariant measurement of persons (i.e., the measurement of persons must be independent of the particular items that happen to be used for the measuring), (b) non-crossing person response functions (i.e., a more able person must always have a better chance of success on an item than a less able person), (c) person-invariant calibration of test items (i.e., the calibration of the items must be independent of the particular persons used for calibration), (d) non-crossing item response functions (i.e., any person must have a better chance of success on an easy item than on a more difficult item), and (e) variable map (i.e., items and person must be simultaneously located on a single underlying latent variable) (Engelhard, 2013).

The Multifaceted Rasch (MFR) model allows for separate parameterization of each facet measured (performers, raters, and items) and displays locations of each along a unidimensional measure of a single latent trait for comparison (Linacre, 2009). The Partial Credit Rasch (PCR) analysis offers insight into response category usage of raters including thresholds showing the distance perceived between available response categories. The PCR analysis also reveals instances in which fewer response categories could be useful for particular items.

An analysis of fit statistics of the rating scales revealed several trends. Fit statistics indicate the degree to which the data fit the ideal Rasch model. Inlier sensitive fit statistics (infit) show the degree of variance from the model for persons whose ability is relatively close to the difficulty of a given item. Outlier sensitive fit statistics (outfit) show the degree of variance for persons whose ability is further from an item, and is therefore more sensitive to unexpected results such as correct guessing on a difficult item by a less able person. Both infit and outfit statistics are reported as a squared value or mean square error (MSE). An infit or outfit MSE value is a measure of the sum of all variances from the model.

Infit MSE scores close to 1.00 indicate good data to model fit and support the validity of the rating scale. In the case of the flute, clarinet, saxophone, trumpet, horn, trombone, and tuba rating scales, infit MSE values all indicated very good data to model fit for persons, items, and raters. Infit MSE values ranged from 0.97 to 1.09 for all scales except the tuba scale in which values ranged from 1.01 (items) to 1.27 (raters). An infit MSE value of 1.27 in this case indicates there was 27% more variation in rater behavior than would be expected by the Rasch model. A review of rater calibration data shows tuba rater 2 with an infit MSE of 2.42, which is by far the largest value demonstrated by any of the raters from any of the panels. Rater 2 was also the most severe with a logit measure of 0.74. Rater 2 is a percussionist with 18 years of classroom experience teaching middle and high school band. Rater 2 also scored performances of other instruments but had much lower MSE values and did not display the same severity of scoring behavior.

Outfit MSE scores were also close to 1.00 for all facets throughout all scales, with the exception of the flute performance rating scale (FPRS). Outfit measures for the FPRS were higher than all other scales in each facet with values of 1.27 for performers, 1.48 for raters, and

1.25 for items. Flute performer 27 showed an outfit MSE value of 5.18 with an observed score of 3.99 and a measure of 6.52 logits. This would indicate that performer 27 is an advanced player and it is possible there were not enough difficult items on the rating scale to accurately measure her ability level. Including more items at a higher level of difficulty, such as items related to expressive choices by the performer, could reduce the likelihood of future performers showing such a high outfit value.

Flute rater 9 showed an outfit MSE of 4.34, which was the largest for any rater on any panel. This rater ranked fourth out of nine in level of severity with a logit measure of 0.40. This rater is a clarinet player with 22 years of classroom experience teaching middle school band. All other flute raters displayed acceptable fit statistics.

There were four items on the FPRS with particularly high outfit MSE values. These included item 3, Placement of Feet (8.26); item 4, Arm Positioning (9.00); item 15, Embouchure Contact with Mouthplate during Inhalation (3.45); and item 61, Tempo Modifications Observed (7.33). In the case of item 61, some of the music examples performed did not include any tempo modifications. Responses for this item could be edited to include an option of "not applicable" or the item could be removed altogether. In the case of items 3, 4, and 15, it is possible that the raters held very different opinions of what constitutes acceptable posture. It is beyond the scope of this study to determine the reasons for this discrepancy.

The reliability of the rating scales can be described in terms of the reliability of separation of the elements within each facet. Reliability of separation indicates the likelihood that the same results would be produced if the study were to be replicated. For instance, a high degree of reliability of separation for persons would suggest that the same ordering and level of separation would result even if those persons were to be assessed using a different set of items or scored by a different set of raters.

The reliability of separation for persons, raters, and items was very high across all rating scales with the exception of the item separation index for both the horn rating scale (0.77) and the tuba rating scale (0.65). Each of these scales had a smaller sample size of performers, which could explain the lower degree of reliability. A high reliability of separation also supports the construct validity of the rating scales.

Research Question 2. How do items vary in difficulty, raters in severity, and performers in achievement? The Rasch data analysis translates observed measures into logit scores to allow for objective comparisons across each facet. It also provides a variable map (i.e., Wright Map) which shows the location of all elements within each facet along the same unidimensional measure of music performance ability. This allows for the fair and objective ordering of elements regardless of who is performing, which items are being scored, or who is rating the performance.

Chapter 4 includes item and person calibration tables for each rating scale as well as variable maps for each scale. Rater calibration tables can be found in Appendix (A?). With the exception of specific items, persons, and raters which fall outside the unidimensional measure as noted in chapter 4, all rating scales were shown to reasonably separate and order the elements within each facet. It can be expected that each of the scales could be used in a classroom setting to accurately measure student performance ability along the intended latent trait.

Item calibration data can also be used to identify a hierarchy of difficulty based on item type. An examination on the item calibration tables shows a clear difference in difficulty between items related to expressive elements of playing in comparison to items related to posture and hand position.

Research Question 3. How does the rating scale structure vary across raters and performers? Rasch analysis provides category diagnostics for each individual item on each rating scale. Specifically, the Partial Credit formulation of the Rasch model allows the distance between response category thresholds to vary between each item and rater. The advantage of using the PC analysis is the opportunity for optimization of response categories for each individual item on a given rating scale. The PC-R analysis describes category usage by frequency and percentage used, average observed logit measures and the average expected logit measure, and outfit mean squares (MSE). Rash-Andrich thresholds show the perceived distance between response categories 1 and 2, 2 and 3, and 3 and 4. An examination of threshold data provides insight into the difficulty of achieving a given score that is unique to each individual item.

Optimization of item response categories involves eliminating categories that are underutilized or show high MSE scores, as well as combining or collapsing adjacent categories that do not have a significant statistical difference between them. Optimization improves the accuracy of scoring and the usefulness of the measurement data (Linacre, 2022). Based on the results discussed in chapter 4, most of the items in each rating scale would be revised. Linacre (2002) recommends eliminating categories with fewer than 10 uses. Due to the smaller sample sizes of some of the rating scales, a percentage of less than 10% was determined as a cut-off point for this study. By this standard most of the items in the FPRS, for example, would have response category 1 eliminated.

Linacre's second recommendation is to eliminate categories with an outfit MSE greater than or equal to 2.00. Using the FPRS again as an example, this applies to several of the response categories, most of which will already be eliminated based on insufficient usage. All four response categories for Item 3, “Placement of Feet,” show a value greater than 2.00, but this item also would be eliminated based on item fit statistics.

Linacre’s third recommendation is to evaluate categories for proper step ordering. In other words, the level of difficulty of achieving each category should increase across responses 1 through 4. As an example, category 2 on FPRS item 10 shows a logit measure of 0.12 which is less than category 1 with a measure of 0.26. Therefore, categories 1 and 2 would be collapsed into one response category.

Limitations

There were several limitations to this study due to the nature of action research. Sample sizes of oboe and euphonium players were insufficient for scale validation. A lack of equipment and/or recording space prohibited the inclusion of bassoon and percussion players. The French horn and tuba scales demonstrated slightly lower degrees of validity than the other scales included in this study, which is possibly a result of small sample sizes. It is possible that further study involving more participants could show increased validity and reliability for these scales. The differential item function analysis for the clarinet and saxophone scales was not valid due to an insufficient number of bass clarinet, tenor saxophone, and baritone saxophone players. Attempts to understand differential item functioning within the clarinet and saxophone scales could be improved with future replications of the study involving more participants.

Discussion

Public school music programs in the United States are largely lacking in research-based classroom assessment tools (Asmus, 1999; Goolsby, 1999; Kelley et al., 2019; Kimpton, 2019; Lehman, 1998; Pellegrino et al., 2015; Tindal & Marston, 1990). While advances in educational measurement have led to the development and validation of more robust assessment tools in high-stakes settings, these same measures have not been widely applied in the field of music performance assessment (Wesolowski et al., 2016). The unique challenges inherent to rater mediated assessment require the researcher to account for factors such as latent variables and rater bias. Utilizing data analysis tools such as the Rasch method provides a way to address these issues in the process of validating assessment measures (Bond & Fox, 2020).

Based on MFR-PC analysis data, the rating scales developed in this study show a high degree of validity and reliability. Response category diagnostics discussed in chapter 4 provide clear steps towards optimization of individual items within each rating scale to be included in future revisions. The result will be a set of instrument-specific research-based rating scales for classroom use at the secondary level. The implementation of these tools in formative and summative performance assessments would allow instrumental music teachers to measure student ability and growth in a way that is fair and objective and clearly communicates level of achievement to student, parents, administrators, and other stakeholders. This addresses a specific problem in the field of music education today. The systems of assessment employed by secondary level music educators in the United States have largely failed to accurately measure the achievement level of the individual student. Burdened by performance schedules and lacking classroom friendly assessment tools, instrumental music teachers find themselves stuck on the "performance treadmill" of continuously introducing and rehearsing ensemble literature while allowing individual assessment to go overlooked (Kimpton, 2019). In many cases, the result is a classroom grading system built on non-achievement related factors such as participation and concert attendance (Goolsby, 1999). The primary goal of this study was to provide a solution to this problem through the development of a research-based, classroom friendly performance assessment tool.

These rating scales are designed to be easily implemented in any secondary level band classroom setting. Considering the performance driven nature of most school band programs, a rating scale focusing on individual performance skills assessment aligns well with the goals and objectives already in place in the classroom. Performance assessments are valued for their authenticity in that students have the opportunity to demonstrate mastery over a concept (Lane, 2013). With the introduction of research-based performance assessment tools to match existing learning objectives, music educators are able to leverage the inherently authentic nature of their classroom for the benefit of individual student growth.

This type of authentic assessment can be beneficial from a standpoint of program accountability and teacher effectiveness measures. School band programs and band directors are often informally assessed by the quality of public performances or the number of students enrolled in the program (Asmus, 1999). Implementing a more robust system of student assessment and communicating student growth data to stakeholders is a necessary step toward changing the criteria used by the public to make value judgments about music programs. Curricular goals and objectives are often communicated through assessment. Without systematic performance assessment, parents and administrators may view the school music program as primarily a form of entertainment or a social organization for students.

Additionally, teacher effectiveness measures throughout the United States rely in part on student growth data (Shuler, 2012). Music educators historically have not been able to produce this type of data from their own classrooms, resulting in the inclusion of core subject testing data on music teacher evaluations. This practice lacks validity and negatively affects teacher self-efficacy and morale. The implementation of systematic individual performance assessment using research-based tools can provide music educators with relevant growth data to offer administrators when it comes to staff performance evaluations (Davidson & Fisher, 2019). Having been validated using the MFR-PC analysis, the assessment tools developed in this study accurately measure student achievement level and can be used to demonstrate individual growth over time.

Beyond the usefulness of the individual rating scales as assessment tools, the item calibrations produced by this study describe a hierarchy of difficulty based on item type. This ordering of difficulty confirms what many music educators know to be true both as performers and as teachers: The most basic and fundamental concepts such as posture and hand position must precede more advanced concepts related to expressive playing. Research data such as the item calibrations in this study can inform curriculum development in the instrumental classroom. This would impact all aspects of teaching from the ordering and pacing of concepts to assessment construction to literature selection.

The potential benefit of instrument-specific rating scales also extends beyond the classroom to region and state level solo music performance assessments (Keene, 2009). The introduction of instrument-specific scoring measures in all-state band auditions or solo and ensemble festivals would likely increase the value of student feedback and increase the accuracy and objectivity of student rankings.

Solo and ensemble festivals provide rich benefits for the individual student musician including increased self-efficacy, peer collaboration, opportunities for feedback from a rater other than their school band director, and the chance to perform chamber music or solo literature (Asmus, 1999; Keene, 2009). Based on the score sheets gathered for this study, solo and ensemble festivals throughout the United States currently do not utilize research-based assessment methods, but instead offer only the most broad categories of feedback. Using the same score sheet for all instrument types contributes to a lack of fairness in ratings due to the effects of differential item functioning, meaning students of similar ability level who play different instruments would not receive similar scores (Wesolowski, 2019). Each instrument included in these festivals presents a unique set of physical demands and technical challenges, and these factors impact the equity of scores and rankings. Unless these differences are accounted for and addressed through instrument specific scoring measures, it is expected that students will receive ratings and feedback that do not accurately and fairly assess their individual level of performance or provide constructive criticism essential to their development. The instrument specific rating scales developed in this study provide a research-based alternative to current scoring methods and have the potential to improve accuracy and equity of scoring. Not only do they inherently address technical issues unique to each instrument, but the MFR-PC data analysis demonstrates a high degree of validity and reliability in order to account for factors common in solo and ensemble events such as rater bias.

Implications

The results of this study offer another step in the development of valid and reliable performance assessment tools for classroom use. Considering the favorable statistics for these scales, in particular the scales with larger sample sizes such as clarinet and trumpet, future replications with larger numbers of performers and raters could offer further evidence of validity and reliability. It is expected that similar results could be achieved with a facet-factorial approach to scale construction and MFR-PC validation for instrument groups which were not included in this study. Replications with larger sample sizes for clarinet and saxophone could also answer questions of differential item functioning for other varieties of these instruments such as bass clarinet or tenor saxophone. An examination of DIF could reveal whether unique rating scales should be developed for these instruments.

There are also potential implications for these scales to be used in middle school and high school feeder programs. Possible topics to investigate include (a) the effect of a common rating scale system on vertical curricular alignment from middle school through high school; (b) long-term achievement outcomes for individual students; (c) the impact on ensemble achievement; and (d) applications of long term-student growth data. For programs offering ability-based ensembles, the implementation of a research-based scoring method can increase equity of ensemble placement for students. Band directors could more effectively defend placement decisions when questions of fairness arise from students or parents, and students may have a greater sense of self-efficacy regarding outcomes if they have greater assurance of accurate and objective rankings.

In today's data driven climate, music educators generally have not faced the same level of accountability for documenting student achievement as their colleagues who teach tested subjects, and therefore have not been motivated to search for or create research-based assessment tools. While a need for such tools has been recognized in the field, much of the research has been disconnected from the public school classroom. The rating scales developed in this study may offer another step towards connecting assessment research to classroom practice.

REFERENCES

Abeles, H. (1971). *An application of the facet-factorial approach to scale construction in the development of a rating scale for clarinet music performance* (Unpublished doctoral dissertation). University of Maryland.

Abeles, H. (1981). Reviewed work: The construction and validation of a scale of trombone performance skills by Robert Lee Kidd III. *Bulletin of the Council for Research in Music Education*(65) 80-83.

Abeles, H. & Custodero, L. (2010). *Critical issues in music education*. Oxford University Press.

Andrade, H. (2013). Classroom assessment in the context of learning theory and research. McMillan, J. Editor, *SAGE Handbook of Research on Classroom Assessment* (17-31) SAGE Publications, Inc.

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

Bergee, M. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education 15*(2) 137-150.

Bergee, M. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education 55*(4), 344-358.

Bergee, M. (1995). Primary and higher-order factors in a scale assessing concert band performance. *Bulletin of the Council for Research in Music Education* (126) 1-14.

Bond, T., & Fox, C. (2020) *Applying the Rasch model: Fundamental measurement in the human sciences.* Lawrence Erlbaum Associates.

Bonner, S. (2013). Validity in classroom assessment: Purposes, properties, and principles. McMillan, J. Editor, *SAGE Handbook of research on classroom assessment.* (87-104) SAGE Publications, Inc.

Boyle, J. (1992). Program evaluation for secondary school music programs. *NASSAP Bulletin,* 76(544), 63-68.

Boyle, J., & Radocy, R. (1987). *Measurement and evaluation of musical experiences.* Schirmer books.

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. (1997). Effects of the classroom assessment environment on mathematics and science achievement. *The Journal of Educational Research 90*(6) 323-330.

Brookhart, S. (2013). Classroom assessment in the context of motivation theory and research. McMillan, J. Editor, *SAGE Handbook of research on classroom assessment*. (35-54) SAGE Publications, Inc.

Brophy, T. S. (1997). An examination of the melodic improvisations of 7 through 11 year old children. In R. Cutietta (Chair). *Symposium on Research in General Music*, University of Arizona, Tucson.

Brown University. (2000). *Action research* [Pamphlet]. Providence, RI. Northeast and islands regional educational laboratory.

Bruning, R. H., Schraw, G. J., Norby, M. M., & Ronning, R. R. (2004). *Cognitive psychology and instruction*, 4th ed. Pearson Prentice Hall, New Jersey.

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

Burrack, F. (2002). Enhanced assessment in instrumental programs. *Music Educators Journal, 88*(6), 27-32.

Burrack, F., & Parkes, K. A. (2019). The development of standards-based assessments in music. Brophy, T. Editor, *The Oxford handbook of assessment policy and practice in music education* (Vol. 1). (651-670) Oxford university press.

Clay, M. M. (1985). *The early detection of reading difficulties*. Auckland, New Zealand: Heinemann.

Cook, W. W. (1951). The functions of measurement in the facilitation of learning. E. F. Lindquist, Ed. *Educational Measurement* (3-46). Washington, D.C.: American Council on Education.

Cohen, D. K. (1995). What is the system in systemic reform? *Educational Researcher, 24*(9), 11-17.

Cooksey, J. M. (1977). A facet-factorial approach to rating high school choral music performance. *Journal of Research in Music Education 25*(6), 100-114.

Cooper, L. C. (2004). *Teaching band and orchestra: Methods and materials*. GIA Publications Inc., Chicago.

Davison, D., & Fisher, R. (2019). Music teacher evaluation and student growth in music. Brophy, T. Editor, *The Oxford handbook of assessment policy and practice in music education* (Vol. 1). (851-870) Oxford university press.

DeCamp, C. B. (1980). *An application of the facet-factorial approach to scale construction in the development of a rating scale for high school band performance*. Dissertation Abstracts International, 41, 1462A.

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

Dressman, M. B. (1990). *The development and validation of a test to evaluate selected wind instrument performance competencies of middle school/junior high school instrumentalists.* [Doctoral dissertation, University of Miami]. ProQuest Dissertations Publishing.

Edwards, K. (2017). *The psychometric development and review of an evaluation system for string ensemble performance using Rasch measurement theory* [Doctoral dissertation, University of Georgia]. University of Georgia Theses and Dissertations.

Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement* (1)1, 19-37.

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

Geringer, J., & Johnson, C. (2007). Effects of excerpt duration, tempo, and performance level on musicians' ratings of wind band performances. *Journal of Research in Music Education* (55)4, 289-301.

Goodman, Y. M. (1985). Kidwatching: Observing children in the classroom. A. Jaggar & M. T. Smith-Burke, Eds. *Observing the Language Learner* (9-18). Newark, DE: International Reading Association and National Council of Teachers of English.

Goolsby, T. (1999). Assessment in instrumental music. *Music Educators Journal 86*(2) 31-50.

Green, S., & Hale, C. (2011). Fostering a lifelong love of music: Instruction and assessment practices that make a difference. *Music Educators Journal 98*(1), 45-50.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan, 80*, 662-666.

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2008). *Standards-based reform in the United States: History, research, and future directions.* Center on Education Policy, Washington, D.C. https://eric.ed.gov/?id=ED503897

Hattie, J. (2009). *Visible learning - A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hogarth, R. M. (1980). *Judgment and choice: The psychology of decision*. Chichester, England. J. Wiley.

Hogarth, R. M. (1987). *Judgment and choice: The psychology of decision* (2nd ed.). John Wiley & Sons.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*(3), 177-182.

Hughes, G., Behuniak, P., Norton, S., Kitmitto, S., Buckley, J., American Institutes for Research (AIR), & National Center for Education Statistics (ED). (2019). NAEP Validity Studies Panel Responses to the Reanalysis of TUDA Mathematics Scores. *American Institutes for Research.*

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

Keene, J. (2009). *A history of music education in the United States.* Glenbridge Publishing Ltd.

Kelly, S. N., Cummings, B., & Gordon, M. G. (2019). The Florida performing fine arts assessment project. Timothy S. Brophy, editor. *The Oxford handbook of assessment policy and practice in music education*. (123-142). Oxford University Press.

Kimpton, P. H., & Kimpton, A. K. (2019). Making assessment meaningful, measurable, and manageable in the secondary music classroom. T. Brophy, Editor, *The Oxford handbook of assessment policy and practice in music education* (Vol. 2). (325-350) Oxford university press.

Koretz, D. M. & Hamilton, L. S. (2006). Testing for accountability in K-12. Robert L. Brennan, editor. *Educational measurement, 4th ed.* (531-578). American Council on Education.

Labuta, J. A., & Smith, D. A. (1997). *Music education: Historical contexts and perspectives*. Prentice-Hall, New Jersey.

Lane, S. (2013). Performance assessment. McMillan, J. Editor, *SAGE Handbook of research on classroom assessment.* (312-329) SAGE Publications, Inc.

Lane, S. & Stone, C. A. (2006). Performance assessment. Robert L. Brennan, editor. *Educational measurement, 4th ed*. (387-431). American Council on Education.

Laprise, R. (2017). Empowering the music educator through action research. *Music Educators Journal 104*(1), 28-33.

Lewis, A. A., Burgess, Y., & Fan, X. (2019). The South Carolina arts assessment program. T. Brophy, Editor, *The Oxford handbook of assessment policy and practice in music education* (Vol. 2). (281-306) Oxford university press.

Lehman, P. (1998). Grading practices in music. *Music Educators Journal 84*(5), 37-40.

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

Masters, G., & Keeves, J. (1999). *Advances in measurement in educational research and assessment.* Pergamon.

McCaffrey, M. & Lovins, L. T. (2019). The status of arts assessment in the United States. T. Brophy, Editor, *The Oxford handbook of assessment policy and practice in music education* (Vol. 2). (58-93) Oxford university press.

McFarland, K. P., & Stansell, J. C. (1993). Historical perspectives. In L. Patterson, C. M. Santa, C. G. Short, & K. Smith (Eds.), *Teachers are researchers: Reflection and action.* Newark, DE: International Reading Association.

Mark, M. (1996). *Contemporary music education*. Schirmer Books.

McMillan, J. (2013). Why we need research on classroom assessment. McMillan, J. Editor, *SAGE Handbook of research on classroom assessment.* (2-16) SAGE Publications, Inc.

McPherson, G. E., & Schubert, E. (2004). Measuring performance enhancement in music. Aaron Williamon, ed. *Musical excellence: Strategies and techniques to enhance performance*. (61-84).

Napoles, J. (2009). The effect of excerpt duration and music education emphasis on ratings of high quality children's choral performances. *Bulletin of the Council for Research in Music Education* (179), 21-32

Ooi, P. S. & Engelhard, G. (2019). Examining rater judgments in music performance assessment using many-facets Rasch rating scale measurement model. *Journal of Applied Measurement 20*(1), 79-99.

Ornstein, A. C., & Levine, D. U. (2000). *Foundations of education, 7th ed*. Houghton Mifflin, Boston.

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

Perrine, W. (2013). Music teacher assessment and race to the top: An initiative in Florida. *Music Educators Journal 100*(1), 39-44.

Radocy, J. D. (1992). Evaluation of music ability. Colwell, R., Ed. *Handbook of research on music teaching and learning.* Schirmer Books, New York.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology, pp. 321–334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV. Berkeley: University of Chicago Press, 1980

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science 1*(3), 181-210.

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

Russell, J. (2011). Assessment and case law: Implications for the grading practices of music educators. *Music Educators Journal 97*(3), 35-39.

Sadler, D. R. (2015). Backwards assessment explanations: Implications for teaching and assessment practices. Lebler, D., Carey, G. & Harrison, S. D., Editors, *Assessment in music education: From policy to practice.* Springer International Publishing, Switzerland.

Saunders, T. C., & Holahan, J. M. (1997). Criteria-specific rating scales in the evaluation of high school instrumental performance. *Journal of Research in Music Education 45*(2), 260-61.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shepard, L. A. (2006). Classroom assessment. Robert L. Brennan, editor. *Educational measurement, 4th ed*. (623-646). American Council on Education.

Shuler, S. C. (2011). Music assessment, part 1: What and why. *Music Educators Journal 98*(2), 10-13.

Shuler, S. C. (2012). Music assessment, part 2: Instructional improvement and teacher evaluation. *Music Educators Journal 98*(3), 7-10.

Shuler, S. C. (2016). Model cornerstone assessments: Clarifying standards, extending capacity, and supporting learning. In T. S. Brophy, J. Marlatt, & G. K. Ritcher (Eds.), *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education* (57-73). Chicago, IL: GIA.

Schneider, M. C., Egan, K. L., & Julian, M. W. (2013). Classroom assessment in the context of high stakes testing. McMillan, J. Editor, *SAGE Handbook of research on classroom assessment* (55-70). SAGE Publications, Inc.

Scott, S. (2002). The development and application of a performance-based measure to assess a creative endeavor by fourth and fifth grade music students. *Contributions to Music Education 29*(1), 29-46.

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice,* 10(1), 7-12.

Stites, R. & Malin, H. (2008). *An unfinished canvas: A review of large-scale assessment in K-12 arts education*. Center for Education Policy, SRI International.

Thompson, S. & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception, 21*, 21-41.

Tindal, G., & Marston, D. (1990). *Classroom-based assessment: Evaluating instructional outcomes.* Merrill Publishing Co.

Toney, B. (2017). *The objective identification of a hierarchy of difficulty of rhythm patterns in the context of high school band students* [Doctoral dissertation, University of Georgia]. University of Georgia Theses and Dissertations.

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.

Watts, H. (1985). When teachers are researchers, teaching improves. *Journal of Staff Development. 6*(2), 118-127.

Wells, R. G., & Shuler, S. C. (2019). Connecticut common music assessments: Collaboratively building capacity and exemplars. Brophy, T. Editor, *The Oxford handbook of assessment policy and practice in music education* (Vol. 2). (95-122) Oxford University Press.

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

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

Wesolowski, B. C. (2019). An examination of differential item functioning in a rubric to assess solo music performance. *Musicae Scientiae*, *2*(23), 1-15.

Wesolowski, B. C., & Wind, S. A. (2017). Investigating rater accuracy in the context of secondary-level solo instrumental music performance. *Musicae Scientiae*. Prepublished June, 13, 2017.

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), 39-47.

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

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2017). Evaluating differential rater functioning over time in the context of solo music performance assessment. *Bulletin of the Council for Research in Music Education*, *212*(Spring 2017), 75-98.

Wesolowski, B. C., Amend, R. M., Barnstead, T. S., Edwards, A. S., Everhart, M., Goins, Q. R., Grogan III, R. J., Herceg, A. M., Jenkins, S. I., Johns, P. M., McCarver, C. J., Schaps, R. E., Sorrell, G. W., & Williams, J. D. (2017). The development of a secondary-level solo wind instrument performance rubric using the Multifaceted Rasch Partial Credit Measurement Model. *Journal of Research in Music Education, 65*(1), 95–119.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 74,* 200-214.

Wiggins, G. & McTighe, J. (2005). *Understanding by design.* Alexandria, VA: Association for Supervision and Curriculum Development.

Willis, J. W., & Edwards, C. (2014). Action research: Models, methods, and examples. *British Journal of Educational Technology, 46*(5), 24-25.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, N.J. Lawrence Erlbaum Associates.

Woolfolk, A. (2001). *Educational psychology, 8th Ed.* Allyn and Bacon, Boston.

Wright, B. & Stone, M. (1979). *Best test design*. Mesa Press.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. Robert L. Brennan, editor. *Educational measurement, 4th ed*. (111-154). American Council on Education.

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.

Zdzinski, S. F. (1991). Measurement of solo instrumental music performance: A review of literature. *Bulletin for the Council of Research in Music Education* (109) 47-58.

APPENDIX A
Flute Performance Rating Scale

Posture				
Upper body position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Head position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Placement of feet	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Arm positioning	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Instrument positioning	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand Position				
Correct balance points	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Angle of flute	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Angle of wrists	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Curvature of fingers	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Finger motion on keys	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right thumb position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Left thumb position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate

Breathing				
Aperture shape during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Embouchure contact with mouth plate during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Fulness of breath	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Timing of breath	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing is relaxed	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Embouchure				
Characteristic shape of embouchure	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Upper and lower teeth alignment	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouth plate placement on chin	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tone hole alignment with aperture	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Oral cavity is open	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Aperture adjustment relative to register (e.g. smaller aperture for upper register, larger for lower register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Jaw adjustment relative to register (e.g. jaw forward for upper register, back for lower register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate

Tone Quality				
Breath support	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic tone quality	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes respond clearly	Almost Never	Rarely	Often	Almost Always
Quality of tone in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic vibrato technique	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Air direction relative to register (e.g. slightly upwards for upper register, downwards for lower register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation				
Coordination of tongue and air	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tongue placement for attacks	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is light	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is quick and efficient	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air flow maintained during articulation	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Articulation markings observed	Almost Never	Rarely	Often	Almost Always
Tongue placement relative to register (e.g. further forward for upper register, further back for lower register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Technique				
Overall note accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Key signature observed	Almost Never	Rarely	Often	Almost Always
Accidentals observed	Almost Never	Rarely	Often	Almost Always
Rapid passages performed with evenness	Almost Never	Rarely	Often	Almost Always
Chromatic fingerings used when appropriate	Almost Never	Rarely	Often	Almost Always
Leaps are played fluidly	Almost Never	Rarely	Often	Almost Always
Rhythm				
Overall rhythmic accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Notes begin on time	Almost Never	Rarely	Often	Almost Always
Notes end on time	Almost Never	Rarely	Often	Almost Always
Accuracy of beat division	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of tied notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of dotted notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Intonation				
Accuracy of intervallic relationships	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Interpretation				
Dynamic markings observed	Almost Never	Rarely	Often	Almost Always
Tempo accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tempo modifications observed	Almost Never	Rarely	Often	Almost Always
Phrasing	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables relative to style	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes released in stylistically appropriate manner	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Expression				
Dynamics follow contour of musical line	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate

Dynamic choices are stylistically appropriate	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamics used to create tension and release	Almost Never	Rarely	Often	Almost Always
Multiple stylistic elements utilized to create character	Almost Never	Rarely	Often	Almost Always
Choice of vibrato depth and speed	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate

APPENDIX B
Oboe Performance Rating Scale

Posture				
Upper body position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Head position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Placement of feet	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Instrument positioning	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand Position				
Hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Angle of wrists	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Curvature of fingers	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Finger motion on keys	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right thumb position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Left thumb position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing				
Embouchure contact with reed during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Fulness of breath	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Timing of breath	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing is relaxed	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Embouchure				
Characteristic shape of embouchure	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Amount of lips covering teeth	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Chin position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Corners of mouth	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Reed placement in embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Pressure on reed from lips	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Oral cavity is open	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tone Quality				
Breath support	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic tone quality	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes respond clearly	Almost Never	Rarely	Often	Almost Always
Quality of tone in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Quality of tone at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation				
Coordination of tongue and air	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tongue placement on reed	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tongue touches reed lightly	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is quick and efficient	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air flow maintained during articulation	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation markings observed	Almost never	Rarely	Often	Almost always
Technique				
Overall note accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Key signature observed	Almost Never	Rarely	Often	Almost Always
Accidentals observed	Almost Never	Rarely	Often	Almost Always
Rapid passages performed with evenness	Almost Never	Rarely	Often	Almost Always
Leaps are played fluidly	Almost Never	Rarely	Often	Almost Always
Correct use of regular vs. forked F	Almost Never	Rarely	Often	Almost Always
Correct use of left vs. right Eb	Almost Never	Rarely	Often	Almost Always

Rhythm				
Overall rhythmic accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Notes begin on time	Almost Never	Rarely	Often	Almost Always
Notes end on time	Almost Never	Rarely	Often	Almost Always
Accuracy of beat division	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of tied notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of dotted notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation				
Accuracy of intervallic relationships	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Interpretation				
Dynamic markings observed	Almost Never	Rarely	Often	Almost Always
Tempo accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tempo modifications observed	Almost Never	Rarely	Often	Almost Always

Phrasing	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables relative to style	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes released in stylistically appropriate manner	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Expression				
Dynamics follow contour of musical line	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamic choices are stylistically appropriate	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamics used to create tension and release	Almost Never	Rarely	Often	Almost Always
Multiple stylistic elements utilized to create character	Almost Never	Rarely	Often	Almost Always
Choice of vibrato depth and speed	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate

APPENDIX C
Clarinet Performance Rating Scale

Posture				
Upper body position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Head position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Placement of feet	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Instrument positioning	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand Position				
Hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Angle of wrists	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Curvature of fingers	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Finger motion on keys	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right thumb position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Left thumb position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing				
Embouchure contact with mouthpiece during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Fulness of breath	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Timing of breath	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing is relaxed	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Embouchure				
Characteristic shape of embouchure	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Upper and lower teeth alignment	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Chin position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Lower lip position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Corners of mouth	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece position in embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Contact between reed and lower lip	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Contact between top teeth and mouthpiece	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Pressure on mouthpiece from lips	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Oral cavity is open	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tone Quality				
Breath support	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic tone quality	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes respond clearly	Almost Never	Rarely	Often	Almost Always

Quality of tone in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation				
Coordination of tongue and air	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tongue placement on reed	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tongue touches reed lightly	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is quick and efficient	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air flow maintained during articulation	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation markings observed	Almost never	Rarely	Often	Almost always
Technique				
Overall note accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Key signature observed	Almost Never	Rarely	Often	Almost Always
Accidentals observed	Almost Never	Rarely	Often	Almost Always
Rapid passages performed with evenness	Almost Never	Rarely	Often	Almost Always
Chromatic fingerings used when appropriate	Almost Never	Rarely	Often	Almost Always

Leaps are played fluidly	Almost Never	Rarely	Often	Almost Always
Rhythm				
Overall rhythmic accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Notes begin on time	Almost Never	Rarely	Often	Almost Always
Notes end on time	Almost Never	Rarely	Often	Almost Always
Accuracy of beat division	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of tied notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of dotted notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation				
Accuracy of intervallic relationships	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Interpretation				
Dynamic markings observed	Almost Never	Rarely	Often	Almost Always

Tempo accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tempo modifications observed	Almost Never	Rarely	Often	Almost Always
Phrasing	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables relative to style	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes released in stylistically appropriate manner	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Expression				
Dynamics follow contour of musical line	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamic choices are stylistically appropriate	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamics used to create tension and release	Almost Never	Rarely	Often	Almost Always
Multiple stylistic elements utilized to create character	Almost Never	Rarely	Often	Almost Always

APPENDIX D
Saxophone Performance Rating Scale

Posture				
Upper body position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Head position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Placement of feet	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Neck strap positioning	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Instrument positioning	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand Position				
Hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Angle of wrists	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Curvature of fingers	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Finger motion on keys	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right thumb position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Left thumb position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing				
Embouchure contact with mouthpiece during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate

Fulness of breath	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Timing of breath	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing is relaxed	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Embouchure				
Characteristic shape of embouchure	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Upper and lower teeth alignment	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Chin position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Lower lip position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece position in embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Contact between reed and lower lip	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Contact between top teeth and mouthpiece	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Pressure on mouthpiece from lips	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Corners of mouth are pushed in toward mouthpiece	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Oral cavity is open	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tone Quality				
Breath support	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Characteristic tone quality	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes respond clearly	Almost Never	Rarely	Often	Almost Always
Quality of tone in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic vibrato technique	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation				
Coordination of tongue and air	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tongue placement on reed	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tongue touches reed lightly	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is quick and efficient	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air flow maintained during articulation	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation markings observed	Almost never	Rarely	Often	Almost always
Technique				
Overall note accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Key signature observed	Almost Never	Rarely	Often	Almost Always

Accidentals observed	Almost Never	Rarely	Often	Almost Always
Rapid passages performed with evenness	Almost Never	Rarely	Often	Almost Always
Chromatic fingerings used when appropriate	Almost Never	Rarely	Often	Almost Always
Leaps are played fluidly	Almost Never	Rarely	Often	Almost Always
Rhythm				
Overall rhythmic accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Notes begin on time	Almost Never	Rarely	Often	Almost Always
Notes end on time	Almost Never	Rarely	Often	Almost Always
Accuracy of beat division	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of tied notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of dotted notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation				
Accuracy of intervallic relationships	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Interpretation				
Dynamic markings observed	Almost Never	Rarely	Often	Almost Always
Tempo accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tempo modifications observed	Almost Never	Rarely	Often	Almost Always
Phrasing	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables relative to style	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes released in stylistically appropriate manner	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Expression				
Dynamics follow contour of musical line	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamic choices are stylistically appropriate	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamics used to create tension and release	Almost Never	Rarely	Often	Almost Always
Multiple stylistic elements utilized to create character	Almost Never	Rarely	Often	Almost Always
Choice of vibrato depth and speed	inappropriate	slightly inappropriate	slightly appropriate	appropriate

APPENDIX E
Trumpet Performance Rating Scale

Posture				
Upper body position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Head position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Arm position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Placement of feet	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand Position				
Bell angle	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Left hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right hand valve motion	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Third valve slide use	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
First valve slide use	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing				
Embouchure contact with mouthpiece during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Fulness of breath	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Timing of breath	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing is relaxed	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air column adjusted relative to register (e.g., faster air for upper register, slower air for lower register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Embouchure				
Characteristic shape of embouchure	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Upper and lower teeth alignment	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Chin position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Corners of embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece placement	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece pressure on embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Oral cavity is open	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tone Quality				
Breath support	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic tone quality	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes respond clearly	Almost Never	Rarely	Often	Almost Always
Quality of tone in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Quality of tone in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic vibrato technique	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation				
Coordination of tongue and air	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tongue placement during attacks	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is light	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is quick and efficient	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air flow maintained during articulation	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation markings observed	Almost never	Rarely	Often	Almost always
Articulation syllables are appropriate relative to register (e.g., toh, tah, tee for low, mid, high register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Technique				
Overall note accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Key signature observed	Almost Never	Rarely	Often	Almost Always
Accidentals observed	Almost Never	Rarely	Often	Almost Always

Rapid passages performed with evenness	Almost Never	Rarely	Often	Almost Always
Third valve slide used when appropriate	Almost Never	Rarely	Often	Almost Always
Leaps are played fluidly	Almost Never	Rarely	Often	Almost Always
Rhythm				
Overall rhythmic accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Notes begin on time	Almost Never	Rarely	Often	Almost Always
Notes end on time	Almost Never	Rarely	Often	Almost Always
Accuracy of beat division	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of tied notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of dotted notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation				
Accuracy of intervallic relationships	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Interpretation				
Dynamic markings observed	Almost Never	Rarely	Often	Almost Always
Tempo accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tempo modifications observed	Almost Never	Rarely	Often	Almost Always
Phrasing	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables relative to style	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes released in stylistically appropriate manner	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Expression				
Dynamics follow contour of musical line	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamic choices are stylistically appropriate	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamics used to create tension and release	Almost Never	Rarely	Often	Almost Always
Multiple stylistic elements utilized to create character	Almost Never	Rarely	Often	Almost Always
Choice of vibrato depth and speed	inappropriate	slightly inappropriate	slightly appropriate	appropriate

APPENDIX F
Horn Performance Rating Scale

Posture				
Upper body position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Head position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Arm position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Placement of feet	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand Position				
Lead pipe angle	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Bell position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Left hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Left hand valve motion	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Use of trigger	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing				
Embouchure contact with mouthpiece during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Fulness of breath	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Timing of breath	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing is relaxed	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air column adjusted relative to register (e.g., faster air for upper register, slower air for lower register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Embouchure				
Characteristic shape of embouchure	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Upper and lower teeth alignment	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Chin position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Corners of embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece placement	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece pressure on embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Oral cavity is open	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tone Quality				
Breath support	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic tone quality	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes respond clearly	Almost Never	Rarely	Often	Almost Always
Quality of tone in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Quality of tone in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation				
Coordination of tongue and air	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tongue placement during attacks	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is light	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is quick and efficient	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air flow maintained during articulation	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation markings observed	Almost never	Rarely	Often	Almost always
Articulation syllables are appropriate relative to register (e.g., toh, tah, tee for low, mid, high register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Technique				
Overall note accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Key signature observed	Almost Never	Rarely	Often	Almost Always
Accidentals observed	Almost Never	Rarely	Often	Almost Always
Rapid passages performed with evenness	Almost Never	Rarely	Often	Almost Always

Trigger used when appropriate	Almost Never	Rarely	Often	Almost Always
Leaps are played fluidly	Almost Never	Rarely	Often	Almost Always
Rhythm				
Overall rhythmic accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Notes begin on time	Almost Never	Rarely	Often	Almost Always
Notes end on time	Almost Never	Rarely	Often	Almost Always
Accuracy of beat division	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of tied notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of dotted notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation				
Accuracy of intervallic relationships	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Interpretation				

Dynamic markings observed	Almost Never	Rarely	Often	Almost Always
Tempo accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tempo modifications observed	Almost Never	Rarely	Often	Almost Always
Phrasing	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables relative to style	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes released in stylistically appropriate manner	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Expression				
Dynamics follow contour of musical line	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamic choices are stylistically appropriate	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamics used to create tension and release	Almost Never	Rarely	Often	Almost Always
Multiple stylistic elements utilized to create character	Almost Never	Rarely	Often	Almost Always

APPENDIX G
Trombone Performance Rating Scale

Posture				
Upper body position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Head position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Arm position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Placement of feet	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand Position				
Bell angle	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Left hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right hand slide motion	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing				
Embouchure contact with mouthpiece during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Fulness of breath	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Timing of breath	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing is relaxed	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Air column adjusted relative to register (e.g., faster air for upper register, slower air for lower register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Embouchure				
Characteristic shape of embouchure	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Upper and lower teeth alignment	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Chin position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Corners of embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece placement	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece pressure on embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Oral cavity is open	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tone Quality				
Breath support	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic tone quality	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes respond clearly	Almost Never	Rarely	Often	Almost Always
Quality of tone in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Quality of tone at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic vibrato technique	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation				
Coordination of tongue and air	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tongue placement during attacks	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is light	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is quick and efficient	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air flow maintained during articulation	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation markings observed	Almost never	Rarely	Often	Almost always
Slur technique	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables are appropriate relative to register (e.g., toh, tah, tee for low, mid, high register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Technique				
Overall note accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Key signature observed	Almost Never	Rarely	Often	Almost Always
Accidentals observed	Almost Never	Rarely	Often	Almost Always
Slide position accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Rapid passages performed with evenness	Almost Never	Rarely	Often	Almost Always
Alternate positions used when appropriate	Almost Never	Rarely	Often	Almost Always
Leaps are played fluidly	Almost Never	Rarely	Often	Almost Always
Rhythm				
Overall rhythmic accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Notes begin on time	Almost Never	Rarely	Often	Almost Always
Notes end on time	Almost Never	Rarely	Often	Almost Always
Accuracy of beat division	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of tied notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of dotted notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation				
Accuracy of intervallic relationships	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Interpretation				
Dynamic markings observed	Almost Never	Rarely	Often	Almost Always
Tempo accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tempo modifications observed	Almost Never	Rarely	Often	Almost Always
Phrasing	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables relative to style	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes released in stylistically appropriate manner	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Expression				
Dynamics follow contour of musical line	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamic choices are stylistically appropriate	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamics used to create tension and release	Almost Never	Rarely	Often	Almost Always
Multiple stylistic elements utilized to create character	Almost Never	Rarely	Often	Almost Always
Choice of vibrato depth and speed	inappropriate	slightly inappropriate	slightly appropriate	appropriate

APPENDIX H
Euphonium Performance Rating Scale

Posture				
Upper body position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Head position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Arm position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Placement of feet	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand Position				
Instrument angle	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Left hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right hand valve motion	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing				
Embouchure contact with mouthpiece during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Fulness of breath	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Timing of breath	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing is relaxed	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Air column adjusted relative to register (e.g., faster air for upper register, slower air for lower register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Embouchure				
Characteristic shape of embouchure	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Upper and lower teeth alignment	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Chin position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Corners of embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece placement	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece pressure on embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Oral cavity is open	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tone Quality				
Breath support	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic tone quality	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes respond clearly	Almost Never	Rarely	Often	Almost Always
Quality of tone in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Quality of tone at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic vibrato technique	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation				
Coordination of tongue and air	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tongue placement during attacks	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is light	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is quick and efficient	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air flow maintained during articulation	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation markings observed	Almost never	Rarely	Often	Almost always
Articulation syllables are appropriate relative to register (e.g., toh, tah, tee for low, mid, high register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Technique				
Overall note accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Key signature observed	Almost Never	Rarely	Often	Almost Always
Accidentals observed	Almost Never	Rarely	Often	Almost Always
Rapid passages performed with evenness	Almost Never	Rarely	Often	Almost Always
4th valve used when appropriate	Almost Never	Rarely	Often	Almost Always

Leaps are played fluidly	Almost Never	Rarely	Often	Almost Always
Rhythm				
Overall rhythmic accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Notes begin on time	Almost Never	Rarely	Often	Almost Always
Notes end on time	Almost Never	Rarely	Often	Almost Always
Accuracy of beat division	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of tied notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of dotted notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation				
Accuracy of intervallic relationships	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Interpretation				
Dynamic markings observed	Almost Never	Rarely	Often	Almost Always

Tempo accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tempo modifications observed	Almost Never	Rarely	Often	Almost Always
Phrasing	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables relative to style	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes released in stylistically appropriate manner	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Expression				
Dynamics follow contour of musical line	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamic choices are stylistically appropriate	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamics used to create tension and release	Almost Never	Rarely	Often	Almost Always
Multiple stylistic elements utilized to create character	Almost Never	Rarely	Often	Almost Always
Choice of vibrato depth and speed	inappropriate	slightly inappropriate	slightly appropriate	appropriate

APPENDIX I
Tuba Performance Rating Scale

Posture				
Upper body position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Head position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Arm position	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Placement of feet	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Hand Position				
Instrument angle	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Left hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right hand position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Right hand valve motion	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing				
Embouchure contact with mouthpiece during inhalation	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Fulness of breath	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Timing of breath	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Breathing is relaxed	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

Air column adjusted relative to register (e.g., faster air for upper register, slower air for lower register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Embouchure				
Characteristic shape of embouchure	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Upper and lower teeth alignment	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Chin position	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Corners of embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece placement	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Mouthpiece pressure on embouchure	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Oral cavity is open	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tone Quality				
Breath support	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic tone quality	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes respond clearly	Almost Never	Rarely	Often	Almost Always
Quality of tone in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Quality of tone at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable

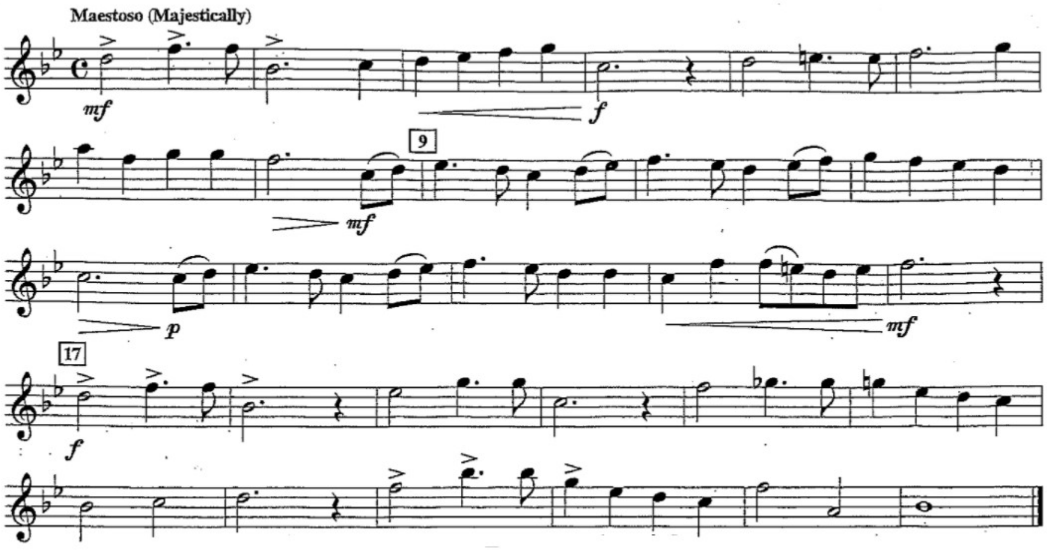
Quality of tone at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Characteristic vibrato technique	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation				
Coordination of tongue and air	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tongue placement during attacks	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is light	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Tonguing is quick and efficient	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Air flow maintained during articulation	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Articulation markings observed	Almost never	Rarely	Often	Almost always
Articulation syllables are appropriate relative to register (e.g., toh, tah, tee for low, mid, high register)	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Technique				
Overall note accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Key signature observed	Almost Never	Rarely	Often	Almost Always
Accidentals observed	Almost Never	Rarely	Often	Almost Always
Rapid passages performed with evenness	Almost Never	Rarely	Often	Almost Always
4th valve used when appropriate	Almost Never	Rarely	Often	Almost Always

Leaps are played fluidly	Almost Never	Rarely	Often	Almost Always
Rhythm				
Overall rhythmic accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Notes begin on time	Almost Never	Rarely	Often	Almost Always
Notes end on time	Almost Never	Rarely	Often	Almost Always
Accuracy of beat division	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of tied notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Accuracy of dotted notes	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation				
Accuracy of intervallic relationships	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in upper register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation in lower register	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at loud dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Intonation at soft dynamics	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Interpretation				
Dynamic markings observed	Almost Never	Rarely	Often	Almost Always

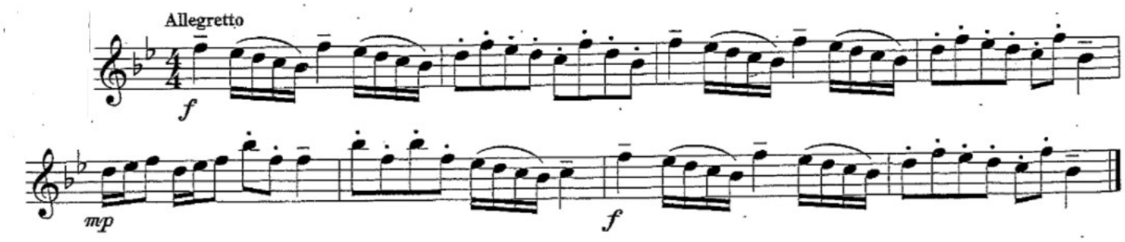
Tempo accuracy	Unacceptable	Slightly Unacceptable	Slightly Acceptable	Acceptable
Tempo modifications observed	Almost Never	Rarely	Often	Almost Always
Phrasing	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Articulation syllables relative to style	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Notes released in stylistically appropriate manner	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Expression				
Dynamics follow contour of musical line	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamic choices are stylistically appropriate	Inappropriate	Slightly Inappropriate	Slightly Appropriate	Appropriate
Dynamics used to create tension and release	Almost Never	Rarely	Often	Almost Always
Multiple stylistic elements utilized to create character	Almost Never	Rarely	Often	Almost Always
Choice of vibrato depth and speed	inappropriate	slightly inappropriate	slightly appropriate	appropriate

APPENDIX J
Flute Music Examples

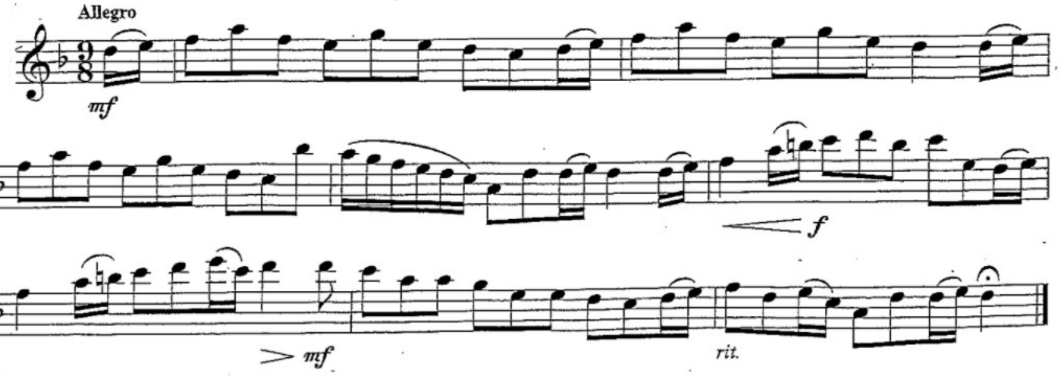
Example 1



Example 2

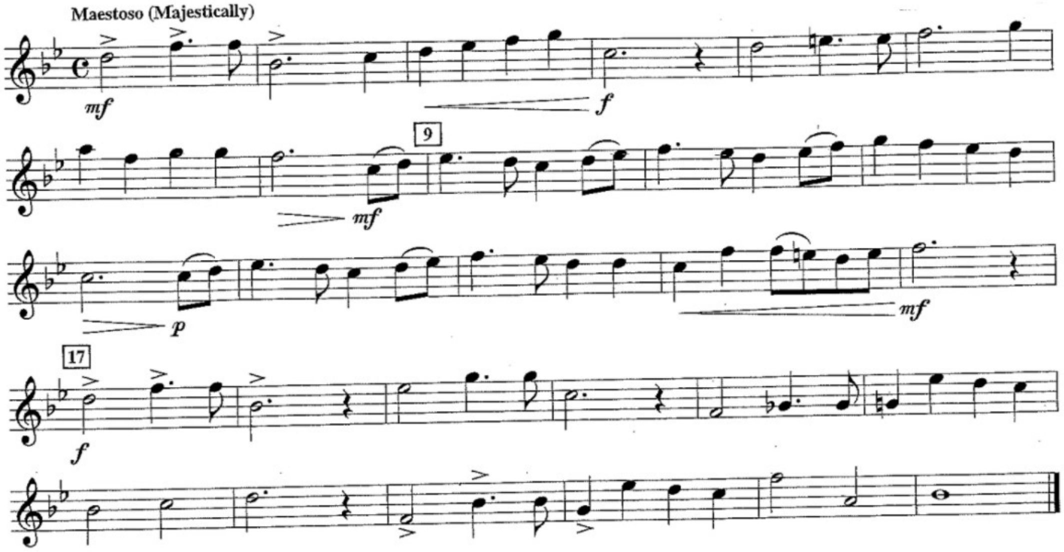


Example 3

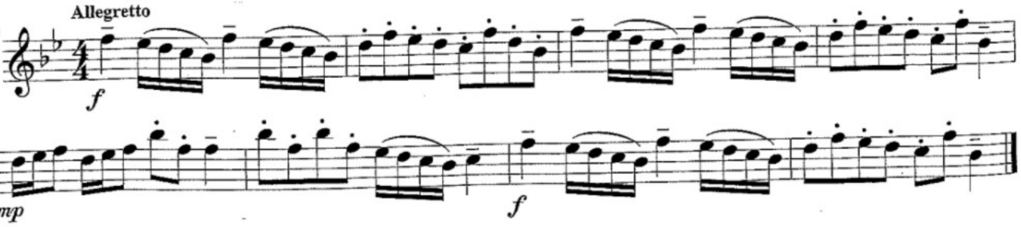


APPENDIX K
Oboe Music Examples

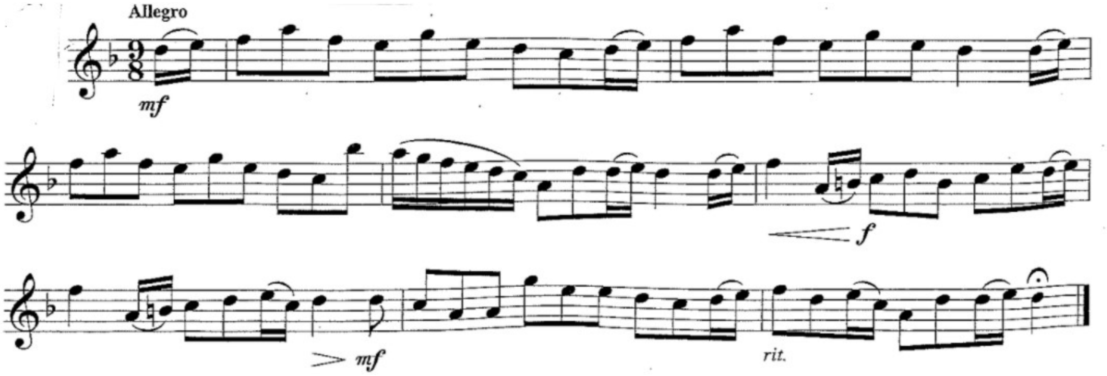
Example 1



Example 2

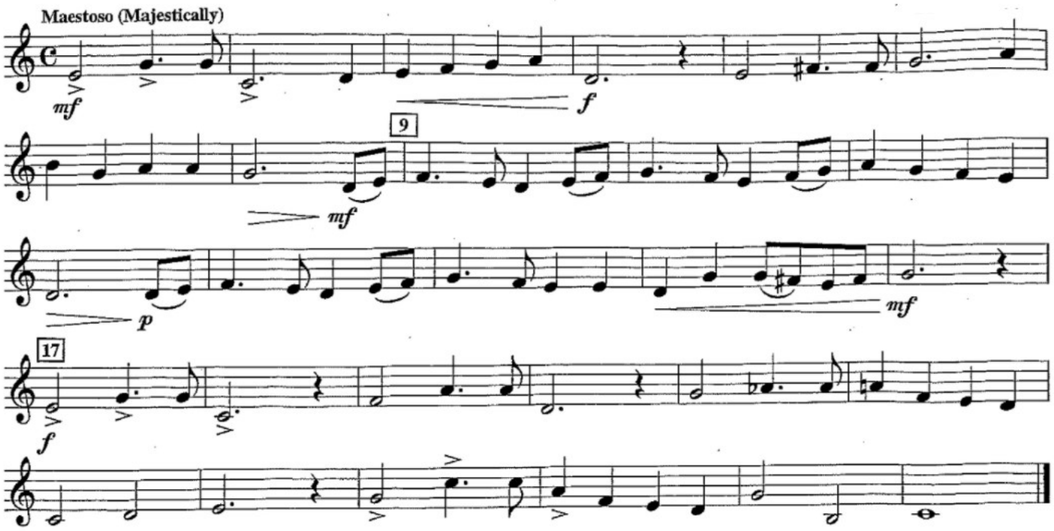


Example 3

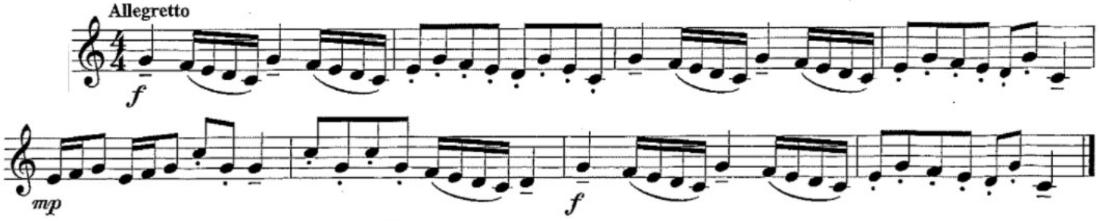


APPENDIX L
Clarinet Music Examples

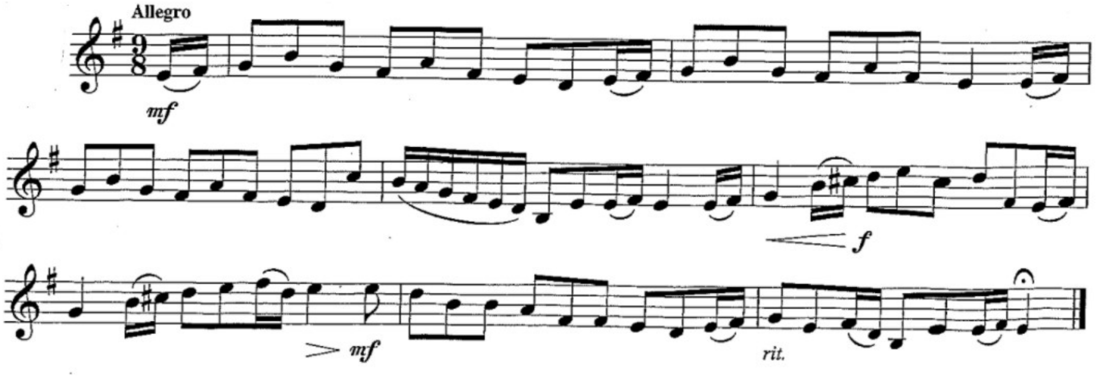
Example 1



Example 2

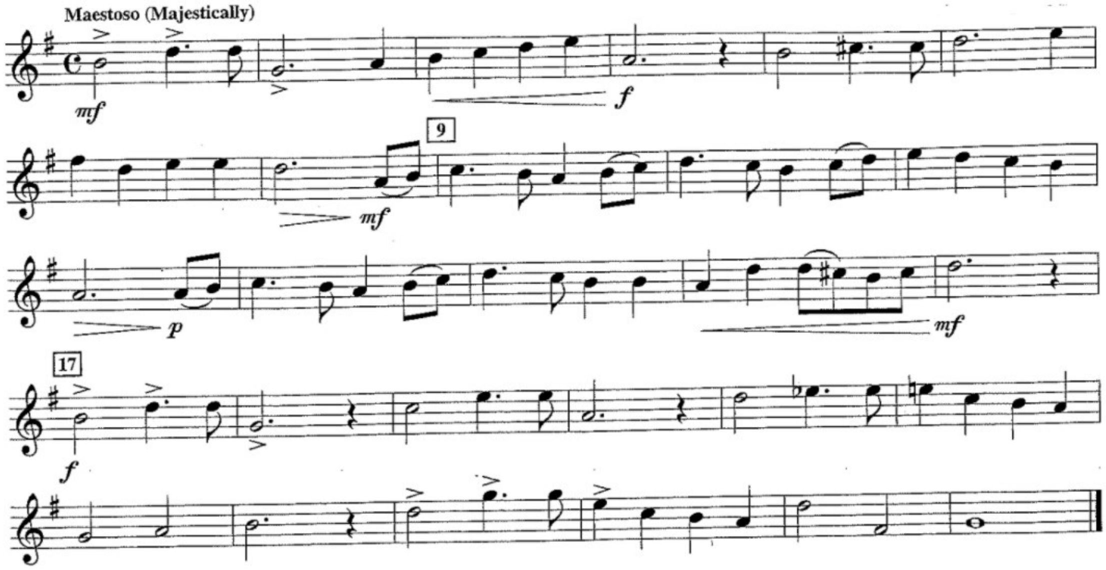


Example 3

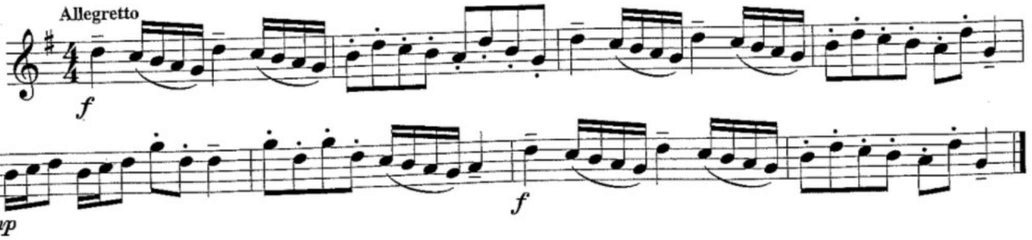


APPENDIX M
Alto Saxophone Music Examples

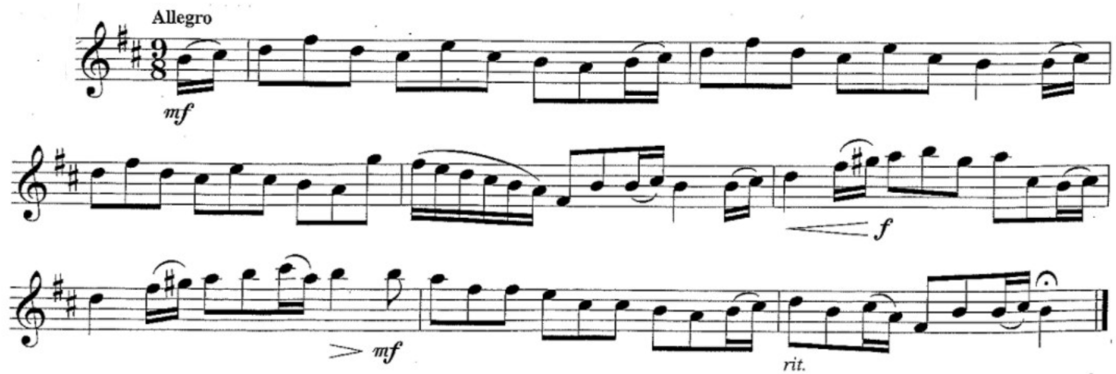
Example 1



Example 2

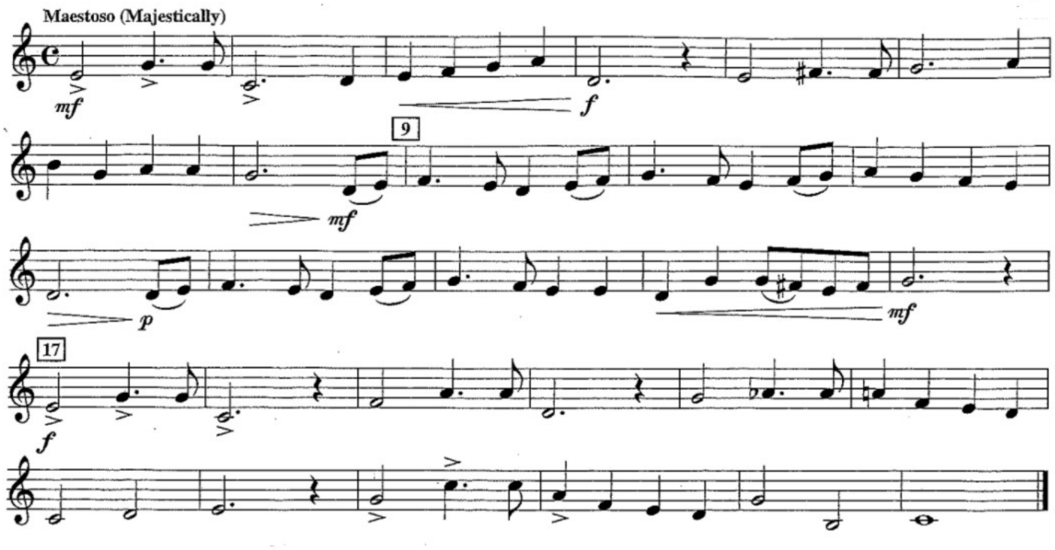


Example 3

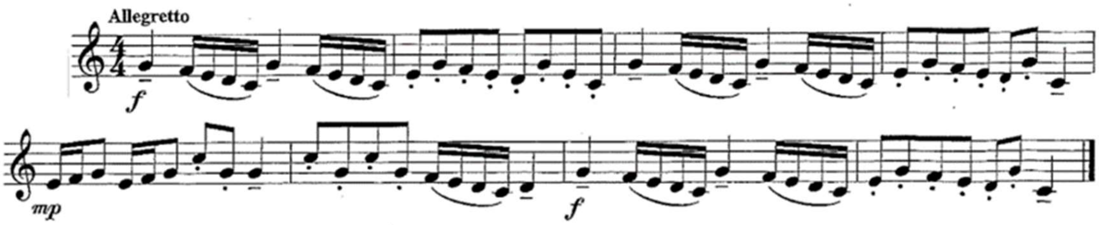


APPENDIX N
Tenor Saxophone Music Examples

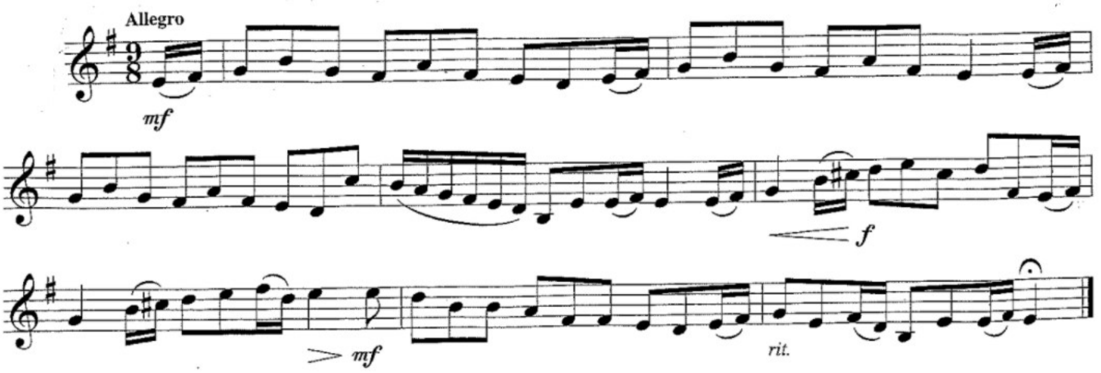
Example 1



Example 2

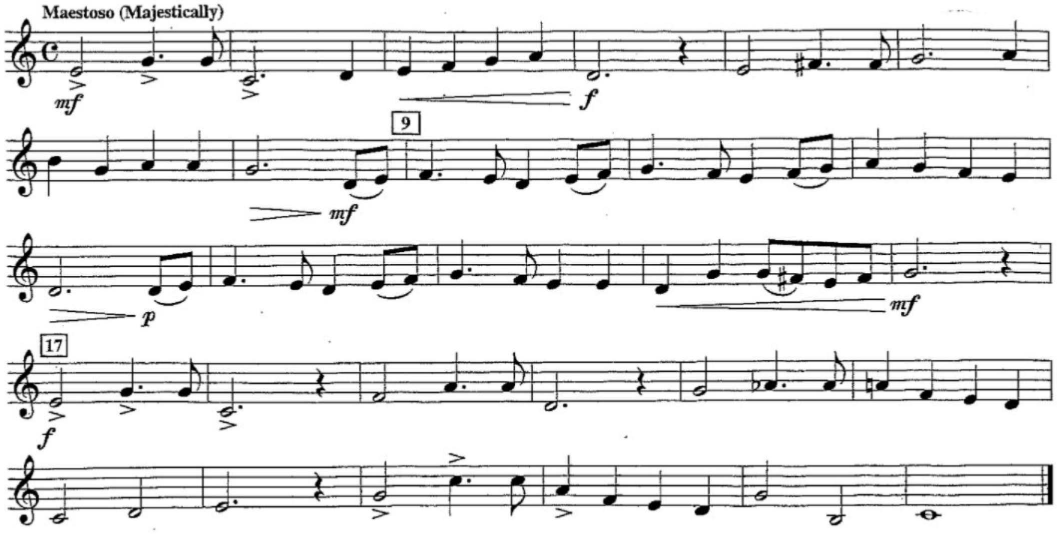


Example 3

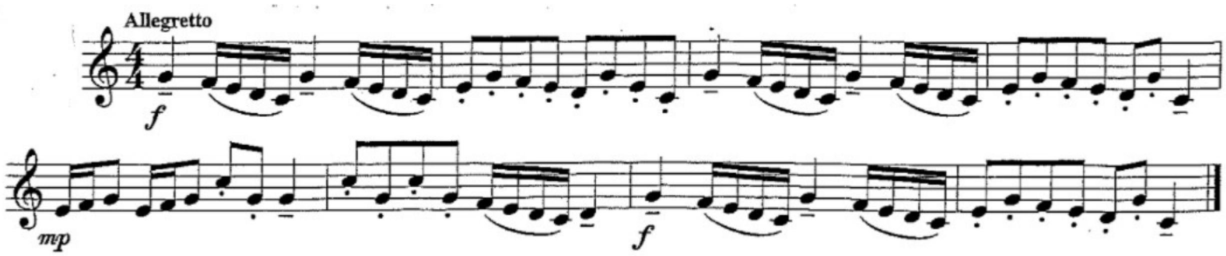


APPENDIX O
Trumpet Music Examples

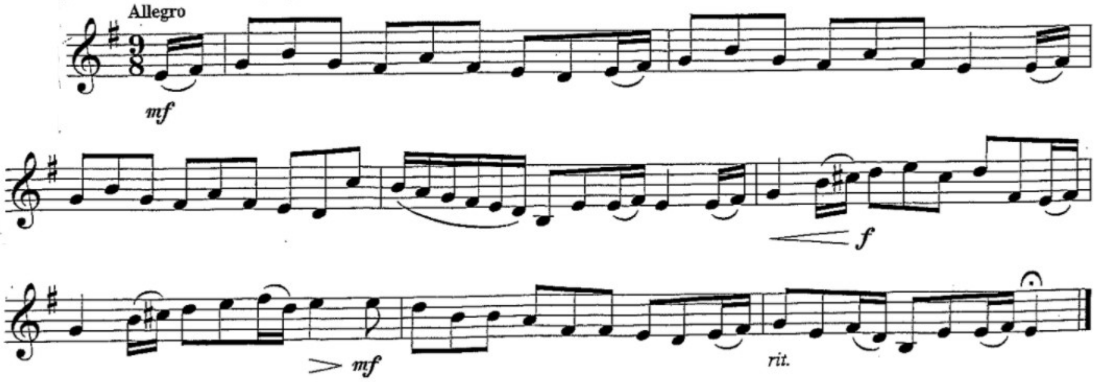
Example 1



Example 2

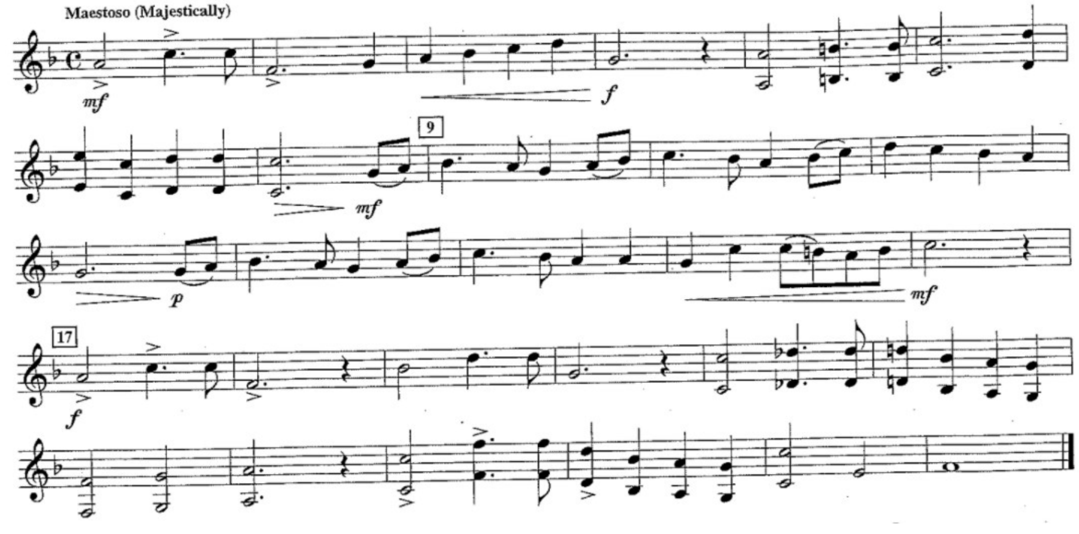


Example 3

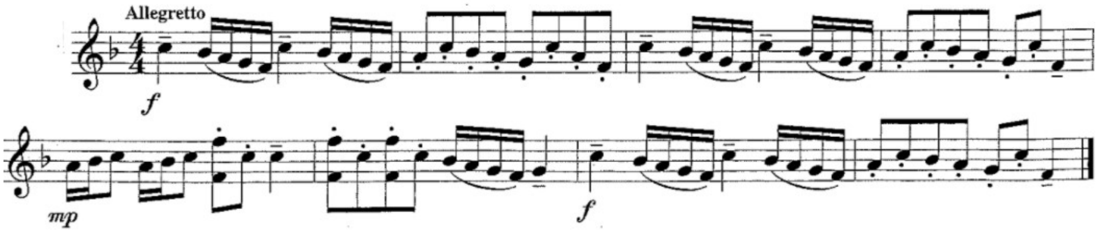


APPENDIX P
French Horn Music Examples

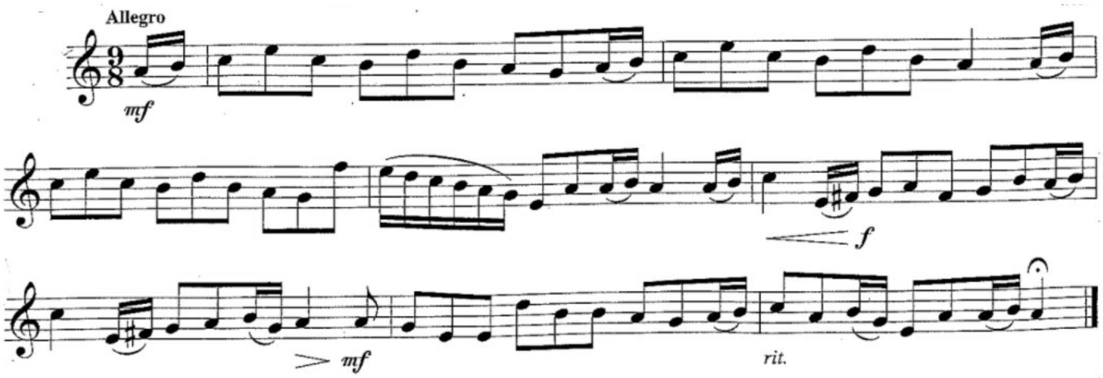
Example 1



Example 2

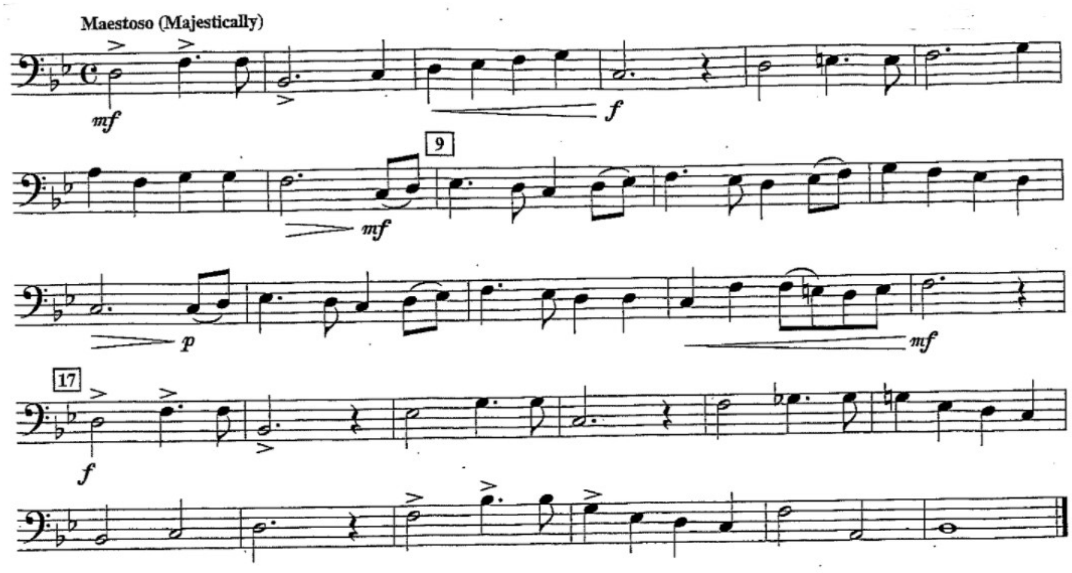


Example 3

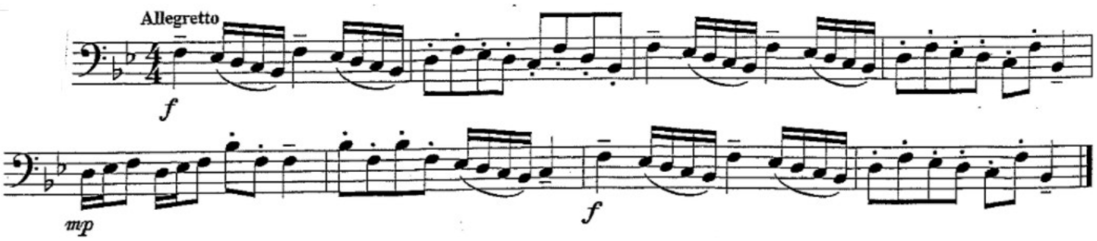


APPENDIX Q
Trombone and Euphonium Music Examples

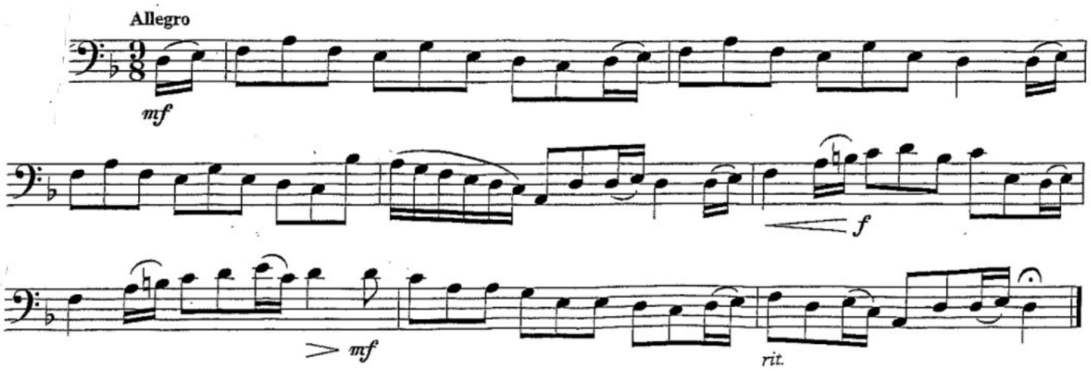
Example 1



Example 2

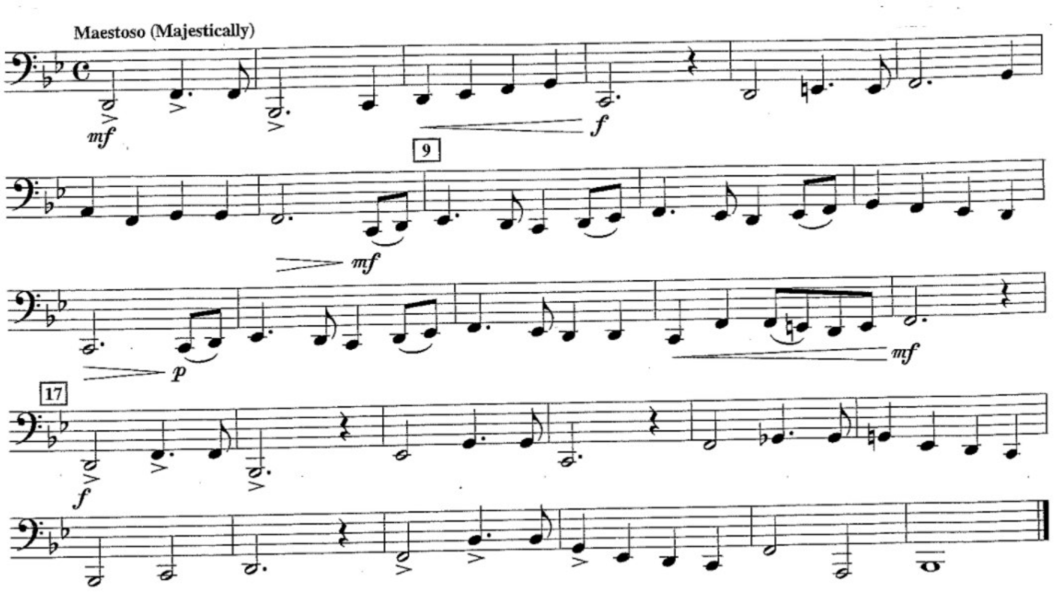


Example 3

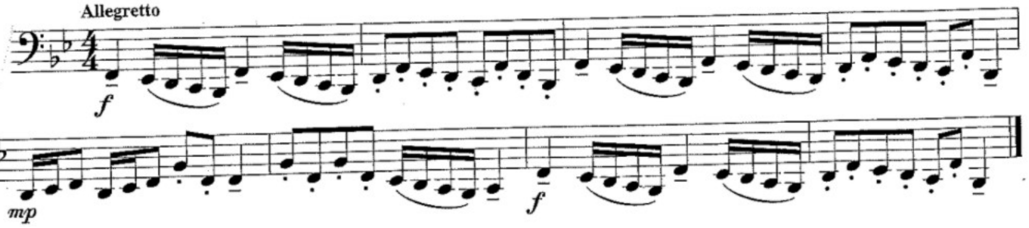


APPENDIX R
Tuba Music Examples

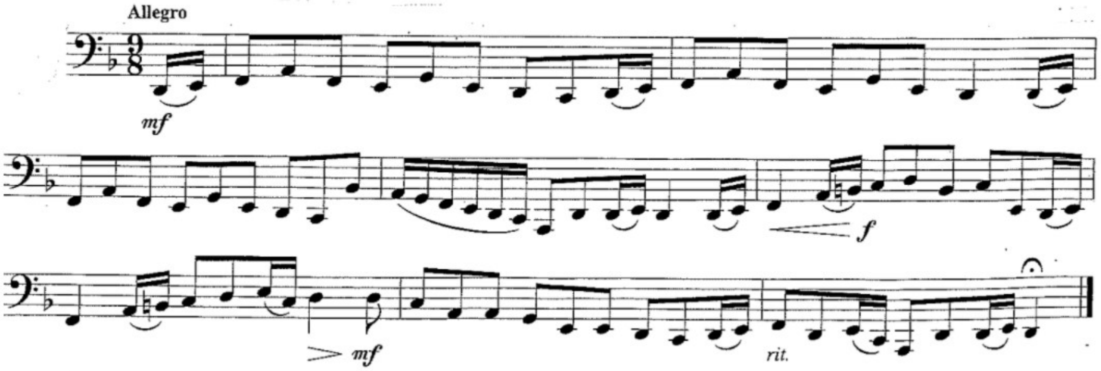
Example 1



Example 2



Example 3



APPENDIX S
Rater Calibration Tables

Calibration of Flute Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
1	2.90	1.20	0.11	0.75	-2.10	0.79	-1.50
8	3.45	0.99	0.09	0.95	-0.40	1.55	1.70
7	2.88	0.86	0.06	1.01	0.10	0.92	-0.60
9	3.50	0.40	0.17	1.36	1.80	4.34	2.10
2	2.76	0.33	0.07	1.03	0.40	1.00	0.00
4	2.81	-0.49	0.06	1.01	0.10	0.97	-0.30
3	3.07	-0.87	0.06	0.80	-3.20	0.92	-1.00
6	3.17	-0.93	0.06	1.06	0.90	1.50	5.00
5	3.30	-1.49	0.07	1.11	1.50	1.37	3.00
Mean	3.09	0.00	0.08	1.01	-0.10	1.48	0.90
SD	0.26	0.91	0.03	0.17	1.60	1.04	2.10

Presented in measure order from most to least severe.

Calibration of Clarinet Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
4	2.65	0.89	0.06	0.81	-3.30	0.82	-2.90
5	2.55	0.83	0.06	0.75	-4.60	0.73	-5.00
1	3.02	0.61	0.08	1.12	1.40	1.31	2.60
9	3.22	0.44	0.07	0.96	-0.50	0.94	-0.60
2	3.03	0.39	0.08	1.11	1.30	1.05	0.40
3	3.11	0.22	0.07	1.20	2.80	1.51	4.20
8	3.39	-0.19	0.08	1.09	1.20	1.11	1.00
7	3.21	-0.19	0.07	1.31	4.00	1.59	4.70
11	3.73	-0.52	0.28	0.94	-0.10	0.94	0.00
12	3.84	-1.18	0.34	0.99	0.00	1.51	1.00
6	3.40	-1.30	0.08	1.00	0.00	1.03	0.30
Mean	3.20	0.00	0.12	1.02	0.20	1.14	0.50
SD	0.38	0.72	0.09	0.15	2.40	0.28	2.70

Presented in measure order from most to least severe.

Calibration of Saxophone Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
2	2.25	0.88	0.06	0.93	-1.30	0.89	-1.40
3	2.62	0.61	0.06	0.93	-1.20	0.96	-0.50
6	3.37	0.11	0.12	0.96	-0.30	0.91	-0.50
5	3.34	-0.21	0.08	1.19	2.40	1.26	2.50
1	2.74	-0.43	0.08	0.87	-1.60	0.89	-1.10
4	3.40	-0.96	0.08	1.05	0.70	1.28	2.60
Mean	2.95	0.00	0.08	0.99	-0.20	1.03	0.30
SD	0.44	0.62	0.02	0.11	1.40	0.17	1.70

Presented in measure order from most to least severe.

Calibration of Trumpet Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
3	2.40	0.73	0.06	0.86	-2.40	0.87	-1.90
2	2.44	0.57	0.06	1.19	3.10	1.14	2.00
4	2.48	0.55	0.06	1.06	0.90	1.08	1.20
1	2.62	0.27	0.06	1.15	2.60	1.27	4.00
7	2.68	-0.09	0.06	0.74	-4.80	0.73	-4.20
5	3.02	-0.20	0.07	0.72	-4.70	0.81	-2.60
8	2.88	-0.25	0.07	0.77	-4.00	0.81	-2.50
6	2.85	-0.44	0.06	1.02	0.30	1.03	0.40
9	2.98	-0.48	0.06	1.33	4.70	1.26	3.20
11	3.24	-0.66	0.15	1.79	4.70	3.08	6.10
Mean	2.76	0.00	0.07	1.06	0.00	1.21	0.60
SD	0.27	0.47	0.02	0.31	3.60	0.65	3.20

Presented in measure order from most to least severe.

Calibration of F Horn Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
2	2.52	1.17	0.08	1.08	0.90	1.09	1.00
3	2.83	0.17	0.06	0.65	-6.60	0.69	-4.80
1	3.56	-0.50	0.21	1.23	1.00	1.65	1.80
4	3.23	-0.84	0.09	1.42	4.10	1.60	4.50
Mean	3.03	0.00	0.11	1.09	-0.10	1.26	0.60
SD	0.39	0.77	0.06	0.28	4.00	0.39	3.40

Presented in measure order from most to least severe.

Calibration of Trombone Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
3	2.50	0.68	0.04	0.83	-4.50	0.88	-2.70
4	2.72	0.59	0.07	0.94	-0.90	0.95	-0.50
1	2.60	0.27	0.12	0.91	-0.80	0.85	-1.20
2	2.38	0.22	0.07	1.30	4.10	1.22	3.00
6	3.30	-0.60	0.10	1.52	4.90	1.64	5.00
5	3.37	-1.16	0.08	0.99	-0.10	1.00	0.00
Mean	2.81	0.00	0.08	1.08	0.40	1.09	0.60
SD	0.39	0.66	0.02	0.25	3.30	0.27	2.60

Presented in measure order from most to least severe.

Calibration of Tuba Rater Facet

Rater Number	Observed Average	Measure	Standard Error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
2	2.55	0.74	0.13	2.42	8.40	2.61	9.00
5	3.14	0.52	0.13	1.40	2.80	1.37	2.50
4	3.13	-0.35	0.08	0.78	-3.10	0.79	-2.90
3	3.01	-0.38	0.08	1.07	1.00	1.09	1.20
1	3.08	-0.54	0.08	0.68	-4.80	0.71	-4.20
Mean	2.98	0.00	0.10	1.27	0.90	1.31	1.10
SD	0.22	0.52	0.03	0.63	4.70	0.69	4.70

Presented in measure order from most to least severe.

APPENDIX T
CPRS Differential Item Function Table

CPRS Differential Item Functioning Statistics

Item	Instrument	Observed Score	Expected Score	Bias + Size	Model SE	t	Infit MSE	Outfit MSE	Square	Measure
55	2	7	6.52	0.47	1.05	0.44	0.30	0.30	110	1.77
64	1	129	126.20	0.10	0.19	0.53	1.00	0.90	127	2.49
4	1	246	245.06	0.04	0.20	0.19	1.50	1.90	7	-0.83
59	1	170	168.76	0.03	0.17	0.21	0.70	0.70	117	1.18
58	1	172	170.89	0.03	0.17	0.19	0.80	0.70	115	1.11
24	1	218	217.25	0.03	0.19	0.14	1.00	1.10	47	-0.65
3	1	223	222.15	0.03	0.18	0.15	2.00	4.10	5	-0.32
47	1	212	211.30	0.02	0.19	0.13	1.10	1.00	93	-0.31
63	1	123	122.38	0.02	0.19	0.12	1.00	0.90	125	2.65
34	1	206	205.38	0.02	0.18	0.11	0.70	0.60	67	0.04
51	1	201	200.41	0.02	0.17	0.10	0.90	1.00	101	0.31
35	1	197	196.50	0.02	0.18	0.09	0.70	0.70	69	0.23
57	1	211	210.13	0.02	0.14	0.12	2.10	1.80	113	-0.10
62	1	138	137.64	0.01	0.18	0.06	1.00	1.00	123	2.08
61	1	135	134.79	0.01	0.18	0.04	1.10	1.10	121	2.19
37	1	184	183.83	0.01	0.18	0.03	0.80	0.80	73	0.76
28	1	181	180.96	0.00	0.18	0.01	0.80	0.80	55	0.85
48	1	213	213.07	0.00	0.16	-0.01	1.80	2.20	95	-0.13
1	1	244	244.06	0.00	0.21	-0.01	1.20	4.70	1	-1.18
42	1	233	233.10	0.00	0.18	-0.02	1.20	0.80	83	-0.52
22	1	236	236.09	0.00	0.19	-0.02	1.20	0.90	43	-0.80
2	1	248	248.07	0.00	0.23	-0.02	1.10	2.10	3	-1.25
9	1	250	250.06	0.00	0.25	-0.02	0.80	0.50	17	-0.91
16	1	238	238.09	0.00	0.21	-0.02	0.90	0.60	31	-0.19
20	1	233	233.11	0.00	0.19	-0.02	1.00	0.90	39	-0.97
18	1	231	231.13	0.00	0.19	-0.02	1.00	0.80	35	-0.70
10	1	237	237.12	-0.01	0.21	-0.02	1.20	1.60	19	-1.04

continued

CPRS DIF Statistics (continued)

Item	Instrument	Observed Score	Expected Score	Bias + Size	Model SE	t	Infit MSE	Outfit MSE	Square	Measure
11	1	227	227.15	-0.01	0.19	-0.03	1.50	1.30	21	-0.63
40	1	214	214.18	-0.01	0.17	-0.03	1.20	1.20	79	-0.18
17	1	225	225.16	-0.01	0.19	-0.03	1.10	1.20	33	-0.59
49	1	225	225.15	-0.01	0.19	-0.03	1.20	1.00	97	-0.82
39	1	230	230.16	-0.01	0.19	-0.03	1.30	1.20	77	-0.54
23	1	227	227.16	-0.01	0.20	-0.03	0.70	0.60	45	0.25
6	1	229	229.16	-0.01	0.20	-0.03	0.90	0.90	11	0.14
21	1	233	233.17	-0.01	0.21	-0.03	0.90	0.80	41	-0.95
15	1	222	222.20	-0.01	0.19	-0.04	1.10	1.00	29	-0.75
5	1	228	228.20	-0.01	0.21	-0.04	0.90	1.20	9	0.14
56	1	200	200.31	-0.01	0.17	-0.05	1.50	1.50	111	-0.43
14	1	211	211.29	-0.01	0.18	-0.05	1.20	1.10	27	-0.58
19	1	233	233.19	-0.01	0.22	-0.04	0.80	0.70	37	-0.18
44	1	204	204.31	-0.01	0.18	-0.06	1.00	0.90	87	0.04
33	1	215	215.30	-0.01	0.18	-0.05	0.70	0.60	65	-0.15
52	1	223	223.27	-0.01	0.20	-0.05	0.60	0.70	103	-0.75
32	1	207	207.36	-0.01	0.18	-0.06	0.80	0.70	63	0.07
50	1	231	231.22	-0.01	0.23	-0.05	0.50	0.50	99	-0.12
54	1	214	214.32	-0.01	0.19	-0.06	0.70	0.80	107	-0.35
7	1	209	209.32	-0.01	0.19	-0.06	1.20	1.70	13	-0.49
38	1	199	199.40	-0.01	0.17	-0.07	1.00	0.90	75	0.27
8	1	203	203.36	-0.01	0.19	-0.07	1.20	1.80	15	-0.38
12	1	212	212.37	-0.01	0.19	-0.07	1.40	1.70	23	-0.30
53	1	215	215.35	-0.01	0.20	-0.07	0.70	0.90	105	-0.59
25	1	198	198.48	-0.01	0.17	-0.08	1.00	1.00	49	-0.36
36	1	201	201.48	-0.02	0.18	-0.09	1.00	1.00	71	-0.40
29	1	212	212.40	-0.02	0.19	-0.08	0.70	0.70	57	-0.29
30	1	204	204.48	-0.02	0.18	-0.09	0.70	0.80	59	0.10
26	1	197	197.55	-0.02	0.18	-0.10	0.80	0.80	51	0.41
41	1	190	190.58	-0.02	0.18	-0.10	0.90	0.90	81	0.53
60	1	167	167.70	-0.02	0.16	-0.11	0.80	0.70	119	1.19

continued

CPRS DIF Statistics (continued)

Item	Instrument	Observed Score	Expected Score	Bias + Size	Model SE	t	Infit MSE	Outfit MSE	Square	Measure
13	1	212	212.47	-0.02	0.20	-0.09	1.10	1.10	25	-0.26
55	1	147	147.56	-0.02	0.18	-0.10	0.90	0.90	109	1.77
46	1	214	214.47	-0.02	0.22	-0.10	0.80	0.80	91	0.68
45	1	218	218.44	-0.02	0.22	-0.10	0.80	0.80	89	0.46
43	1	187	187.73	-0.02	0.17	-0.13	1.10	1.10	85	-0.16
31	1	196	196.74	-0.03	0.19	-0.14	0.80	0.80	61	0.22
27	1	202	202.75	-0.03	0.21	-0.15	0.80	0.80	53	-0.01
28	2	7	7.10	-0.16	1.25	-0.13	0.50	0.50	56	0.85
61	2	6	6.28	-0.24	0.90	-0.26	0.10	0.10	122	2.19
62	2	6	6.43	-0.35	0.88	-0.40	0.10	0.10	124	2.08
37	2	7	7.22	-0.36	1.23	-0.29	0.50	0.50	74	0.76
63	2	5	5.68	-0.54	0.89	-0.61	0.10	0.10	126	2.65
35	2	7	7.55	-0.94	1.13	-0.83	1.00	1.10	70	0.23
51	2	7	7.64	-1.21	1.15	-1.05	0.40	0.40	102	0.31
58	2	6	7.17	-1.27	0.93	-1.36	1.10	1.10	116	1.11
34	2	7	7.66	-1.27	1.15	-1.11	1.00	1.10	68	0.04
59	2	6	7.31	-1.36	0.87	-1.56	2.30	2.30	118	1.18
47	2	7	7.73	-1.47	1.11	-1.32	0.30	0.30	94	-0.31
24	2	7	7.79	-1.68	1.10	-1.52	0.30	0.30	48	-0.65
3	2	7	7.87	-2.07	1.04	-2.00	0.30	0.30	6	-0.32
57	2	7	7.90	-2.24	0.95	-2.37	0.20	0.20	114	-0.10
64	2	3	5.87	-2.64	1.20	-2.21	0.40	0.40	128	2.49
4	2	7	7.95	-3.00	0.97	-3.10	0.80	1.00	8	-0.83
1	2	8	7.95	-2.22<	1.36	-1.63	0.30	0.20	2	-1.18
2	2	8	7.94	-2.26<	1.50	-1.51	0.30	0.30	4	-1.25
5	2	8	7.83	-1.16<	1.51	-0.77	0.30	0.30	10	0.14
6	2	8	7.86	-1.32<	1.47	-0.90	0.30	0.30	12	0.14
7	2	8	7.72	-0.60<	1.46	-0.41	0.30	0.30	14	-0.49
8	2	8	7.68	-0.46<	1.43	-0.32	0.30	0.30	16	-0.38
9	2	8	7.95	-2.32<	1.48	-1.57	0.30	0.30	18	-0.91
10	2	8	7.90	-1.69<	1.49	-1.14	0.30	0.30	20	-1.04

continued

CPRS DIF Statistics (continued)

Item	Instrument	Observed	Expected	Bias +	Model	t	Infit	Outfit	Square	Measure
		Score	Score	Size	SE		MSE	MSE		
11	2	8	7.88	-1.41<	1.44	-0.98	0.30	0.30	22	-0.63
12	2	8	7.67	-0.48<	1.54	-0.31	0.30	0.30	24	-0.30
13	2	8	7.57	-0.19<	1.59	-0.12	0.30	0.30	26	-0.26
14	2	8	7.75	-0.75<	1.49	-0.50	0.30	0.30	28	-0.58
15	2	8	7.83	-1.11<	1.46	-0.76	0.30	0.30	30	-0.75
16	2	8	7.92	-1.83<	1.40	-1.31	0.30	0.20	32	-0.19
17	2	8	7.87	-1.34<	1.43	-0.94	0.30	0.30	34	-0.59
18	2	8	7.89	-1.56<	1.45	-1.07	0.30	0.30	36	-0.70
19	2	8	7.83	-1.22<	1.56	-0.78	0.30	0.30	38	-0.18
20	2	8	7.91	-1.69<	1.39	-1.22	0.30	0.20	40	-0.97
21	2	8	7.86	-1.39<	1.54	-0.90	0.30	0.30	42	-0.95
22	2	8	7.92	-1.85<	1.40	-1.32	0.30	0.20	44	-0.80
23	2	8	7.87	-1.33<	1.43	-0.93	0.30	0.30	46	0.25
25	2	8	7.57	-0.18<	1.54	-0.11	0.30	0.30	50	-0.36
26	2	8	7.49	0.02<	1.57	0.01	0.30	0.30	52	0.41
27	2	8	7.30	0.46<	1.61	0.29	0.30	0.30	54	-0.01
29	2	8	7.64	-0.39<	1.56	-0.25	0.30	0.30	58	-0.29
30	2	8	7.56	-0.15<	1.57	-0.09	0.30	0.30	60	0.10
31	2	8	7.31	0.42<	1.59	0.26	0.30	0.30	62	0.22
32	2	8	7.68	-0.50<	1.53	-0.33	0.30	0.30	64	0.07
33	2	8	7.73	-0.72<	1.55	-0.47	0.30	0.30	66	-0.15
36	2	8	7.56	-0.16<	1.55	-0.11	0.30	0.30	72	-0.40
38	2	8	7.65	-0.38<	1.49	-0.25	0.30	0.30	76	0.27
39	2	8	7.86	-1.40<	1.53	-0.92	0.30	0.30	78	-0.54
40	2	8	7.85	-1.15<	1.37	-0.84	0.30	0.20	80	-0.18
41	2	8	7.47	0.06<	1.52	0.04	0.30	0.30	82	0.53
42	2	8	7.92	-1.80<	1.42	-1.26	0.30	0.30	84	-0.52
43	2	8	7.33	0.36<	1.55	0.24	0.30	0.30	86	-0.16
44	2	8	7.73	-0.62<	1.44	-0.43	0.30	0.30	88	0.04
45	2	8	7.59	-0.25<	1.60	-0.16	0.30	0.30	90	0.46
46	2	8	7.57	-0.19<	1.59	-0.12	0.30	0.30	92	0.68

continued

CPRS DIF Statistics (continued)

Item	Instrument	Observed Score	Expected Score	Bias + Size	Model SE	t	Infit MSE	Outfit MSE	Square	Measure
48	2	8	7.95	-1.79<	1.20	-1.50	0.20	0.20	96	-0.13
49	2	8	7.87	-1.34<	1.41	-0.95	0.30	0.20	98	-0.82
50	2	8	7.80	-1.06<	1.58	-0.67	0.30	0.30	100	-0.12
52	2	8	7.76	-0.85<	1.55	-0.55	0.30	0.30	104	-0.75
53	2	8	7.69	-0.55<	1.54	-0.36	0.30	0.30	106	-0.59
54	2	8	7.72	-0.65<	1.52	-0.43	0.30	0.30	108	-0.35
56	2	8	7.74	-0.64<	1.40	-0.45	0.30	0.20	112	-0.43
60	2	8	7.37	0.24<	1.45	0.17	0.30	0.30	120	1.19

APPENDIX U
SPRS Differential Item Function Table

SPRS Differential Item Functioning Statistics

Item	Instrument	Observed Score	Expected Score	Bias + Size	Model SE	t	Infit MSE	Outfit MSE	Square	Measure
45	2	15	13.29	1.56	1.14	1.37	0.90	0.70	90	0.28
35	2	15	13.49	1.36	1.13	1.21	0.90	0.70	70	0.21
28	2	15	14.07	0.86	1.10	0.78	0.60	0.50	56	-0.40
56	2	14	13.08	0.74	0.94	0.79	1.00	1.00	112	0.47
61	2	14	12.90	0.74	0.88	0.84	0.90	0.90	122	0.76
52	2	15	14.44	0.58	1.11	0.52	0.80	0.70	104	-0.13
43	2	14	13.25	0.54	0.89	0.61	0.90	0.90	86	0.71
44	2	15	14.48	0.54	1.11	0.48	1.20	1.40	88	0.35
36	2	14	13.50	0.37	0.90	0.42	0.40	0.40	72	0.33
40	2	15	14.71	0.31	1.09	0.28	0.60	0.50	80	-0.67
25	2	14	13.53	0.28	0.81	0.35	0.60	0.60	50	-0.15
23	2	15	14.72	0.28	1.06	0.27	1.10	1.40	46	-0.35
30	2	15	14.70	0.26	0.99	0.26	0.40	0.30	60	-0.10
38	2	14	13.65	0.24	0.84	0.28	0.70	0.70	76	0.43
24	2	14	13.71	0.22	0.89	0.25	1.40	1.50	48	-0.03
55	2	14	13.73	0.21	0.90	0.23	0.90	0.90	110	0.25
58	1	96	92.94	0.18	0.24	0.73	1.40	1.60	115	0.20
47	2	15	14.85	0.16	1.08	0.15	0.80	0.70	94	-0.45
14	1	93	90.40	0.16	0.25	0.65	0.80	0.70	27	0.24
59	1	121	119.46	0.16	0.33	0.49	2.00	2.20	117	-0.84
65	1	64	62.53	0.13	0.29	0.44	1.00	1.00	129	2.16
51	1	107	104.57	0.13	0.23	0.55	1.00	1.20	101	-0.31
2	1	117	115.42	0.12	0.27	0.42	1.00	1.50	3	-1.12
4	1	118	116.37	0.11	0.27	0.42	0.80	0.80	7	-0.94
33	1	49	47.72	0.11	0.29	0.38	0.80	0.50	65	2.05
67	1	48	47.00	0.09	0.30	0.30	0.90	0.50	133	2.16
63	1	67	65.91	0.09	0.28	0.31	1.10	1.20	125	0.82

continued

SPRS DIF Statistics (continued)

Item	Instrument	Observed Score	Expected Score	Bias + Size	Model SE	t	Infit MSE	Outfit MSE	Square	Measure
64	1	70	68.96	0.07	0.26	0.27	1.10	1.20	127	1.73
3	1	109	107.92	0.07	0.25	0.27	1.30	1.30	5	-0.47
15	1	97	95.88	0.07	0.24	0.27	1.00	1.00	29	0.11
31	1	89	88.06	0.07	0.26	0.25	0.80	0.80	61	0.43
49	1	101	100.14	0.05	0.24	0.20	0.90	0.80	97	-0.20
26	1	95	94.24	0.05	0.25	0.19	1.20	1.30	51	0.22
57	1	67	66.45	0.04	0.27	0.15	1.30	1.50	113	1.64
66	1	74	73.39	0.04	0.25	0.15	0.70	0.70	131	1.41
1	1	117	116.45	0.04	0.27	0.15	1.00	1.10	1	-0.93
53	1	94	93.47	0.04	0.26	0.14	0.70	0.70	105	0.18
13	1	86	85.37	0.03	0.23	0.15	1.00	1.00	25	0.51
62	1	85	84.52	0.03	0.26	0.13	0.80	0.80	123	0.82
37	1	95	94.65	0.03	0.28	0.10	1.00	1.00	73	0.06
29	1	90	89.61	0.03	0.25	0.10	1.00	0.90	57	0.30
27	1	98	97.57	0.03	0.24	0.10	0.90	0.90	53	-0.07
46	1	105	104.68	0.02	0.23	0.07	0.90	0.80	91	-0.23
8	1	106	105.79	0.01	0.24	0.05	1.20	1.40	15	-0.54
60	1	82	81.83	0.01	0.26	0.04	0.80	0.70	119	1.05
54	1	96	95.91	0.01	0.25	0.02	0.70	0.80	107	0.02
32	1	86	85.98	0.00	0.25	0.01	0.70	0.70	63	0.71
32	2	13	13.01	-0.01	0.82	-0.01	1.30	1.30	64	0.71
47	1	104	104.15	-0.01	0.25	-0.04	0.70	0.70	93	-0.45
55	1	93	93.26	-0.02	0.25	-0.06	0.80	0.80	109	0.25
30	1	97	97.29	-0.02	0.23	-0.07	0.80	0.80	59	-0.10
23	1	101	101.28	-0.02	0.25	-0.07	0.80	0.80	45	-0.35
38	1	89	89.34	-0.02	0.24	-0.08	0.80	0.80	75	0.43
40	1	104	104.28	-0.02	0.26	-0.07	1.10	1.00	79	-0.67
24	1	95	95.28	-0.02	0.26	-0.07	1.20	1.20	47	-0.03
44	1	93	93.51	-0.02	0.22	-0.11	1.40	1.60	87	0.35
10	1	111	111.47	-0.03	0.24	-0.11	1.60	2.20	19	-0.63
21	1	108	108.52	-0.03	0.23	-0.12	0.50	0.60	41	-0.44

continued

SPRS DIF Statistics (continued)

Item	Instrument	Observed Score	Expected Score	Bias + Size	Model SE	t	Infit MSE	Outfit MSE	Square	Measure
50	1	127	127.19	-0.03	0.39	-0.07	1.40	2.20	99	-1.77
5	1	113	113.50	-0.03	0.25	-0.12	1.20	1.80	9	-0.70
36	1	91	91.49	-0.03	0.25	-0.12	1.00	1.10	71	0.33
11	1	116	116.43	-0.03	0.27	-0.12	1.00	0.90	21	-1.15
25	1	91	91.46	-0.03	0.27	-0.12	0.70	0.70	49	-0.15
52	1	99	99.56	-0.03	0.25	-0.14	0.50	0.50	103	-0.13
41	1	117	117.45	-0.03	0.28	-0.13	1.20	1.10	81	-1.19
42	1	115	115.50	-0.04	0.27	-0.14	1.30	1.10	83	-1.12
17	1	112	112.55	-0.04	0.26	-0.15	0.80	1.00	33	-1.01
22	1	112	112.55	-0.04	0.26	-0.15	0.80	0.80	43	-1.01
12	1	110	110.64	-0.04	0.24	-0.16	1.30	1.50	23	-0.59
20	1	106	106.69	-0.04	0.24	-0.17	1.00	1.40	39	-0.45
43	1	85	85.74	-0.04	0.24	-0.18	1.00	1.00	85	0.71
19	1	106	106.80	-0.05	0.24	-0.19	0.90	1.20	37	-0.45
6	1	109	109.83	-0.05	0.25	-0.21	1.00	1.10	11	-0.67
18	1	103	104.02	-0.06	0.24	-0.24	1.00	1.10	35	-0.33
56	1	89	89.91	-0.06	0.26	-0.24	0.80	0.80	111	0.47
54	2	14	14.09	-0.06	0.85	-0.07	0.70	0.70	108	0.02
61	1	83	84.09	-0.06	0.24	-0.27	0.80	0.70	121	0.76
48	1	101	102.14	-0.07	0.24	-0.27	0.90	0.90	95	-0.26
28	1	97	97.92	-0.07	0.27	-0.25	0.80	0.80	55	-0.40
7	1	107	108.04	-0.07	0.26	-0.27	1.50	1.70	13	-0.59
34	1	98	99.56	-0.09	0.24	-0.38	0.80	0.80	67	-0.13
35	1	91	92.50	-0.10	0.26	-0.39	1.00	1.00	69	0.21
9	1	101	102.55	-0.10	0.26	-0.40	1.10	1.20	17	-0.36
16	1	101	102.55	-0.10	0.26	-0.40	1.00	1.10	31	-0.36
45	1	90	91.71	-0.12	0.26	-0.45	0.50	0.50	89	0.28
60	2	12	12.16	-0.12	0.86	-0.14	1.40	1.60	120	1.05
39	1	84	87.05	-0.20	0.26	-0.78	0.50	0.50	77	0.55
57	2	10	10.54	-0.22	0.64	-0.34	1.20	1.20	114	1.64
29	2	13	13.38	-0.22	0.75	-0.29	1.20	1.10	58	0.30

continued

SPRS DIF Statistics (continued)

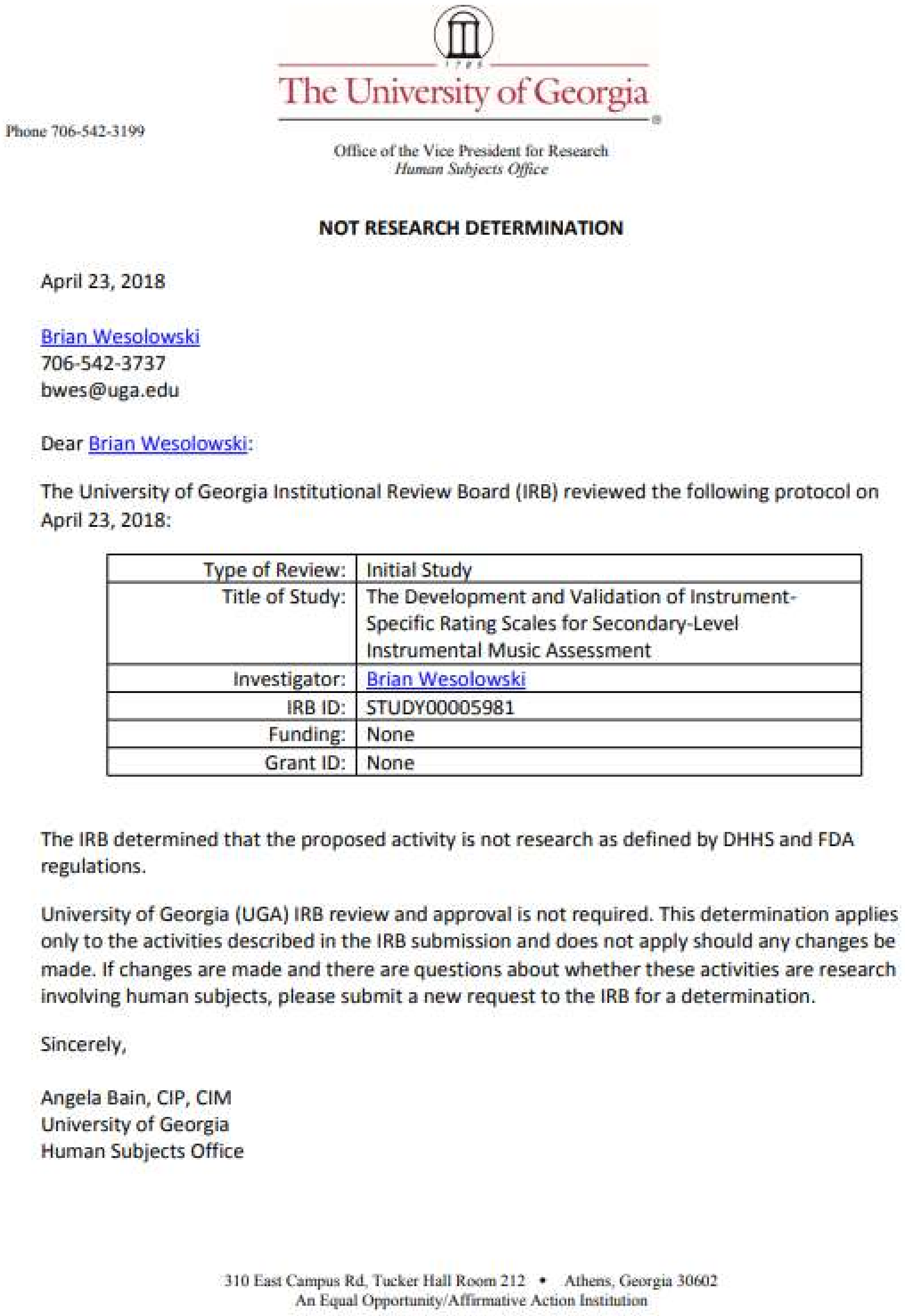
Item	Instrument	Observed	Expected	Bias +	Model	t	Infit	Outfit	Square	Measure
		Score	Score	Size	SE		MSE	MSE		
8	2	15	15.20	-0.22	1.00	-0.22	0.50	0.30	16	-0.54
37	2	13	13.34	-0.27	0.88	-0.30	0.30	0.40	74	0.06
13	2	13	13.62	-0.28	0.65	-0.43	1.40	1.50	26	0.51
27	2	14	14.42	-0.30	0.82	-0.37	0.60	0.70	54	-0.07
66	2	11	11.60	-0.30	0.71	-0.43	1.00	1.10	132	1.41
62	2	12	12.47	-0.33	0.83	-0.40	1.30	1.40	124	0.82
53	2	13	13.52	-0.37	0.83	-0.45	0.30	0.30	106	0.18
46	2	15	15.32	-0.38	1.01	-0.38	0.70	0.60	92	-0.23
67	2	7	7.99	-0.39	0.64	-0.61	1.30	1.20	134	2.16
33	2	7	8.27	-0.48	0.63	-0.77	1.30	1.20	66	2.05
64	2	10	11.02	-0.51	0.70	-0.73	2.10	2.10	128	1.73
26	2	13	13.75	-0.55	0.82	-0.67	0.30	0.30	52	0.22
31	2	12	12.93	-0.57	0.77	-0.74	0.80	0.80	62	0.43
49	2	14	14.85	-0.61	0.76	-0.80	0.80	0.70	98	-0.20
63	2	9	10.08	-0.73	0.79	-0.93	1.70	1.60	126	0.82
15	2	13	14.11	-0.76	0.77	-0.98	0.20	0.30	30	0.11
1	2	15	15.55	-0.90	1.07	-0.84	0.80	0.70	2	-0.93
65	2	8	9.45	-0.92	0.85	-1.08	1.00	1.00	130	2.16
3	2	14	15.07	-1.02	0.85	-1.21	1.00	1.10	6	-0.47
14	2	11	13.59	-1.31	0.69	-1.91	0.90	0.90	28	0.24
51	2	13	15.42	-1.47	0.59	-2.48	2.10	1.70	102	-0.31
58	2	11	14.05	-1.52	0.65	-2.36	0.50	0.50	116	0.20
2	2	14	15.58	-1.58	0.75	-2.12	1.50	1.20	4	-1.12
4	2	14	15.63	-1.74	0.73	-2.38	0.80	1.00	8	-0.94
59	2	14	15.53	-1.78	0.85	-2.08	0.30	0.30	118	-0.84
5	2	16	15.50	-0.01	< 1.47	0.00	0.20	0.10	10	-0.70
6	2	16	15.17	0.58	< 1.49	0.39	0.20	0.10	12	-0.67
7	2	16	14.96	0.88	< 1.50	0.58	0.20	0.10	14	-0.59
9	2	16	14.44	1.43	< 1.51	0.95	0.20	0.10	18	-0.36
10	2	16	15.53	-0.07	< 1.40	-0.05	0.20	0.10	20	-0.63
11	2	16	15.57	-0.14	< 1.44	-0.10	0.20	0.10	22	-1.15

continued

SPRS DIF Statistics (continued)

Item	Instrument	Observed	Expected	Bias +	Model	t	Infit	Outfit	Square	Measure
		Score	Score	Size	SE		MSE	MSE		
12	2	16	15.36	0.27	< 1.47	0.19	0.20	0.10	24	-0.59
16	2	16	14.44	1.43	< 1.51	0.95	0.20	0.10	32	-0.36
17	2	16	15.44	0.10	< 1.41	0.07	0.20	0.10	34	-1.01
18	2	16	14.98	0.79	< 1.46	0.54	0.20	0.10	36	-0.33
19	2	16	15.19	0.52	< 1.46	0.35	0.20	0.10	38	-0.45
20	2	16	15.30	0.33	< 1.41	0.24	0.20	0.10	40	-0.45
21	2	16	15.47	0.05	< 1.40	0.03	0.20	0.10	42	-0.44
22	2	16	15.44	0.10	< 1.41	0.07	0.20	0.10	44	-1.01
34	2	16	14.44	1.38	< 1.49	0.93	0.20	0.10	68	-0.13
39	2	16	12.94	2.55	< 1.50	1.70	0.20	0.10	78	0.55
41	2	16	15.55	-0.11	< 1.47	-0.08	0.20	0.10	82	-1.19
42	2	16	15.50	0.01	< 1.45	0.01	0.20	0.10	84	-1.12
48	2	16	14.86	0.91	< 1.46	0.62	0.20	0.10	96	-0.26
50	2	16	15.81	-1.03	< 1.48	-0.70	0.20	0.10	100	-1.77

APPENDIX V
IRB Approval



The University of Georgia

Phone 706-542-3199

Office of the Vice President for Research
Human Subjects Office

NOT RESEARCH DETERMINATION

April 23, 2018

Brian Wesolowski
706-542-3737
bwes@uga.edu

Dear Brian Wesolowski:

The University of Georgia Institutional Review Board (IRB) reviewed the following protocol on April 23, 2018:

Type of Review:	Initial Study
Title of Study:	The Development and Validation of Instrument-Specific Rating Scales for Secondary-Level Instrumental Music Assessment
Investigator:	Brian Wesolowski
IRB ID:	STUDY00005981
Funding:	None
Grant ID:	None

The IRB determined that the proposed activity is not research as defined by DHHS and FDA regulations.

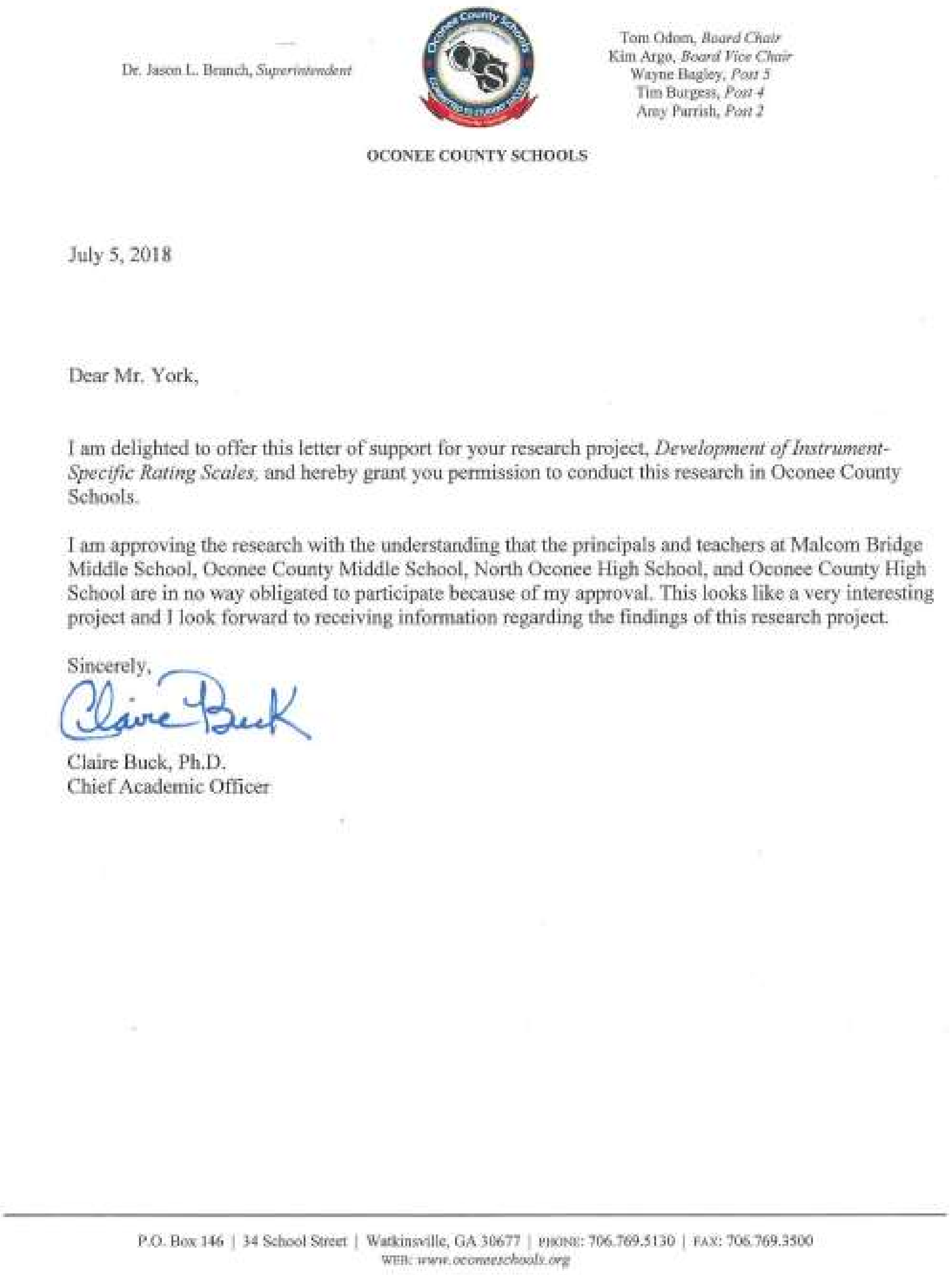
University of Georgia (UGA) IRB review and approval is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities are research involving human subjects, please submit a new request to the IRB for a determination.

Sincerely,

Angela Bain, CIP, CIM
University of Georgia
Human Subjects Office

310 East Campus Rd, Tucker Hall Room 212 • Athens, Georgia 30602
An Equal Opportunity/Affirmative Action Institution

APPENDIX W
School System Approval



Dr. Jason L. Branch, *Superintendent*

Tom Odom, *Board Chair*
Kim Argo, *Board Vice Chair*
Wayne Bagley, *Post 5*
Tim Burgess, *Post 4*
Amy Parrish, *Post 2*

OCONEE COUNTY SCHOOLS

July 5, 2018

Dear Mr. York,

I am delighted to offer this letter of support for your research project, *Development of Instrument-Specific Rating Scales,* and hereby grant you permission to conduct this research in Oconee County Schools.

I am approving the research with the understanding that the principals and teachers at Malcom Bridge Middle School, Oconee County Middle School, North Oconee High School, and Oconee County High School are in no way obligated to participate because of my approval. This looks like a very interesting project and I look forward to receiving information regarding the findings of this research project.

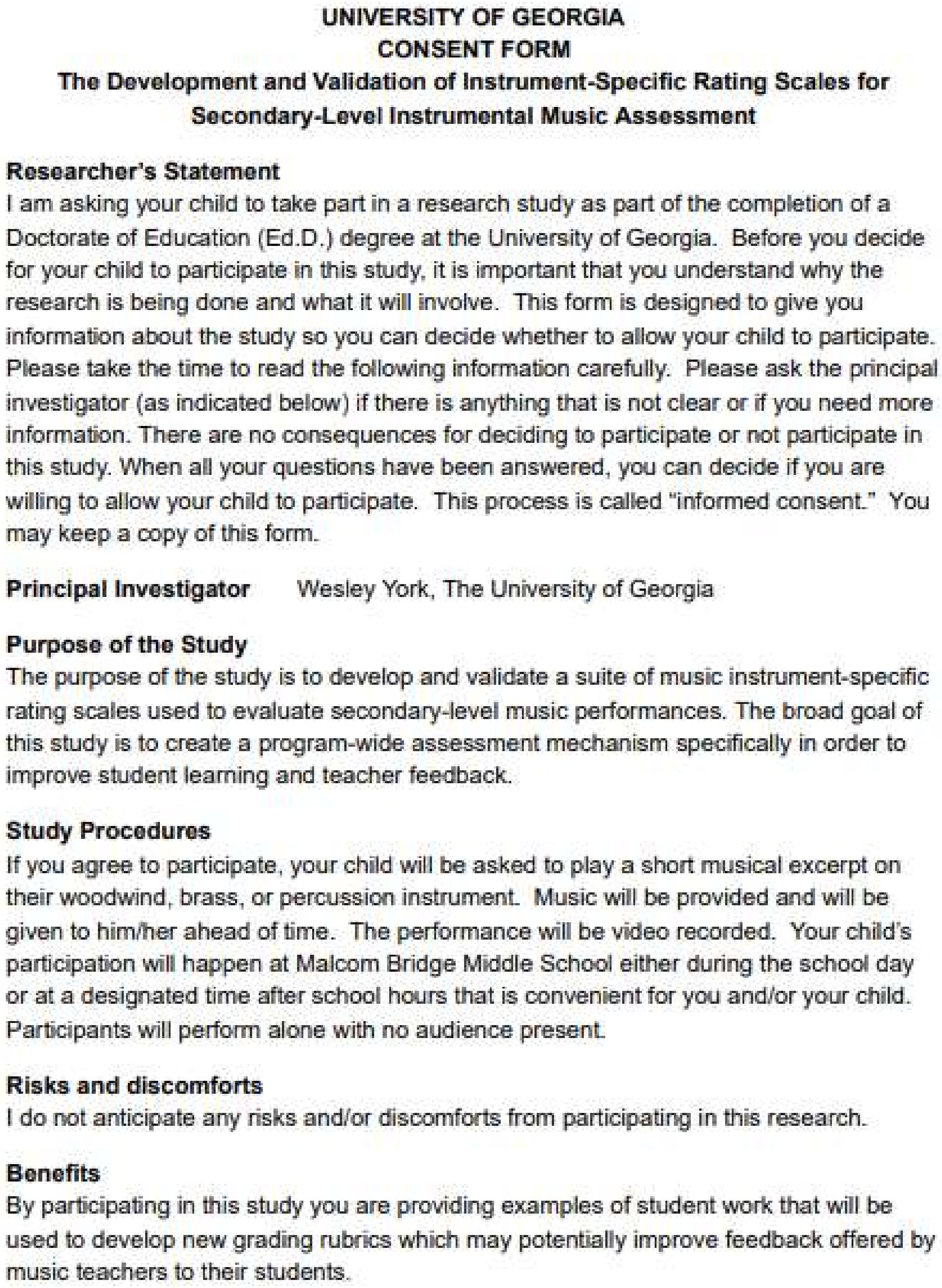
Sincerely,

Claire Buck

Claire Buck, Ph.D.
Chief Academic Officer

P.O. Box 146 | 34 School Street | Watkinsville, GA 30677 | PHONE: 706.769.5130 | FAX: 706.769.3500
WEB: *www.oconeeschools.org*

APPENDIX X
Student Consent Form



UNIVERSITY OF GEORGIA
CONSENT FORM
The Development and Validation of Instrument-Specific Rating Scales for Secondary-Level Instrumental Music Assessment

Researcher's Statement
I am asking your child to take part in a research study as part of the completion of a Doctorate of Education (Ed.D.) degree at the University of Georgia. Before you decide for your child to participate in this study, it is important that you understand why the research is being done and what it will involve. This form is designed to give you information about the study so you can decide whether to allow your child to participate. Please take the time to read the following information carefully. Please ask the principal investigator (as indicated below) if there is anything that is not clear or if you need more information. There are no consequences for deciding to participate or not participate in this study. When all your questions have been answered, you can decide if you are willing to allow your child to participate. This process is called "informed consent." You may keep a copy of this form.

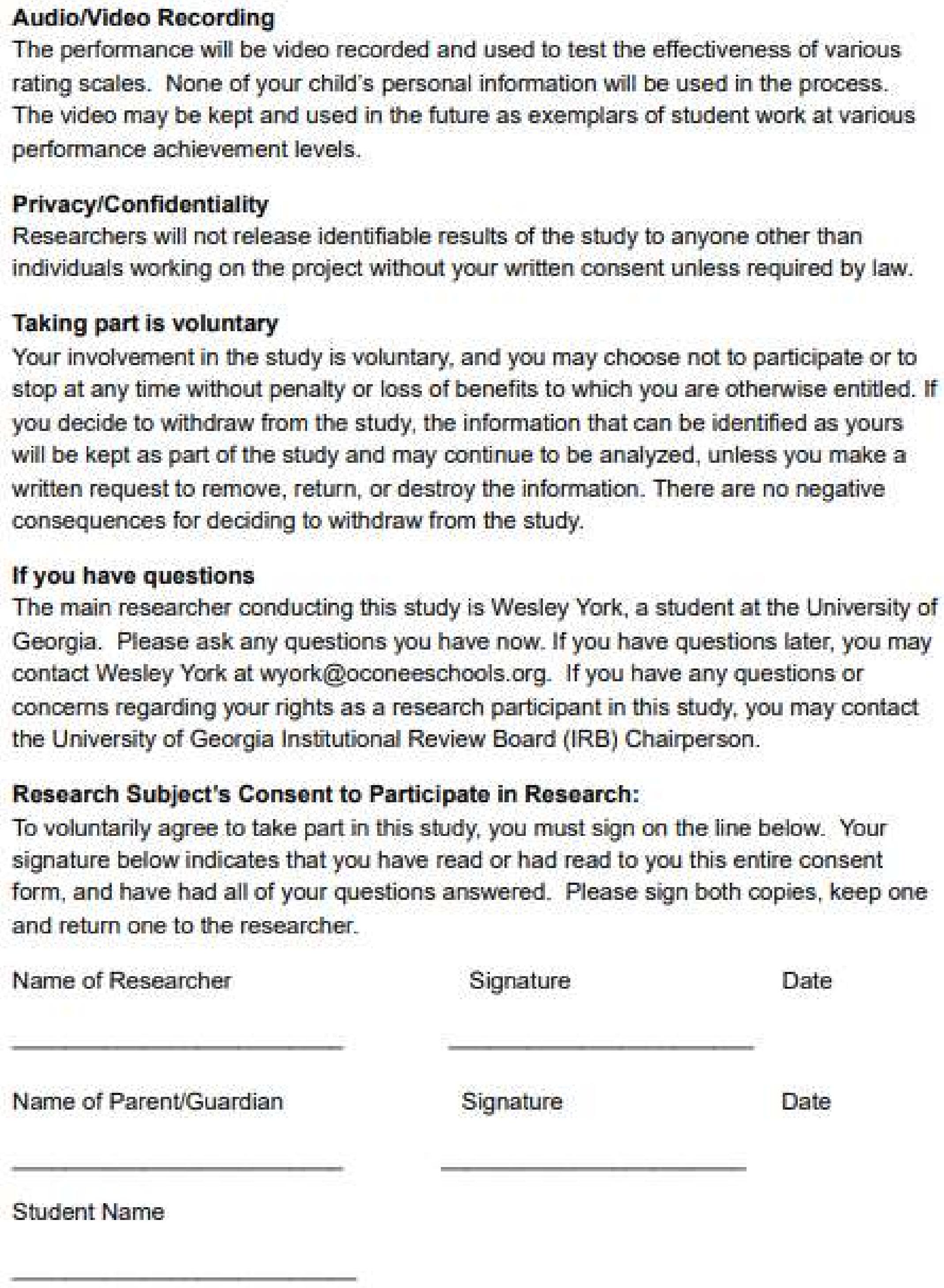
Principal Investigator Wesley York, The University of Georgia

Purpose of the Study
The purpose of the study is to develop and validate a suite of music instrument-specific rating scales used to evaluate secondary-level music performances. The broad goal of this study is to create a program-wide assessment mechanism specifically in order to improve student learning and teacher feedback.

Study Procedures
If you agree to participate, your child will be asked to play a short musical excerpt on their woodwind, brass, or percussion instrument. Music will be provided and will be given to him/her ahead of time. The performance will be video recorded. Your child's participation will happen at Malcom Bridge Middle School either during the school day or at a designated time after school hours that is convenient for you and/or your child. Participants will perform alone with no audience present.

Risks and discomforts
I do not anticipate any risks and/or discomforts from participating in this research.

Benefits
By participating in this study you are providing examples of student work that will be used to develop new grading rubrics which may potentially improve feedback offered by music teachers to their students.



Audio/Video Recording

The performance will be video recorded and used to test the effectiveness of various rating scales. None of your child's personal information will be used in the process. The video may be kept and used in the future as exemplars of student work at various performance achievement levels.

Privacy/Confidentiality

Researchers will not release identifiable results of the study to anyone other than individuals working on the project without your written consent unless required by law.

Taking part is voluntary

Your involvement in the study is voluntary, and you may choose not to participate or to stop at any time without penalty or loss of benefits to which you are otherwise entitled. If you decide to withdraw from the study, the information that can be identified as yours will be kept as part of the study and may continue to be analyzed, unless you make a written request to remove, return, or destroy the information. There are no negative consequences for deciding to withdraw from the study.

If you have questions

The main researcher conducting this study is Wesley York, a student at the University of Georgia. Please ask any questions you have now. If you have questions later, you may contact Wesley York at wyork@oconeeschools.org. If you have any questions or concerns regarding your rights as a research participant in this study, you may contact the University of Georgia Institutional Review Board (IRB) Chairperson.

Research Subject's Consent to Participate in Research:

To voluntarily agree to take part in this study, you must sign on the line below. Your signature below indicates that you have read or had read to you this entire consent form, and have had all of your questions answered. Please sign both copies, keep one and return one to the researcher.

Name of Researcher Signature Date

________________________ ________________________

Name of Parent/Guardian Signature Date

________________________ ________________________

Student Name

____