Gardner-Webb University

Digital Commons @ Gardner-Webb University

Doctor of Education Dissertations College of Education

Spring 2022

A Weighted Individual Performance-Based Assessment for Middle School Orchestral Strings: Establishing Validity and Reliability

Kevin Ward
Gardner-Webb University, kward14@gardner-webb.edu

Follow this and additional works at: https://digitalcommons.gardner-webb.edu/education-dissertations

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation

Ward, Kevin, "A Weighted Individual Performance-Based Assessment for Middle School Orchestral Strings: Establishing Validity and Reliability" (2022). *Doctor of Education Dissertations*. 100.
https://digitalcommons.gardner-webb.edu/education-dissertations/100

This Dissertation is brought to you for free and open access by the College of Education at Digital Commons @ Gardner-Webb University. It has been accepted for inclusion in Doctor of Education Dissertations by an authorized administrator of Digital Commons @ Gardner-Webb University. For more information, please see Copyright and Publishing Info.

A WEIGHTED INDIVIDUAL PERFORMANCE-BASED ASSESSMENT FOR MIDDLE SCHOOL ORCHESTRAL STRINGS: ESTABLISHING VALIDITY AND RELIABILITY

By
Kevin Ward

A Dissertation Submitted to the
Gardner-Webb University College of Education
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Education

Gardner-Webb University
2022

Approval Page

This dissertation was submitted by Kevin Ward under the direction of the persons listed below. It was submitted to the Gardner-Webb University College of Education and approved in partial fulfillment of the requirements for the degree of Doctor of Education at Gardner-Webb University.

Prince Bull, PhD
Chair/Dean, College of Education

Date

Mitch Porter, PhD
Committee Member

Date

Fred Spano, PhD
Committee Member

Date

Kelsey Greer, EdD
Committee Member

Date

Jennifer Putnam, EdD
Associate Dean, College of Education

Date

Acknowledgements

Reflecting on my parents' memory, I give thanks for all their love, support, and guidance throughout their time here on earth. I made a promise to my mother to earn my doctoral degree. After her passing in 2018, I wanted to give honor to her life by finally starting this journey. This path presented many obstacles as well as many celebrations. At my lowest point, I stalled and was ready to give up this goal. Through many talks with my father, he motivated me that giving up was not an option. Unfortunately, my father contracted COVID-19 and lost the battle in 2020. Our very last conversation ended with him asking for clarification about the location and details of my future graduation. With his passing, I had to honor his memory by finishing this journey. Giving me renewed strength every day, my parents' spirits continue to guide my life's journey of making a profound, everlasting difference. When it is my time to join them, I will rejoice and give thanks for all of life's opportunities, experiences, and moments.

I want to express my sincerest appreciation and thanks to my amazing dissertation committee: Dr. Prince Bull, Dr. Mitch Porter, Dr. Fred Spano, and Dr. Kelsey Greer. Another pivotal person in my dissertation journey, Anna Coats, spent many hours reading, analyzing, and constructing feedback on my work. Without her dedication and friendship, this lifelong goal would not be possible.

Last and certainly never least, I want to thank Melissa for all her love, support, compassion, and understanding. While finishing up her master's degree, teaching full-time, and taking care of our family, she continued to be my cheerleader and motivated me to accomplish this lifelong goal. I am blessed to have her by my side as we dedicate our lives to the field of education. She is truly my rock, soulmate, and muse. I love you.

Abstract

A WEIGHTED INDIVIDUAL PERFORMANCE-BASED ASSESSMENT FOR MIDDLE SCHOOL ORCHESTRAL STRINGS: ESTABLISHING VALIDITY AND RELIABILITY. Ward, Kevin, 2022: Dissertation, Gardner-Webb University.

The study established the validity and reliability of a weighted individual performance-based assessment tool within the utility scope of middle school orchestral strings. The following research questions guided this study:

1. What specific string-playing behaviors and corresponding criteria validate a weighted individual performance-based assessment tool for middle school orchestral strings?
2. What are the psychometric properties of the weighted individual performance-based assessment tool in authentic situations?

For Research Question 1, the expert panel and I were able to 100% mutually agree on 10 string-playing behaviors: tempo, rhythm, tone, pitch, intonation, technique, bowing, dynamics, phrasing, and posture that created the DISAT. Being interdependent, these string-playing behaviors are relevant because they encompass every necessary facet of orchestral string performance (Zdzinski & Barnes, 2002). According to Zdzinski and Barnes (2002), an orchestral string performance assessment must evaluate each facet of a participant's playing ability to rate the overall musicianship. Bergee and Rossin (2019) stated in their research that it is important to have various aspects of a performance utilized in a musical assessment.

The DISAT obtained reliability of 0.872 by having enough variance between raters in the authentic situation. Linacre (2015) stated that reliability greater than 0.8 is acceptable to distinguish separation between raters. Combined with the expert panel's 100% mutual agreement on content validity, this proved the DISAT to be a valid and reliable assessment tool for individual performance-based orchestral strings assessment (AERA, APA, & NCME, 2014).

The DISAT can be utilized by districts and middle school orchestral string music teachers in North Carolina. Being a consistent, objective tool, the DISAT can standardize our approach to middle school orchestral string music education assessment (AERA, APA, & NCME, 2014). The data collected by the DISAT could easily track the musical progression of students while giving opportunities for constructive, purposeful feedback.

Keywords: middle school, music education, orchestral strings, assessment, assessment tool, weighted individual assessment, validity, reliability

Table of Contents

Page

Appendices

Tables

Figures

Chapter 1: Introduction

Music is an important aspect of people's lives. Burkholder et al. (2019) revealed the integral factor of instrumental music in society. It appears that the culture of people passing on music trends and techniques has fostered the development of instrumental music (Burkholder et al., 2019). Schools have taken on the study of instrumental music intending to enable learners to use it to improve their academic abilities, social skills, and physical abilities (Hamlin, 2018). Through studying and playing instrumental music, Hamlin (2018) established learners can develop lifelong skills that can assist them in developing into dedicated and intelligent leaders.

Instrumental music is broken down into different categories based on its genre, purpose, and means of sound production. The two main ensemble categories of instrumental music are band and symphony orchestra (Burkholder et al., 2019). According to Burkholder et al. (2019), band ensembles primarily contain instruments that require air to produce sound, and symphony orchestra ensembles mostly contain bowed string instruments with a small section of band instruments within their ensemble.

The symphony orchestra has a subgroup within their ensemble called a string orchestra. String orchestras only have instruments that require their strings to be bowed or plucked to produce sound (Burkholder et al., 2019). According to Burkholder et al. (2019), these groups mainly focus on music that predates the development of band instruments. According to Hamlin (2018), beginning ensemble string instrumentalists start in string orchestra before moving into a symphony orchestra setting, to gain foundational string performance skills. A majority of beginning string orchestra instrumentalists start in middle school to develop that foundation (Wu et al., 2016).

According to Wu et al. (2016), this foundation requires an appropriate assessment to gauge the overall performance skills and highlight areas of needed growth.

Assessment of string instrumentation tends to fall into two main divisions: technique and expression. According to Wu et al. (2016), technical aspects of assessment usually consider the production of sound mechanics. The assessment contemplates the coordination of movement and fluidity and the proficiency of technique in performance (Wu et al., 2016). On the other hand, the expressive skills reveal themselves in the musical expression or interpretation of the piece as presented by the student (Mazur & Łaguna, 2017).

The two components of string assessment have their strengths and weaknesses when it comes to evaluation. According to Wu et al. (2016), technical skills can be objectively defined. Due to the assessor's personal opinion and musical taste, the performance's expressive elements are subjectively perceived and more difficult to measure (Wu et al., 2016). According to Wu et al., the positive evaluation of the expressive elements of the performance of a musician depends on whether the evaluator shares their interpretation of the musical composition and whether the interpretation is persuasive. The assessment of expressive competencies is usually overlooked and oftentimes dismissed by scholars (Meissner, 2017). In instrumental music learning, more distinguished and specific performance qualities are note accuracy, articulation, rhythmic accuracy, appropriate sense of trend, or effective dynamics (Wu et al., 2016).

Statement of the Problem

Middle school orchestral music teachers need valid assessment tools to provide evidence of student growth and achievement. With current educational reforms, teachers are expected to cite evidence of student achievement (Mazur & Łaguna, 2017; Simones, 2017). Since mandatory standardized testing does not exist for instrumental music, teachers must create their own assessments that communicate student growth (Mazur & Łaguna, 2017). Due to the absence of an effective assessment strategy, instrumental music programs lack teacher objectivity, produce unreliable data, and result in the crippling loss of funding (Simones, 2017). To save instrumental music education programs from being eliminated from the curriculum, valid and reliable assessment tools need to be in place to collect and communicate student achievement data (Mazur & Łaguna, 2017; Simones, 2017).

There are no state-mandated standardized performance level indicators in instrumental education assessment. For instance, it is unclear what constitutes a rhythmically correct performance (Wu et al., 2016). Criteria-particular rating scales for music performance are regarded as beneficial diagnostic tools in assessing achievement in playing an instrument (Meissner, 2017). The evaluators use them to clarify the criteria provided, which describes the performance level (Mazur & Łaguna, 2017). In this manner, the evaluators observe and note their experience of the performance, what impressed or dissatisfied them, or the level to which they disagreed or agreed that the execution was closer to an unspecified standard (Mazur & Łaguna, 2017). According to Mazur and Łaguna (2017), multidimensional assessment rubrics are effective performance assessments for two reasons: They incorporate each skill level’s description, and they allow for a general assessment of playing a certain instrument while considering numerous elements of the performance. While the scales incorporate criteria for assessing the expressive and technical aspects of the performance, rubrics allow evaluators to examine the performance more reliably (Mazur & Łaguna, 2017). The rubrics aid in performance evaluations of a diverse set of instruments at many levels of education (Mazur & Łaguna, 2017).

A critical weakness in current assessment methods is the impact of the assessor's opinion on the rating. Mazur and Łaguna (2017) documented the common tools used to evaluate performance are grounded entirely on the evaluator's personal impression as to the character or quality of the performance. This approach results in the evaluator proclaiming whether the overall skills of the performer are above average, average, poor, or below average (Mazur & Łaguna, 2017). According to Mazur and Łaguna, the assessment may encompass the general impression of the whole performance or certain aspects of the dimensions such as the impressions on the intonation or technique. Evaluators can at times use a Likert scale in assessing musical performance because it allows the judges to have some level of agreement concerning the various categories of the performance to be evaluated (Mazur & Łaguna, 2017). Using these scales, judges rate on a continuum in demonstrating their agreement level on certain specific performance aspects, such as rhythmic accuracy (Mazur & Łaguna, 2017). According to Mazur and Łaguna, the scales are usually attributed by a validity degree, but there are misgivings regarding their relevance. These scales do not provide particular criteria descriptions or weighted importance when assessing skill (Mazur & Łaguna, 2017). According to Mazur and Łaguna, weighted importance involves understanding that some components are more crucial than others in fundamental skill mastery.

Purpose of the Study

The purpose of this study developed, validated, and tested a weighted individual performance-based assessment tool that provided objectivity in the assessment process for middle school orchestral strings. The assessment tool objectively measured skill level with specific observable items. This provided both the teacher and student with valid documented achievement data. That documented data provided specific feedback, developed opportunities, and created a consistent process to track individual student growth. Analyzing the data, teachers identified trends and adjusted classroom instruction to meet the students' needs. This reflective practice identified areas needed for teacher professional development and growth. Utilizing the data, teachers provided consistent achievement reports with all stakeholders and advocated for resources and materials to strengthen areas of instructional weakness.

Research Questions

The study established the validity and reliability of a weighted individual performance-based assessment tool within the utility scope of middle school orchestral strings. The following research questions guided this study:

1. What specific string-playing behaviors and corresponding criteria validate a weighted individual performance-based assessment tool for middle school orchestral strings?
2. What are the psychometric properties of the weighted individual performance-based assessment tool in authentic situations?

Methodological Approach

This study utilized a quantitative approach, methods, and research strategies (Creswell & Creswell, 2018). The quantitative approach utilized the Polytomous Rasch Model. This model's objective measured ability through analyzing responses to constructs that are scored with successive numbers (rating scale).

Quantitative methods were utilized to help construct the assessment tool and tested its validity and reliability. Quantitative research strategies in this study obtained interrater reliability and analyzed data from a pilot study and final item pool study.

Definition of Terms

The study utilized various technical terms. Defining these terms helped clarify the intent and purpose of their usage. The following important terms were included in this study:

COVID-19 Pandemic

A worldwide pandemic caused by the coronavirus disease that impacted schools by limiting in-person learning opportunities for students (World Health Organization, 2021).

Expert Panel

Panel responsible for the construction and validity of the weighted individual assessment tool (Creswell & Creswell, 2018).

Individual Assessment

Evaluation of a singular participant's skills that is not impacted by another singular participant's evaluation data (Loughran & O'Neill, 2016).

Pilot Assessors

Responsible for the reliability of the weighted individual assessment tool by administering the validated assessment tool and participating in structured interviews (Creswell & Creswell, 2018).

Reliability

An evaluation tool to generate the same results over a given amount of time (Wesolowski & Wind, 2019). According to Wesolowski and Wind (2019), the definition emphasizes the standards in replications that mirror certain interpretations and uses of test scores.

Validity

A fundamental aspect in designing and evaluating assessments. Validity suggests the level to which theory and evidence support the interpretations of the scores from an assessment for the proposed assessment's use (Wesolowski & Wind, 2019).

Weighted

Analytic assessment approach where individual concepts are judged more heavily compared to others (Loughran & O'Neill, 2016).

Organization

This dissertation is organized into five chapters: Introduction, Literature Review, Methodology, Results, and Discussion. The first chapter includes the statement of the problem, purpose of the study, research questions, methodological approach, definition of terms, and organization. The second chapter includes an alignment of various sources, which creates the foundation for this study. The third chapter includes a detailed three-step procedure that supports answering the research questions. The fourth chapter provides the collection and analysis of data and alignment to the research questions. The fifth and final chapter includes the summary of findings, recommendations for research, and the conclusion.

Chapter 2: Literature Review

Overview

Instrumental music is an important aspect of human life and is utilized in different important life events including graduations, weddings, and funerals. Whether it is conducted as a tradition or for the comfort it provides, music is present and accompanies people in all life stages (Tan, 2017). Music, as an academic discipline, is available in schools through which students simultaneously nurture physical abilities, social skills, and academic capacities, which is particularly seen in instrumental music classrooms (Tan, 2017). According to Tan (2017), instrumental music education refers to any musical learning scenario where instruments are the primary music-making medium.

Instrumental music education comes in many forms, encompassing a variety of instruments and group settings. The lessons may occur within or outside school, and the common instruments taught in western schools comprise keyboard instruments (accordion, piano, organ), string instruments (violin, guitar, cello, harp), and wind instruments (flute, saxophone, clarinet, horn, tuba, bassoon, etc.; Montemayor et al., 2018). According to Montemayor et al. (2018), teaching and learning instrumental music can occur in an orchestra, a band, or another musical ensemble.

Given that each of these groups encompasses different instruments, instructors must customize instruction using applied tools while guiding the entire class. Primarily, students usually reveal their psychomotor and cognitive learning by way of performance (Johnson & Fautley, 2017). Hamlin (2018) attributed the development of cognitive abilities to extensive learning and engaging in music. Bowie (2018) also noted that learning music is akin to learning a new language. To demonstrate the learning of students in instrumental music, assessment is necessary for the music classroom (Bowie, 2018; Hamlin, 2018). Developing assessment systems that allow the students to reveal their development and motivate skill development mastery is incredibly significant to learners, educators, parents, and school administrators (Bowie, 2018; Hamlin, 2018).

By playing instrumental music in ensembles, students gain many lifelong skills which enable them to become dedicated and intelligent leaders (Tan, 2017). According to Tan (2017), the student's cognitive abilities expand and grow when they study music for long periods of time. A study has shown improved test scores compared to those who do not participate (Montemayor et al., 2018). Tan also supported those students learn a new language by learning and reading music, which is more complex than any other language. According to Tan, they learn the significance of working in a group from their peers as much as they do from the teachers. Their fine motor control also improves by refining their aptitudes in playing instruments like the piano (Tan, 2017). Students also learn dedication, organization, and teamwork as they develop interactive awareness of the happenings around them (Tan, 2017).

History of String Orchestra

String orchestras are an arrangement of musical instruments from the string family set up so they can all create music through the guidance of a conductor. Such instruments include the violin, viola, cello, and double bass (Radice, 2012). Radice (2012) stated violins are usually grouped into two sections, which play different musical parts. Violas, cellos, and double bass are singular groupings (Radice, 2012). The string instruments are characteristically similar in structure but differ in size, allowing for the calibration of the desired pitch (Bukofzer, 1949). In their arrangement, the instruments are set in sections with section leaders to allow for the best quality of sound from the stage to the listening audience (Ma & Hall, 2018). According to Ma and Hall (2018), the science of it all is in how they are arranged to create specific sounds that complement each other into the musicality desired by a conductor.

The baroque era provided an opportunity for practicing methods, which provided the foundation for the future of music. Baroque music was the name given to music composed during this era which, though unclear, spanned from 1600 to 1750 (Bukofzer, 1949). According to Radice (2012), it was the era in which composers achieved a scintillatingly great level of success in how they came up with and delivered music. A characterization of exaggerated movement and detail was used to the ostentation of drama and a measure of exuberance in art forms, particularly music (Bukofzer, 1949). Baroque composers fixated on musical arrangements that created sonic tranquility through the use of the pitches created by these string instruments (Radice, 2012). According to Bukofzer (1949), the baroque sonata was an ideal form for composers to practice techniques for making compositions. Music in the baroque era first gave way to renaissance music, followed by classical music’s contribution to the present-day string orchestra (Bukofzer, 1949). The baroque era was characteristic of great developments that were foundational to the later development of classical music (Radice, 2012).

String Instruments

There are four bowed string instruments: violin, viola, cello, and double bass. The concept of the musical application of a vibrating string was known long into the annals of history (Nelson, 2003). According to Laird (2004), attention is drawn to ancient civilizations that had long had a variety of stringed instruments used to produce music and make melodies. These included things such as the lyre and the harp (Laird, 2004). Paintings that date back to the 13th century reveal what seems to be a violin or an evolution of the byzantine-Greek lira (Laird, 2004). Nelson (2003) pointed to the fact that it was known, to the Egyptians for instance, that using strings of varied length on one instrument would yield a pitch range that is also varied. According to Nelson, many manuscripts denote the presence and existence of plucked and/or struck stringed instruments such as harps. Bowed instruments, Nelson stated, appear to have been a development that came much later.

The Violin. The violin is the smallest and highest pitched (soprano) instrument of the bowed string family. Elements, such as tuning pegs, were utilized in rudimentary form long before the official advent of the violin (Nelson, 2003). Around the 16th century, the violin in its present form begins to appear even as the viola is still present in the baroque period (Nelson, 2003). According to Nelson (2003), the violin's structure has equally gone through a morphing of sorts. The wood paneling of the half pear-shaped body of the violin was later found to give better resonance if constructed of many strips of wood glued together (Nelson, 2003).

Before arriving at its present-day form, the violin had a history of revisions. The violin was played over wide expanses in the European region including Germany, Italy, and France (Nelson, 2003). Nelson (2003) stated the violin was used primarily for dancing in England. These dances were done in huge central courts (Nelson, 2003). Court masques in the monarchy were a prevalent form of distraction where violins were a prominent spectacle among the bands (Lindley, 1984). The violin remains an important member of a stringed orchestra, contributing a great number by composition (Lindley, 1984; Nelson, 2003). The final, most common form of the violin is between 28 and 36 cm long with four strings, tuning pegs, and is played on the shoulder, primarily with a bow (Nelson, 2003).

The Viola. The viola is the next to highest pitched (alto) instrument of the bowed string family. A similar-looking instrument to the violin, the viola is recorded to have existed in the 14th century (Nelson, 2003). They serve the purpose of filling in the middle space between violins and the lower strings (Campbell & Campbell, 2010). According to Campbell and Campbell (2010), violas are the tenors in the violin family and existed in two forms throughout history, much like the other instruments in the violin family. Viola da braccio was an arm viola and viola da gamba was a leg viola (Campbell & Campbell, 2010).

While the viola da braccio and the viola da gamba had a significant difference in the posture in which they were played, there were other notable differences as well. According to Campbell and Campbell (2010), the construction of either differed from the other with the arm viola being fretless at the fingerboard with a rounded backplate and characteristic low ribs. It had an f-shaped sound hole and its neck, bridge, and scroll, allowing its players to independently bend each one of its strings (Campbell & Campbell, 2010). According to Campbell and Campbell, the viola da gamba had five or seven strings with high ribs and a straightened back. Its uncarved bridge and frets made it possible for its players to play with two or more strings simultaneously (Woodfield, 1988). Woodfield (1988) went on about how the viola da gamba produced a mellow soft sound when played, while the da braccio created a more powerful sound. The music in the 17th century demanded a powerful range from instruments that allowed for better volume, and it was in this respect that the viola da braccio overtook the viola da gamba (Campbell & Campbell, 2010; Woodfield, 1988).

Conscious efforts were made to improve the viola. At the turn of the 19th century, the viola received several important changes that have defined its sound and structure to this very day (Campbell & Campbell, 2010). According to Woodfield (1988), heavier strings were used to make violas. An increased tension was achieved by wrapping gut strings with silver and other metals and effectively making for a better projection of sound (Woodfield, 1988). According to Holman (2013), the length was added to the neck of the viola, and it was tilted at a slight backward angle to embrace its body. Holman said an improvement in the strength of the viola was done by fortifying the body, brass bar, and bridge. The size of the viola was later reduced in response to players' needs for a manageable stance while playing. In the 18th century, violas doubled cellos in an orchestra and only obtained distinctive roles under specific composers (Holman, 2013). Over time, they gradually assumed an heir of independence in the role they played (Campbell & Campbell, 2010; Holman, 2013). The final, most common form of the viola is between 37 and 43 cm long with four strings, tuning pegs, and is played on the shoulder, primarily with a bow (Holman, 2013).

The Cello. The cello is the next to lowest pitched (tenor) instrument of the bowed string family. Italian Andrea Amati, the inventor of the cello in the 16th century, began with the cello considerably larger than the present-day instrument (Laird, 2004). The progress towards smaller sized cellos is traced back to 1700 (Laird, 2004). Laird (2004) noted the development of smaller sized cellos made it easier for cellists to play by significantly reducing the tension needed in their left hand.

The cello's purpose has evolved over its history, from that of an accompanying instrument to one that can take center stage. According to Laird (2004), cellos were played as accompanying instruments. For most of the 17th century, cellists were able to hang the instrument around their necks and play it while standing and during processions (Laird, 2004). At the tail end of the 17th century, the cello began to gain solo pieces in their repertoire (Laird, 2004). According to Laird, this was because they had a rich, crisp sound that had a wide range to explore. The cello was affected by cultural shifts that occurred during the time it was undergoing development as an instrument (Laird, 2004). The French Revolution, for instance, at the turn of the century, shifted the focus of sound created for exclusive peers to sound created for large audiences (Laird, 2004). Subsequent changes reflected the need for an improved volume, clarity, and responsiveness (Laird, 2004). The cello enjoys a wide scale of enthusiasm from a great number of people. (Laird, 2004). According to Laird, the cello's alluring versatility ensures people will continue to perform great works of art on it for years to come. The final, most common form of the cello is between 69 and 76 cm long with four strings, tuning pegs, and is played with a floor end pin, primarily with a bow (Laird, 2004).

The Double Bass. The double bass is the largest and lowest pitched (bass) instrument of the bowed string family. Whether played by bow or finger, the double bass is an integral member of a sizeable number of genres of music (Planyavsky, 1998). Such genres include classical music, jazz, bluegrass, and country (Askenfelt & Jansson, 1992). According to Planyavsky (1998), the exact origin of the double bass is a disputed question whose answer remains unclear. Many alterations have characterized the instrument's rich history, and placing a finger on an exact beginning has proven to be difficult (Askenfelt & Jansson, 1992). According to Askenfelt and Jansson (1992), centuries of change have affected the double bass tuning, dimensions, and design to the point where it has never undergone a complete standardization in terms of either shape or construction. This has made the shape and appearance of the double bass a widely irregular phenomenon (Askenfelt & Jansson, 1992).

The double bass is a versatile instrument, with varying numbers of strings, methods of playing, and open string tuning. Historically, there were up to 50 different tunings available for the double bass and a string arrangement that went up to six numbered strings (Planyavsky, 1998). According to Planyavsky (1998), this is a great departure from the commonplace four- and three-stringed double basses. Throughout the history of the instrument, a bow is primarily utilized in classical settings (Askenfelt & Jansson, 1992). According to Askenfelt and Jansson (1992), many popular present-day playing techniques take the plucking and/or slapping form. Slapping is characterized by a pulling away of the strings from the fingerboard thus allowing them to bounce off (Planyavsky, 1998). According to Planyavsky, this method gives a beat and pitch to the performance. Artists in the early 20th century came to know of the beat and percussive aspect of the instrument (Chevan, 1989). According to Chevan (1989), it was established that the style helped produce a stronger baseline and a more projected sound, and artists used this to alleviate the very real shortfalls of recording equipment at the time.

Even though the history of the bass is disputed and heavily overlapped with other members of the violin family, it does not negate its present-day dominance as an instrument of powerful range and importance (Chevan, 1989; Stowell, 2001). It has been a hallmark of versatility in the instrumentation of music over a wide range of styles and genres of music and will remain so for a long time to come (Stowell, 2001). The final, most common form of the double bass is between 95 and 115 cm long with four strings, machined heads and gears for tuning, and is played with a floor end pin, primarily with a bow or finger (Stowell, 2001).

Considerations of Sound Production

Stringed instruments depend purely on the strings hoisted over their carefully crafted body stratum to produce sounds of different pitches. The strings on these instruments are made from a variety of materials ranging from plain gut to nylon (Woodhouse & Lynch-Aird, 2019). According to Woodhouse and Lynch-Aird (2019), a stringed instrument is as important to its player as a paper and pen is to an author. These two are the reasons either can tell stories through music and writing (Woodhouse & Lynch-Aird, 2019). Unlike writing, stringed instruments need to be played by their owners for a considerable amount of time before they become accustomed to their instruments (Vaiedelich & Fritz, 2017). String assessment is about determining when to adjust and complete changes to the strings of an instrument (Hopkins, 2014).

Visually inspecting the strings can tell the musician a lot about their playability. According to Hopkins (2014), fraying strings are a direct indication of the need for a change. It is characteristic of very noticeable degeneration of the strings, particularly at the edges (Hopkins, 2014). Though rarely the first sign to occasion changing strings, its appearance is an indicator of a languishing state (Schemmann et al., 2020). Continued use will lead to a higher probability of injury on the player's side (Schemmann et al., 2020). According to Schemmann et al. (2020), this usually involves an extremity.

Beyond physical wear on the string, the condition of the string can also impact the quality of the tone of the instrument. Hopkins (2014) mentioned that sound quality is a more subtle way of picking up a required assessment of a string. Subtle changes in the quality of sound over a noted period are a great indication that some changes need to be made (Hopkins, 2014). According to Hopkins (2014), it could be a tuning issue or a replacement issue. To be able to pick up such changes, the player of the instrument needs to have a good ear and a connection with the instrument through constant playing and touch (Schelleng, 1973). According to Schelleng (1973), these subtle changes are often recognizable through many physical constraints. If, for instance, a particular resonance is becoming harder to achieve or if a bit more pressure has to be employed to obtain a particular tonality to a sound, changes need to be made (Hopkins, 2014; Schelleng, 1973). Tuning difficulties where they were not presently experienced are also an indicator of underlying issues in the strings (Hopkins, 2014). These manifest as an inability to remain tuned for a long period. From this point onward, the quality of the string and its performance begin to deteriorate (Hopkins, 2014; Schelleng, 1973).

The type of string one chooses to use will go a long way in determining the sound achieved. Gut core strings are made from sheep intestines and are known for the warmth in their rich tones (Schelleng, 1973). According to Schelleng (1973), they were a favorite before the advent of the synthetic string. Boasting an unparalleled tonal depth, gut core strings have no problem creating full volumes from their tension levels (Schelleng, 1973). According to Hopkins (2014), they are menacingly susceptible to changes in humidity and temperature in their surroundings and therefore require constant retuning; they are also ridiculously expensive.

An alternative to gut core strings is steel core strings. They were the only other alternative to gut core strings before the synthetic strings (Hopkins, 2014). According to Hopkins (2014), they provided great pitch stability and were not affected by atmospheric conditions. Their ease of tuning made them a favorite among jazz, country, and bluegrass players (Schelleng, 1973). According to Schelleng (1973), steel core strings lacked warmth and richness in tone that were paramount to genres such as classical music and were not used for them.

Synthetic core strings bring a careful mixture of the best of both worlds between gut core and steel core strings. According to Hopkins (2014), these strings stay in tune for a long time, and they bring warmth and prolonged pitch to music. Tension and gauge of strings matter in curating specific sounds and pitches (Hopkins, 2014). While there are no specific standards for specific sounds, it is important to experiment until a desired gauge and tension works to produce specific sound tonalities that are desirable (Hopkins, 2014; Schelleng, 1973).

Early Tuning Instruction

There is a gap in research related to acquisition of rudimentary musical skills, as studies have focused on string technique and pedagogy. Hopkins (2013) examined the experiences of teachers while teaching the tuning of stringed instruments in elementary and middle school group classes. Hopkins (2013) noted that independent tuning skills are important and fundamental for string players; therefore, teachers should be focused on teaching their students to tune independently if they are to be successful in any music ensemble (Hopkins, 2013). Hopkins (2013) emphasized the need for string players to learn instrument tuning to a pitch standard within a group class; however, they face difficulties related to tuning the remaining strings to perfect fifth intervals or unison intervals using harmonics (Hopkins, 2013). Hopkins (2013) believed these complexities in the group class make it an almost impossible and slow task to teach students instrument tuning. Hopkins (2013) found that teaching students tuning skills is complex and takes time, which requires aural and physical skills. Since teachers vary in the activities and time spent on instruction, students in school orchestra programs develop tuning independence (Hopkins, 2013). The string class instructors often refrain from teaching physical tuning skills in the first year and prefer waiting for the second or third year, choosing instead to teach aural skills initially (Hopkins, 2013). According to Hopkins (2013), instructors also prefer using verbal instruction to teach tuning. Overall, Hopkins (2013) found a gap between common practice and literature on effective teaching for developing students' tuning independence.

Students tend to learn tuning independently rather than expecting their instructors to guide them in tuning their instruments. Hopkins (2013) applied a questionnaire method to assess teacher practices and beliefs about teaching stringed instrument tuning. Hopkins (2013) developed the tool after reviewing pedagogical and research literature and consulting a panel of three elementary and middle school orchestra teachers, social research and questionnaire design university experts, and the music education faculty of a university. Consequently, the resulting survey tool is relevant and appropriate and sets the right parameters for evaluating the research question (Hopkins, 2013). Hopkins (2013) believed involving the teachers lends credibility to the methodology because the questions designed will collect accurate and detailed information about their beliefs and experiences. Hopkins (2013) indicated he strengthened content and constructed validity of the questionnaire items by involving external review by string education experts. The research yields valuable information from experts and instrument tuning teachers (Hopkins, 2013). According to Hopkins (2013), there are important lessons about why the instructors delay teaching physical tuning skills while emphasizing aural skills during the initial stages of learning.

Students need to understand the approach their teachers take in their instruction of aural and physical skills related to instrument tuning. Hopkins (2013) believed they need to be patient for the year the instructors find it appropriate to instruct them about physical tuning or adapt their independent learning by involving external experts. Moreover, the instructors can learn how various techniques work for their students when teaching them about instrument tuning (Hopkins, 2013). Hopkins (2013) especially stated that instructors' teaching approaches are different, and it is crucial to learn, which is more effective. According to Hopkins (2013), experts can also learn from the research to develop and propose programs and standard curriculums that can inform when teachers begin instructing physical and aural tuning skills.

Instrumental Music Performance Assessment

Musical performance can be considered the phase in the musical process in which previously identified ideas are transmitted to the audience. Music performance is often viewed as an interpretive art (Bergee, 1994). According to Bergee (2007), the performer has a duty to engage in various activities which tend to determine the components of the music they are performing. These activities can include melody, rhythm, sound, and expression (Bergee, 2007). Moreover, music performance tends to offer an extensive repertoire of both motor and cognitive skills (Bergee, 2004, 2015). For a musical performance to be effectively assessed, an emphasis must be placed on aspects that contribute to the formation of a conceptual interpretation, retrieval from memory of the musical structures, and the transformation into the right motor actions (Bergee, 2007). According to Bergee (2006), both structural and emotional aspects regarding the performer's conceptual interpretations are often put into consideration during a musical performance assessment; thus, it may be argued that there are perceptual consequences of music performance such as effective interpretation communication, structural ambiguities resolution, and meeting audience expectations (Bergee, 1994, 2006, 2015).

Music encompasses learning the use of diverse instruments. In a typical United States school setting, students are often assigned instruments and are required to work together with their classmates to create music. Consequently, the students get a chance to adopt behavioral skills such as leadership, teamwork, dedication, organization, and interactive awareness (Johnson & Fautley, 2017). Individual student assessment is distinct from what might be anticipated in a typical music classroom (Music, 2019). Different countries have different approaches to pedagogic assessments in instrumental music education, which correspond with their respective national systems and standards. For instance, the U.S. implements protocols for individual student assessment in instrumental music (Johnson & Fautley, 2017), thus the significance of differences in instrumental music educational approaches cannot be understated.

In instrumental music education, assessment is a significant component, but this proves to be a challenging aspect for numerous instrumental music instructors. Faced with restrictive instructional time, minimal or little training in assessment, and large sizes of classes, instrumental music assessment continues to face numerous challenges (Music, 2019). Organizations implement measures to enhance student learning and foster achievement in instrumental music through better practices of assessment (Johnson & Fautley, 2017).

The debate about instructional music educational assessment has become an important area of focus in the academic realm in the latest years. Notwithstanding the existing academic culture where assessment and data-based instruction are at the core of the philosophy of many education leaders, little empirical data are carried out to assess instrumental music education (Mazur & Łaguna, 2017). Many journals discussed assessment, but most of this discussion focused on the procedures of evaluating teachers or corporatizing student assessment on English and math (Simones, 2017). According to Simones (2017), the assessment of student achievement in instrumental music is a relatively new discipline with few academic inquiries approaching the topic systematically. Various scholars in the music education field produced assessments on instrumental music by addressing future measurement concerns and evaluating music experiences (Dayal, 2017; Hallam, 2019; Mazur & Łaguna, 2017; Simones, 2017). The research is often motivated by numerous trends, countertrends, and future trends, which stimulate theory.

Best practices drive the instrumental music assessment process. St. Pierre and Wuttke's (2017) study focused on the standards grounded on the grading activities among practicing music educators. In their results, the scholars documented a fairly balanced distribution of school population, school size, grading practices, the experience of the director, and the social-economic environment of the school. The music educators used grading criteria comprised of participation and attendance of performance, performance-grounded tests (91%), and daily attendance of rehearsal (82.1%; St. Pierre & Wuttke, 2017). Similarly, the study by Simones (2017) established that band directors in the study believed the available amount of classroom time was a great issue affecting assessment strategies. Nevertheless, the directors also believed that the most significant issues affecting their practice of assessment were in their perspectives of music education and the general set class objectives (Simones, 2017). Overall, the most critical aims of assessments were to recognize the needs of the students, deliver feedback, and have more awareness of the general instructional and program direction (Simones, 2017; St. Pierre & Wuttuke, 2017).

Comprehensive music skills assessments, standardized cumulative testing, are not widely used in the music education field. Wesolowski et al. (2016a) reported that string teachers usually use the teacher-provided verbal critique, student evaluations, and teacher-rated rubrics as the most common methods of assessment. The study also documented that string educators infrequently used comprehensive music skills in assessment. These include music history, composition, interdisciplinary assignments, improvisation, and portfolios. Hopkins et al. (2017) similarly established that in successful string programs, educators usually use student reflections, written assessments, rubrics rated by the teachers, learner evaluations, history assignments, portfolios, music theory, and student-rated rubrics.

The approach to assessment is different between instrumental and vocal music education. Brockmann-Bauser et al. (2018) documented some fundamental differences between instrumental and vocal practices of assessment. The two assessments are usually grounded on performing in huge ensembles and are influenced by almost similar circumstances like performance expectations, size of the class, and so on. Nonetheless, Simones (2017) showed that more practices of every type of music may have more influence. For example, middle school music directors provided substantially more weight to the music knowledge written assessment than the middle school’s instrumental director, while Simones found no significant dissimilarity in the amount of attention high school instrumental directors give to music awareness.

Strategies

Assessment is essential in today's classrooms. More than ever before, designing assessment systems that allow students to demonstrate growth and motivate skill development mastery is incredibly significant to teachers, administrators, students, and parents (Carey, 2017). Carey (2017) presented the latest practices of assessment used by teachers to evaluate instrumental music at different sites in highly rated state festival ensembles.

Research has been done to determine the types of assessments used as well as the perceptions of these assessments. Carey (2017) sought to discover the types of assessments band teachers use to guide student achievement and provide sufficient feedback for growth. Using a learning for mastery framework, the study also considered the way band directors perceive these assessments and found that different types of assessments are normally used, which provided feedback to students and assisted to provide an educational program for students grounded on data and were appropriate to skills taught in class (Carey, 2017). Carey established that evaluators and administrators can develop practices for individual assessment, which motivates students to grow and eventually show the growth of students to stakeholders.

Teachers must practice their assessment strategies to gain the proper perspective of their students. Meissner (2017) discussed the strategies of teachers in instructional music learning and reported about the views of secondary and elementary school teachers on the idea of assessment and its significance in classrooms. According to Meissner, practice had a direct link to the perspectives of the teacher on assessment. The practice of the teacher or school on assessment influenced their perspectives on the use of assessment. Juchniewicz (2018) revealed the difference in the accreditation of student and instructor accountability. Juchniewicz provided some insights on evaluation in the instrumental music classrooms. Given that the perspective of a teacher on the assessment directly links to their currently used assessment practice, it would be sensible to extrapolate that the teacher’s assessment practices can be linked to the way they were assessed as a student (Juchniewicz, 2018).

Implementation of assessments varies, constituting a need for further study. According to Carey (2017), how assessments are carried out influences the learning outcomes for pupils. Carey compared assessment tactics to the learner's educational outcomes in the literacy objectives from the fourth grade to high school. Literacy skills in language arts directly relate to the skills needed in the music classes, because similar decoding competencies exist in both subjects (Vaughan, 2019). Carey established that teachers in the United States spent more energy and time using traditional pencil and paper evaluative strategies than their counterparts in England, who were predisposed to using more oral practices to examine student learning. Vaughan (2019) also reported on the differences in the learning outcomes for students by positing that United States instructors also preferred the multiple-choice alternative of assessment like their overseas counterparts. Overall, the results generated by Vaughan were inconclusive concerning the types of strategies used in assessment, with suggestions calling for more studies in this area. Music assessments should model the different kinds of assessments as they would apply to the nature of learning, which the students would encounter (Carey, 2017). According to Carey, these would not just be successful in the music classroom but could also be similarly used in other instrumental music. Multiple-choice testing tactics can be used to translate student comprehension concerning the historical background of the arts under study (Vaughan, 2019). Oral and performance strategies are used virtually every day in successful instrumental music classes and are effective measures of student success if an appropriate grading criterion and rubric are in place for teachers and students to comprehend the quality of the presented work (Simones, 2017). The use of true quality evaluations in the music classroom will guarantee the excellence and prosperity of students.

Instrumental music education assessment strategies differ from country to country. A global comparative study by Johnson and Fautley (2017) sought to examine the difference in the assessment of classroom instrumental music learning in the U.S. and the UK. Similar to the study by Loughran and O’Neill (2016), Johnson and Fautley established that the context of assessment differs from country to country. For instance, in the United States context, music education pays attention to growing the elementary understandings of the students as well as their interaction with melody, harmony, rhythm, timbre, form, texture, and dynamics (music elements), and the instrumental performance ensemble medium. Johnson and Fautley indicated that the focus of instrumental music education is the performance as the driver for student engagement in studying music. Loughran and O’Neill also asserted that the corresponding objectives and student learning outcomes vary, but teachers of instrumental music often endeavor to facilitate lasting musicianship and positive social experiences, musical achievement through performance, and autonomous musicianship. On the other hand, in England, the situation is quite different compared to schooling in the United States (Johnson & Fautley, 2017). In England, the entire learning of class instrumental music occurs in elementary schools under the scheme of “wider opportunities” and is also identified by names like "first access," "whole-class ensemble teaching," and "whole class instrumental vocal teaching" (Johnson & Fautley, 2017). For simplicity purposes, England’s whole class teaching ensemble is a prevalent activity, which has been occurring for some years in some places, but its use in pedagogic assessment has increased because of the National Plan for Music Education (Johnson & Fautley, 2017). The National Plan for Music Education developed the music hubs concept in the UK, which is a local area grounded group of organizations.

In the UK, several answers have been given to the query of for whom the evaluation is intended. One strand of this answer is the intention to assist with student improvement (Johnson & Fautley, 2017). Nonetheless, in the UK, the presence of league tables documented in the local and national media suggests that schools are apprehensive that their public-facing evaluations are often at their highest to facilitate league table achievement (Johnson & Fautley, 2017). According to Johnson and Fautley (2017), present music does not figure much in these; head teachers know that it could be detrimental to have away time from English and math, the core subjects. This could decrease the accessibility of opportunities for music learning for the involved students (Johnson & Fautley, 2017). In the United States, superintendents of schools also track academic grades and link success measures like the state test scores (Moss et al., 2019);

however, concerning instrumental music, the intensity of scrutiny for academic ratings differs from that of England. In fact, in many United States schools, music is non-examined which contrasts English, math, and other academic disciplines (Moss et al., 2019). On the question of the effect of lacking non-musical classes on educational attainment, Johnson and Fautley found that participating in “pull-out” programs to permit learning instrumental music does not adversely impact student education. Most often, music educators in the United States in charge of instrumental ensembles pay attention to festive scores and MPA ratings because these provide a significant external impetus in the form of trophies as well as other extrinsic motivations.

Verification of theoretical frameworks has been completed using exploratory and confirmatory factor analysis of high school concert band performance. To achieve credible results, Bergee (2015) implemented the exploratory as well as confirmatory factor analysis to be able to effectively verify the theoretical structure. From the exploratory factor analysis, Bergee (2015) was able to identify that a structure is comprised of three interlinked fundamental factors. The components did not have a similar factor loading across the two performances (Bergee, 2015). In the first performance, the fundamental aspects that accounted for most of the variance were tone quality, rhythm, and intonation (Bergee & Rossin, 2019). On the second performance, the fundamental aspects that accounted for most of the variance were intonation, expression, and rhythm (Bergee & Rossin, 2019). In the exploratory factor analysis, both performance frameworks portrayed a robust second-order general aspect (Bergee & Rossin, 2019). Moreover, the confirmatory factor analysis illustrated those models comprising of the three interlinked primary order factors and one-second order tend to fit both sets of data (Bergee, 1995; Bergee & Rossin, 2019).

Bergee continued to study to build a more robust understanding of music performance assessment. In another study, Bergee (2006) aimed to develop an understanding regarding the theoretical model of selected musical variable abilities to be able to effectively explain solos and small ensemble festival ratings. Bergee (2006) utilized logistic regression as the basis for the model-building approach where the binomial logistic regression was used to analyze the 2004 rating data from a large midwestern state's solo and small ensemble music festival. The modeling approach within the two studies significantly coincided. The model variables were processed through an external (cross) validation through the application of both 2002 and 2003 data festivals (Bergee, 2006). Despite minimal variance being identified, the results from the study illustrated there is an acceptable fit between the provided data sets (Bergee, 2006). Regarding the internal validity of 50 samples, approximately 25% of the 2004 data set were randomly drawn and issued to the binomial logistic regression analysis (Bergee, 2006). According to Bergee (2006), the coefficient findings illustrated that an estimated coefficient indicated that there was consistency as well as limited biases; however, the results also pointed out that there was a case of inefficiency among the various estimates provided, hence asserting that the evidence attainment was under specificity (Bergee, 2006). Based on the research results, soloists who portrayed appropriate stage deportment such as a confident entrance, proper body alignment and weight distribution, as well as cue towards the pianist, stood a chance to receive a high-performance rating, compared to the soloist who portrays a more casual deportment (Bergee, 2006). According to Bergee (2006), the length of time the soloist performed also had a significant impact on the audience’s performance rating. It also may be argued that the initial impression may have been paramount to the raters having diverse expectations on the subsequent musical performances, hence contributing to their ability to diversely evaluate the excerpts (Bergee, 2006).

Beyond identifying the stage components that support performance, a theoretical structure could support musical theory knowledge. Furthermore, Bergee (2004) claimed that the establishment of a theoretical structure within the musical performance evaluation would significantly aid in the identification of various factors that tend to influence the attainment of the jazz theory knowledge. This may aid to understand the complex process within the aspect of music such as music listening, extramusical influences, sight-reading, and evaluation of musical performance (Bergee, 2006). Identifying the hypothetical paradigm indicates that the process of attaining knowledge on jazz theory assists the students in achieving an understanding regarding the improvisation of jazz art (Bergee, 2004). Also, the knowledge offers essential information to music educators on issues associated with learning as well as teaching jazz improvisation (Bergee, 2007).

There are assessment inconsistencies in the field of music education. According to Myers (2021), this is caused by schools allowing teachers the freedom to cover content standards of their choice. Music courses remain as electives in many schools, especially at the secondary school level, where some students may graduate without any fine arts credits (Myers, 2021). The schools that prioritize music programs still grant teachers the freedom to cover content standards of their choosing based on district, state, or national level benchmarks (Myers, 2021). Myers believed there is no oversight for music instructors, which allows them to get away with failing to compile in-class assessment data on the students; hence, there is a lack of a standard approach to ensure students acquire similar skills and training (Myers, 2021). Myers noted that many schools give passing grades to students just for attending ensembles and maintaining good behavior; therefore, to address this problem, there is a need for assessments as they provide formative development where choral students learn skills in music and performance (Myers, 2021). If teachers commit to regular individualized assessments to address the national anchor standards, Myers believed the students can also gain more well-rounded musical experiences that will benefit them in the present and help them serve the world in the future.

Many schools have gaps in their elective music courses. Myers (2021) adopted a more analytical approach to the problem to establish the status of music assessment in school. Myers took the readers on a historical journey to remind them of events such as the 1990s Goals 2000, which entrenched music courses in the academic core in the American educational system; however, this move did not receive much recognition as even secondary schools do not prioritize their music programs. Myers noted that teachers express freedom and creativity through the repertoire programmed for their choirs; however, the only way students can gain similar skills and training is when the teachers incorporate standard-based assessment strategies. According to Myers, there is a clear comparative analysis of the existing approaches used by secondary school teachers and the standard tools meant for improving their students' skills. Myers provided a review of national music standards history, assessment purposes, common assessment trends, and suggestions for assessing music, and recommended the implementation approach for local school districts; therefore, Myers systematically unpacked the complexities of music assessment for any interested audience. According to Myers, this could also be beneficial for secondary school teachers to learn how to integrate standard assessment to enhance their teaching. Music educators can also obtain valuable insight on developing well-defined curriculums that expressly guide teachers on how to instruct their students about music (Myers, 2021).

Adjudicator. Music performance presents variability within the evaluation process. Bergee (2007) identified that the generalizability theory was a key aspect of music performance. Generalizability theory is a framework to determine a performance assessment's reliability. In this study, performers recorded three audio excerpts from their solos, leading to the development of an occasion variable (Bergee, 2007). The results of the study indicated that the utilization of the generalizability coefficient was an essential criterion for the five hypothetical rates to meet the .80 benchmark reliability (Bergee, 2007). According to Bergee (2007), music contests tend to utilize adjudicators to evaluate the music performances, thus the exploration of the configuration of the adjudicator panels needs to be conducted. Fundamentally, this approach contributes to the facilitation of a fair assessment and, in turn, the ability to control the possible influence of biases in the judgment or evaluation (Bergee, 2007). According to Bergee (2007), this adjudicator panel operates similarly to an Olympic-style panel. The dispersion of the score within the provided rating scales asserts that the evidence scores within music performance are quite high (Bergee, 1994). Bergee (1994) affirmed that the high scores among music performances were mainly contributed by substantive measurement errors among the adjudicators or raters.

While there are many benefits to using an adjudicator panel, limitations exist as well. Bergee (2007) argued that the technique of calculating music scores using the Olympic-style panel is quite an effective assessment approach but tends to be quite expensive because it requires a significant number of judges or adjudicators. Music score validity can be improved through a technique in which cues are utilized (Bergee, 2015). In this method, performance measurements define the achievement as the raters' appropriate interpretations in addition to the use of the cues (Bergee, 2015). In situations where numerous adjudicators independently assess the same musical performance, the likelihood of perfect agreement is minimal (Bergee, 2007). The variability of the adjudicator scores is comprised of both the probabilistic and systematic components; hence, to be able to understand each adjudicator's score, each of the quantitative features needs to be reviewed prior to use to be able to promote a valid assessment practice (Bergee, 2004).

Diverse, methodological strategies such as casual comparison, quasi-experimental research, and surveys have been developed to appropriately examine the adjudicator events. Rossin and Bergee (2020) completed a research study aimed at attaining a conceptual understanding of school band performance and establishing a music performance evaluation. Resulting from this study, Rossin and Bergee postulated that through the utilization of a cross-validation approach and a rating scale for school band performance, the consistency of the previous scale was unidimensional. This scale involved one robust second-order aspect and three distinct primary factors: rhythm and technique, musicianship and expressiveness, and tone quality and intonation (Bergee, 1995).

The adjudicators applied the School Band Performance Rating Scale (SBPRS) online version while on a field test. Rossin and Bergee (2020) revealed the 25-item SBPRS validity was established using musical performances at both middle and high school levels. Being consistent with previous research, the SBPRS presented itself to be unidimensional (Rossin & Bergee, 2020). Regarding this approach, the SBPRS demonstrated to have effective and efficient internal consistency (Rossin & Bergee, 2020). According to Rossin and Bergee, this enabled them to be able to view the ratings from other adjudicators in a real-time approach, hence appropriate interrater agreement was attained. The research identified the SBPRS reflected on the conceptual framework of the school band performance (Rossin & Bergee, 2020). The SBPRS may be applied in a more advantageous approach to be able to effectively serve both the adjudicators and the school band ensembles (Bergee, 2006).

Concert band performance is comprised of a three-level judgment hierarchy: fundamental level with basic elements, intermediate level with a limited number of interrelated fundamental factors, and the highest level with an overarching higher order. Bergee (1995) asserted that three primary factors, namely tone quality/intonation, musicianship/expressiveness, and rhythm/articulation, loaded robust to only a single higher-order factor. To be able to attain a complex understanding of the hypothesis, Bergee (1995) utilized an adjudicator panel that was required to utilize the revised band performance rating scale. To determine the criteria of validity of the rating scale, global categorical rating performances were utilized (Bergee, 1995). The attained results illustrated that the validity and interrelated coefficients were uniformly high, with ranges from approximately 0.84 to 0.99 (Bergee, 1995). The research focused mainly on pointing out the need for critical mass evaluators, as it would allow achieving a valid musical performance rating (Bergee, 1995). On the other hand, Rossin and Bergee (2020) also argued that a large number of adjudicators is not required every time. The generalizability theory framework may aid in determining the minimum evaluators' population for a particular musical performance or event type (Rossin & Bergee, 2020).

Portfolio. A student could map out their own growth while at school through the development of a portfolio. A portfolio is defined as a dedicated collection of the efforts, growths, and attainments of a student (Rowley & Dunbar-Hall, 2017). According to Denis (2018), a portfolio-grounded assessment can be used in different disciplines to demonstrate student development and mastery of skills and techniques. Developing the skills of students is critical to their growth and development processes in the learning environment (Denis, 2018). Using portfolios can assist students in demonstrating their attainment levels on different materials and objectives (Silveira et al., 2017). Silveira et al. (2017) described the factors influencing the use of portfolio evaluation in secondary school environments as a strong evidence-based assessment strategy. Similarly, Denis assessed the challenges educators who wish to nurture these kinds of assessments faced. The study established that the main hurdle was the time needed for teachers to develop the assessments (Denis, 2018). Denis went further to state that portfolio-grounded evaluation can be used together with numerous styles of assessment to paint a picture of the knowledge base and growth of students. This nature of assessment can be implemented in all subjects to demonstrate attainment to the administrators and the public and can be implemented in instrumental music education as an important assessment strategy (Denis, 2018).

Portfolio assessment requires teachers and students to have a solid understanding of the process. Carey’s (2017) work on the latest practices of instrumental music assessment corroborated the results of Denis (2018) on assessments in music education. Carey went further to state that a portfolio-based assessment can also be used in assessing teachers. Assessing teachers qualitatively establishes superior instrumental music educational experiences and environments for learners and motivates high-quality learning to occur in classes (Carey, 2017; Denis, 2018). Dayal (2017) described the connection between teacher behaviors and beliefs in portfolio evaluation. Dayal determined that it was inconclusive concerning the relationship between the two, and without extra study, they could not prove the existence of such a correlation. Students gain from teachers and develop a positive view towards portfolio evaluation (Dayal, 2017). Likewise, Silveira et al. (2017) supported the notion that listing of objectives and behaviors by students or teachers needs to be completed to make them aware of the processes of their assessments. Such types of practical performance-grounded assessments are important to the education process and are particularly valuable in assessing non-concrete topics like the art of learning and similarly performing art (Silveira et al., 2017). Teachers and pupils must read on the same page in issues of assessment to guarantee effective results (Dayal, 2017; Silveira et al., 2017).

To be effective, portfolios must be conscientiously implemented. According to Dayal (2017), meaningful portfolios are designed when educators consciously embrace what the portfolio intends to achieve. Using a portfolio assessment in a classroom setting with achieving objective sets is easy to attain (Dayal, 2017). The idea of maintaining the growth goals and seeking ways for every child to show their own personal development needs to be sustained in the entire education process to sustain its validity. Denis (2018) supported a similar notion by positing that the portfolio itself has to show a precise picture of the student's achievement level as it concerns their classroom setting. The quantity of guidance given before implementing such portfolios is important to their success (Denis, 2018). Educators and students all need to be educated on using this kind of assessment and its meaning (Dayal, 2017; Denis, 2018).

Teachers must possess skills in all phases of portfolio implementation for the portfolio to be an effective method of assessment. According to Music (2019), for valid portfolio assessments, teachers need skills in designing, evaluating, and maintaining student portfolios. Allowing to practice these tactics, teachers are taught to engage learners meaningfully in the portfolio assessment (Music, 2019). Portfolio evaluation is not focused solely on assessing the abilities of the students but on their competencies and outcomes as well (Mitchell, 2020). According to Mitchell (2020), these documents have to be considered as continually growing in the entire course. When passed on throughout the educational experience of the students, these documents guide the curriculum and instruction (Mitchell, 2020). If created properly, the information contained in the documents is beneficial in establishing educational objectives for the next year's instruction (Music, 2019). Skill development is supported and demonstrated in the entire life of the portfolio (Mitchell, 2020). This kind of assessment encourages all children to contribute to their education and, if well used, encourages them to continue growing at their own speed during their entire educational experiences in elementary and secondary schools (Mitchell, 2020; Music, 2019).

Portfolios, different from other forms of assessment, have positive attributes that encourage their use. According to Music (2019), the high stakes and level of fear entailed with standardized tests are not found in portfolio evaluation because the objective is demonstrating growth, as opposed to mastery. This is a very challenging concept for teachers and students to comprehend because mastery has typically been the sole objective (Music, 2019). Teacher and student experiences have a great effect on the value of portfolio evaluation in all subjects. In instrumental music, portfolios are progressively popular as they enable learners to gather recordings, artifacts, and other important items that assist to guide their individual instruction in the arts (Dayal, 2017). The arts are a personal experience in growth terms because students have diverse skill sets and, in an effective setting, they set their personal growth objectives and assist to develop a plan to attain these goals (Dayal, 2017; Music, 2019).

Portfolios cannot only be physical collections of artifacts but must be web based as well. According to Wan and Gregory (2018), using a web-based portfolio can assist in communicating with both students and parents about their progress. It can also establish a more long-lasting portfolio owing to the storage of the artifacts for a longer time (Wan & Gregory, 2018). Wan and Gregory used two distinct groups of learners selected from a secondary school computer class. They concluded that using a web-based portfolio had an obvious impact on the group that used them. According to Wan and Gregory, the inquiry was intended toward a secondary school computer application class. Nonetheless, if a web-grounded portfolio was created for courses in other subjects like instrumental music, the same conclusion could be found (Wan & Gregory, 2018).

Using web-grounded portfolios facilitates the easier use of peer assessments and enhances the process of self-assessment. In instrumental music, students can be expected to conduct their own assessment and assessments of others (Hallam, 2019). The feedback the other students receive from the peer assessment can be beneficial, and they can relate to one another better than they can take the teacher’s suggestions (Hallam, 2019). According to Hallam (2019), web-grounded portfolio evaluation also allows the raising of student resumes, which they can carry to college and beyond. It can include the recordings of the performances of students in the entire years, and they can be easily used during the audition process at college (Wan & Gregory, 2018). According to Wan and Gregory (2018), this can establish a virtual scrapbook of the successes and growth of a child during their study. Integrating peer assessments is a priceless tool for a teacher, as it enables them to know if their evaluation matches that of their pupils (Hallam, 2019). Where the assessments match, it could be an indicator for a case for greater validity in the evaluation itself (Wan & Gregory, 2018). Hallam provided numerous suggestions for implementing portfolio evaluation for science teachers.

Portfolio assessment not only tracks the growth a student makes but can provide information on how to improve a music program for the future. According to Hallam (2019), many educators throughout the last few decades have used portfolio assessment. Implementing a program that gathers the records of instrumental music assessments from high school can greatly assist instructors in obtaining data that guide future instruction (Hallam, 2019). Collecting these data is integral and can assist to ensure that the students have a beneficial educational experience (Wan & Gregory, 2018). Rawlings (2016) added that individual learner assessment can also be integrated into their e-portfolio to include the student’s own perspectives on their development in the process. If maintained and reviewed periodically, these kinds of portfolios are valuable to all subjects (Hallam, 2019, Rawlings, 2016; Wan & Gregory, 2018).

There are methods available for improving the portfolio process as well as the program to which the students are creating portfolios. According to Rowley and Dunbar-Hall (2017), student artifacts demonstrate to others the benefits of their programs within their schools. Portfolios and artifacts of this kind are very important in developing the curriculum and creating student development objectives in the future (Rowley & Dunbar-Hall, 2017). According to Rowley and Dunbar-Hall, there are numerous ways portfolios could be used in the musical class to mirror student development. A part of the portfolio process has been considered as the use of recordings students make and eventually assessing the quality of their own performance (Giraldo et al., 2019; Rowley & Dunbar-Hall, 2017). To improve the portfolio, educators could use the SmartMusic^{TM} computer programs to deliver instant objective feedback to the students (Giraldo et al., 2019). According to Giraldo et al. (2019), the students are awarded a score as a percentage as soon as they finish the assignment grounded on the correct number of rhythms and notes they perform. The program also gathers these data and keeps all submissions' recordings for later review by teachers and students (Giraldo et al., 2019). In music, portfolios can be a powerful way of assessing the skills of students in objective segments like pitch and rhythm (Giraldo et al., 2019; Rowley & Dunbar-Hall, 2017).

Computer. Technology is useful and beneficial in instrumental music evaluation comprising the analysis of musical creativity, listening skills and knowledge, and techniques of performance. According to Loughran and O’Neill (2016), the emergence of computer technologies, telecommunications, distance education, and television is said to affect the accuracy and speed of delivering information to all learners in the process of education. The computer will enhance the capacity of self-education for non-musicians and musicians in almost every element of music (Waddell & Williamon, 2019). According to Waddell and Williamon (2019), implemented tools develop traditional evaluation, transform traditional uses, and facilitate methods concerning the evaluation of student learning.

Computerized instrumental evaluations provide a variety of methods for assessment. Educators can use many computer and software programs in assessing the knowledge and understanding of the student (Loughran & O’Neill, 2016). According to Loughran and O’Neill (2016), Google Docs allows teachers to collaboratively edit and administer assessments. Finale, a music notation software, can create notation-grounded and audio-visual assessments (Loughran & O’Neill, 2016). There are available applications for developing electronic surveys and online quizzes (Waddell & Williamon, 2019). According to Waddell and Williamon (2019), many of the applications have provisions for different types of questioning combinations. By gathering instant and organized data, educators can save a lot of time utilizing these tools (Loughran & O’Neill, 2016).

Various studies have explored the computer-based formative evaluation and learner behavior towards feedback as well as their motivation and beliefs. For instance, Van Groen and Eggen (2019) explored the suitability of computer grounded student evaluation and behavior of students towards feedback and their motivations and beliefs. They analyzed the frequency at which students solicited feedback and the period of time they used to process the feedback. Their conclusion was it was difficult to observe some of the aspects because of the timeline of the study (Van Groen & Eggen, 2019).

Alternative and Self-Assessments. Alternative and self-assessments give educators the flexibility to customize the assessment process to meet the needs of their students. Wise (2016) reported on the use of alternative evaluation and the learning that occurred since 2003. Alternative evaluations are differentiated approaches to collecting needed data (Wise, 2016). Wise examined existing literature concerning alternative student evaluation forms. According to Wise, the alternate evaluations seek to serve a particular population. Nonetheless, the concepts they raise concerning alternative evaluation certainly apply to the non-standardized subject disciplines (Wise, 2016). Cranmore and Wilhelm (2017) also mentioned different types of alternative evaluation, performance, checklist, and portfolio, for secondary school teachers. From the observations above, more research has to be carried out on the impact of alternative evaluation on the child and raise some issues relating to federal policies on student assessment (Cranmore & Wilhelm, 2017).

In the educational setting, a huge obstacle faced is consistency in the time taken to assess the efficiency of any learning initiative, which is too long to be seen by one administration, and unfortunately, with every new state and federal-level government, the pendulum is inclined to swing. In a study on different aspects that affect performance, Diaz (2018) stated that significant gaps exist in research to provide any individual with evidence on the effectiveness of alternate assessments. This is unsurprising given the amount of time it takes in evaluating the ability of a child to succeed (Diaz, 2018). At the very least, the child’s educational career is 13 years long, and it is impossible to evaluate any initiative within a limited time (Cranmore & Wilhelm, 2017; Diaz, 2018).

It is a challenge to steer ensemble classes towards a learner-centered approach.

Scruggs (2009) investigated the nurturing role of learner-centered instrumental music education classroom environment for musical growth and independence. Scruggs provided background information indicating how American schools offer instrumental classes that emphasize public performance preparation. Teachers have adopted rehearsal rather than a learning model in instrumental classes as they organize them as performing ensembles (Scruggs, 2009). The teacher often takes a conductor rather than an educator character as teacher-centered rehearsal paradigms are widespread (Scruggs, 2009). Based on the data and analysis in the current study, Scruggs noted that public schools encourage uniformity to serve the greater social good and to meet the qualities bureaucratic associations desire. This becomes an impediment as teachers must adhere to the approved curriculum, use textual instruction, and control their students (Scruggs, 2009). According to Scruggs, teachers themselves are also resistant to change and are unable to adopt learner-centered teaching that opposes societal expectations from public schools. Consequently, the students fail to develop their leadership, problem-solving, and creative thinking as they participate in a factory model of instruction (Scruggs, 2009). They desire flexible classrooms that encourage their independence but are limited by the learning environment where the teacher is a conductor rather than an educator (Scruggs, 2009).

There is a limited instructional environment for music students. According to Scruggs (2009), there is obvious criticism about how music learning has been structured without any concern for the student interests. Public schools and teachers are focused on meeting the demands of bureaucratic associations and societal expectations (Scruggs, 2009). The teachers take a conductor role in the classroom environment instead of teaching music and performance to the students (Scruggs, 2009); hence, Scruggs believed learner-centered education has been abandoned, which exposes the students to career failures. They are unable to express themselves musically and show their creativity because the teacher restricts them to textual teaching and exercises control over all activities (Scruggs, 2009). Scruggs criticized public schools and music associations for not prioritizing the students' needs and interests. Scruggs emphasized that there is a rigid curriculum for the teachers to follow, which limits the music students from acquiring playing skills. Scruggs utilized scientific data to shine light on the barriers in music education for the learners, adequately addressed the problem, and offered a starting point for reform in the music curriculum.

Music education must be student-centered and focused. Scruggs (2009) critiqued the instructional and pedagogical methods students are treated to in learning music. Scruggs seemed to be urging the instructors to be more of educators rather than conductors and performers because theoretical knowledge is instrumental in skills development. Scruggs advocated for the interests of the music students as they seem to have been abandoned by everyone who is supposed to be safeguarding their needs. There is nothing better than an evidence-based argument against existing policies and practices as it can provoke action to correct the limitations (Scruggs, 2009). Scruggs believed policy makers and curriculum developers have not adapted to the students' needs. These are the entities most influential against the instructors, and they need to realize students are required to exercise their creativity and independence if they are to become skillful music players (Scruggs, 2009).

Tools

A variety of assessment tools can be used to evaluate students, including rubrics, rating scales, and technology tools. In educational settings, it is important to have valid and reliable tools for evaluation to support students in improving their skill set (Mazur & Łaguna, 2017; Stambaugh & Demorest, 2010; Uygun & Kilinçer, 2017). These skills include the technical components of playing an instrument (Loughran & O’Neill, 2016), expression (Uygun & Kilinçer, 2017), improvisation (Stambaugh & Demorest, 2010), quality/intonation, musicality, and sight-reading (Mazur & Łaguna, 2017). Each assessment has strengths and limitations and at present no comprehensive assessment tool exists that can address all these aspects of music performance (Stambaugh & Demorest, 2010; Uygun & Kilinçer, 2017).

Rubrics. Rubrics incorporate criteria for examining the expressive and technical aspects of the performance. Regarded as the most useful tool in assessing instrument playing achievement, rubrics are rating scales of measuring music performance (Loughran & O’Neill, 2016). In this manner, Loughran and O’Neill (2016) noted what was heard in the performance and not what they disliked or liked or the level to which they disagreed or agreed that the performance was of an unknown standard. Multidimensional evaluation rubrics can also be used to carry out performance assessments (Loughran & O’Neill, 2016). According to Loughran and O’Neill, rubrics assist judges in assessing the performance with more reliability. They assist in assessing individuals’ performances while playing diverse instruments at various stages of education (Loughran & O’Neill, 2016).

Rating Scales. Various kinds of rating scales can be applied in the measurement of the assessment of playing diverse instruments reliably. For example, the Brass Rating Scale measures the performance level of playing brass instruments, while the Clarinet

Scale of Performance is used to measure the performance in clarinet playing (Mazur & Łaguna, 2017). Some scales evaluate the attainment of playing string instruments. The String Performance Assessment is a valid and reliable rating scale developed to assess overall string performance (Mazur & Łaguna, 2017). A study by Souza et al. (2017) also found that various factors could account for the performance of students in string instruments, such as articulation/tone, musical/interpretation effect, vibration, intonation, and tempo/rhythm. The Jazz Improvisation Scale is also a tool developed for measuring the achievement of a jazz performance improvisation (Stambaugh & Demorest, 2010). The Kaleńska-Rodzaj scale can be used to assess instrumental music performance when expression is not being evaluated (Uygun & Kilinçer, 2017). According to Uygun and Kilinçer (2017), a factor omitted in the performance evaluation includes expression. The Zdzinski Performance Rating Scale Supplement is designed as an auxiliary instrument that includes criteria for the subjective aspects of performance, including expression (Uygun & Kilinçer, 2017). Mazur and Łaguna (2017) echoed that the scale is used in evaluating the performance aspects, which are not considered by approaches grounded on an objective conversion system point, namely tone quality/intonation, musicality, or technique. These tools are not used globally, but they have been adopted by different countries and translated into psychometric tools for measuring performance in instrument playing (Uygun & Kilinçer, 2017).

Rating scales are diagnostic tools for musical achievement evaluation that are intended to measure particular kinds of skills. Mazur and Łaguna (2017) stated a performance assessment scale targets particular musical skills to obtain data for specific purposes. For example, the Watkins-Farnum scale measures sight-reading competencies and technical elements of performance (Stambaugh & Demorest, 2010). According to Stambaugh and Demorest (2010), the scale does not just incorporate criteria for assessing playing instruments but also provides suggestions for the musical pieces, which can be played by the study's participants. Stambaugh and Demorest studied the student performance whose assessment can be more accurate in notes and rhythm and are more objective measures. According to Stambaugh and Demorest, the method enables the computation of the number of errors made in the dimensions of a given performance seen by the listener in different parts of the music text, such as errors that happened in every bar measure of the performance.

Rating scales in music do not account for the aesthetics of the performance. From the studies, rating scales are concerned with items that greatly influence the ability of listeners to assess consistently and accurately, such as articulation, tone, interpretation, and rhythmic articulation (Stambaugh & Demorest, 2010; Mazur & Łaguna, 2017; Uygun & Kilinçer, 2017). According to Uygun and Kilinçer (2017), current assessment tools do not give information on indications of growth and performance level. The nature of the student's performance that led to the judges or examiners assessing performance as average or above average remains unknown (Stambaugh & Demorest, 2010; Uygun & Kilinçer, 2017).

Additional research was aimed at developing a rating scale for the midlevel band performance as well as the validation of a theoretical structure for the scale. Bergee and Rossin (2019) posited various aspects such as the qualities that described an excellent band performance were components within the Midlevel Band Performance Rating Scale. The Midlevel Band Performance Rating Scale was applied to illustrate validity in a musical assessment (seventh and eighth grade; Bergee & Rossin, 2019). Through a mixed methods approach, the researchers identified there were 27 musical and technique components appropriate for the rating scale; and after an analysis was conducted, all components indicated sustainability for a Midlevel Band Performance Rating Scale (Bergee & Rossin, 2019).

While internal consistency existed, rater behavior impacted scoring. The internal consistency of all 27 elements was appropriate for both rating sets (Bergee & Rossin, 2019). This assessment approach demonstrated that the two rating sets indicated validity as they confirmed their underlying frameworks were consistent with the research conducted previously (Bergee, 2007). There must be various approaches used to understand the rater effect when conducting a music performance assessment, including rater behavior strategy (Bergee & Rossin, 2019). According to Bergee (2006), this strategy aims to understand the ecological content of human judgment and it may be categorized into four specific fields: (a) extramusical impact linked to the performer such as variations of expression and body movement; (b) extramusical impact linked to the evaluation context such as communication within ensemble performance, acoustic, social aspects, and the support of the audience; (c) rater-centered impacts like memory, mood, first impressions, musical preference, and repertoire familiarity; (d) the nonmusical impact that involves stereotyping, order of performance time evaluation, teaching level, primary instrument, and musical expression facets. Nevertheless, research has identified that the main disadvantage of the rater-centered strategies in the assessment of music performance protocols is that the scores observed through the raters tend to be reported free from psychometric considerations of the behavior of the rater (Bergee, 2006). The study aimed at demonstrating the application of the modern measurement methods within the context of musical performance evaluation contributed to a contrasting factor on most of the current practices involved during the assessment of music performances (Bergee, 2006). Particularly, the research within the music evaluation is controlled by the classical test theory, with other research studies using generalizability hypothesis to examine raters (Bergee, 2006, 2007).

In determining correlations for scores, several factors are identified as weighing heavily on scores. Among the various variables added, they all eliminated the high collinearity owing except for the geographical location (Bergee, 2004). Also, the results demonstrated that the type of event through a performing medium interaction was identified as a significant outcome rating predictor (Bergee & McWhirter, 2005). The research confirmed that the afternoon scheduling, high expenditure school, and performing as a soloist or vocalist significantly predicted a higher rating (Bergee & McWhirter, 2005). Based on this notion, performance is a significant approach for evaluating music as it is of prime importance regarding music development as well as the ability to motivate the audience (Bergee & McWhirter, 2005). Performance evaluation is a subjective endeavor as research has identified factors such as the size of the school, time of the day, event type, and expenditure level regarding daily attendance as significant festival score predictors (Bergee, 1994; Bergee & McWhirter, 2005).

Technological. Technology has become a more common tool to assess instrumental music performance. Waddell and Williamon (2019) described several technological tools that many instrumental music examiners use in the assessment of students. These comprise simple tools such as spreadsheets and audio-recording gadgets to complex ones like software specifically intended for instrumental music classes. The tools have the advantage of saving a lot of class time. Students can use technology outside the class allowing educators to concentrate class time on musical progress and enable more individualized evaluation. Waddell and Williamon established that 32% of high school educators used out-of-school recordings in assessing their pupils. Similarly, Loughran and O'Neill (2016) established similar outcomes; 33% of educators conducted their assessment by asking students to self-record themselves. Assessing every student individually using the approach can save classroom time (Loughran & O'Neill, 2016). It can often be time-consuming to give feedback to every student who hands over a copy of their playing (Loughran & O'Neill, 2016; Waddell & Williamon, 2019).

There have been attempts made to measure the performance attainments objectively in playing musical pieces comprising the use of computerized programs. Using computer technology is highly dependable in measuring performance but it is restricted to just several elements of the performance (Dunbar, 2018). The most frequently considered factors are rhythm accuracy and pitch. Some musical teachers have used computer-aided programs in assessing their learners (Dunbar, 2018). Waddell and Williamon (2019) established that 5.1% of programs apply for generic computer-aided assessments, whereas 13.1% of programs use a particular program called SmartMusic^{TM}. The program can "listen" to the students play lines from their method books and evaluate their rhythm and pitch's accuracy to an accepted standard (Dunbar, 2018). SmartMusic^{TM} is an assessment software tool that can be used for practice and teaching (Dunbar, 2018; Waddell & Williamon, 2019). According to Dunbar (2018), the software enables the demonstration of the audio and visual content on a screen and also captures recorded performances to a video or audio file. The external or built-in microphone is utilized to record sounds (Wu et al., 2016). *MakeMusic* has developed this function to enable educators to provide feedback and corrections automatically to students (Dunbar, 2018; Waddell & Williamon, 2019). The attributes of SmartMusic^{TM} not only provide instant feedback but can also be used in various forms of assessments, including formative and summative (Dunbar, 2018). Reflecting on their strategies, approaches, and materials of teaching, educators can assess student learning outcomes and cope with the function of evaluating every student within a huge ensemble (Waddell & Williamon, 2019). According to Dunbar, there are cautions concerning computer-grounded evaluations. Concerning its use, SmartMusic^{TM} can present several data limitations (Wu et al., 2016). For instance, the software cannot surpass the intersection of math and microphones (Dunbar, 2018). Illustratively, the SmartMusic^{TM} program can assess the rhythm and pitch accuracy of the performer but is unable to determine the humanistic side of music, including tuning, intonation, tone, and phrasing (Dunbar, 2018; Waddell & Williamon, 2019).

Various studies have ventured to examine the effectiveness of SmartMusic^{TM} as an assessment program (Dunbar, 2018; Waddell & Williamon, 2019; Wu et al., 2016). Dunbar (2018) evaluated the effectiveness of SmartMusic^{TM} as a tool of assessment and established that it was valuable for etude performances, particularly for technical passages and less for lyrical passages. However, technology use faces some hurdles. Most technology is an expense to the capital of students and their parents. The expense of subscribing to an evaluation program, internet connection to connect to the program, and the required hardware for using the program may be prohibitive to numerous families who might have probably already used the money on a purchase or an instrumental rental (Waddell, & Williamon, 2019). Using technology could be very valuable but needs support to guarantee equal access to technology by all learners.

Investigating the impacts of the SmartMusic^{TM} program, Shih (2018) analyzed the experience of new band students in reference to instruction and time on the performance capacity. The study computed Cronbach's alpha in estimating reliability and validity. For the reliability assessment, three students of the eighth grade were tasked to play the test piece thrice each to explore the scoring model of the SmartMusic^{TM} evaluation. With a coefficient alpha (a=.91), it showed the reliability of SmartMusic as a testing tool. According to Shih, the reliability for tests in the SmartMusic^{TM} program would show diverse patterns and the reliability would be considered acceptable. Assessing the validity of the SmartMusic^{TM} evaluation, Shih compared the scoring of three music teachers to the scores of the SmartMusic^{TM} program. The data revealed a correlation of r=.93 (high) between the three panel judges and the comparison of the score composites of the judges to the software, r=.91. In another study, a panel of four judges sought to establish the reliability and validity of SmartMusic^{TM} assessments; the examiners measured inter-rater reliability and recorded a high correction from r=.87-.97 (Pati et al., 2018). These rates considered that the program's validity is acceptable.

The use of audio recording can support consistency within scoring. According to Bergee (1994), there is a significant invariance degree between evaluator sets. Regarding generality, Performance Approach 1 did not produce similar results as Performance Approach 2 (Bergee, 1994). According to Bergee (1994), various approaches may be utilized in the facilitation of the self-assessment, but the utilization of recording is quite important when evaluating a musical performance. Often, the contest adjudicators, as well as teachers, are thoroughly trained to evaluate various music performances effectively and efficiently through the use of audio recordings (Bergee, 2007). Audio recordings are also used when educating the students on approaches of self-evaluations, hence promoting the advantages as well as popularity of using audio recordings (Bergee, 2007, 2015).

When creating audio recordings for assessment, technical considerations must be considered. One consideration, Bergee (1994) asserted, is the essential nature of practitioners engaging in communication when students are asked to conduct a self-recording. Beyond this communication, practitioners must provide practice on the procedures to counteract the challenges presented by the utilization of technology in music assessment (Bergee, 1994). Though music performance evaluation is a popular practice among the adjudicator, raters, and judges, the audience tends to experience challenges linked to the music performance evaluation practice (Bergee, 2004). There tends to be a complex system that consists of many aspects as well as interrelated influences that need to be appropriately understood by the various parties within the musical performance evaluation process (Bergee, 2004; Bergee & Rossin, 2019).

Timbre is influential in recognizing sound sources. Lee and Müllensiefen (2020) highlighted the scarcity in published literature about tests measuring individual differences in perceiving musical timbre. Lee and Müllensiefen focused on describing the development of the Timbre Perception Test (TPT). People’s ability to perceive timbre qualities like an instrument’s color or texture enables them to discriminate musical pieces played by different instruments (Lee & Müllensiefen, 2020). According to Lee and

Müllensiefen, the multidimensional and complicated nature of timbre renders it a poorly understood auditory attribute. Lee and Müllensiefen applied multidimensional scaling (MDS) of timbre (dis)similarity ratings to identify the perceptual timbre space dimensionality. Lee and Müllensiefen identified attack time and spectral centroid as the most salient timbral properties; hence, after Lee and Mullensiefen developed the TPT, they measured amplitude envelope, spectral centroid, and spectral flux. Lee and Müllensiefen established that TPT has internal consistency per common standard and good test-retest reliability. There was also a significant correlation between the TPT composite score and PROMS battery timbre test, which supported its validity (Lee & Müllensiefen, 2020); hence, Lee and Müllensiefen confirmed TPT as a promising tool to measure timbre perception ability.

There are practical ways to assess musical auditory characteristics for individuals. Lee and Müllensiefen (2020) explored a previously under-researched area concerning tests for measuring musical timbre perception; hence, Lee and Müllensiefen explored a new phenomenon and expanded knowledge about music. The development of the TPT reveals how individuals can differentiate sounds produced by different instruments (Lee & Müllensiefen, 2020); therefore, Lee and Müllensiefen provided crucial information not only for experts in judging musical aspects but also the average person who enjoys music and instruments. Lee and Müllensiefen provided a valuable analysis and description of timbre, a misunderstood perceptual attribute of music, thus students can utilize the source to learn the basics of music features because the researchers approached the topic from a simplified viewpoint. Lee and Müllensiefen provided important theoretical knowledge for music educators and teachers who are teaching their students to assess musical characteristics; therefore, they can enhance their expertise and contribute meaningfully to training skilled players and performers who can distinguish the important features of what they are playing.

Judges can expand their ability to detect sound characteristics using the TPT model. They can now gain more expertise on TPT as a tool for distinguishing sound qualities and differentiating musical pieces in an ensemble (Lee & Müllensiefen, 2020). According to Lee and Müllensiefen (2020), there is an adequate definition of various features of timbre, including spectral centroid and attack time. These details are important to consider, especially for judges looking to assess the performances of different music players (Lee & Müllensiefen, 2020). Lee and Müllensiefen also indicated the use of non-acoustical instruments in modern commercial music. This sound-processing technology poses a significant challenge to modern timbre perception experts as they may not have their hearing attuned to the combined string instruments (Lee &Müllensiefen, 2020). Luckily, their ability to hear fine sound attributes gives them an advantage over non-trained individuals (Lee & Müllensiefen, 2020). Lee and Müllensiefen believed this can be useful for music educators and their students as the content can be integrated into secondary school music courses.

Feedback

Effective feedback in instrumental music education increases student achievement on assessments. Hallam (2019) stated it is incredibly difficult to observe student motivation and it can change depending on the task being executed or the subject being studied. Even in a subject the students enjoy, there are topics where a student might or might not have any intrinsic motivation (Hallam, 2019). Van Groen and Eggen (2019) established that students who sought feedback regularly performed better in formative evaluations. The process with which learners solicit feedback could include their motivation for learning (Van Groen & Eggen, 2019). Nevertheless, the research by Hallam did not consider that as a contributing issue. According to Hallam, the idea of both providing and getting frequent feedback is especially important to instrumental music because students who seek feedback regularly from their teacher record a higher success rate. Using the feedback concept is important to the process of student evaluation and can be used in portfolio assessment to establish students who actively contribute to their education (Hallam, 2019). According to Hallam, learners can integrate the feedback from their teachers into their goal setting and the assessment of the goals. Denis (2018) argued for creating and using computer-interactive evaluations in music. Using computer assessments will ensure that data sharing is more efficient and reliable among students in the music community (Denis, 2018); however, there are some challenges identified as problematic in assessing music education within the confines of the higher education environment (Denis, 2018; Van Groen & Eggen, 2019).

As a student progresses into higher education, it becomes increasingly difficult to assess. Denis (2018) described the difficulty of assessment in the higher education setting within the confines of the music studio. Denis focussed on applied music education that is often one-on-one and sought to describe the evaluation of student attainment using quality ways. According to Denis, written assignments do not apply to the instrumental music studio. For this reason, Denis considerd it important to implement strategies designed by their counterparts to design quality evaluations in their studios (Denis, 2018). Denis presented a rubric that both the teacher and the student can use in performance assessment. To expand the rubric uses, recordings can be integrated to allow students to assess themselves to guarantee validity between the teacher and student evaluation (Loughran & O’Neill, 2016). Using rubrics in assessing their performance will also demonstrate whether the learner truly comprehends the concepts they learn in the studio through their lessons (Loughran & O’Neill, 2016). The students need to understand the terminology in the rubric, and they will have the ability to demonstrate that understanding whether or not they can evaluate themselves adequately (Denis, 2018). Denis highlighted the possibility of implementing tactics from other subject areas into successfully learning instrumental music. Doing this ensures that data can be easily shared with non-performing artists (Loughran & O’Neill, 2016).

Student evaluations provide crucial data for stakeholder awareness and teacher reflective practice. Silvey and Springer (2020) stated the most significant reason educators evaluate is to augment student levels in solo performances. The primary aim of assessing is to enhance student education, enhance teaching, develop better programs, and inform stakeholders (Silvey & Springer, 2020). Through increasing the awareness of stakeholders of the outcomes of the music program, educators will gain more support from the public that will allow for a better network of support whenever the program faces danger (Denis, 2018; Silvey & Springer, 2020). Silvey and Springer presented the notion that for students to improve, they require precise feedback on what they are doing well and what is needed to be addressed to advance further. Silvey and Springer concentrated on the way educators can use the evaluation to improve themselves. Silveira and Gavin (2016) stated all good educators discover the significance of assessment and make it their duty to enhance the quality of work of the students. According to Silveira and Gavin, reflective practice allows teachers to look at aspects to ensure a meaningful educational experience. According to Millican and Forrester (2019), it is essential for teachers to consider consistent evaluations and use the data to drive the instruction. Grading is not the most significant reason for evaluation, but rather student achievement (Millican & Forrester, 2019; Silvey & Springer, 2020).

Assessments are currently based on a variety of methods. Using a pedagogical knowledge structure, the study assessed the complexities of teaching, knowledge teaching, and assessment for learners (Millican & Forrester, 2019). Millican and Forrester (2019) showed how assessments can be created based on the competencies, which the teacher would like learners to gain after the culmination of a given time. Focusing on the end outcome enables the educators to keep the instruction on track and eventually serves the best interests of the student (Millican & Forrester, 2019). Millican and Forrester proceeded to posit that effective instruction includes assessments in the entire instructional process and permits students to evaluate themselves (Millican & Forrester, 2019). Millican and Forrester concluded educators should consider assessments as progressive and skills-based to have the most impact on the student. There are numerous ways to evaluate students in instrumental music, which can be beneficial to both students and teachers (Millican & Forrester, 2019; Silvey & Springer, 2020). Equally, Azzara (2016) asserted the importance of assessments to the instruction process, which delivers useful information to teachers and students. According to Azzara, three factors are inherent in all music instruction: content and process, progressive evaluation, and results of instruction. Evaluating the outcomes of student instruction is an aspect that can be concentrated on to permit non-subjectivity (Azzara, 2016; Millican & Forrester, 2019).

Developing and Validating Music Performance Assessments

Validity and Reliability

Validity. Validity connotes the level to which theory and evidence support the interpretations of the scores from a test for proposed tests uses; therefore, validity is a fundamental aspect in designing and evaluating tests (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Wesolowski & Wind, 2019). The validation process entails the accumulation of pertinent evidence to deliver a thorough basis for the suggested score interpretations. According to Wesolowski and Wind (2019), validity encompasses sophisticated summaries of calibrating values of interest and correlating them to preexisting established criterion standards The validity metric seeks the measurement of the skill level; and where there are high and positive correlations, there is evidence of the existence of criterion validity (Wesolowski & Wind, 2019).

There are different types of validity. According to Hallam (2019), intrinsic validity is an accepted measure where achievement in instrumental music is accepted based on the opinion of music content specialists. To determine the validity of instrumental music assessments, music evaluators usually consider the subjective and technical skills of a performance (Hallam, 2019). The subjective performance usually happens when used in an evaluation principle comprising of the performance's overall general impression (Hallam, 2019). Under the situation, it is a challenge to determine the subject and criteria of assessment. Many evaluators lack the awareness of what establishes their judgments (Hallam, 2019). When attempting to evade such hurdles, Hallam recommended the success of the assessment be based on certain musicians’ respective expressive and technical skills.

Construct and content validity assesses whether an instrument adequately covers all needed facets and correctly measures for its intended purpose. Construct validity assesses whether the instrument adequately measures what it claims (AERA, APA, & NCME, 2014). Constructs represent items that are hypothetical and can be expressed through the measurement process (AERA, APA, & NCME, 2014). The measurement process can utilize a written test, a performance, or another assessment through examination (AERA, APA, & NCME, 2014). This examination will generate variables that can be appraised and analyzed (AERA, APA, & NCME, 2014). Content validity assesses whether the instrument adequately addresses all facets of the construct (AERA, APA, & NCME, 2014). Constructs represent items that are hypothetical and can be expressed through the measurement process (AERA, APA, & NCME, 2014). The measurement process aims to ensure that all relevant parts of a subject are included in the assessment (AERA, APA, & NCME, 2014). This examination will generate variables that will provide evidence to whether the purpose of the assessment has been met (AERA, APA, & NCME, 2014).

Orchestra directors assess individual and group performance when they meet with an ensemble. Smith and Barnes (2007) conducted a study to evaluate musical performance based on videotapes of festival performances. Smith and Barnes explored the underlying factor structure of orchestral performance, the individual items best representing the identified factors, and the reliability or validity of an Orchestra Performance Rating Scale (OPRS). The goal is to develop a factor-derived assessment of orchestra performance achievement and test its validity and reliability as a tool for evaluating secondary school orchestras (Smith & Barnes, 2007). According to Smith and Barnes, the evaluation entails informal statements and a formal structure that grades individuals or rates groups. Smith and Barnes determined from their findings that they can accurately replicate group rankings and Music Educators National Conference festival ratings; hence, OPRS-2 can be an equivalent alternative tool to the more global measurement approaches (Smith & Barnes, 2007). Meanwhile, Smith and Barnes's results show subjectivity to the factor weighting approach in the total score, which does not result in a valid total score. OPRS-2 scores can also be adapted and combined with additional items and commentary to suit situations (Smith & Barnes, 2007).

Adjudicators can fairly assess orchestra performance from both positive and negative angles. Smith and Barnes (2007) provided a comprehensive evidence pool for assessing orchestra performance because it examines existing rating scales and develops numerous items to describe music performance. Importantly, Smith and Barnes collected both positive and negative statements of each item through the two forms, an approach that helps assess the possibility of judges reacting differently to particular statements. The randomization of the item order also ensures there is no bias or prioritization of the researchers' preferred items (Smith & Barnes, 2007). Moreover, there is better objectivity from utilizing two inverse forms to obtain equal numbers of responses (Smith & Barnes, 2007). Smith and Barnes believed this is an honest way to judge music players, as it leaves out no details. Smith and Barnes indicated how their study was limited by factor weighting in the total score, which made the research subjective and unable to yield a valid total score; hence, they give valuable insight to future researchers to consider alternate methods of item weighting and scoring to improve on the present results (Smith

& Barnes, 2007).

Existing assessment tools can adapt some of the most relevant and common items to enhance the fairness of their rating scales. Smith and Barnes's (2007) research can be used to improve the orchestra performance assessment tools, especially since it developed an item pool of statements to describe aspects of music playing. Teachers and adjudicators could be the main beneficiaries since their work will have been made easier as the research comprehensively addresses aspects of orchestra performance (Smith & Barnes, 2007). Smith and Barnes involved a substantial number of judges to review and evaluate performance, and their responses can help judge performances in real concerts; therefore, the results of the research can be used as a reference point for teachers and adjudicators in an actual music setting. According to Smith and Barnes, players can also refer to the ratings on the study to understand the areas they can improve to attract a better rating from the judges. Items like ensemble, position, rhythm, tempo, and presentation are properly researched, which can be a point of information for members of an orchestra seeking to enhance their skills (Smith & Barnes, 2007). Moreover, Smith and Barnes believed music educators can use the research as part of their curriculum to train their students on various performance factors to meet the expectations of judges and adjudicators.

Performance measurement is complicated because of the subjective nature of assessment. Zdzinski and Barnes (2002) argued that assessing musical performance is crucial in the instructional process in string education. Zdzinski and Barnes recognized how performance assessment occurs in many instructional situations, including seating auditions and ensemble placement, rehearsals, festivals, and concerts; however, there is an undeniable problem in judging music performance, with numerous researchers indicating challenges related to low judge consistency, even for the most experienced experts (Zdzinski & Barnes, 2002). Zdzinski and Barnes suggested solutions, such as using a panel of judges or training them to reduce inconsistent judging. In most cases, Zdzinski and Barnes believed that measurement tools are general impressions of musical performance, with each judge utilizing their internal criteria of assessing an individual performance based on the scale. The results of the present study indicate the presence of five factors to assess string performance, which is different from what other researchers found (Zdzinski & Barnes, 2002). According to Zdzinski and Barnes, there are significant differences in areas of articulation, intonation, tone, and technique, while musical effect and rhythm/tempo are common; therefore, Zdzinski and Barnes discovered five factors that can be used to enhance music performance evaluation because they show inter-rater reliability at high levels, and they are moderately high in criterion-related validity.

Evidentiary backing is a crucial way to measure the reliability and validity of the assessment scales. Zdzinski and Barnes (2002) did not shy from pointing out the existing limitations. Zdzinski and Barnes gave crucial recommendations of how to enhance performance assessment to show an accurate measurement of the quality of music and the players’ skills. Zdzinski and Barnes sought to improve musical standards by developing more objective tools for measuring performance. Zdzinski and Barnes enhanced the strength of their arguments by referencing numerous past studies that emphasized the limitations of existing assessment scales in music playing. Importantly, Zdzinski and Barnes’s study is relevant for the directors, music educators and teachers, and curriculum developers of music education programs. Zdzinski and Barnes made a comparison between the factor analysis from their research and other past studies from other researchers. The strategy also promotes the credibility of the authors because they do not introduce new concepts or knowledge that is not already familiar to the discipline (Zdzinski & Barnes, 2002). According to Zdzinski and Barnes, teachers should target improving their competence in the factors that show consistency on the rating scales of several researchers and judges.

Rasch Analysis. Rasch analysis improves precision in constructing instruments. This technique helps researchers think critically about the constructs they need to measure (Linacre, 2015; Randall & Engelhard, 2010; Smith, 2004). According to Smith (2004), techniques can also help document and assess the functional measurement of instruments. Through this assessment process, researchers are able to create alternate forms of the instrument (Smith, 2004). These alternate forms allow for change and student growth (Smith, 2004). According to Randall and Engelhard (2010), the data produced from using this technique will help explain the meaning of test scores and direct researchers to adapt the instrument to improve efficiency. Rasch analysis provides a careful measurement of an instrument's quality and avoids mathematical errors common to other techniques (Linacre, 2015). This technique provides for better communication of findings and provides evidence of constructs that need adjustment (Linacre, 2015). Rasch analysis utilizes raw data from test scores and rating scales to create linear performance measures of participants (Linacre, 2015; Smith, 2004).

Reliability. Reliability is an evaluation tool to generate the same results over a given amount of time. According to Wesolowski and Wind (2019), the definition emphasizes the standards in replications that mirror a certain interpretation and uses of test scores. In the music arena, the traditional metrics for rater behavior assessment include agreement estimations of inter-rater consistency and reliability projections (Wesolowski & Wind, 2019). The limitation of using the indices in assessments when testing rater behavior is that the observed measures might be exaggerated in circumstances of varied leniency or severity rates for learners with similar abilities (Wind & Wesolowski, 2018). According to Wesolowski and Wind, the effect can also result in a skewed demonstration of what comprises an “accurate,” “good,” and “fair” rater from “unfair’, “inaccurate,” and “bad” rater.

The other method of examining rater behavior in the music evaluation context uses empirically driven statistical indices that are used in the measurement processes. According to Wesolowski and Wind (2019), rater variability could stem from the compliance level of the raters with the measurement instrument, the approach to interpreting the criteria of the raters in practical scoring events, the severity or clemency level demonstrated, the awareness of raters of the categories of the rating scale of the measurement tool, and the consistency level of the ratings across the scoring criteria, examinees, and activities of performance. The Rasch Model is effective in evaluating musical performance (Wesolowski et al., 2016b). According to Wesolowski et al. (2016b), the Rasch Model’s major benefit is the observation of a proper fit of the model culminates in the attainment of invariant measurement. Whenever the data fit the Rasch Model's requirements, the measurement of the rater-invariant of performances is attained (Wesolowski et al., 2016b).

Playing music with others in curricular instrumental ensembles bears meaningful ramifications for the performing groups and their directors. Latimer et al. (2010) aimed to investigate a performance assessment rubric in a large group festival setting. The purpose was to examine the rubric's reliability and perceived level of pedagogical utility (Latimer et al., 2010). According to Latimer et al., adjudicated music festivals usually utilize performance assessment protocols involving adjudicators using a performance assessment tool. Latimer et al. also underscored the subjectivity of human judgment as it is based on human impressions, but they are useful in successfully evaluating a sound's musical worth. The rubric is moderately good in its reliability as a measurement tool, with its performance in the mid-to-upper range of previously investigated rubric-type assessment tools (Latimer et al., 2010). According to Latimer et al., the "other" dimension of the tool shows unreliability in the present and previous studies, meaning it should be omitted. Latimer et al. concluded that the comments offered for improving the rubric suggest the need for a more integrated assessment approach for the different performance dimensions.

Assessment tool creditability is achieved by utilizing real-world data. Latimer et al.'s (2010) method collected duplicates of all completed assessment rubrics and performed a statistical analysis of the copies. This also included requests for copies of follow-up data, including questionnaires and surveys about the new rubric (Latimer et al., 2010); therefore, Latimer et al. utilized data from Kansas State High School Activities Association. Latimer et al. utilized existing data, which saved time for the research. The information was also broad and offered sufficient evidence to investigate the topic of interest (Latimer et al., 2010). Furthermore, there was also data credibility because all the copies requested were from site managers and the rubrics had been completed during actual festivals (Latimer et al., 2010). Latimer et al.'s robust method assessed the research question from a real-world point of view involving actual data from numerous festivals. Latimer et al. adequately answered the question about the reliability of several dimensions of a performance assessment rubric. This method offers a practical way for adjudicators to judge music performance in a real-world setting (Latimer et al. 2010). The follow-up questionnaires ensure all the information captured is accurate, which enhances the reliability of the results and the replicability of the study (Latimer et al., 2010).

Music performance rubrics must integrate relevant dimensions; hence, committees in charge of music activities and festivals can obtain crucial knowledge from this research as it contains the deliberations of other members in their capacity (Latimer et al., 2010). According to Latimer et al. (2010), adjudicators and judges can also gain invaluable information about how to assess music performance because they have clearly defined dimensions from this study. Latimer et al. underlined the strengths and limitations of the existing tool that helped develop a more fitting rubric suited for assessment events. The dimensions contained in the tool are crucial for music performance groups because they get to see how the adjudicating team gauges performance by each group (Latimer et al, 2010). Latimer et al.'s study contained real data from festivals, which the participants can access and compare with the ratings of the adjudicators; therefore, they can learn to improve in the dimensions they feel they might still lack. Ultimately, the audience enjoying the music can also gain with improved performances from bands playing the music of their choice (Latimer et al., 2010).

String Performance Assessment

String performances come in two types: ensemble or individual. In an ensemble performance, there is a range of players, from a duet to a full symphony orchestra with approximately 135 total personnel (Roesner, 2018). With this spectrum, each player has a role and responsibility for the overall outcome, and the goal is to work together to create a balanced sound that complements each instrument (Pitts, 2016). When moving beyond an ensemble to an individual performance, the solo player chooses all components of their production and has complete control and responsibility for the outcome (Cohen, 2017).

Performance as part of an ensemble or solo in a competition provides a single summative evaluation of student ability to effectively understand the piece and an individual instrument. According to Bergee (2007), teachers provide numerous evaluations in different settings and enable them to mitigate various aspects identified regarding the performance evaluation subjectivity. Bergee (2007) suggested significant theoretical as well as practical approaches regarding festival management, thus policy suggestions have been pointed out on how the state may be able to promote the support of art festivals through paying more attention to cultural festivals like operatic or musical festivals (Bergee, 2007).

There are various influences on solo and small ensemble festival ratings. Bergee and McWhirter (2005) identified there was a major statistical difference within the three components: the main impact being of time of day and then the event type and school size making an impact when it comes to developing a valid assessment practice. This study also pointed out that there was a significant difference in performing medium (Bergee & McWhirter, 2005). To be able to achieve a valid assessment on music performance variables such as geographical location as well as district-level average, daily attendance was added (Bergee & McWhirter, 2005).

The Ensemble. An ensemble is a conglomeration of more than one player and instrument. Roesner (2018) said that performance is a tapestry of experience between conductor, player, instrument, and audience. It is a rich atmosphere of human connectivity that transcends the realms of verbal communication nuanced by sound and directed music motions (Roesner, 2018). According to Roesner, there is an embodiment of some sort of desired goal and the inherent reality that all the players involved are aiming to achieve this goal or objective in a communal and deeply complementary way. Such is the characterization of an ensemble (Pitts, 2016; Roesner, 2018). According to Morrison and Selvey (2014), the instruments and their players cue each other in the pursuit of musicality or crescendo of sorts. The ensemble brings a need to adjust internal time clocks into asynchrony that relays itself in the intonation and dynamism encompassed in a performance (Morrison & Selvey, 2014; Roesner, 2018). Auditory and visual feedback is paramount in having a loop of information between members of an ensemble to the end of them achieving their music goals (Roesner, 2018). By and large, string quartets are conductor-less sets of four performers with a mastery, evident or implied, of dividing their attention between shaping their performance and keeping alignment of sorts with the other members of the ensemble (Roesner, 2018). According to Roesner, it all makes for a true test of cognitive and anticipatory prowess that is refined through practice and instruction. Interdependence on things such as tempo and movement is hard to account for and often is corrected through a stimulus and response process that follows no written laws (Pitts, 2016; Roesner, 2018).

Students who are members of noncompetitive bands score significantly higher on music aptitude tests than their counterparts from competitive bands. Mick and Pope

(2018) sought to identify how school level, festival level, music classification, and instrumentation influence the overall performance ratings of orchestral performances at large ensemble festivals. There exists criticism of the competitive nature of adjudicated festivals, with directors being focused too much on winning while sacrificing student improvement and learning (Mick & Pope, 2018). Moreover, there has been controversy over using ranking versus rating evaluation systems (Mick & Pope, 2018). According to Mick and Pope, one side of the debate prefers ranking to produce higher music standards and better performances, but the other school concludes ratings are the best option. All in all, directors use the criterion scores and comments of adjudicators to assess the effectiveness of their teaching methods and for monitoring the musical progress of their students (Mick & Pope, 2018). Mick and Pope believed students can utilize adjudicators' feedback to assess their performance strengths and weaknesses, thereby helping them improve future performance. The data also indicate high overall performance ratings for orchestra performances (Mick & Pope, 2018). The orchestras were also predominantly assigned a 1 (superior) or 2 (excellent) rating for district-level performances in the sight-reading rooms (Mick & Pope, 2018). Mick and Pope confirmed the skewed results of festivals toward the highest ratings.

An adjudicator's assessment can provide an inaccurate picture of a concert performance. Mick and Pope (2018) believed adjudicators can learn from the experiences of their fellow experts about what to do or not do when evaluating the quality of music. There is a likelihood that adjudicators exaggerate the performances of orchestras on the assessment forms (Mick & Pope, 2018). Mick and Pope believed it is important to understand existing discrepancies in adjudicators' set performance standards for concerts.

In this way, future directors and judges of adjudicated festivals can judge players more objectively and avoid giving inaccurate ratings (Mick & Pope, 2018). Besides, players can also learn not to judge the quality of their playing based on the concert assessment forms as they may not represent their skill levels (Mick & Pope, 2018). According to Mick and Pope, music educators can learn from the study findings so they can develop more standardized performance rating systems to avoid exaggerated assessments. This will enhance the quality of music and improve the players' skills as they strive to meet the standards set for the overall industry (Mick & Pope, 2018).

Reflective practice with real-world data helps to reveal weaknesses in performance rating systems. Mick and Pope (2018) reviewed authentic, real-world result data from the Florida Music Education Association all district- and state-level concert festival. Mick and Pope felt this makes it easy for the researchers to examine the rating and assessment criteria for festival performances because they do not have to collect data anew. Mick and Pope believed it is possible to have a more reflective investigation into the validity and fairness of how adjudicators and judges might have rated the skill levels of orchestras. Students seeking to improve their own abilities can look at the data to gauge their status as players of music and whether they meet industry expectations (Mick & Pope, 2018).

The Individual. An individual player is bound by no set of rules. According to Morrison and Selvey (2014), they have no secret language they need to coordinate with another performer except for ensuring they pick the cues, musical freedom, they set for themselves within their performance. Individual performance is heavily self-reliant in that performers must plan out their entire engagement by themselves and follow through every step of the planned process (Cohen, 2017). A slight gap or mistake will leave them vulnerable without any fallback to cover for their missed bases (Morrison & Selvey, 2014). Performances done by individuals are rich in individual exploration and allow the performer to experiment with the strengths of their chosen instruments in ways an ensemble does not allow (Jack et al., 2017). According to Cohen (2017), individual performers take up a role to entice an entire audience by themselves. For an ensemble, this is a smaller load to lift (Pitts, 2016). By the singularity of their instrument choice, the individual performer only performs at the strength of one instrument (Cohen, 2017; Morrison & Selvey, 2014; Pitts, 2016).

Chapter 3: Methodology

Overview

This study developed, validated, and tested a weighted individual performance-based assessment tool to improve the assessment process for middle school orchestral strings. The assessment tool provides clear alignment of skill level with specific observable items. The assessment provides both the teacher and student with valid documented achievement data. Providing specific feedback opportunities, the documented data provide a consistent process to track individual student growth. Through data analysis, teachers identified trends and adjusted classroom instruction to meet the students' needs. The data provide an opportunity for reflection, supporting identification of professional needs and opportunities for professional growth. To be the most significant outcome, teachers provided consistent achievement reports with all stakeholders and advocated for resources and materials to strengthen areas of instructional weakness. This chapter includes the research questions, design, steps, and limitations of the study.

Research Questions

The study established the validity and reliability of a weighted individual performance-based assessment tool within the utility scope of middle school orchestral strings. The following research questions guided this study:

1. What specific string-playing behaviors and corresponding criteria validate a weighted individual performance-based assessment tool for middle school orchestral strings?
2. What are the psychometric properties of the weighted individual performance-

based assessment tool in authentic situations?

Research Design

This study utilized a quantitative approach, methods, and research strategies (Creswell & Creswell, 2018). The quantitative approach utilized the Polytomous Rasch Model. This model's objective measured ability through analyzing responses to constructs that are scored with successive numbers (rating scale). Quantitative methods were utilized to help construct the assessment tool and test its validity and reliability. Quantitative research strategies in this study obtained interrater reliability and analyzed data from a pilot study and final item pool study.

Step 1: Item and Scale Development

Participants. Utilizing the North Carolina Music Educators Association database, participants were recruited by email (Appendix A). An expert panel (*N*=5) was selected based on the first five responses that met the study's expert panel membership criteria of being either a college professor, retired orchestral music teacher, or an active National Board-certified teacher not participating in the pilot study. Due to the amount of needed communication and collaboration in this study, I recruited the most eager, active participants through convenience sampling. These members are experts in the field of orchestral strings and completed informed consent forms to participate in this study (Appendix B). Table 1 introduces the expert panel, their total years of experience, their current educator status, gender, and race.

Table 1

Expert Panel Roster

Identification code	Years	Educator status	Gender	Race
EP1	32	Retired orchestral music teacher	Female	White
EP2	34	College professor	Male	Asian
EP3	12	National board-certified teacher	Male	Black
EP4	8	National board-certified teacher	Female	White
EP5	20	Retired orchestral music teacher	Female	White

Note. Identification code system used to keep member's identity confidential.

Due to the impact of the COVID-19 pandemic (World Health Organization, 2021), the panel utilized email, phone calls and text, Zoom, Google Forms, and Google Sheets to work safely in a completely virtual environment. The expert panel was responsible for the construction and validity of the weighted individual assessment tool.

Procedure. I convened the expert panel (*N*=5) to collaborate in the item and scale development of a weighted individual assessment tool for middle school orchestral strings to align and answer Research Question 1 of this study. All expert panel's item and scale development tasks establish specific string-playing behaviors, and corresponding criteria validate a weighted individual performance-based assessment tool for middle school orchestral strings. I tasked the expert panel to collaboratively select 10 string-playing behaviors to comprehensively assess a middle school string musician's playing ability. Through discussion and 100% mutual agreement, expert panel members selected 10 string-playing behaviors to be included in the assessment item pool. Table 2 introduces the final list of 10 assessment items.

Table 2

Assessment Item Pool

Assessment item
1: Tempo
2: Rhythm
3: Tone
4: Pitch
5: Intonation
6: Technique
7: Bowing
8: Dynamics
9: Phrasing
10: Posture

Note. Items not prioritized.

Limiting the pool to 10 items, we structured this tool's potential to be delivered individually during one class period without omitting quality items for a comprehensive performance assessment. After 100% mutual agreement was achieved, I was tasked to construct descriptors for each assessment item.

I created descriptors for each level of proficiency from 1 (unsatisfactory) to 5 (exemplary) for each assessment item. It was important for each descriptor to describe string-playing behaviors fully across this performance scale. Each assessment item entered an agreement cycle to be approved by the expert panel. The panel rated each descriptor for agreement aligned with their proficiency level. Terminology and actions were adjusted through expert panel discussion to establish clarity for assessment facilitators. The panel had to reach 100% agreement on each descriptor before ending an assessment item agreement cycle. We repeated this process for each item in the pool (*N*=10). At the end of all agreement cycles, the assessment tool included 10 assessment items with five descriptors each (*N*=50) that established proficiency from 1 (unsatisfactory) to 5 (exemplary).

Once the descriptors for each assessment item (*N*=50) were 100% mutually agreed upon, the panel individually rated the list of assessment items (*N*=10) by importance from 1 (least important) to 10 (most important). This prioritization developed specific quality points for obtaining levels of proficiency through each descriptor of that assessment item. The expert panel's prioritization scores of each assessment item were averaged and used as a multiplier for that assessment item (higher rate importance equals a higher score multiplier) Each assessment item multiplier would add to a participant's performance level (1-5) in those specific categories to emphasize the importance of those string-playing behaviors. For example, a performer scoring 5 of 5 with a multiplier of 9 would score 45 raw points for that particular assessment item.

After assessment item multipliers were identified, the expert panel totaled raw scores for a participant scoring all performance levels of 1s (unsatisfactory), 3s (average), and 5s (exemplary). Based on these total possible raw scores, the panel determined score ranges for a participant's total proficiency in the areas of below standard, meets standard, and exceeds standard. Table 3 introduces the score range proficiency levels.

Table 3

Score Range Proficiency

Below standard	Meets standard	Exceeds standard
55-179	180-246	247-275

Note. Based on total possible raw score.

These score ranges reflect the total proficiency of participants based on their performance of all 10 assessment items.

After completion of expert panel tasks, I inputted the 10 assessment items with their five descriptors each into a Google Form to create the Digital Individual String Assessment Tool (DISAT; Appendix C). The DISAT was created in digital format to provide instant assessment calculations and disaggregation of data as well as provide a safe, green individual assessment option for facilitators. I linked a Google Sheet to the form to collect responses. Within this sheet, I inputted the individual assessment item multipliers and coded the sheet to provide assessment data calculations. These calculations provided a participant's individual assessment score, each assessor's combined responses in each category, their class average, and the total population data. The DISAT was the instrument utilized for the rest of the study to collect data to assess validity and reliability.

Step 2: Pilot Study

Participants. Utilizing the North Carolina Music Educators Association database, participants were recruited by email (Appendix A). The pilot assessor team (*N*=5) was selected on the first five responses that met the study's pilot assessor team membership criteria of being a licensed full-time middle school orchestra teacher and having access to the total number of students needed to collect pilot data (*N*=200) for Polytomous Rasch Model analysis. Due to the amount of needed communication and collaboration in this study, I recruited the most eager, active participants through convenience sampling. These members are active educators in the field of orchestral strings and completed informed consent forms to participate in this study (Appendix D). Table 4 introduces the pilot assessor team, their total years of experience, current educator status, the number of orchestral strings students, gender, and race.

Table 4

Pilot Assessor Team Roster

Identification code	Years	Educator status	Number of students	Gender	Race
PA1	28	National Board-certified teacher	47	Male	White
PA2	16	National Board-certified teacher	31	Female	Black
PA3	9	Licensed full-time teacher	43	Male	White
PA4	32	National Board-certified teacher	50	Female	Asian
PA5	4	Licensed full-time teacher	37	Female	Black

Note. Identification code system used to keep member's identity confidential.

Due to the impact of the COVID-19 pandemic (World Health Organization, 2021), the panel utilized email, phone calls and text, Zoom, Google Forms, and Google Sheets to work safely in a completely virtual environment. The pilot assessor team was responsible for the reliability of the weighted individual assessment tool by administering the validated assessment tool and participating in team training sessions. The expert panel members (*N=5)* were also participants in the pilot study to approve the tool's ability to be utilized in the final item pool study.

Procedure. I convened the pilot assessor team (*N*=5) for a pilot study of the DISAT in their middle school orchestral string programs to align and answer Research Question 2 of this study. All pilot assessor's pilot study tasks collect and analyze psychometric properties of the weighted individual performance-based assessment tool in authentic situations. Before implementing the tool, the pilot assessor team trained on DISAT assessment procedures to ensure a proper and consistent approach. For uniformity, the training session helped each assessor approach the assessment with the same lens. After completing our training session, the pilot assessor team utilized the DISAT for a standardized individual assessment of their students (*N*=208). This sample size was needed to test the initial function of the tool and highlight any areas needing revisions before a full final study.

Assessment data were automatically submitted after each individual assessment with no student personal identifiable data (Appendix E). The data were organized and sorted to display the comprehensive item selection. Assessment data submissions were also organized and sorted for each individual pilot assessor (Appendix F). Since the sole purpose of this study was to create, validate, and test the reliability of the DISAT, individual pilot assessor data were only disseminated back to each assessor for their data records and pedagogical reflection. These data were not analyzed in comparison with other pilot assessors.

I used the Polytomous Rasch Model to analyze the DISAT pilot data. This method allowed me to assess how constructs were functioning in this common performance scale among different raters to obtain reliability and validity. The first step of my Polytomous Rasch process was to assess the person test reliability of the pilot data. Utilizing the formula $\frac{OV-EV}{OV}$, I took the observed variance (OV) of my pilot assessors' subtracted mean of squared standard errors (EV) and divided by the observed variance (Linacre, 2015). Person reliability gauges the separation between raters and whether the tool is sensitive enough to distinguish between high- and low-performing constructs (Linacre, 2015). This sensitivity provides reliability to the tool's function. According to Linacre (2015), a person reliability greater than 0.8 is acceptable to distinguish that separation between raters.

The next step was to analyze the pilot data correlation matrix to see the impact constructs had on each other during the assessment process. I looked for significantly strong positive and negative correlations. Strong positive correlations (0.7 – 1.0) are observed when items consistently move together, while strong negative correlations (-0.7 – -1.0) are observed when items consistently move apart. For example, if one correlation item scores higher and the other item scores lower consistently, you observe a negative correlation. The rate of this correlation strength signifies the movement's dependency on one another. Significantly strong positive and/or negative correlations can skew performance data.

After assessing assessment item correlations, I analyzed each assessment item in the partial credit model. The formula, $\ln\left[\frac{P_{ni(xi=k)}}{P_{ni(xi=k-1)}}\right] = \theta_n - \delta_i - \tau_{ik}$, provides estimates concerning item difficulty and their thresholds (Wind & Hua, 2021). I notated any observable measures of partial credit being awarded to any constructs through the five ability levels. These data inform me if the linear performance scale has proper sequential function. For example, if 5 is the highest achievable score, 4 does not show the highest ability level.

After analyzing the partial credit model, I took a closer look at the individual assessment item Fit statistics and assessed their functionality and validity. I analyzed each assessment item's infit statistics, $\frac{\sum_{v=1}^{n} R_{vi}^{2}}{\sum_{v=1}^{n} VAR(X_{vi})}$, to notate any misfit readings among the pilot data (Müller, 2020). The infit statistics allowed me to assess the pattern of targeted assessment responses and misfit readings that distort the data's picture. These data signify whether items are performing properly for the people for whom the item was targeted (response patterns). This statistic has the most impact on the tool's validity. For example, if an assessment item consistently scores 5 between different raters in a sample population, the item is not functioning properly to assess range of ability between the population. This would be expressed as a misfit (mean square value of greater than 2.0).

After compiling all analysis pieces, I met with the pilot assessors. During the meeting, I first thanked them for their dedication to the study and proceeded to give an analysis overview of the pilot study data. After the overview, I informed them of their next steps for the upcoming final item pool study. Concluding the pilot assessor meeting, I met with the expert panel. I presented the analysis overview and informed them the pilot assessors would convene again for the final item pool study.

Step 3: Final Item Pool Study

I convened the pilot assessor team (*N*=5) for a final item pool study of the DISAT in their middle school orchestral string programs to align and answer Research Question 2 of this study. All pilot assessor's final item pool study tasks collect and analyze psychometric properties of the weighted individual performance-based assessment tool in authentic situations. Before completing the final item pool assessments, I met with the pilot assessor team. The pilot assessor team reviewed DISAT assessment procedures to ensure a proper and consistent approach during the final item pool study. For uniformity, the review session helped each assessor approach the assessment with the same lens.

After completing our session, the pilot assessor team utilized the DISAT for a standardized individual assessment of their students (N=222). The final item pool assessments were collected and combined with the original pilot assessments (N=430). Assessment data were automatically submitted after each individual assessment with no student personal identifiable data (Appendix G). The data were organized and sorted to display the comprehensive item selection.

Assessment data submissions were also organized and sorted for each individual pilot assessor (Appendix H). Since the sole purpose of this study was to create, validate, and test the reliability of the DISAT, individual pilot assessor data were only disseminated back to each assessor for their data records and pedagogical reflection. These data were not analyzed in comparison with other pilot assessors.

I used the Polytomous Rasch Model to analyze the DISAT final item pool study data. This method allowed me to assess how constructs were functioning in this common performance scale among different raters to obtain reliability and validity. Based on their assessment responses, the first step of my Polytomous Rasch process was to assess the person test reliability. Utilizing the formula $\frac{OV-EV}{OV}$, I took the observed variance (OV) of my pilot assessors subtracted mean of squared standard errors (EV) and divided by the observed variance (Linacre, 2015). This reliability gauges the separation of responses between assessors and whether the tool is sensitive enough to distinguish between high-

and low-performing constructs. This sensitivity provides reliability to the tool’s function.

The next step was to analyze the pilot data correlation matrix to see the impact constructs had on each other during the assessment process. For this research study, data points between -0.7 and 0.7 are within the standard (Linacre, 2015). I looked for significantly strong positive and negative correlations. Strong positive correlations (0.7 – 1.0) are observed when items consistently move together, while strong negative correlations (-0.7 – -1.0) are observed when items consistently move apart. For example, if one correlation item scores higher and the other item scores lower consistently, you observe a negative correlation. The rate of this correlation strength signifies the movement’s dependency on one another. Significantly strong positive and/or negative correlations can skew performance data.

After assessing assessment item correlations, I analyzed each assessment item in the partial credit model. The formula, ln $\left[\frac{P_{ni(xi=k)}}{P_{ni(xi=k-1)}}\right] = \theta_n - \delta_i - \tau_{ik}$, provides estimates concerning item difficulty and their thresholds (Wind & Hua, 2021). I notated any observable measures of partial credit being awarded to any constructs through the five ability levels. These data inform me if the linear performance scale has proper sequential function. For example, if 5 is the highest achievable score, 4 does not show the highest ability level.

After analyzing the partial credit model, I took a closer look at the individual assessment item Fit statistics and assessed their functionality and validity. I analyzed each assessment item’s infit statistics, $\frac{\sum_{v=1}^{n} R_{vi}^{2}}{\sum_{v=1}^{n} VAR(X_{vi})}$, to notate any misfit readings among the pilot data (Müller, 2020). These data signify whether items are performing properly for the people for whom the item was targeted (response patterns). This statistic has the most impact on the tool's validity. For example, if an assessment item consistently scores the same proficiency level between different raters in a sample population, the item is not functioning properly to assess range of ability between the population.

Since this was the final item pool study, I took a closer look at the data by analyzing the Wright Map. A Wright Map is divided into two vertical columns of data. The left side presents participant data, and the right side presents assessment item data. Participant data focuses on organizing the ability levels from most able at the top down to least able at the bottom. Assessment item data focuses on organizing item difficulty from most difficult at the top down to the least difficult at the bottom. The Wright Map charted all 10 items' difficulties and expressed the separation between constructs. These data identify string-playing behaviors that may be prioritized in the classroom as well as those behaviors that need more instructional attention.

For the final data analysis piece, I focused on each individual assessment construct's Item Response Function (IRF). The formula, $P_{ij}\left(\theta_j, b_i\right) = \frac{\exp\left(\theta_j - b_i\right)}{1+\exp\left(\theta_j - b_i\right)}$, expressed the probability of a correct response based on the individual's ability level (Linacre, 2015). Each assessment item was compared to the Rasch Model prediction from inputted final item pool data. Based on these data, the Rasch Model charts out a unique curve that symbolizes the expected score of participants in each ability level range for that particular assessment item. Simply, this curve represents the expected score based on a participant's ability. The final item pool study data are plotted on these IRF graphs for each individual assessment item. I analyzed each construct graph to make note of any ability group separations from the expected score curve.

After compiling all analysis pieces, I met with the expert panel and pilot assessor team. During the meeting, I gave a data analysis overview of the final item pool study data. After the data analysis, I debriefed them on their participation in the study and thanked them for their time and effort.

Chapter 4: Results

Overview

In collaboration with many educational professionals, we developed, validated, and tested a weighed individual performance-based assessment tool for middle school orchestral strings with a three-step plan. The expert panel collaborated in Step 1 for the individual performance-based assessment tool's item and scale development. Reflecting on the impact and challenges of COVID-19 within our schools, the team was in mutual agreement for a digital platform for our tool. Utilizing a Google Form and coded Google Sheets, we built and titled this tool the DISAT (Appendix C). Respectively in Steps 2 and 3, the DISAT was tested and validated through a pilot study and final item pool study by the pilot assessor team. To give a complete 360-degree view of this study's findings, I align my results to the research questions that guided this study with corresponding process steps.

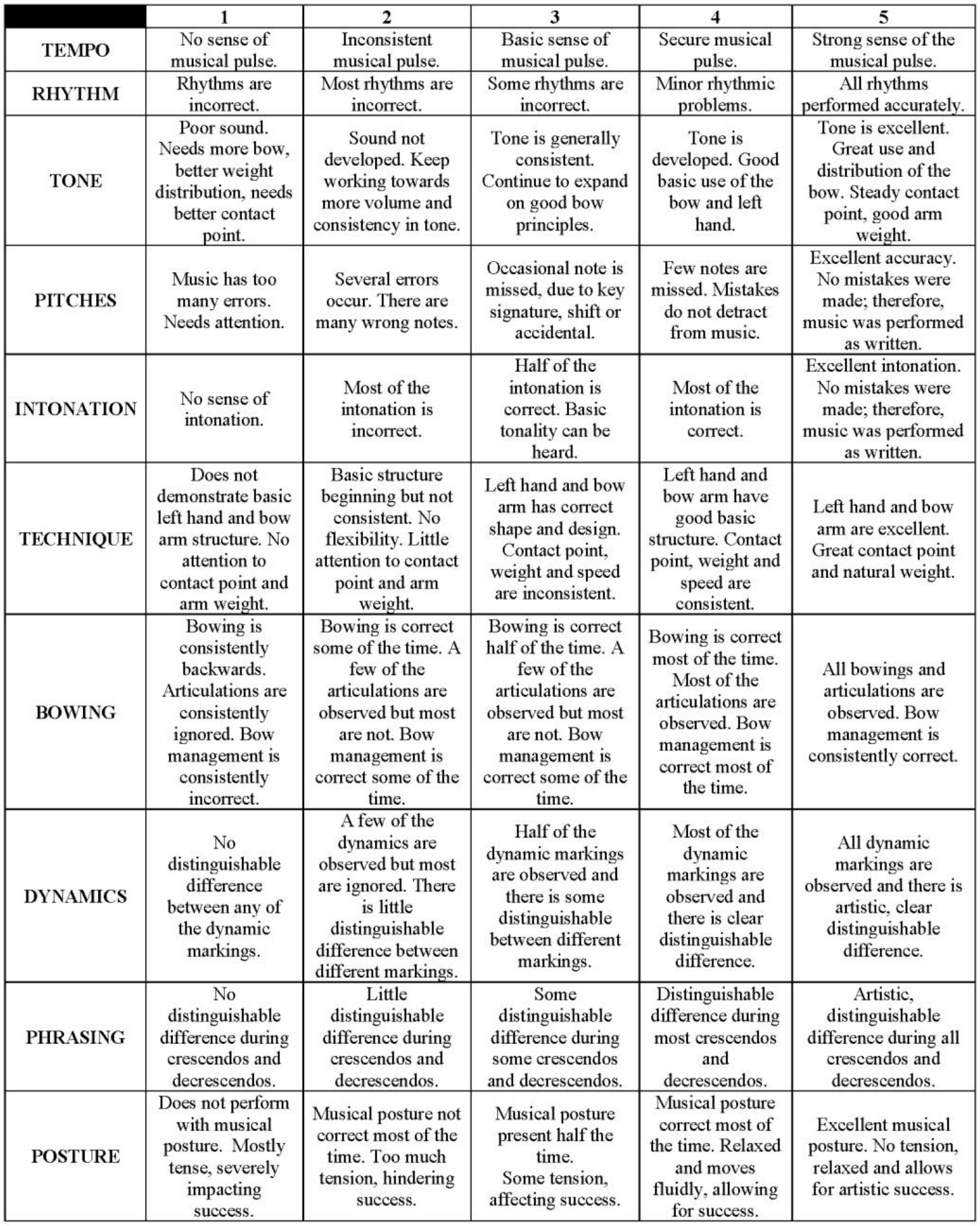
Research Question 1

What Specific String-Playing Behaviors and Corresponding Criteria Validate a Weighted Individual Performance-Based Assessment Tool for Middle School Orchestral Strings?

Step 1: Item and Scale Development. The expert panel's (n=5) mutually agreed upon assessment item pool from Table 2 identified tempo, rhythm, tone, pitch, intonation, technique, bowing, dynamics, phrasing, posture as the top 10 string-playing behaviors that were important to be part of a performance-based assessment. Once the assessment item pool was selected, the expert panel participated in agreement cycles to adopt proficiency descriptors for each assessment item of the final pool (*N*=50). Figure 1 represents the mutually agreed upon assessment item proficiency descriptors.

Figure 1

Assessment Item Proficiency Descriptors



	1	2	3	4	5
TEMPO	No sense of musical pulse.	Inconsistent musical pulse.	Basic sense of musical pulse.	Secure musical pulse.	Strong sense of the musical pulse.
RHYTHM	Rhythms are incorrect.	Most rhythms are incorrect.	Some rhythms are incorrect.	Minor rhythmic problems.	All rhythms performed accurately.
TONE	Poor sound. Needs more bow, better weight distribution, needs better contact point.	Sound not developed. Keep working towards more volume and consistency in tone.	Tone is generally consistent. Continue to expand on good bow principles.	Tone is developed. Good basic use of the bow and left hand.	Tone is excellent. Great use and distribution of the bow. Steady contact point, good arm weight.
PITCHES	Music has too many errors. Needs attention.	Several errors occur. There are many wrong notes.	Occasional note is missed, due to key signature, shift or accidental.	Few notes are missed. Mistakes do not detract from music.	Excellent accuracy. No mistakes were made; therefore, music was performed as written.
INTONATION	No sense of intonation.	Most of the intonation is incorrect.	Half of the intonation is correct. Basic tonality can be heard.	Most of the intonation is correct.	Excellent intonation. No mistakes were made; therefore, music was performed as written.
TECHNIQUE	Does not demonstrate basic left hand and bow arm structure. No attention to contact point and arm weight.	Basic structure beginning but not consistent. No flexibility. Little attention to contact point and arm weight.	Left hand and bow arm has correct shape and design. Contact point, weight and speed are inconsistent.	Left hand and bow arm have good basic structure. Contact point, weight and speed are consistent.	Left hand and bow arm are excellent. Great contact point and natural weight.
BOWING	Bowing is consistently backwards. Articulations are consistently ignored. Bow management is consistently incorrect.	Bowing is correct some of the time. A few of the articulations are observed but most are not. Bow management is correct some of the time.	Bowing is correct half of the time. A few of the articulations are observed but most are not. Bow management is correct some of the time.	Bowing is correct most of the time. Most of the articulations are observed. Bow management is correct most of the time.	All bowings and articulations are observed. Bow management is consistently correct.
DYNAMICS	No distinguishable difference between any of the dynamic markings.	A few of the dynamics are observed but most are ignored. There is little distinguishable difference between different markings.	Half of the dynamic markings are observed and there is some distinguishable between different markings.	Most of the dynamic markings are observed and there is clear distinguishable difference.	All dynamic markings are observed and there is artistic, clear distinguishable difference.
PHRASING	No distinguishable difference during crescendos and decrescendos.	Little distinguishable difference during crescendos and decrescendos.	Some distinguishable difference during some crescendos and decrescendos.	Distinguishable difference during most crescendos and decrescendos.	Artistic, distinguishable difference during all crescendos and decrescendos.
POSTURE	Does not perform with musical posture. Mostly tense, severely impacting success.	Musical posture not correct most of the time. Too much tension, hindering success.	Musical posture present half the time. Some tension, affecting success.	Musical posture correct most of the time. Relaxed and moves fluidly, allowing for success.	Excellent musical posture. No tension, relaxed and allows for artistic success.

Note. 1 (unsatisfactory) to 5 (exemplary).

These descriptors identified corresponding criteria to assess a participant's proficiency from 1 (unsatisfactory) to 5 (exemplary) and targeted specific observable characteristics of each string-playing behavior. Reflecting on this chart, the expert panel rated each assessment item from 1 (least important) to 10 (most important) to prioritize the assessment item pool. Table 5 represents the expert panel's prioritization on the assessment items (*N*=10).

Table 5

Expert Panel Assessment Item Prioritization

Assessment item	EP1	EP2	EP3	EP4	EP5	Average
1: Posture	10	10	9	10	9	9.6
2: Tone	9	8	10	9	8	8.8
3: Tempo	7	9	8	8	7	7.8
4: Rhythm	8	7	6	7	10	7.6
5: Pitch	6	6	7	5	4	5.6
6: Bowing	4	5	5	6	6	5.2
7: Technique	5	4	4	4	5	4.4
8: Intonation	2	3	2	3	3	2.6
9: Dynamics	3	2	3	2	1	2.2
10: Phrasing	1	1	1	1	2	1.2

Note. 1 (least important) to 10 (most important).

Their ratings were averaged to produce assessment item multipliers for the weighted system. Posture was identified as the most important item gaining a 9.6 assessment multiplier with a 0.8 separation from the next assessment item. Phrasing was identified as the least important item gaining a 1.2 assessment multiplier with a 1.0 separation from the next assessment item. These elements were combined to create the DISAT.

Research Question 2

What Are the Psychometric Properties of the Weighted Individual Performance-Based Assessment Tool in Authentic Situations?

Step 2: Pilot Study. The pilot assessors (*N*=5) trained on DISAT assessment procedures and utilized the tool for a combination of 208 assessments. Table 6 represents the pilot assessor team's comprehensive item selection (*N*=208).

Table 6

Pilot Comprehensive Item Selection

Assessment item	Proficiency levels					Proficiency average
	1	2	3	4	5	
1: Tempo	6	24	25	56	97	4.03
2: Rhythm	12	34	31	47	84	3.75
3: Tone	7	9	24	39	129	4.32
4: Pitch	6	24	46	65	67	3.78
5: Intonation	35	40	46	44	43	3.10
6: Technique	7	25	47	76	53	3.69
7: Bowing	3	8	19	52	126	4.39
8: Dynamics	16	16	25	46	105	4.00
9: Phrasing	63	45	28	38	34	2.69
10: Posture	5	10	22	61	110	4.25

Note. Completed 208 individual assessments.

After compiling pilot assessments (*N*=208), I completed a Polytomous Rasch Model analysis of pilot assessor (PA) data. Obtaining a person test reliability of 0.878, the study had enough variance between raters to establish reliability. The separation variance proves the tool is sensitive enough to distinguish between high- and low-performing constructs. After assessing the person test reliability, I analyzed the correlation among assessment items. The correlation matrix identified two strong positive correlations. Table 7 and Table 8 present the pilot data's Q3 Correlation Matrix.

Table 7

Part 1 Pilot Q3 Correlation Matrix

	Tempo	Rhythm	Tone	Pitch	Intonation
1: Tempo	__				
2: Rhythm	0.775	__			
3: Tone	-0.353	-0.261	__		
4: Pitch	-0.194	-0.105	0.120	__	
5: Intonation	-0.167	-0.152	-0.051	0.497	__
6: Technique	-0.104	-0.148	0.010	-0.137	0.061
7: Bowing	-0.216	-0.241	-0.114	-0.162	-0.079
8: Dynamics	-0.245	-0.298	-0.152	-0.267	-0.241
9: Phrasing	-0.241	-0.297	-0.062	-0.206	-0.249
10: Posture	-0.201	-0.226	-0.084	-0.173	-0.200

Note. First five assessment item constructs of the pilot study.

Table 8

Part 2 Pilot Q3 Correlation Matrix

	Technique	Bowing	Dynamics	Phrasing	Posture
6: Technique	__				
7: Bowing	0.171	__			
8: Dynamics	-0.381	0.101	__		
9: Phrasing	-0.289	-0.189	0.281	__	
10: Posture	0.014	-0.088	-0.077	0.016	__

Note. Last five assessment item constructs of the pilot study.

The strongest positive correlation of 0.775 was between tempo and rhythm assessment items. This strong positive correlation can be explained because music is simply noise organized (rhythm) by time (tempo). Rhythmic accuracy is dependent on the participant's ability to internalize the proper tempo of the selected exercise and execute the correct subdivision and/or augmentation of beats. The other strong positive correlation of 0.497 was between the pitch and intonation assessment items. This strong positive correlation can be explained because a note (pitch) must also be played with correct tuning (intonation). Melodic accuracy is dependent on the participant's ability to play the correct pitch as well as to ensure the pitch is adjusted (tuned) for its harmonic intended purpose. Ranging between -0.353 and 0.171, correlations between all other assessment items were not significantly positive and/or negative.

I observed sequential movement from the left threshold (1) to the right threshold (5) in the pilot's partial credit model. Table 9 presents the pilot data's Partial Credit Model.

Table 9

Pilot Partial Credit Model

		Threshold				
	Measure	1	2	3	4	5
1: Tempo	-9.94	-11.53	-3.65	-2.3470	-1.397	0.0340
2: Rhythm	-9.02	-11.02	-3.14	-1.8355	-0.886	0.5454
3: Tone	-9.96	-12.00	-4.29	-2.9895	-2.040	-0.6086
4: Pitch	-9.25	-11.07	-3.19	-1.8866	-0.937	0.4943
5: Intonation	-8.30	-9.94	-2.05	-0.7530	0.197	1.6279
6: Technique	-9.18	-10.90	-3.02	-1.7185	-0.769	0.6624
7: Bowing	-10.11	-12.00	-4.49	-3.1922	-2.243	-0.8112
8: Dynamics	-9.81	-11.47	-3.59	-2.2896	-1.340	0.0913
9: Phrasing	-7.76	-9.27	-1.39	-0.0856	0.864	2.2955
10: Posture	-9.96	-12.00	-4.14	-2.8368	-1.887	-0.4557

Note. The Thurstonian threshold for a score category is defined as the ability at which the probability of achieving that score or higher reaches 0.50.

No observations were made of partial credit being awarded to any constructs. This finding confirmed that all assessment item descriptors were functioning properly to assess a participant’s item proficiency between unsatisfactory (1) and exemplary (5).

There were no random responses by low performers (outliers) and items are performing properly for the people of whom the items are targeted (response patterns). Table 10 presents the pilot data’s item statistics (Rating Scale Model).

Table 10

Pilot Item Statistics (Rating Scale Model)

	Measure	S.E. Measure	Infit	Outfit
1: Tempo	-3.78	0.117	0.965	0.956
2: Rhythm	-3.27	0.112	1.114	1.106
3: Tone	-4.42	0.127	1.405	1.412
4: Pitch	-3.32	0.112	0.597	0.593
5: Intonation	-2.18	0.108	0.758	0.760
6: Technique	-3.15	0.111	0.752	0.747
7: Bowing	-4.62	0.131	1.029	1.032
8: Dynamics	-3.72	0.116	1.486	1.491
9: Phrasing	-1.52	0.110	1.220	1.223
10: Posture	-4.27	0.124	1.136	1.152

Note. Infit=Information-weighted mean square statistic; Outfit=Outlier-sensitive means square statistics. Person reliability of 0.878.

Being productive for measurement, all assessment item outfit mean squares and infit mean squares are within the acceptable parameter of 0.5 to 1.5 (Linacre, 2002). The acceptable outfit mean squares signify there were no random responses by low performers (outliers). The acceptable infit mean squares signify the items are performing properly for the people for whom the items are targeted (response patterns). Infit mean squares have the greatest impact on measurement validity, and I believe it is important to note assessment items that reside near the muted (lower) and noisy (upper) end of the acceptable range. The pitch assessment item infit mean square of 0.597 is slightly less stable on the statistical muted (lower) end of the acceptable range. The dynamics assessment item infit mean square of 1.486 is slightly less stable on the statistical noisy (higher) end of the acceptable range. According to Linacre (2002), infit mean square measurements below 0.5 are less productive and may produce misleadingly good reliabilities and separations. Infit mean square measurements between 1.5 and 2.0 are unproductive for construction of measurement but not degrading. Infit measurements greater than 2.0 distort or degrade the measurement system. The pilot study data showed the DISAT was reliable and validated to continue further testing without adjustments from the expert panel.

Step 3: Final Item Pool Study. The pilot assessors (*N*=5) again trained on DISAT assessment procedures and utilized the tool for a total study combination of 430 assessments. Table 11 represents the pilot assessor team's final study comprehensive item selection (*N*=430).

Table 11

Final Study Comprehensive Item Selection

Assessment item	Proficiency levels					Proficiency average
	1	2	3	4	5	
1: Tempo	9	46	61	125	189	4.02
2: Rhythm	18	82	65	97	168	3.73
3: Tone	15	22	51	92	250	4.26
4: Pitch	13	46	100	136	135	3.78
5: Intonation	72	86	99	80	93	3.08
6: Technique	14	45	108	156	107	3.69
7: Bowing	4	27	41	108	250	4.33
8: Dynamics	25	34	49	116	206	4.03
9: Phrasing	135	96	56	74	69	2.64
10: Posture	8	15	42	136	229	4.31

Note. Completed 430 individual assessments.

After compiling all assessments (*N*=430), I again completed a Polytomous Rasch Model analysis of pilot assessor data. Obtaining a person test reliability of 0.872, the study again had enough variance between raters to establish reliability. The separation variance again proves the tool is sensitive enough to distinguish between high- and low-performing constructs. After assessing the person test reliability, I analyzed the correlation among assessment items. The correlation matrix again identified two strong positive correlations. Table 12 and Table 13 present the final item pool's Q3 Correlation Matrix.

Table 12

Part 1 Final Item Pool Q3 Correlation Matrix

	Tempo	Rhythm	Tone	Pitch	Intonation
1: Tempo	__				
2: Rhythm	0.796	__			
3: Tone	-0.382	-0.248	__		
4: Pitch	-0.271	-0.239	0.229	__	
5: Intonation	-0.325	-0.345	0.068	0.550	__
6: Technique	-0.106	-0.147	-0.095	-0.052	0.108
7: Bowing	-0.190	-0.198	-0.180	-0.202	-0.151
8: Dynamics	-0.230	-0.258	-0.079	-0.296	-0.174
9: Phrasing	-0.221	-0.286	-0.134	-0.196	-0.178
10: Posture	-0.153	-0.183	-0.066	-0.129	-0.188

Note. First five assessment item constructs of the final item pool study.

Table 13

Part 2 Final Item Pool Q3 Correlation Matrix

	Technique	Bowing	Dynamics	Phrasing	Posture
6: Technique	__				
7: Bowing	0.121	__			
8: Dynamics	-0.335	0.064	__		
9: Phrasing	-0.233	-0.124	0.224	__	
10: Posture	-0.088	-0.066	-0.075	0.010	__

Note. Last five assessment item constructs of the final item pool study.

The correlation matrix again identified two strong positive correlations. The strongest positive correlation of 0.796 was between tempo and rhythm assessment items. This strong positive correlation can be explained because music is simply noise organized (rhythm) by time (tempo). Rhythmic accuracy is dependent on the participant’s ability to internalize the proper tempo of the selected exercise and execute the correct subdivision and/or augmentation of beats. The other strong positive correlation of 0.550 was between the pitch and intonation assessment items. This strong positive correlation can be explained because a note (pitch) must also be played with correct tunning (intonation). Melodic accuracy is dependent on the participant’s ability to play the correct pitch as well as to ensure the pitch is adjusted (tuned) for its harmonic intended purpose. It is important to note that both strong positive correlations were also identified in the pilot study, but their positive correlations are slightly stronger in the final item pool study. Ranging between -0.382 and 0.229, correlations between all other assessment items were not significantly positive or negative.

I again observed sequential movement from the left threshold (1) to the right threshold (5) in the final item pool's partial credit model. Table 14 presents the final item pool data's Partial Credit Model.

Table 14

Final Item Pool Partial Credit Model

		Threshold				
	Measure	1	2	3	4	5
1: Tempo	-11.32	-11.17	-3.59	-2.2876	-1.354	0.0752
2: Rhythm	-10.33	-10.66	-3.08	-1.7729	-0.840	0.5899
3: Tone	-10.74	-11.66	-4.08	-2.7771	-1.844	-0.4145
4: Pitch	-11.28	-10.73	-3.15	-1.8476	-0.914	0.5150
5: Intonation	-9.95	-9.63	-2.05	-0.7455	0.188	1.6173
6: Technique	-11.08	-10.59	-3.01	-1.7030	-0.770	0.6596
7: Bowing	-12.18	-11.85	-4.26	-2.9596	-2.026	-0.5968
8: Dynamics	-11.26	-11.20	-3.61	-2.3100	-1.377	0.0526
9: Phrasing	-9.03	-8.92	-1.34	-0.0378	0.895	2.3248
10: Posture	-12.12	-11.79	-4.21	-2.9029	-1.969	-0.5401

Note. The Thurstonian threshold for a score category is defined as the ability at which the probability of achieving that score or higher reaches 0.50.

No observations were made of partial credit being awarded to any constructs. This finding again confirmed that all assessment item descriptors were functioning properly to assess a participant's item proficiency between unsatisfactory (1) and exemplary (5).

Again, there were no random responses by low performers (outliers) and items are performing properly for the people for whom the items are targeted (response patterns).

Table 15 presents the final item pool's item statistics (Rating Scale Model).

Table 15

Final Item Pool Statistics (Rating Scale Model)

	Measure	S.E. Measure	Infit	Outfit
1: Tempo	-3.67	0.0799	0.986	0.983
2: Rhythm	-3.15	0.0763	1.244	1.240
3: Tone	-4.16	0.0851	1.318	1.312
4: Pitch	-3.23	0.0767	0.582	0.581
5: Intonation	-2.12	0.0744	0.845	0.851
6: Technique	-3.08	0.0760	0.724	0.732
7: Bowing	-4.34	0.0876	1.185	1.194
8: Dynamics	-3.69	0.0801	1.280	1.293
9: Phrasing	-1.42	0.0763	1.245	1.257
10: Posture	-4.28	0.0868	1.059	1.062

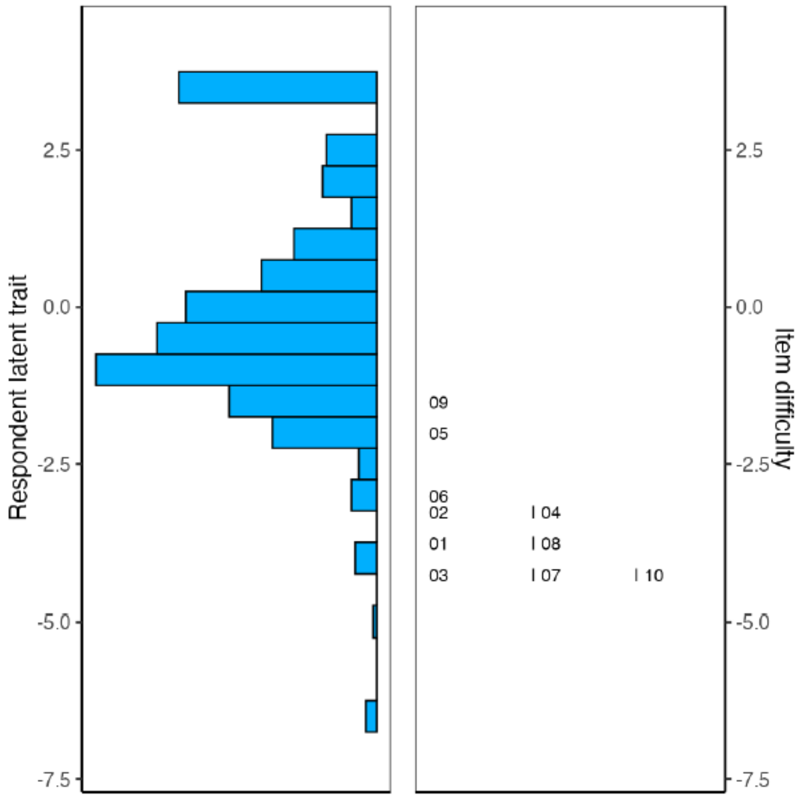
Note. Infit= Information-weighted mean square statistic; Outfit= Outlier-sensitive means square statistics. Person Reliability of 0.872.

Being productive for measurement, all assessment item outfit mean squares and infit mean squares are within the acceptable parameter of 0.5 to 1.5 (Linacre, 2002). The acceptable outfit mean squares signify there were no random responses by low performers (outliers). The acceptable infit mean squares signify the items are performing properly for the people for whom the items are targeted (response patterns). Infit mean squares have the greatest impact on measurement validity, and I believe it is important to note assessment items that reside near the muted (lower) and noisy (upper) end of the acceptable range. The pitch assessment item infit mean square of 0.582 is slightly less stable on the statistical muted (lower) end of the acceptable range. Infit mean square measurements below 0.5 are less productive and may produce misleadingly good reliabilities and separations. The tone assessment item infit mean square of 1.318 is slightly less stable on the statistical noisy (upper) end of the acceptable range. Infit mean square measurements between 1.5 and 2.0 are unproductive for construction of measurement but not degrading (Linacre, 2002). Infit measurements greater than 2.0 distort or degrade the measurement system. It is important to note the dynamics assessment item in the pilot study was the construct closest to the noisy (upper) end of the acceptable range (1.486) but reduced to 1.280 in the final item pool study.

The Wright Map data express assessment item difficulty in an order that flows with a standard string orchestra course of study. Figure 2 presents the final item pool data's Wright Map.

Figure 2

Final Item Pool Wright Map

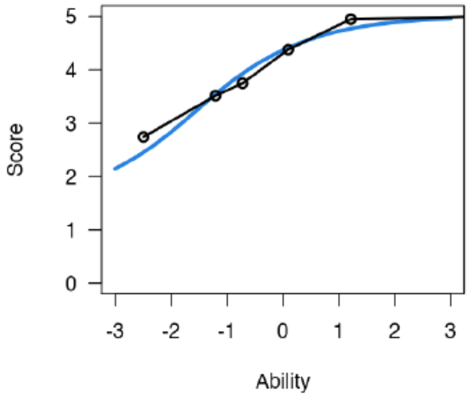


Analyzing the final item pool Wright Map, I was able to determine that phrasing, the ninth assessment item, was the most difficult followed by intonation, the fifth assessment item. It is important to note there is a significant item difficulty separation between those assessment constructs and the rest of the pool. Three assessment constructs ranked the lowest in item difficulty: tone, bowing, and posture (third, seventh, and 10th assessment constructs). It is important to note that posture and tone were the top two prioritized assessment items by the expert panel.

Since this was the final item pool study, I analyzed the expected score curves for each assessment item. Figure 3 presents the expected scores curve for the tempo assessment item.

Figure 3

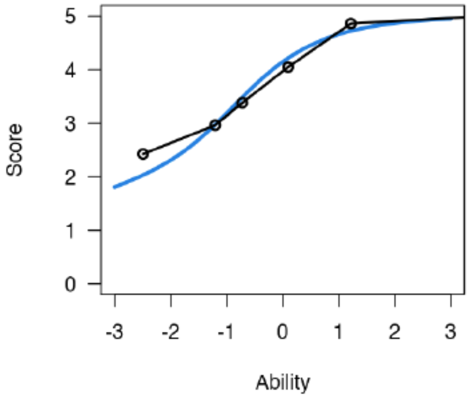
Tempo Expected Scores Curve



Analyzing Figure 3, I found the group with the ability level between -2 and -3 had a higher residual separation score than the Polytomous Rash Model expectation. This was the most notable finding from the tempo analysis. Other ability groups either met Polytomous Rasch Model expectations or showed very little positive or negative residual separation from the expectation. Moving on to the next assessment item, Figure 4 presents the expected scores curve for the rhythm assessment item.

Figure 4

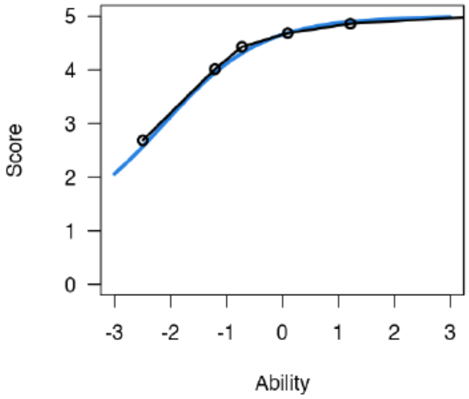
Rhythm Expected Scores Curve



Analyzing Figure 4, I found the group with the ability level between -2 and -3 had a higher residual separation score than the Polytomous Rash Model expectation. It is important to note this finding was also noticed in the tempo assessment item analysis. This was also the most notable finding from the rhythm analysis. Other ability groups either met Polytomous Rasch Model expectations or showed very little positive or negative residual separation from the expectation. Moving on to the next assessment item, Figure 5 presents the expected scores curve for the tone assessment item.

Figure 5

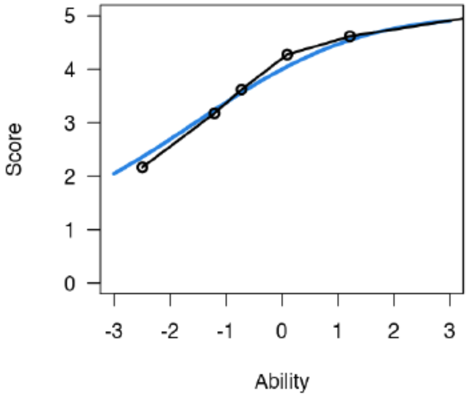
Tone Expected Scores Curve



Analyzing Figure 5, I found that all groups either met Polytomous Rasch Model expectations or showed very little positive residual separation from the expectation. This shows that all ability groups during the final item pool study scored statistically as expected based on the item difficulty. Moving on to the next assessment item, Figure 6 presents the expected scores curve for the pitch assessment item.

Figure 6

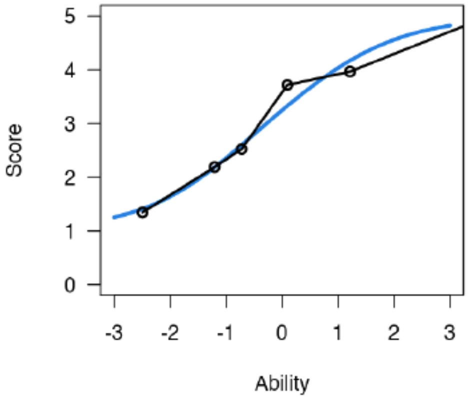
Pitch Expected Scores Curve



Analyzing Figure 6, I found two ability groups that showed very slightly positive and negative residual separations from the expectation. Changing from the previous trends, the ability group between -2 and -3 now presented a slightly negative residual separation. Also worth noting, the ability group slightly above 0 presented a slightly positive residual separation from the expectation. These were the most notable findings from the pitch analysis. Other ability groups either met Polytomous Rasch Model expectations or showed very little positive or negative residual separation from the expectation. Moving on to the next assessment item, Figure 7 presents the expected scores curve for the intonation assessment item.

Figure 7

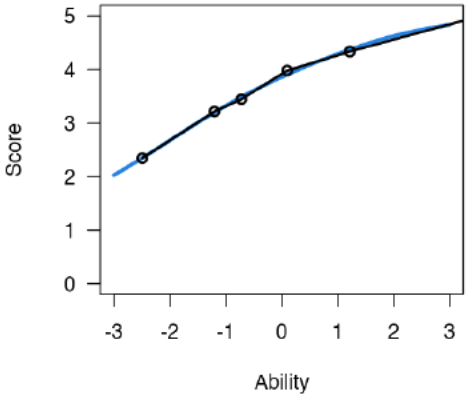
Intonation Expected Scores Curve



Analyzing Figure 7, I found the group with the ability level slightly above 0 had a higher residual separation score than the Polytomous Rash Model expectation. The ability group above 1 had a very slight negative residual separation from the expectation. Other ability groups either met Polytomous Rasch Model expectations or showed very little positive or negative residual separation from the expectation. Moving on to the next assessment item, Figure 8 presents the expected scores curve for the technique assessment item.

Figure 8

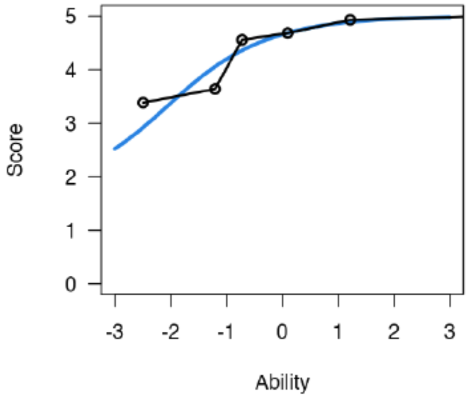
Technique Expected Scores Curve



Analyzing Figure 8, I found that all groups either met Polytomous Rasch Model expectations or showed very little residual separation from the expectation. This shows that all ability groups during the final item pool study scored statistically as expected based on the item difficulty. It is important to note these data mirror the tone assessment item's data expectation. Moving on to the next assessment item, Figure 9 presents the expected scores curve for the bowing assessment item.

Figure 9

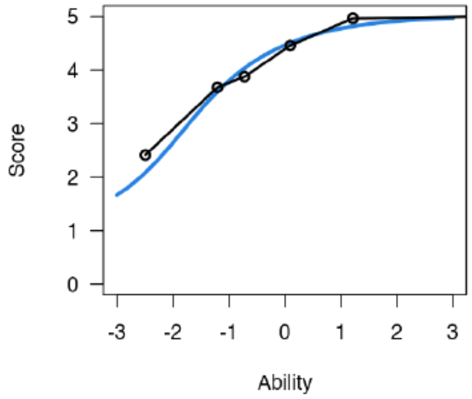
Bowing Expected Scores Curve



Analyzing Figure 9, I found two ability groups that showed positive and negative residual separations from the expectation. Again, the ability group between -2 and -3 presented a positive residual separation from the expectation. For the first time, the ability group slightly below -1 showed a residual separation from the expectation, which contrasted from the -2 to -3 ability group by being negative. These were the most notable findings from the bowing analysis. Other ability groups either met Polytomous Rasch Model expectations or showed very little positive or negative residual separation from the expectation. It is important to note the bowing assessment item had the most significant and drastic residual separations from the expectation. Moving on to the next assessment item, Figure 10 presents the expected scores curve for the dynamics assessment item.

Figure 10

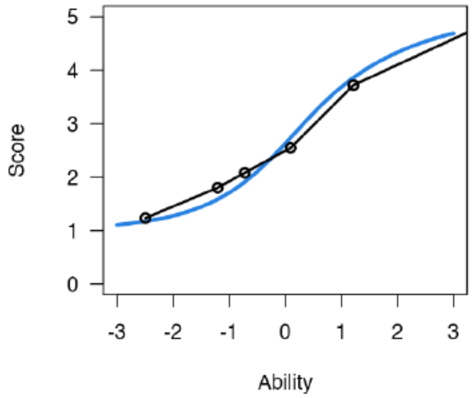
Dynamics Expected Scores Curve



Analyzing Figure 10, I found the group with the ability level between -2 and -3 had a higher residual separation score than the Polytomous Rash Model expectation. It is important to note this finding mirrors the data from the tempo and rhythm assessment item. This was also the most notable finding from the dynamics analysis. Other ability groups either met Polytomous Rasch Model expectations or showed very little positive or negative residual separation from the expectation. Moving on to the next assessment item, Figure 11 presents the expected scores curve for the phrasing assessment item.

Figure 11

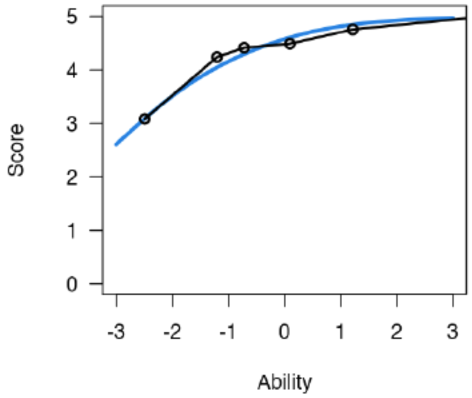
Phrasing Expected Scores Curve



Analyzing Figure 11, I found the group with the ability level below -1 had a slightly positive residual separation score than the Polytomous Rash Model expectation. This was the most notable finding from the phrasing analysis. Other ability groups either met Polytomous Rasch Model expectations or showed very little positive or negative residual separation from the expectation. Moving on to the next assessment item, Figure 12 presents the expected scores curve for the posture assessment item.

Figure 12

Posture Expected Scores Curve



Analyzing Figure 12, I found the group with the ability level below -1 had a more slightly positive residual separation score than the Polytomous Rash Model expectation. It is important to note these data mirror the residual expectation difference from the phrasing assessment item. This was the most notable finding from the posture analysis. Other ability groups either met Polytomous Rasch Model expectations or showed very little positive or negative residual separation from the expectation. Reflecting on the assessment of all individual assessment items, I found no residual separations significant enough to impact the purpose of the DISAT. All assessment items functioned properly based on Polytomous Rasch Model expectations.

All score curves showed little to no significant residual separation from the Polytomous Rasch Model expectation. It is important to note a few observations from analyzing the expected score curves. Six of the 10 assessment items had the lowest ability level score higher than the expectation. The greatest positive separation appeared in the next to highest ability group in the intonation assessment item. This ability group almost scored the same as the highest ability group. This observation was also made possible by a negative separation by the highest ability group. This finding shows that similar intonation proficiency skills are present in the upper two ability groups. The greatest negative separation appeared in the next to lowest ability group in the bowing assessment item. This ability group almost scored the same as the lowest ability group. This observation was also made possible by a positive separation by the lowest ability group. This finding shows that similar bowing proficiency skills are present in the lower two ability groups. Even though a few slight outliers were found, the separations were not significant enough to impact the DISAT’s assessment item validity. The final item pool data showed me the DISAT is a reliable and valid assessment tool.

Chapter 5: Discussion

Overview

The purpose of this study was to develop, validate, and test a weighted individual assessment tool to provide objectivity in the assessment process for middle school orchestral strings. An assessment tool objectively measured skill levels with specific observable items (Bergee, 1994, 2006, 2015). Johnson and Fautley (2017) stated that assessment tools would provide both the teacher and student with valid documented achievement data. Analyzing the assessment data, teachers can identify trends and adjust classroom instruction to meet the students' needs (Simones, 2017). Hopkins et al. (2017) stated this reflective practice can identify areas needed for teacher professional development and growth. Utilizing the data, teachers can provide consistent achievement reports with all stakeholders and advocate for resources and materials to strengthen areas of instructional weakness (Carey, 2017; Loughran & O'Neill, 2016).

Research Questions

The study established the validity and reliability of a weighted individual performance-based assessment tool within the utility scope of middle school orchestral strings. The following research questions guided this study:

1. What specific string-playing behaviors and corresponding criteria validate a weighted individual performance-based assessment tool for middle school orchestral strings?
2. What are the psychometric properties of the weighted individual performance-based assessment tool in authentic situations?

For Research Question 1, the expert panel and I were able to 100% mutually agree on 10 string-playing behaviors: tempo, rhythm, tone, pitch, intonation, technique, bowing, dynamics, phrasing, and posture that created the DISAT. Being interdependent, these string-playing behaviors are relevant because they encompass every necessary facet of orchestral string performance (Zdzinski & Barnes, 2002). According to Zdzinski and Barnes (2002), an orchestral string performance assessment must evaluate each facet of a participant's playing ability to rate the overall musicianship. Bergee and Rossin (2019) stated in their research that it is important to have various aspects of a performance utilized in a musical assessment.

DISAT Performance Descriptors

Each string-playing behavior had corresponding performance ability level descriptors from 1 (unsatisfactory) to 5 (satisfactory) to provide the narrative of observations during the assessment (Myers, 2021).

Tempo. The descriptor rates the ability of the participant to keep a steady pulse. For example, participants scoring 5 (exemplary) display a strong sense of the music pulse; however, participants scoring 1 (unsatisfactory) display no sense of musical pulse.

Rhythm. The descriptor scale rates the ability of the participant to perform notated rhythms. For example, participants scoring 5 (exemplary) perform all rhythms accurately; however, participants scoring 1 (unsatisfactory) perform all rhythms incorrectly.

Tone. The descriptor scale rates the ability of the participant to create a characteristic sound on their instrument. For example, participants scoring 5 (exemplary) display a steady contact point with their bow and good arm weight; however, participants scoring 1 (unsatisfactory) display needing more bow usage and weight.

Pitch. The descriptor scale rates the ability of the participant to perform the correct notes on their instrument. For example, participants scoring 5 (exemplary) perform without any note errors; however, participants scoring 1 (unsatisfactory) perform with errors on most notes.

Intonation. The descriptor scale rates the ability to perform pitches that are harmonically appropriate for the chord structure. For example, participants scoring 5 (exemplary) perform all pitches accurately within their given chord structure; however, participants scoring 1 (unsatisfactory) perform pitches with no intention to their harmonic relevance.

Technique. The descriptor scale rates the positions of their left hand and bow arm (right). For example, participants scoring 5 (exemplary) display excellent positions with great contact; however, participants scoring 1 (unsatisfactory) display no attention to position or contact point.

Bowing. The descriptor scale rates the participant's bow usage and direction. For example, participants scoring 5 (exemplary) display all bowings correctly, and management is consistent; however, participants scoring 1 (unsatisfactory) display mostly incorrect bowings with poor bow management.

Dynamics. The descriptor scale rates the ability of the participant to perform marked volume levels. For example, participants scoring 5 (exemplary) perform all volume level markings with a clear, distinguishable difference; however, participants scoring 1 (unsatisfactory) do not perform any volume level markings.

Phrasing. The descriptor scale rates the ability of the participant to perform increases and decreases of sound. For example, participants scoring 5 (exemplary) perform artistically appropriate increases and decreases in sound; however, participants scoring 1 (unsatisfactory) do not perform any increases or decreases in sound.

Posture. The descriptor scale rates the sitting or standing position of the participant. For example, participants scoring 5 (exemplary) display no tension with a relaxed approach to the instrument; however, participants scoring 1 (unsatisfactory) perform with great tension.

For Research Question 2, the analysis provided information on how assessment item constructs were functioning in authentic situations. These authentic situations contributed to the psychometric property findings of the study (Linacre, 2015). Posture, the highest prioritized item by the expert panel, was found to be the second easiest item construct during the performance assessment. This psychometric property stayed consistent between both the pilot and final item pool study. Phrasing, the lowest prioritized item by the expert panel, was found to be the most difficult item on the performance assessment. This psychometric property stayed consistent in both the pilot and final item pool study. Assessment item difficulties in the study align with the standard course of study for a middle school orchestral string musician. Consistent, disciplined training in their music program over time will allow students to score higher on the DISAT.

In the third step, the DISAT data showed that all assessment item constructs were functioning properly in authentic situations (Linacre 2015; Wesolowski et al., 2016b). The DISAT obtained reliability of 0.878 by having enough variance between raters in the authentic situation. Linacre (2015) stated that reliability greater than 0.8 is acceptable to distinguish separation between raters. Combined with the expert panel's 100% mutual agreement on content validity, this proved the DISAT to be a valid and reliable assessment tool for individual performance-based orchestral strings assessment (AERA, APA, & NCME, 2014).

Limitations

There were several limitations to the study. The expert panel and pilot assessor team were selected through convenience sampling (Zdzinski & Barnes, 2002). Based on this limitation, Zdzinski and Barnes (2002) suggested the study needed detailed and acceptable requirements for participant selection. These requirements ensured that the expert panel and pilot assessor team had the knowledge, experience, and skills to effectively participate in the study. A future study should allow more time for collecting interested participants and select a random sampling from participants who meet qualifications.

Orchestral string assessment, no matter the instrument, can still present variance between assessors, due to the subjective opinion of proficiency. According to Zdzinski and Barnes (2002), this subjective opinion comes from teacher bias, which is based on their background, education, and experience. With music performance assessment, teachers sometimes utilize previous data perceptions (Wesolowski & Wind, 2019; Zdzinski & Barnes, 2002). For example, a student could earn forgiveness for a performance assessment error based on the teacher's perception of that student's musicianship. A future study should include an evaluation of the pilot assessor's assessment skills.

Since this study developed and tested a middle school orchestral strings assessment, pilot assessors were all middle school orchestra teachers who assessed their own sixth- through eighth-grade orchestral strings students. A future study should develop a comprehensive tool to accommodate any grade-level musician. This would allow for greater diversity among pilot assessors and students.

Delimitations

There were several delimitations in this study. Being important to note, I was in control of the participant selection process of the expert panel and pilot assessor team through convenient sampling. I was also heavily involved in the construction of the weighted individual assessment tool. Considering my involvement in the construction, the expert panel was in place to ensure instrument validity and functionality (Smith & Barnes, 2007). Having some previous professional connections with participants who volunteered for this study, I communicated professionally throughout the study to ensure the study's progress and success (Latimer et al., 2010).

Recommendations

The DISAT can be utilized by districts and middle school orchestral string music teachers in North Carolina. Being a consistent, objective tool, the DISAT can standardize our approach to middle school orchestral string music education assessment (AERA, APA, & NCME, 2014). The data collected by the DISAT could easily track the musical progression of students while giving opportunities for constructive, purposeful feedback. Being proven as a valid and reliable assessment tool, the DISAT can also give credibility to music programs and provide data needed to advocate for additional resources (Millican & Forrester, 2019; Silvey & Springer, 2020; Wesolowski & Wind, 2019).

The DISAT has several benefits for districts, teachers, and students.

1. Being easy to utilize, this digital approach enables the assessor to complete a

simple Google Form that collects data into a Google Sheet.

2. Teachers do not have to complete training, follow a guide, or navigate complex software because it is a simple rating scale from 1 (unsatisfactory) to 5 (satisfactory). Being direct, clear, and concise, this rating scale has corresponding descriptors to provide a narrative of observations they should witness to select each proficiency level.
3. After the teacher submits their Google Form assessment, the DISAT will automatically score student performances and disaggregate the data into a customizable database utilizing a query formula.
4. Teachers will be able to easily track student progress and achievement.
5. Since the rating scale has descriptors to describe the observable string-playing behaviors, students will be able to reflect on their performance and identify required behaviors to acquire higher proficiency.
6. Teachers will be able to reflect on whole class assessment data to pinpoint areas of needed reinforcement.
7. School districts will be able to unify their approach to evaluating teacher and student performance.
8. School districts will be able to utilize the data to direct needed funds, resources, and tools to teachers and students.

Future Studies

The DISAT should be utilized in a study to collect student maturation assessment data. Even valid and reliable, the tool should be evaluated with the student achievement lens throughout several assessments with appropriate teacher-driven feedback and instruction (Bergee, 2004; Bergee & Rossin, 2019). According to Van Groen and Eggen (2019), direct feedback engages and motivates learners to achieve higher rates of performance success. This study would enable stakeholders to view whether the assessment constructs in the DISAT are impactful on student progress as well as creating opportunities for targeted achievement growth. After validating student maturation assessment data, I would like to target full implementation by one district to gain more authentic data before complete public consumption.

This research concept and design can easily be repeated for studies focusing on middle school band and chorus. Keeping the same study structure, the expert panel would need to adjust certain assessment items. These adjustments are needed because middle school band and chorus do not utilize a bow in their instrumental performance. Middle school band and chorus also have observations pertaining to posture and technique. This study also creates a driving force to take a closer look at assessment tools for high schools and higher education institutions. According to Denis (2018), many of these assessment items are relevant at higher educational levels. The assessment item proficiency descriptors would need to be rewritten to focus on the advanced accomplishment of these string-playing behaviors (Smith & Barnes, 2007). Even though assessment items mirror the middle school level tool, the musical excerpts, being assessed at higher levels, will be more difficult and require specific observable traits in their descriptors.

Focusing specifically on this study’s parameters, research can be conducted to focus on the orchestra teachers’ perspectives in the assessment process. According to Wesolowski and Wind (2019), data could be collected to not only analyze a participant pool but also focus on an orchestra teacher’s focus, reaction, and reflection. Gaining curiosity through my initial data analysis, I noticed trends from certain pilot assessors who scored a specific item construct higher or lower consistently throughout their assessments. Whether it was harsh or lenient, these data could document an orchestra teacher’s assessment of a particular string-playing behavior compared to their colleagues. Further qualitative data could represent their reasoning and justifications.

Research can be conducted utilizing the tool with a targeted sample size. The participants can be tracked by their individual growth through pre- and post-assessments over the course of an instruction semester. If you collect and align data from different assessors of that targeted sample size, you can research different pedagogical concepts, processes, and tools to assess best practices for instructional middle string-playing advancement. This study could lead to collecting best practices for other middle school level music content areas as well as leading up to high school and higher education music instruction.

Conclusion

This study developed, validated, and tested a weighted individual assessment tool called the DISAT. The DISAT provides objectivity in the assessment process and measures skill levels with specific observable items for middle school orchestral strings. Since the DISAT utilizes a Google Form and Sheet, data are automatically aggregated, disaggregated, and organized based on teacher preference Middle school orchestra teachers can analyze these data and identify trends to adjust classroom instruction to meet the students’ needs. This reflective practice can also identify areas needed for teacher professional development and growth (Hallam, 2019; Van Groen & Eggen, 2019).

Teachers can use the data to provide consistent achievement reports to all stakeholders and advocate for resources and materials to strengthen areas of instructional weakness.

The DISAT is a reliable and valid assessment tool for middle school orchestra teachers. Using the DISAT, middle school orchestra teachers will have evidence of student growth and achievement to meet any upcoming local, state, or federally mandated data requirements (AERA, APA, & NCME, 2014). Not investing extra time creating tools, middle school orchestra teachers will gain the assurance of a reliable and valid 360-degree assessment of their students' string-playing behaviors (Bergee, 2004; Bergee & Rossin, 2019). According to Silvey and Springer (2020), instrumental music programs can gain teacher objectivity, produce reliable data, and result in program sustainability and growth. I look forward to promoting this tool's utilization and continuing my work in developing, validating, and testing other assessments. The DISAT is only the start of my movement to engage, motivate, and inspire assessment growth and advancement in the field of music education.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. AERA.

Askenfelt, A., & Jansson, E. V. (1992). On vibration sensation and finger touch in stringed instrument playing. *Music Perception*, *9*(3), 311-349. https://doi.org/10.2307/40285555

Azzara, C. D. (2016). Developing musicianship in instrumental music. *Engaging Musical Practices: A Sourcebook for Instrumental Music*, 181.

Bergee, M. J. (1994). *Developing a formal construct: Structuring concert band performance via scale assessment* [Poster presentation]. Proceedings of International Society for Music Education 21st World Conference, Tampa, FL, United States.

Bergee, M. J. (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.

Bergee, M. J. (2004). *Principal-component instability in concert band performance structure* [Poster presentation]. Proceedings of MENC: The National Association for Music Education’s 59th National Biennial In-Service Conference. Minneapolis, MN, United States.

Bergee, M. J. (2006). Validation of a model of extramusical influences on solo and small-ensemble festival ratings. *Journal of Research in Music Education, 54*(3), 244–256. https://doi.org/10.1177/002242940605400307

Bergee, M. J. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education, 55*(4), 344–358. https://doi.org/10.1177/0022429408317515

Bergee, M. J. (2015). A theoretical structure of high school concert band performance. *Journal of Research in Music Education, 63*(2), 145–161. https://doi.org/10.1177/0022429415585959

Bergee, M. J., & McWhirter, J. L. (2005). Selected influences on solo and small-ensemble festival ratings: Replication and extension. *Journal of Research in Music Education, 53*(2), 177–190. https://doi.org/10.1177/002242940505300207

Bergee, M. J., & Rossin, E. G. (2019). Development and validation of a scale assessing midlevel band performance: A mixed methods study. *Journal of Research in Music Education, 67*(2), 214–232. https://doi.org/10.1177/0022429418825144

Bowie, A. (2018). *Aesthetics and subjectivity: From Kant to Nietzsche.* Manchester University Press.

Brockmann-Bauser, M., Bohlender, J. E., & Mehta, D. D. (2018). Acoustic perturbation measures improve with increasing vocal intensity in individuals with and without voice disorders. *Journal of Voice, 32*(2), 162-168.

Bukofzer, M. (1949). Music in the Baroque era. *The Musical Times, 90*(1276), 191. https://doi.org/10.2307/934235

Burkholder, P. J., Grout, D. J., & Palisca, C. V. (2019). *A history of western music.* W.W. Norton & Company.

Campbell, M., & Campbell, P. (2010). Viols and other historic bowed string instruments. *The Science of String Instruments*, 301-315. https://doi.org/10.1007/978-1-4419-7110-4_17

Carey, S. E. (2017). *Instrumental music assessment: current practices in individual music assessment* [Doctoral dissertation, Northeastern University]. Northeastern Digital Repository Service. http://hdl.handle.net/2047/D20248325

Chevan, D. (1989). The double bass as a solo instrument in early jazz. *The Black Perspective in Music, 17*(1/2), 73. https://doi.org/10.2307/1214744

Cohen, M. (2017). Musical gestures: Conceptualising, communicating and collaborating in performance. *Sydney Conservatorium of Music*. http://hdl.handle.net/2123/16538

Cranmore, J., & Wilhelm, R. (2017). Assessment and feedback practices of secondary music teachers: A descriptive case study. *Visions of Research in Music Education*, *29*.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

Dayal, H. C. (2017). *Teachers developing and implementing portfolio assessment in mathematics: A case study in Fiji* [Unpublished Ph.D. thesis]. Suva, Fiji: The University of the South Pacific.

Denis, J. M. (2018). Assessment in music: A practitioner introduction to assessing students. *Update: Applications of Research in Music Education*, *36*(3), 20-28.

Diaz, F. M. (2018). Relationships among meditation, perfectionism, mindfulness, and performance anxiety among collegiate music students. *Journal of Research in Music Education, 66*(2), 150-167.

Dunbar, L. (2018). Using technology to assess in the music classroom. *General Music Today*, *32*(1), 38-40.

Giraldo, S., Waddell, G., Nou, I., Ortega, A., Mayor, O., Perez, A., & Ramirez, R. (2019). Automatic assessment of tone quality in violin music performance. *Frontiers in Psychology*, *10*, 334.

Hallam, S. (2019). The influence of assessment on learning and teaching. *The Oxford Handbook of Philosophical and Qualitative Assessment in Music Education*, 167.

Hamlin, J. M. (2018). *Community music for human development: A brain-centered approach to outreach* [University honors thesis, Portland State University]. Portland State University PDXScholar. https://doi.org/10.15760/honors.536.

Holman, P. (2013). *Life after death: The viola da gamba in Britain from Purcell to Dolmetsch.* The Boydell Press.

Hopkins, M. T. (2013). Teachers' practices and beliefs regarding teaching tuning in elementary and middle school group string classes. *Journal of Research in Music Education*, *61*(1), 97-114.

Hopkins, M. T. (2014). Collaborative composing in high school string chamber music ensembles. *Journal of Research in Music Education, 62*(4), 405-424. https://doi.org/10.1177/0022429414555135

Hopkins, M., Provenzano, A. M., & Spencer, M. S. (2017). Benefits, challenges, characteristics and instructional approaches in an El Sistema inspired after-school string program developed as a university–school partnership in the United States. *International Journal of Music Education*, *35*(2), 239-258.

Jack, R., Stockman, T., & McPherson, A. (2017). Rich gesture, reduced control. *Proceedings of the 4th International Conference on Movement Computing–MOCO '17*. https://doi.org/10.1145/3077981.3078039

Johnson, D., & Fautley, M. (2017). Assessment of whole-class instrumental music learning in England and the United States of America: An international comparative study. *Education 3-13*, *45*(6), 701-709.

Juchniewicz, J. (2018). An examination of music student teaching practices across institutions accredited by the National Association of Schools of Music. *Bulletin of the Council for Research in Music Education, Summer 2018*(217), 27-44.

Laird, P. R. (2004). *The baroque cello revival: An oral history*. Scarecrow Press.

Latimer Jr., M. E., Bergee, M. J., & Cohen, M. L. (2010). Reliability and perceived pedagogical utility of a weighted music performance assessment rubric. *Journal of Research in Music Education*, *58*(2), 168-183.

Lee, H., & Müllensiefen, D. (2020). The Timbre perception test (TPT): A new interactive musical assessment tool to measure timbre perception ability. *Attention, Perception, & Psychophysics*, *82*(7), 3658-3675.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2015). *Facets Rasch measurement* (Version 3.71.4). Winsteps.

Lindley, D. (1984). *The court masque.* Manchester University Press.

Loughran, R., & O’Neill, M. (2016). *Generative music evaluation: Why do we limit to “human?”* NCRA, University College.

Ma, J., & Hall, R. (2018). Learning a part together: ensemble learning and infrastructure in a competitive high school marching band. *Instructional Science, 46*(4), 507-532. https://doi.org/10.1007/s11251-018-9455-3

Mazur, Z., & Łaguna, M. (2017). Assessment of instrumental music performance: Definitions, criteria, measurement. *Edukacja*. https://doi.org/10.24131/3724.170508

Meissner, H. (2017). Instrumental teachers’ instructional strategies for facilitating children’s learning of expressive music performance: An exploratory study. *International Journal of Music Education*, *35*(1), 118-135.

Mick, J. P., & Pope, D. A. (2018). An analysis of ratings and interrater reliability of adjudicated orchestra festivals. *Bulletin of the Council for Research in Music Education*, *218*, 49-68.

Millican, J. S., & Forrester, S. H. (2019). Pedagogical content knowledge and core practices for music teaching. *The Oxford Handbook of Preservice Music Teacher Education in the United States*, 377.

Mitchell, A. (2020). Implementing group teaching in music performance. In J. Encarnacao, & D. Blom (Eds.), *Teaching and evaluating music performance at university: Beyond the conservatory model* (pp. 95-110). Routledge.

Montemayor, M., Coppola, W. J., & Mena, C. (2018). *World music pedagogy, Volume IV: Instrumental music education*. Routledge.

Moss, K., Benham, S., & Pellegrino, K. (2019). Assessment practices of American orchestra directors. In T. S. Brophy (Ed.), *The Oxford handbook of assessment policy and practice in music education* (Volume 2, p. 401). Oxford University Press.

Müller, M. (2020). Item fit statistics for Rasch analysis: Can we trust them? *Journal of Statistical Distributions and Applications, 7*(5). https://doi.org/10.1186/s40488-020-00108-7

Music, E. G. (2019). Assessment practices in American elementary general music classrooms. *The Oxford Handbook of Assessment Policy and Practice in Music Education, 2,* 423.

Myers, M. J. (2021). Standards-based assessment for secondary choral ensembles: A framework to document student learning. *Arts Education Policy Review*, 1-12.

Nelson, S. M. (2003). *The violin and viola: History, structure, techniques*. Dover Publications.

Pati, K. A., Gururani, S., & Lerch, A. (2018). Assessment of student music performances using deep neural networks. *Applied Sciences*, *8*(4), 507.

Pitts, S. (2016). *Coughing and clapping*. Taylor & Francis.

Planyavsky, A. (1998). *The baroque double bass violone*. Scarecrow Press.

Radice, M. A. (2012). *Chamber music: An essential history*. University of Michigan Press.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*(7), 1372-1380.

Rawlings, J. (2016). The use of e-portfolios in music teacher education programs: 2003-2013. *Contributions to Music Education*, 53-69.

Roesner, D. (2018). Found and framed. A conversation with composer and designer Mathis Nitschke. *Theatre and Performance Design*, *4*(3), 204-221. https://doi.org/10.1080/23322551.2018.1523512

Rossin, E. G., & Bergee, M. J. (2020). Cross-validation and application of a scale assessing school band performance. *Journal of Research in Music Education, 69*(1), 24–42. https://doi.org/10.1177/0022429420951789

Rowley, J., & Dunbar-Hall, P. (2017). ePortfolios in a music faculty: Student differentiations in expectations, applications, and uses. In J. Rowley (Ed.), *ePortfolios in Australian Universities* (pp. 83-98). Springer, Singapore.

Schelleng, J. C. (1973). The bowed string and the player. *The Journal of the Acoustical Society of America*, *53*(1), 26-41. https://doi.org/10.1121/1.1913322

Schemmann, H., Rensing, N., & Zalpour, C. (2020). *Medical problems of performing artists - medical problems of performing artists.* https://www.sciandmed.com/MPPA/journalviewer.aspx?issue=1220&article=2222

Scruggs, B. B. (2009). *Learning outcomes in two divergent middle school string orchestra classroom environments: A comparison of a learner-centered and a teacher-centered approach*. Georgia State University.

Shih, Y. J. (2018). *Evaluation of music performance: Computerized assessment versus human judges* [Doctoral dissertation, University of Hawai‘i at Mānoa]. University of Hawai‘i at Mānoa Scholarspace. http://hdl.handle.net/10125/62574

Silveira, J. M., Beauregard, J., & Bull, T. (2017). Development of the processfolio: Promoting preservice music teacher reflection through authentic assessment. *Journal of Music Teacher Education*, *27*(1), 11-23.

Silveira, J. M., & Gavin, R. (2016). The effect of audio recording and playback on self-assessment among middle school instrumental music students. *Psychology of Music*, *44*(4), 880-892.

Silvey, B. A., & Springer, D. G. (2020). The role of accompaniment quality in band directors' evaluations of solo instrumental performance. *Journal of Research in Music Education*, *67*(4), 481-493.

Simones, L. L. (2017). Beyond expectations in music performance modules in higher education: Rethinking instrumental and vocal music pedagogy for the twenty-first century. *Music Education Research*, *19*(3), 252-262.

Smith, B. P., & Barnes, G. V. (2007). Development and validation of an orchestra performance rating scale. *Journal of Research in Music Education*, *55*(3), 268-280.

Smith, R. M. (2004). Fit analysis in latent trait models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). JAM Press.

Souza, A. C. D., Alexandre, N. M. C., & Guirardello, E. D. B. (2017). Psychometric properties in instruments evaluation of reliability and validity. *Epidemiologia e Serviços de Saúde, 26*, 649-659.

St. Pierre, N. A., & Wuttke, B. C. (2017). Standards-based grading practices among practicing music educators: Prevalence and rationale. *Update: Applications of Research in Music Education, 35*(2), 30-37.

Stambaugh, L. A., & Demorest, S. M. (2010). Effects of practice schedule on wind instrument performance: A preliminary application of a motor learning principle. *Update: Applications of Research in Music Education, 28*(2), 20-28.

Stowell, R. (2001). *The early violin and viola: A practical guide*. Cambridge University Press.

Tan, L. (2017). Concept teaching in instrumental music education: A literature review. *Update: Applications of Research in Music Education*, *35*(2), 38-45.

Uygun, M. A., & Kilinçer, Ö. (2017). Developing a scale for strategies used during the practice and learning of instrumental music. *Educational Research and Reviews*, *12*(8), 518-530.

Vaiedelich, S., & Fritz, C. (2017). Perception of old musical instruments. *Journal of Cultural Heritage*, *27*, S2-S7. https://doi.org/10.1016/j.culher.2017.02.014

Van Groen, M. M., & Eggen, T. J. (2019). Educational test approaches: The suitability of computer-based test types for assessment and evaluation in formative and summative contexts. *Journal of Applied Testing Technology*, *21*(1), 12-24.

Vaughan, C. J. (2019). Assessment practices of American band directors. In T. S. Brophy (Ed.)*, The Oxford handbook of assessment policy and practice in music education* (Volume 2, p. 351). Oxford University Press.

Waddell, G., & Williamon, A. (2019). Technology use and attitudes in music learning. *Frontiers in ICT*, *6*, 11.

Wan, L. A., & Gregory, S. (2018). Digital tools to support the motivation of music students for instrumental practice. *Journal of Music, Technology & Education*, *11*(1), 37-64.

Wesolowski, B. C., & Wind, S. A. (2019). Validity, reliability, and fairness in music testing. In T. S. Brophy (Ed.), *The Oxford Handbook of Assessment Policy and Practice in Music Education* (Volume 1, p. 437). Oxford University Press.

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

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

Wind, S., & Hua, C. (2021). *Rasch measurement theory analysis in R: Illustrations and practical guidance for researchers and practitioners.* Routledge.

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

Wise, S. (2016). Changes in direction: Alternative pathways in senior school music assessment. In D. Forrest, & L. Godwin (Eds.), *32nd world conference on music education* (p. 319). International Society for Music Education.

Woodfield, I. (1988). *The early history of the viol*. Cambridge University Press.

Woodhouse, J., & Lynch-Aird, N. (2019). Choosing strings for plucked musical instruments. *Acta Acustica United With Acustica*, *105*(3), 516-529. https://doi.org/10.3813/aaa.919333

World Health Organization. (2021, 13 May). *Coronavirus disease (COVID-19)*. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19

Wu, C. W., Gururani, S., Laguna, C., Pati, A., Vidwans, A., & Lerch, A. (2016, July). Towards the objective assessment of music performances. In *Proc. of the International Conference on Music Perception and Cognition (ICMPC;* pp. 99-103).

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, *50*(3), 245-255.

Appendix A

Study Recruitment Email

Dear Music Educators,

My name is Kevin Ward, and I am a doctoral student under the supervision of Dr. Prince Bull in the Doctor of Education in Curriculum and Instruction program at Gardner-Webb University. I am conducting a research study to develop, validate, and test a weighted individual performance-based assessment tool for middle school orchestral strings. This study will create a valid and reliable tool to objectivity collect performance-based assessment data. There are two options to which you can participate in this study.

I am recruiting individuals for an **Expert Panel** that meet at least ONE of these criteria:

- College Professor
- Retired Orchestral Music Teacher
- Active National Board-Certified Teacher

I am recruiting individuals for a **Pilot Assessor Team** that meet ALL these criteria:

- Middle School Orchestra Teacher
- Professional Teaching License
- Full-Time Teacher

Participation in this study is voluntary.

The privacy of all participants will be maintained throughout the study. All identifiable information will be removed from data sources, and I will be the only one with data access. I will be responsible for the secured storage of the data on my personal password protected computer, to which only I have the password.

If you are interested in participating in this study, please respond to this email by **June 1st, 2021** Please include the capacity to which you would like to participate: **Expert Panel** or **Pilot Assessor Team**

I may be contacted at XXXXX or by email at XXXXX to answer any further questions about this study.

Thank you so much for your time!

Musically yours,

Kevin Ward

Appendix B

Expert Panel Informed Consent Form

Gardner-Webb University IRB
Expert Panel Informed Consent Form

Title of Study: *A Weighted Individual Performance-Based Assessment for Middle School Orchestral Strings: Establishing Validity and Reliability*

Researcher: *Kevin Ward: Doctoral Candidate—School of Education, Gardner-Webb University*

Purpose

The purpose of the research study is: *to develop, validate, and test a weighted individual performance-based assessment tool for middle school orchestral strings.*

Procedure

What you will do in the study: In this study, you will collaborate in an expert panel to create a weighted individual assessment tool for middle school orchestral strings. You will create a comprehensive list of assessment tool items. Through expert panel agreement, you will narrow the list down to ten items. You will create five descriptors for each item describing proficiency from 1 (unsatisfactory) to 5 (explementary). Once completed, you will prioritize the list of ten items to develop weights for each category.

Time Required

It is anticipated the study will require about *6 hours* of your time over the course of four weeks.

Voluntary Participation

Participation in this study is voluntary. You have the right to withdraw from the research study at any time without penalty. If you choose to withdraw, you may request that any of your data which has been collected be destroyed unless it is in a de-identified state.

Confidentiality

I will collect data through the expert panel's work session documents and files and the completed assessment tool. All identifying information will be redacted from documents and files. Documents and files will be stored on my personal password protected computer, to which only I have the password. Three years after the study, the data will be permanently deleted from my personal computer.

Data Linked with Identifying Information

The information that you give in the study will be handled confidentially. Your information will be assigned a unique code. The list connecting your name to this unique code will be kept on my password protected computer, to which only I have the password. When the study has been completed and the data sets have been analyzed, this list will be destroyed. Your name will not be used in any report.

Risks

There are no anticipated risks in this study.

Benefits

There are no direct benefits associated with participation in this study. The study may help us to create a valid and reliable assessment tool for practical use in the middle school orchestral strings classroom. This information may inform professional development for orchestral music educators in the future. The Institutional Review Board at Gardner-Webb University has determined that participation in this study poses minimal risk to participants.

Payment

You will receive no payment for participating in the study.

Right to Withdraw from the Study

You have the right to withdraw from the study at any time without penalty. If you choose to withdraw from the study, you may request that any of your data which has been collected be destroyed unless it is in a de-identified state.

How to Withdraw from the Study

To withdraw from the study, you can notify me of your intent through email. To withdraw after completion of the study, you can notify me of your intent through email. You have the right to request any identifiable data to be destroyed immediately.

If you have questions about the study, contact:

Kevin Ward
EdD Candidate
School of Education, Gardner-Webb University
XXXXX
XXXXX

Dr. Prince Bull
Faculty Research Advisor
School of Education, Gardner-Webb University
704.406.4402
pbull@gardner-webb.edu

If the research design of the study necessitates that its full scope is not explained prior to participation, it will be explained to you after completion of the study. If you have concerns about your rights or how you are being treated, or if you have questions, want more information, or have suggestions, please contact the IRB Institutional Administrator listed below.

Dr. Sydney K. Brown
IRB Institutional Administrator
Gardner-Webb University
Telephone: 704-406-3019

Email: skbrown@gardner-webb.edu

Voluntary Consent by Participant

I have read the information in this consent form and fully understand the contents of this document. I have had a chance to ask any questions concerning this study and they have been answered for me. I agree to participate in this study.

____ Date: ____________________

Participant Printed Name

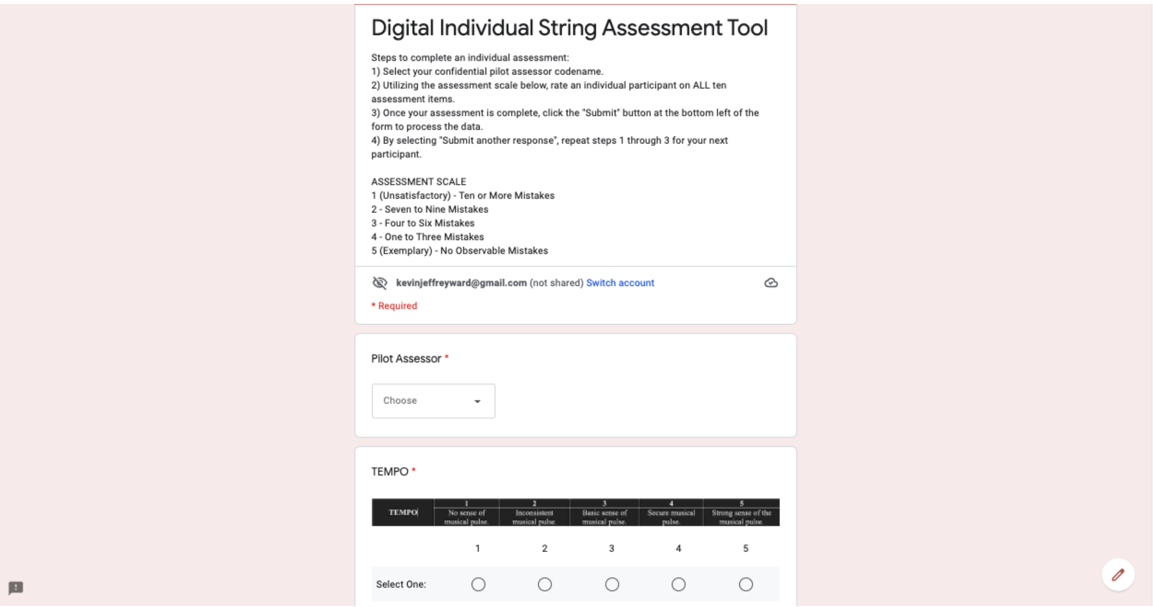
____ Date: ____________________

Participant Signature

You will receive a copy of this form for your records.

Appendix C

Digital Individual String Assessment Tool



Digital Individual String Assessment Tool

Steps to complete an individual assessment:
1) Select your confidential pilot assessor codename.
2) Utilizing the assessment scale below, rate an individual participant on ALL ten assessment items.
3) Once your assessment is complete, click the "Submit" button at the bottom left of the form to process the data.
4) By selecting "Submit another response", repeat steps 1 through 3 for your next participant.

ASSESSMENT SCALE
1 (Unsatisfactory) - Ten or More Mistakes
2 - Seven to Nine Mistakes
3 - Four to Six Mistakes
4 - One to Three Mistakes
5 (Exemplary) - No Observable Mistakes

kevinjeffreyward@gmail.com (not shared) Switch account

* Required

Pilot Assessor *

Choose ▾

TEMPO *

TEMPO	1 No sense of musical pulse.	2 Inconsistent musical pulse.	3 Basic sense of musical pulse.	4 Secure musical pulse.	5 Strong sense of the musical pulse.

	1	2	3	4	5
Select One:	○	○	○	○	○

Appendix D

Pilot Assessor Team Informed Consent Form

Gardner-Webb University IRB
Pilot Assessor Team Informed Consent Form

Title of Study: *A Weighted Individual Performance-Based Assessment for Middle School Orchestral Strings: Establishing Validity and Reliability*

Researcher: *Kevin Ward: Doctoral Candidate—School of Education, Gardner-Webb University*

Purpose
The purpose of the research study is: *to develop, validate, and test a weighted individual performance-based assessment tool for middle school orchestral strings.*

Procedure

What you will do in the study: In this study, you will participate in a training session for a weighted individual assessment tool for middle school orchestral strings. You will utilize the assessment tool for a common assessment pilot study. After the assessment, you will submit all data to me. You will participate in another training session to review calibration changes to the assessment tool. You will utilize the assessment tool for a common assessment final item pool study. After the assessment, you will submit all data to me.

Time Required
It is anticipated the study will require about *5 hours* of your time over the course of six weeks.

Voluntary Participation
Participation in this study is voluntary. You have the right to withdraw from the research study at any time without penalty. You also have the right to refuse to answer any question(s) for any reason without penalty. If you choose to withdraw, you may request that any of your data which has been collected be destroyed unless it is in a de-identified state.

Confidentiality
I will collect data through the pilot assessor team’s assessment tool utilization and pilot assessor team training sessions. All identifying information will be redacted from documents and files. This includes references to other participants and places. Documents and files will be stored on my personal password protected computer, to which only I have the password. Three years after the study, the data will be permanently deleted from my personal computer.

Data Linked with Identifying Information
The information that you give in the study will be handled confidentially. Your information will be assigned a unique code. The list connecting your name to this unique code will be kept on my password protected computer, to which only I have the

password. When the study has been completed and the data sets have been analyzed, this list will be destroyed. Your name will not be used in any report.

Risks
There are no anticipated risks in this study.

Benefits
There are no direct benefits associated with participation in this study. The study may help us to create a valid and reliable assessment tool for practical use in the middle school orchestral strings classroom. This information may inform professional development for orchestral music educators in the future. The Institutional Review Board at Gardner-Webb University has determined that participation in this study poses minimal risk to participants.

Payment
You will receive no payment for participating in the study.

Right to Withdraw from the Study
You have the right to withdraw from the study at any time without penalty. You also have the right to refuse to answer any question(s) for any reason without penalty. If you choose to withdraw from the study, you may request that any of your data which has been collected be destroyed unless it is in a de-identified state.

How to Withdraw from the Study
To withdraw from the study, you can notify me of your intent through email. To withdraw after completion of the study, you can notify me of your intent through email. You have the right to request any identifiable data to be destroyed immediately.

If you have questions about the study, contact:
Kevin Ward
EdD Candidate
School of Education, Gardner-Webb University
XXXXX
XXXXX

Dr. Prince Bull
Faculty Research Advisor
School of Education, Gardner-Webb University
704.406.4402
pbull@gardner-webb.edu

If the research design of the study necessitates that its full scope is not explained prior to participation, it will be explained to you after completion of the study. If you have concerns about your rights or how you are being treated, or if you have questions, want more information, or have suggestions, please contact the IRB Institutional Administrator listed below.

Dr. Sydney K. Brown
IRB Institutional Administrator
Gardner-Webb University
Telephone: 704-406-3019
Email: skbrown@gardner-webb.edu

Voluntary Consent by Participant
I have read the information in this consent form and fully understand the contents of this document. I have had a chance to ask any questions concerning this study and they have been answered for me. I agree to participate in this study.

____ Date: __________________

Participant Printed Name

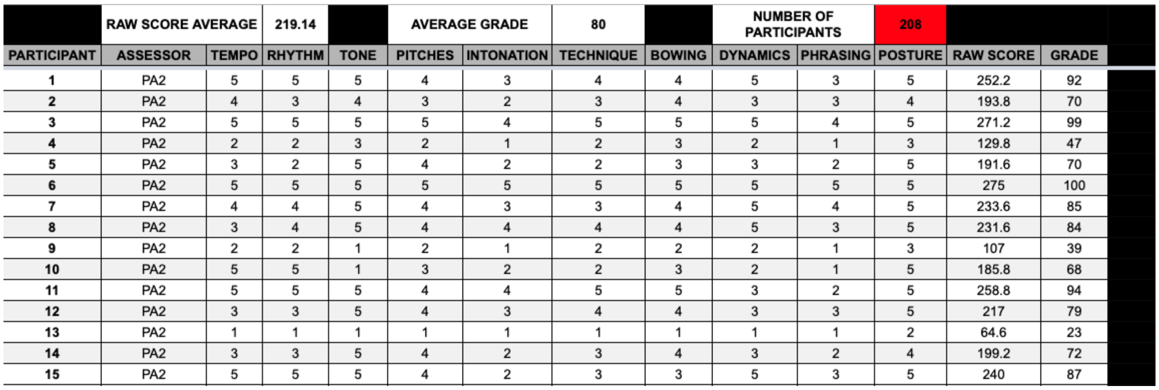
____ Date: __________________

Participant Signature

You will receive a copy of this form for your records.

Appendix E

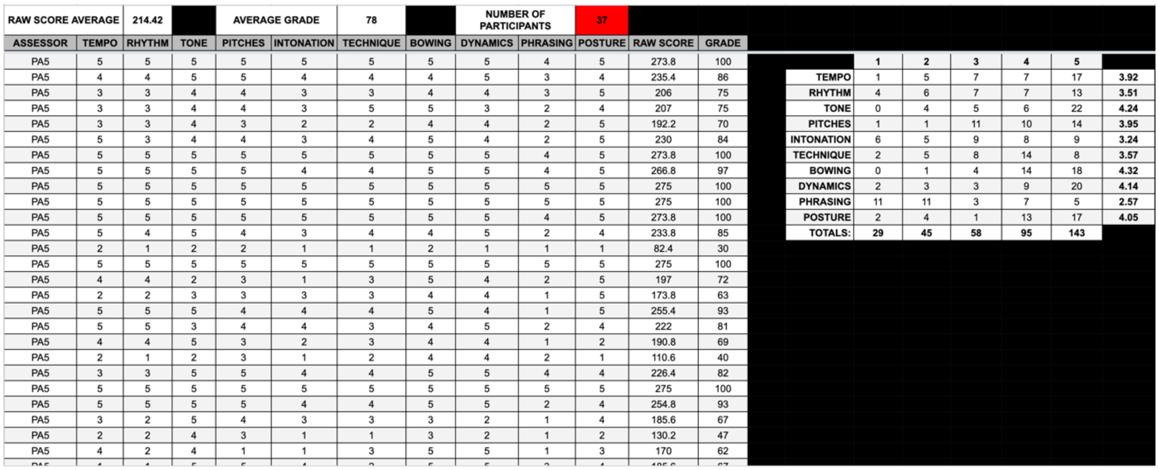
Pilot Assessment Data Collection



	RAW SCORE AVERAGE		219.14		AVERAGE GRADE		80		NUMBER OF PARTICIPANTS		208		
PARTICIPANT	ASSESSOR	TEMPO	RHYTHM	TONE	PITCHES	INTONATION	TECHNIQUE	BOWING	DYNAMICS	PHRASING	POSTURE	RAW SCORE	GRADE
1	PA2	5	5	5	4	3	4	4	5	3	5	252.2	92
2	PA2	4	3	4	3	2	3	4	3	3	4	193.8	70
3	PA2	5	5	5	5	4	5	5	5	4	5	271.2	99
4	PA2	2	2	3	2	1	2	3	2	1	3	129.8	47
5	PA2	3	2	5	4	2	2	3	3	2	5	191.6	70
6	PA2	5	5	5	5	5	5	5	5	5	5	275	100
7	PA2	4	4	5	4	3	3	4	5	4	5	233.6	85
8	PA2	3	4	5	4	4	4	4	5	3	5	231.6	84
9	PA2	2	2	1	2	1	2	2	2	1	3	107	39
10	PA2	5	5	1	3	2	2	3	2	1	5	185.8	68
11	PA2	5	5	5	4	4	5	5	3	2	5	258.8	94
12	PA2	3	3	5	4	3	4	4	3	3	5	217	79
13	PA2	1	1	1	1	1	1	1	1	1	2	64.6	23
14	PA2	3	3	5	4	2	3	4	3	2	4	199.2	72
15	PA2	5	5	5	4	2	3	3	5	3	5	240	87

Appendix F

Individual Pilot Assessor Data

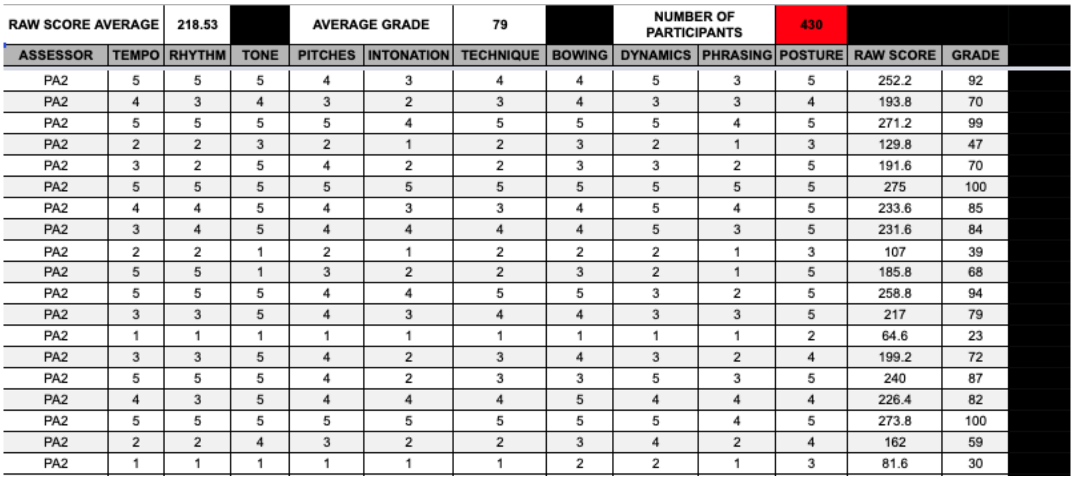


RAW SCORE AVERAGE		214.42		AVERAGE GRADE		78		NUMBER OF PARTICIPANTS		37		
ASSESSOR	TEMPO	RHYTHM	TONE	PITCHES	INTONATION	TECHNIQUE	BOWING	DYNAMICS	PHRASING	POSTURE	RAW SCORE	GRADE
PA5	5	5	5	5	5	5	5	5	4	5	273.8	100
PA5	4	4	5	5	4	4	4	5	3	4	235.4	86
PA5	3	3	4	4	3	3	4	4	3	5	206	75
PA5	3	3	4	4	3	5	5	3	2	4	207	75
PA5	3	3	4	3	2	2	4	4	2	5	192.2	70
PA5	5	3	4	4	3	4	5	4	2	5	230	84
PA5	5	5	5	5	5	5	5	5	4	5	273.8	100
PA5	5	5	5	5	4	4	5	5	4	5	266.8	97
PA5	5	5	5	5	5	5	5	5	5	5	275	100
PA5	5	5	5	5	5	5	5	5	5	5	275	100
PA5	5	5	5	5	5	5	5	5	4	5	273.8	100
PA5	5	4	5	4	3	4	4	5	2	4	233.8	85
PA5	2	1	2	2	1	1	2	1	1	1	82.4	30
PA5	5	5	5	5	5	5	5	5	5	5	275	100
PA5	4	4	2	3	1	3	5	4	2	5	197	72
PA5	2	2	3	3	3	3	4	4	1	5	173.8	63
PA5	5	5	5	4	4	4	5	4	1	5	255.4	93
PA5	5	5	3	4	4	3	4	5	2	4	222	81
PA5	4	4	5	3	2	3	4	4	1	2	190.8	69
PA5	2	1	2	3	1	2	4	4	2	1	110.6	40
PA5	3	3	5	5	4	4	5	5	4	4	226.4	82
PA5	5	5	5	5	5	5	5	5	5	5	275	100
PA5	5	5	5	5	4	4	5	5	2	4	254.8	93
PA5	3	2	5	4	3	3	3	2	1	4	185.6	67
PA5	2	2	4	3	1	1	3	2	1	2	130.2	47
PA5	4	2	4	1	1	3	5	5	1	3	170	62
PA5	4	4	5	5	4	3	5	5	2	4	185.6	67

	1	2	3	4	5	
TEMPO	1	5	7	7	17	3.92
RHYTHM	4	6	7	7	13	3.51
TONE	0	4	5	6	22	4.24
PITCHES	1	1	11	10	14	3.95
INTONATION	6	5	9	8	9	3.24
TECHNIQUE	2	5	8	14	8	3.57
BOWING	0	1	4	14	18	4.32
DYNAMICS	2	3	3	9	20	4.14
PHRASING	11	11	3	7	5	2.57
POSTURE	2	4	1	13	17	4.05
TOTALS:	29	45	58	95	143	

Appendix G

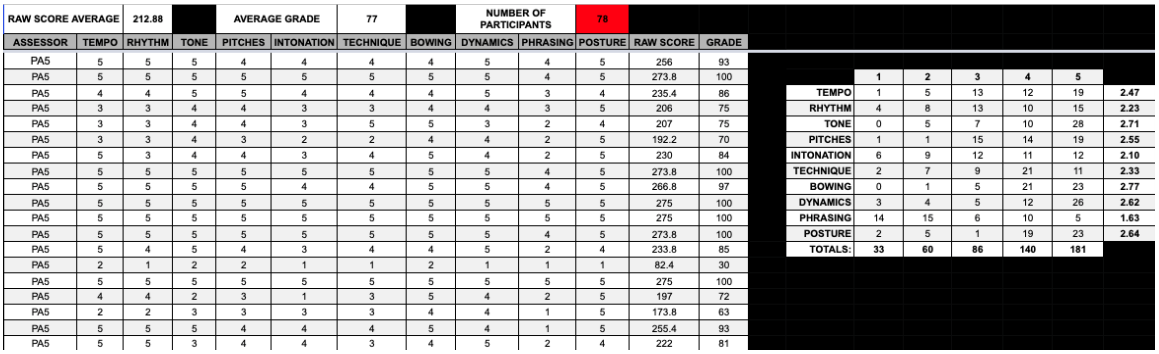
Final Item Pool Assessment Data Collection



RAW SCORE AVERAGE		218.53		AVERAGE GRADE		79		NUMBER OF PARTICIPANTS		430		
ASSESSOR	TEMPO	RHYTHM	TONE	PITCHES	INTONATION	TECHNIQUE	BOWING	DYNAMICS	PHRASING	POSTURE	RAW SCORE	GRADE
PA2	5	5	5	4	3	4	4	5	3	5	252.2	92
PA2	4	3	4	3	2	3	4	3	3	4	193.8	70
PA2	5	5	5	5	4	5	5	5	4	5	271.2	99
PA2	2	2	3	2	1	2	3	2	1	3	129.8	47
PA2	3	2	5	4	2	2	3	3	2	5	191.6	70
PA2	5	5	5	5	5	5	5	5	5	5	275	100
PA2	4	4	5	4	3	3	4	5	4	5	233.6	85
PA2	3	4	5	4	4	4	4	5	3	5	231.6	84
PA2	2	2	1	2	1	2	2	2	1	3	107	39
PA2	5	5	1	3	2	2	3	2	1	5	185.8	68
PA2	5	5	5	4	4	5	5	3	2	5	258.8	94
PA2	3	3	5	4	3	4	4	3	3	5	217	79
PA2	1	1	1	1	1	1	1	1	1	2	64.6	23
PA2	3	3	5	4	2	3	4	3	2	4	199.2	72
PA2	5	5	5	4	2	3	3	5	3	5	240	87
PA2	4	3	5	4	4	4	5	4	4	4	226.4	82
PA2	5	5	5	5	5	5	5	5	4	5	273.8	100
PA2	2	2	4	3	2	2	3	4	2	4	162	59
PA2	1	1	1	1	1	1	2	2	1	3	81.6	30

Appendix H

Individual Final Item Pool Assessor Data



RAW SCORE AVERAGE	212.88		AVERAGE GRADE	77		NUMBER OF PARTICIPANTS	78

ASSESSOR	TEMPO	RHYTHM	TONE	PITCHES	INTONATION	TECHNIQUE	BOWING	DYNAMICS	PHRASING	POSTURE	RAW SCORE	GRADE
PA5	5	5	5	4	4	4	4	5	4	5	256	93
PA5	5	5	5	5	5	5	5	5	4	5	273.8	100
PA5	4	4	5	5	4	4	4	5	3	4	235.4	86
PA5	3	3	4	4	3	3	4	4	3	5	206	75
PA5	3	3	4	4	3	5	5	3	2	4	207	75
PA5	3	3	4	3	2	2	4	4	2	5	192.2	70
PA5	5	3	4	4	3	4	5	4	2	5	230	84
PA5	5	5	5	5	5	5	5	5	4	5	273.8	100
PA5	5	5	5	5	4	4	5	5	4	5	266.8	97
PA5	5	5	5	5	5	5	5	5	5	5	275	100
PA5	5	5	5	5	5	5	5	5	5	5	275	100
PA5	5	5	5	5	5	5	5	5	4	5	273.8	100
PA5	5	4	5	4	3	4	4	5	2	4	233.8	85
PA5	2	1	2	2	1	1	2	1	1	1	82.4	30
PA5	5	5	5	5	5	5	5	5	5	5	275	100
PA5	4	4	2	3	1	3	5	4	2	5	197	72
PA5	2	2	3	3	3	3	4	4	1	5	173.8	63
PA5	5	5	5	4	4	4	5	4	1	5	255.4	93
PA5	5	5	3	4	4	3	4	5	2	4	222	81

	1	2	3	4	5	
TEMPO	1	5	13	12	19	2.47
RHYTHM	4	8	13	10	15	2.23
TONE	0	5	7	10	28	2.71
PITCHES	1	1	15	14	19	2.55
INTONATION	6	9	12	11	12	2.10
TECHNIQUE	2	7	9	21	11	2.33
BOWING	0	1	5	21	23	2.77
DYNAMICS	3	4	5	12	26	2.62
PHRASING	14	15	6	10	5	1.63
POSTURE	2	5	1	19	23	2.64
TOTALS:	33	60	86	140	181	