UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

# DEVELOPMENT, IMPLEMENTATION, AND LONGITUDINAL ANALYSIS OF THE OKLAHOMA ALL-STATE BAND AUDITION EVALUATION MEASURE

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

CHRISTOPHER SCOTT BARBER
Norman, Oklahoma
2025

DEVELOPMENT, IMPLEMENTATION, AND LONGITUDINAL ANALYSIS OF THE OKLAHOMA ALL-STATE BAND AUDITION EVALUATION MEASURE

A DISSERTATION APPROVED FOR THE

SCHOOL OF MUSIC

BY THE COMMITTEE CONSISTING OF

Dr. Casey Gerber, Chair

Dr. Eric Pennello, Co-Chair

Dr. Matt Stock

Prof. Karie Antell

**Acknowledgments**

First and foremost, I would like to thank my committee co-chair, Dr. Eric Pennello. I would not have reached the finish line with this document without his guidance and support. Further thanks go to my committee chair, Dr. Casey Gerber, for truly going the extra mile. Pennello and Gerber are consummate professionals, exceptional educators, and wonderful people. They never gave up on me and helped me navigate numerous challenges to get here, for which I am forever thankful. I would also like to thank my committee members, Dr. Matt Stock and Ms. Karie Antell, for their support and patience. Additionally, special thanks to Dr. Chris Baumgartner for his assistance in ensuring there was always a path through the forest for me during these past few years.

I would also like to acknowledge all of the educators I have encountered throughout my life. Whether you were a co-worker, colleague, or teacher, I applaud you for dedicating yourself to one of the most challenging yet vital and rewarding careers in the world. Each of you has left a lasting impression on me. I would also like to thank the Oklahoma Music Educators Association and its band division for trusting me with this project and for their desire to solve a problem by asking the right questions and taking an academic approach.

A huge, heartfelt thanks goes to my circle of friends who have carried me on their backs these last ten years. There are too many to name here, but you know who you are, and I feel so fortunate to have you in my life. I absolutely could not have reached this goal without you. A special thanks goes to my family, who believed in me when I didn't believe in myself. Completing this project is as important to them as it is to me; I greatly appreciate them. Lastly, I thank my two children, Dora and Huey, who fill my cup daily. I hope the completion of this project shows them that anything is possible, no matter what life throws at you.

## Table of Contents

## List of Tables

## List of Figures

**Abstract**

In 2012, the Oklahoma Music Educators Association (OkMEA) band division embarked on a mission to research and address some of the reliability issues it encountered with its existing audition adjudication measurement tool. In this study, I utilized the McPherson and Thompson (1998) Process Model for Assessing Music Performance as a framework for analyzing the OkMEA audition structure. From this analysis, I developed a general audition measurement rating scale for its all-state band and orchestra wind and percussion auditions. The OkMEA implemented the Barber General Audition Measure (BGAM) at the auditions in 2014. This study examined five years (2015–2019, $N = 2699$, $K = 425$) of total scores for all final round wind and percussion audition results, categorized by instrument groups, audition year, and total results. The reliabilities were measured using Cronbach's alpha, the intraclass coefficients (ICC), and 95% confidence intervals (CI) based on a mean rating ($k = 5$) in a consistency, 2-way mixed-effects model. The analysis results showed strong overall consistency and reliability by individual group, total year, and the five-year block using the BGAM.

# Chapter 1: Introduction

Every year, thousands of high school instrumentalists nationwide audition for possible placement in regional and statewide honor bands and orchestras (Elliott et al., 2000; Owen, 1969; Vasil, 1973). Acceptance to an honor group not only validates a student's hard work and ability but also provides a high-level musical experience that typically cannot be duplicated in their educational setting (Geringer, 2009). Additionally, all-state ensemble recognition is often a crowning achievement in one's secondary music career and can be an essential supplement to college and scholarship applications (Barnes & McCashin, 2002; DeCarbo et al., 1990; Elliott, 1995a; Lien & Humphreys, 2001). At the most basic level, achieving placement in an all-state ensemble requires dedicated work, extra time outside of music classes, and usually intensive private lessons on the part of the student (DeCarbo et al., 1990; Hickman, 2015). Students audition for recognition, the opportunity to perform with colleagues at an advanced level, and to open the doors for a possible career in the music field (Oldham, 1966). For whatever intrinsic or extrinsic reasons students have for auditioning, they should expect a reliable, efficient, and fair audition (Geringer, 2009). Research in music performance and measurement has been thorough and prevalent for 50 years. However, a standardized and globally accepted adjudication system for all-state auditions does not exist, and state organizations must find what works best for their unique situations.

## Development of Honor Ensembles

The early public-school band movement of the 1920s in the United States positioned band education as an essential institution of school districts, athletics events, and community culture (Welker, 1997). The emergence of national competitions, complete with distinguished honors and cash prizes for students and band participants, buoyed this movement. The first

nationwide band contest, sponsored by various instrument manufacturers, took place in Chicago, Illinois, in 1923, followed in 1926 by the first official National Band Contest sponsored by the Music Supervisors National Convention (Oldham, 1966).

As band programs grew and national competitions evolved, the contests became less competitive and more celebratory. During the early half of the twentieth century, highly competitive band contests were embedded in school curricula and culture as an accepted extra-curricular activity (Oldham, 1966; Welker, 1997). Notwithstanding the success of the national contest movement, music teachers, organizations, and stakeholders felt that high-stakes competitions might not be the best motivator for students as only a few winners were awarded. As a result, the once-standard culminating event of crowning individual students and school rankings began to evolve into combining the entirety of the participating students for a thrilling "Mass Band" final concert. The intent was to recognize the achievements of all the student participants in one enormous and electrifying ensemble performance. This practice eventually overshadowed the band-versus-band component of the competitions and transferred more importance to the communal act of students at rival schools performing together.

As public-school band programs grew throughout the United States, the prominence of festivals became less national and more localized (Burdett, 1985). Factors such as disagreements of the rules governing ratings and rankings, travel arrangements, lengthy absences, high costs, school district liability issues, and the scarcity of room and board resources affected participation in massive, far-off national contests (Burdett, 1985; Owen, 1969).

Coincidentally, the national contest movement directly stimulated the growth of statewide and regional chapters of the Music Educators National Conference (MENC) and

spawned music-focused subdivisions of state teacher organizations (Burdett, 1985; Welker, 1997). As the local chapters grew, more music teachers, musicians, and businesses gained the capacity to administrate local and regional festivals. Eventually, honoring top students from a geographic region became the emphasis, and regional mass bands and honor bands overtook the national events (McDow, 1989; Oldham, 1966; Welker, 1997). By 1937, the existing national contests were pared down into ten regional competitions (Owen, 1969).

As the localized mass band festivals supplanted the national contests, these festivals eventually gave way to regional and all-state bands (Burdett, 1985; McDow, 1989; Welker, 1997). From this natural progression, the term "All-State" began to describe the accomplished student musicians, chosen through a selective process, to represent their school in a state-wide "honor band" experience (Neil, 1944; Welker, 1997). The first all-state honor band was realized in Indiana in 1922 to expand the curriculum for students with exceptional skills in music performance (Oldham, 1966). By 1938, all-state bands, or statewide clinic bands, existed in Iowa, Kentucky, Nebraska, New Jersey, Ohio, Oklahoma, Montana, New York, Pennsylvania, Texas, and Utah (Campbell, 1955; McDow, 1989; Oldham, 1966). The first Oklahoma All-State Band was established in 1929. It was organized and conducted by Eddie Gambill, a music merchandiser from Tulsa who had experience organizing and orchestrating the mass band festivals in Tulsa for many years (McDow, 1989). By the beginning of World War II, statewide and sub-state honor band festivals had entirely superseded the national band contests of the first half of the twentieth century (Burdett, 1985; McDow, 1989). Today, all 50 states have versions of selective all-state and all-district music ensembles, as well the existence of the National Honor Band of America, organized and produced by the Music for All/Bands of America organization (Barnes & McCashin, 2002; Music for All, n.d.; Welker, 1997).

### *All-State Ensembles and Audition Practices*

Being selected into an all-state honor group is not only one of the most sought-after achievements for a high school band student (Elliott, 1995a), but it is also recognized as defining the best students within music programs (DeCarbo et al., 1990). With this level of significance, it might seem unexpected that a standard evaluation method for placing student musicians in all-state bands does not exist. Nonetheless, research investigating all-state band selection processes is limited (Lien & Humphreys, 2001). The selection of musicians for the early all-state bands was simply through teacher recommendations (McDow, 1989). In 1966, Oldham found that 82.5% of all-state bands had no formal written policies regarding selecting student personnel; however, 72.5% of the existing all-state bands selected students by in-person audition or with an audition paired with another identifying means such as teacher recommendation. Studying audition procedures in the fifteen Texas band regions, Owen (1969) found that all the areas employed scoring sheets that provided point ranges for adjudicators to rate auditions. However, audition requirements, point ranges for scored areas (musicality, prepared etude, scales, etc.), and the itemized weighting of the sub-areas creating overall scores varied from district to district, displaying the lack of a unified standard performance measurement tool.

Near the turn of the 21st century, all-state audition procedures for bands, orchestras, and choirs (Barnes & McCashin, 2002; Elliott, 1995a; Welker, 1997; Wine, 1996) were examined. By surveying all-state band coordinators in 50 states, Elliott (1995a) found:

- Eighteen states had multiple auditions or "rounds" for placement.
- Most states employed live auditions.
- No states used director recommendations for selection.

- Most states used anonymous auditions instead of open auditions.
- All but nine states chose judges from participating teachers, while nine states claimed to use only outside, unaffiliated judges.
- Most states utilized etudes or selected audition material instead of music excerpts.
- Seven states denoted lawsuits or threatened lawsuits regarding all-state band auditions.

Additionally, Welker (1997) found that adjudication duties in 30 states were shared with college professors, with a few states employing retired band directors, private instructors, and college students. It is not stated if Elliott (1995a) considered professors and retired band directors as teachers, therefore not listing them separately; furthermore, neither Elliott (1995a) nor Welker (1997) clarified if states used a mix of adjudicator types (i.e., professional musicians paired with teachers) for each event. Neither study surveyed the types of adjudication measurements employed in the various state's audition adjudication, with the latter study receiving a few samples of scoring sheets from respondents. Welker (1997) noted similar verbiage on the obtained adjudication sheets. As for the audition setting, most states studied utilized anonymous auditions (Elliot, 1995a, Welker, 1997). However, Welker (1997) found that the few states with open, non-screened auditions commonly employed just one round of auditions. All in all, despite no standard format for selective honor band auditions, there were commonalities nationwide in all-state band audition procedures.

**Music Performance Measurement Models**

The two common approaches to performance measurement are a) criteria-based objective measures based on specifically defined factors or criteria and b) holistic, subjective judgment measures that place a global score on a performance or performance area. In criteria-based music performance, measurement is often achieved by employing segmented checklists,

rating scales, and rubrics (Barry, 2009). Whereas global or holistic scoring consists of raters placing a single value judgment on a performance in the form of a point scale or numeric ranking based on their individual cognitive processes (Stanley, 2002).

### *Criterion-Based Assessment*

Criterion-based music assessments come in the form of rubrics, checklists, and rating scales utilizing measurable criteria (Barry, 2009). In 1954, the Watkins-Farnum Performance Scale (WFPS) became one of the first widely accepted criterion-based music assessment systems (Levinowitz, 1985; Haley, 1999; Wrigley, 2005). Based on objective criteria such as marking correct rhythms and pitches, the WFPS has consistently shown strong reliability and has been implemented for many years to measure student music performance ability levels and improvement (Haley, 1999; Wrigley, 2005). However, the lack of subjective criteria, such as interpretation, phrasing, and tone, has caused doubt about the validity of the WFPS as a vital performance measure (Bergee, 1987; Wrigley, 2005; Zdzinski & Barnes, 2002). Later studies (Cooksey, 1977; Fiske, 1977; Lien & Humphreys, 2001; Saunders & Holahan, 1997; Stegman, 2009) were administered to develop and test dependable performance evaluation tools to assist adjudicators in producing reliable and formative results. Studies as such have formed a body of research guiding music performance assessment intended to provide meaningful and substantive feedback to students to facilitate musical growth (Bergee, 2003).

In summary, criteria-based music assessment is a reliable measurement tool to for supporting musical growth. However, its limiting components spark debate when evaluating subjective artistry and ranking performances. Johnson (1997) states, “Criteria that emphasise technical achievement strongly exacerbate this problem. Analysis of technical achievement

should normally follow and not precede the assessment itself: its primary function is diagnostic" (p. 279).

### *Holistic Assessment*

On the surface, holistic scoring is much simpler in design than detailed rubrics or criteria-focused rating scales. Raters make a judgment on their impression of a performance formed by their own music experience and cognition (Radocy, 1986). Because of holistic scoring's subjectivity and simplicity, it challenges identifying and quantifying two fundamental concepts of evaluation: reliability and validity (Barry, 2009). Without the presence of measurable criteria, empirical studies on global measurement tend to focus on either the effects of nonmusical factors on adjudicator response, scoring/ranking (Glesjer & Heyndels, 2001; Hewitt, 2007; Coimbra et al., 2001; Lien & Humphreys, 2001; Sloboda, 1991; Thompson, 2006; Thompson et al., 2007; Tsay, 2014) or studies comparing global scoring to criterion-based measurements (Eysenck, 1939; Fiske, 1975; Mills, 1991; Smith, 2004) for validity. Although testing the validity of holistic scoring is ambiguous, research shows there is a place for using holistic measures when paired with some guiding criteria for scoring. Bergee (2003; also see Fiske, 1983) notes that in most cases of holistic scoring, adjudicators often have little evidence to support how they arrived at the score. Nonetheless, Mills (1991) argues that mixing and balancing the two methods does show strong reliability. In all, holistic scoring is based on an adjudicator's ability to accurately rate an event based on the whole of the performance; therefore, it is subject to all the musical and nonmusical influences present at the time.

### *Reliability in Criterion-Based and Holistic Evaluation Practices*

Reliability in performance measurement tools has long been a challenge in adjudication (Shih, 2018). Although limited, a stream of dialogue exists that supports holistic judgment by

centering on concerns that music evaluation is much more than just a performance evaluation (Johnson, 1997; Mills, 1991; Radocy, 1986; Stanley et al., 2002). Many commenters argue that employing an objective, criteria-based judgment with high reliability and validity comes with pitfalls concerning the lack of accounting for the artistic aspects and nonmusical factors when quantifying specific musical traits. The subjective nature of various immeasurable musical and nonmusical factors encompassed in producing a musical performance is hard to quantify (Radocy, 1986). Smith (2004) debates that quantifying musical performance evaluation implies that absolute standards for 'musical merit' exist and are shared by all and that all scores are based on this 'musical merit' and are free of bias. Thompson and Williamon (2003) theorized that objectifying music performance reliably assumes that placing one performance over another means that a listener is aware that a measurable difference exists, how to define those differences, and that all listeners will agree.

At the most basic level, evaluating musical performances is assigning value to a fabric of complex and subjective events grounded in artistry (Williams & Chapman, 2020). Radocy (1986) posits that when using a criteria-based quantitative measurement of human performance, the decisions are still ultimately subjective in nature. In further explanation, Johnson (1997) wrote:

> The listing of requisites for each level of grading also causes problems for the examiner. For example, given that criteria include playing accurately and in tune, the conscientious examiner should ask how 'accurate' and 'in tune' the performance needs to be for a determined grade. Since this cannot be answered objectively, the criterion fails, and such cases can only be resolved by judgment and common sense informed by a good deal of experience (p. 271–272).

Although agreed-upon criteria-based assessment models have shown to be highly reliable for capturing diagnostic formation (Abeles, 1973a; Bergee, 1987; Saunders et al., 1997), the subjective versus objective debate is omnipresent when measuring artistic human behaviors.

**Accountability in Music Performance Evaluation**

Skepticism about consistency and reliability in music performance measurement and honor-group adjudication is evident throughout the last half of the 20th century. Through survey data, Wine (1996) found that most state association directors often questioned the reliability of their all-state choir audition results. Dugger (1997) investigated the reliabilities of the Oklahoma All-State Band's practice of using an Olympic-style scoring system within its holistic measurement approach to tabulate results, and Fiske (1975) questioned the reliability of different judge groups applying performance rating scales and holistic rankings. Similarly, Mills (1991) examined the reliability of holistic scoring evaluations versus the segmented criteria-specific rating scales developed in the prior decades.

In 1995a, Elliott found that most all-state band organizations frequently questioned their audition practices and were commonly concerned with making improvements or adaptations to their audition procedures. During the study, Elliot also discovered that seven states had recently been threatened with litigation regarding audition results and suggested that all-state organizations should periodically adopt policies to review their practices in an open forum. Elliot concluded that organizations should work to ensure that their audition process is a fair and objective measure for awarding its participants the all-state honor. Such disagreements are not limited to band auditions. Wine (1996) accounted for 22 states showing dissatisfaction with the fairness of all-state chorus auditions. And later, Barnes and McCashin (2002) found that

similar concerns about a lack of fairness in all-state orchestra auditions nationwide were also routine among instructors.

Ironically, with all of the recorded skepticism and concern, there are few studies focused on developing the best honor-group evaluation practices. Comparable to the current research, Owen (1969) sought to develop a specific all-state audition measurement tool unique to the Texas area all-region band practices and eliminate as many inconsistencies as possible when quantifying subjective judgment. Alford (1985), under the guidance of the Percussive Arts Society, produced a suggested model of best practices for all-state percussion auditions. Based on survey results, Alford's model recommends a standardized context in which percussionists should be auditioned regarding the use of specific instruments, literature, sightreading, anonymity, and so forth. However, a recommended measure for performance evaluation of the auditions is missing from the model. Saunders and Holahan (1997) studied the reliability of employing a criteria-based rating scale for defining ability levels and predicting the 1994 Connecticut All-State Band results.

At the turn of the 21st century, legislative emphasis on the "No Child Left Behind" (NCLB) educational policy ushered in a strong trend of high-stakes accountability in school districts and regional accreditation agencies, thus fostering a desire for accurate assessment instruments designed to measure real-world tasks in an educational setting (Barry, 2009; Raymond, 2009). As a result, a culture of employing reliable data collection instruments to inform data-driven decisions at all levels of public education became commonplace; assessment in music education was not spared (Barry, 2009; Smith & Barnes, 2007). For example, in 2009, Stegman reported on the quest of the Michigan School Vocal Association's revision of its performance adjudication system to provide more meaningful feedback and

consistent and fair adjudication. Newman (2008) examined how Disney partnered with MENC and the National Federation of High Schools after becoming concerned about the inconsistency of adjudication at their festivals and co-created a judge-training seminar featuring real-live practice scenarios to improve festival ratings.

**A Model for Assessing Music Performance**

In 1998, McPherson and Thompson derived a "Process Model for Assessing Music Performance" (PMAMP). The intent was to aid in grouping the salient factors affecting many types of music performance assessment contexts and to organize and guide research on music performance measurement. McPherson and Thompson's model (Figure 1.1) was based on a similar occupational performance rating model derived by Landy and Farr (1980), which considered the rater, the ratee, and outside factors in making value judgments of employees. McPherson and Thompson's adaptation of their model to music assessment established that the performance context of a specific event drives the evaluative approach to be considered in music performance appraisal. The researchers conceived that determining the performance context be guided by the following:

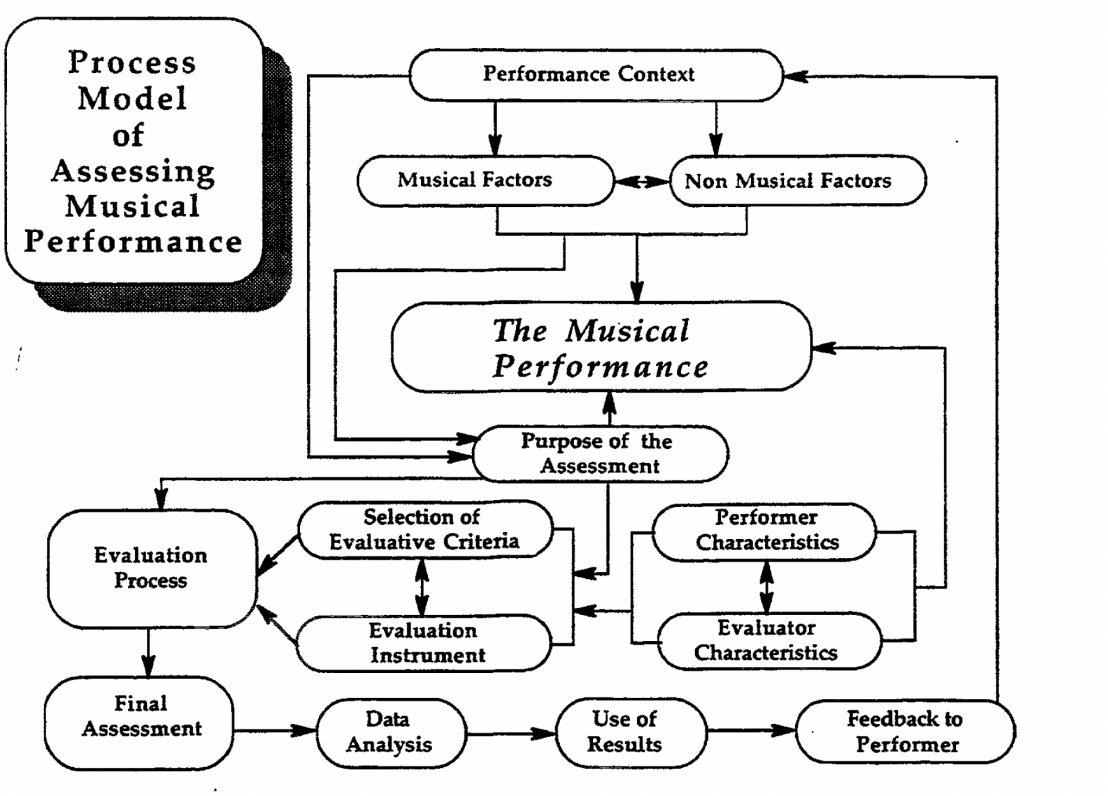
- purpose of assessment: whether the performer is auditioning for an ensemble, playing a solo, large-group festival performance, playing for semester jury, and so forth. McPherson and Thompson (1998) conceived that the purpose of a specific music performance "strongly influences the way a judge will listen to, and therefore, evaluate a musical performance" (p. 12).
- type of performance: McPherson (1995) established that the five types of performance are playing rehearsed literature, sightreading, playing from memory, playing by rote, and improvisation. Each of these areas requires specific skills to make accurate, critical

evaluations. The type of performance can also be susceptible to bias, which can affect assessment (McPherson & Thompson, 1998; Radocy, 1976).

- performance proportions: whether the performance is a solo, chamber, or large-ensemble performance. Each category has a host of unique nonmusical factors that can affect adjudication (McPherson & Thompson, 1998).
- performance environment: acoustics, performance space, screens or barriers for anonymity, equipment available to the performer, and so forth. For instance, technical, acoustical, or audio issues can have a noticeable effect on adjudication.

**Figure 1.1**

*Process Model of Assessing Musical Performance*




From "Assessing music performance: issues and influences," by G. McPherson, & W.F. Thompson, 1998, *Research Studies in Music Education*, *10*(1), p. 13 (https://doi.org/10.1177/1321103X9801000102). Copyright 1998 by SAGE Publications. Reprinted with permission.

Once a performance context is established, McPherson and Thompson (1998) suggest the two influential factors driving judgment can be divided into specific musical factors (e.g., instrument, repertoire, ensemble size and ability, accompanists) and nonmusical factors (e.g., order of appearances, gender, adjudicator experience). Musical factors are often objective, measurable, and controlled variables in an audition, such as specific audition excerpts used,

instrument groupings, metronome markings, and so forth. Although many musical items are quantifiable, bias can often present issues regarding interpretation and preference. The presence of nonmusical factors is the most difficult to control and frequently offers entry points for bias in the evaluation process.

Even though musical factors can be quantifiable with criterion-based assessment, partiality still presents issues in performance adjudication in areas of preference and interpretation. Radocy (1986) argued that although some musical factors might provide a more reliable point of agreement, many judgments stand alone as subjective points of measurement based on an adjudicator's personal cognitive processes and perception.

Evidence supports that the effects of nonmusical factors must be considered for accurate performance assessment (Bergee & Westfall, 2005; Welch, 1994). Uncontrolled nonmusical factors regularly provide a gateway for preconception throughout the evaluation process (McPherson & Thompson, 1998). The effects of implicit and explicit bias in performance evaluation have been investigated heavily in music performance adjudication. For instance, the consequences of nonmusical factors such as adjudicator prowess (Hedden & Johnson, 2008; Hewitt & Smith, 2004; Kinney, 2009), audition order (Bruine de Bruin & Keren, 2003; Bruine de Bruin, 2005, 2006; Glesjer & Heyndels, 2001, Wind & Wesolowski, 2018), physical appearance (Goldin & Rouse, 2000; Peynircioğlu et al. 2018; Tsay, 2014; Waddell & Williamon, 2017; Wapnick et al., 1997), and so forth, have been found to have an overwhelming influence on audition results.

The research of nonmusical effects on music performance adjudication is thorough. However, Pope and Barnes (2015) noted that most research in this area involves analyzing the evaluating capabilities of professional musicians, music majors, and active music teachers as

adjudicators in controlled music performance settings. Only a few studies (Glesjer & Heyndels, 2001; Hickman, 2015; Lien & Humphreys, 2001; Saunders & Holahan, 1997) have focused on musical and nonmusical variables of the entire body of all-state audition results.

### *Summary*

McPherson and Thompson (1998) created the PMAMP based on existing research in performance measurement. In their model, the designation of the performance context, followed by understanding an event's musical and nonmusical factors, is paramount when focusing on an evaluation process. Their framework acts as an informative, organizing launchpad for reviewing necessary literature and scope when customizing a research-based music performance measurement instrument.

## Oklahoma All-State Band and Orchestra Winds and Percussion Auditions

Being selected into an all-state honor group is not only one of the most sought-after accomplishments for students (Elliott, 1995a), but it is also recognized as one of the highest honors in most states (DeCarbo et al., 1990). Dugger (1997) asserted that acceptance into the OkMEA All-State band is among the highest honors in the State of Oklahoma. Conversely, with such importance, there is no record of the development of the measurement tool used before this study, the length of time it was used, nor a record of any prior measurement systems that may have existed.

### *Call to Reform*

At the annual meeting of the Oklahoma Music Educators Association (OkMEA) Band Division in 2011, the member band directors debated the reliability of its traditional all-state adjudication measure (OkMEA: Band Division, 2011). There were many musical and nonmusical variables that directors found vexing. Questions arose about whether time slots

affect audition scores and why some students who had made all-state before the current year did not make it past the first round of their senior years. They discussed how to quantify "tone." Other variables ranging from inexperienced adjudicators, the current holistic scoring system needing to be more effective, and the effectiveness of the Olympic scoring system were deliberated. Many directors supported their arguments with experiential stories, which caused doubts about the reliability of the existing system. In support of their current system, a strong, almost folkloric, common theme of "as long as a judge's scoring stays consistent, there shouldn't be a problem" acted as a collective reasoning mechanism throughout the discussion. After debate and a vote of the directors in attendance, a committee was created of current band directors to research and propose a revised guiding rubric for judges, and if approved by vote, ultimately implemented in future all-state auditions. For purposes of disclosure, I was present at this meeting as a band director member of OkMEA, which occurred before entering graduate school and seeded this study.

### *Performance Context and Other Factors of OkMEA Auditions*

McPherson and Thompson (1998) identified that forming the performance context is possible by means of establishing the purpose of the assessment, the type of performance, the performance proportions, and the performance environment. Additionally, establishing the performance context is paramount in understanding the assessment process. In the case of the OkMEA All-State audition process, the sole purpose of the assessment was to rank high school musicians from best to worst in each instrument group. The only diagnostic feedback provided to the students was the ranking of the minority of students selected as members of the Oklahoma All-State Band or Orchestra. The type of performance utilized was a randomized, anonymous audition consisting of playing rehearsed excerpts from assigned music etudes and a

non-rehearsed sightreading excerpt provided in the audition. The performance proportion in this audition was a solo performance without assistance from a conductor, recording, or accompanist. Finally, the audition environment utilized for the audition was a typical high school classroom with added screening to protect the identities of the players and adjudicators. Outside the audition room, the school environment was distractingly crowded, energetic, and loud. Most of those present were mainly parents, teachers, and hundreds of students primarily practicing, waiting in cue for their audition, playing games with friends, or waiting for their audition results to be posted.

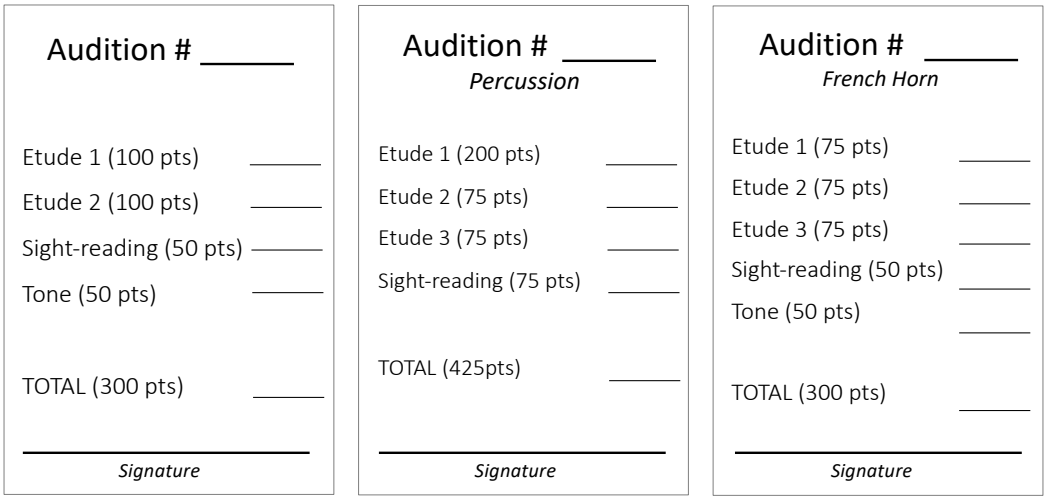
**Musical Factors.** When applying the McPherson and Thompson (1998) PMAMP to the Oklahoma All-State Band and Orchestra Auditions (OASBOA), many musical factors affecting the OkMEA audition process were controlled within the procedure manual. For instance, each instrument group was auditioned separately from other instruments. Each group was assigned a specific repertoire to prepare for the audition posted on the same day. Furthermore, each student performed the same selected excerpts from the audition material and the same sightreading audition etude. It was assumed that students were using the correct editions of posted audition etudes; therefore, the composer's tempo and expression markings were allegedly identical for each performer. All music excerpts were supplied to the adjudication panel. Except for percussion, musical instruments were not provided for students to use. The use of provided percussion instruments was optional, with no records of the usage rate.

But even with the controlled repertoire and matching audition experience for everyone involved, no specific criteria or segmented rating scales were provided to adjudicators to assess the individual auditions. This is visible in the replication of the global adjudication measure

employed before this study, shown in Figure 1.2. All decisions by adjudicators were holistic and utilized varied, undefined point ranges.

**Figure 1.2**

*Sketch of the Adjudication Ballot Form Used by OkMEA (circa 1985–2012)*




*Note*. There is no known copy of the original OkMEA Final Round scoring ballot shown here. The sketch above is a rendering from memory and cross-referenced with Duggar's (1997) description of the measure.

**Nonmusical factors.** The presence of nonmusical factors and their combinations with other factors are complex and have been shown to affect adjudication (Lien & Humphreys, 2001). OkMEA's use of an anonymous audition scenario controls for the effects of nonmusical factors such as race, gender, and deportment and is also a preventive measure for cheating. Other safeguards such as a uniform audition performance environment, releasing the audition music on a uniform date, using the same excerpts for all auditioners, uniform allotment of preparation time during the sightreading component, auditioning specific grade-level students,

and using an unchanging adjudication panel for each set of auditions are in place to prevent or limit the possible effects of numerous other nonmusical factors.

However, endogenous variables can appear to some degree, anywhere and anytime, and measuring their influence empirically is a perplexing mission (Bergee & McWhirter, 2005). Factors such as player anxiety, student/adjudicator preparation time, and experience are impossible to control. A robust amount of research has shown nonmusical variables such as school size, student access to private instruction, distance traveled, etc., to reliably predict adjudicator scores and ratings (Bergee & McWhirter, 2005; Bergee & Platt, 2003; Bergee & Westfall, 2005; Lien & Humphreys, 2001). With most all-state selections awarded to students from suburban metro areas, one can assume that OkMEA is not immune to such variables.

Nevertheless, one nonmusical variable possibly influencing OSBOA audition scores centers on the adjudicators themselves. Variables such as an adjudicator's experience, familiarity with excerpts, familiarity with instruments, musical expertise, or even fatigue are loosely controlled regarding the OASBOA adjudicator pool. In addition, another nonmusical variable indirectly related to adjudicator experience and fatigue is "order effect." Even though participants are randomized and kept anonymous, the influence of order effects on adjudication is always a factor (Bruine de Buin, 2005).

**Selecting Adjudicators**. Choosing adjudicators for the Oklahoma All-State Band audition was a mixture of oral tradition and loose procedures. Similar to national trends (Elliot, 1995), the primary resource used for judge selection was a living list of participating band directors and their known instrument(s) of strength. This list was edited and passed down yearly to each audition committee chairperson and used to help guide judging assignments. The OkMEA Band committee chairperson often relied on their institutional knowledge and

networking capabilities for securing the potential judges before the event through various forms of communication. The common practice was to position adjudicators for the second round of OkMEA All-State auditions from the pool of band directors required to be present at the audition location as they had participating students. Outside adjudicators were often hired in advance to fill additional needs predicted by the OkMEA audition staff. The additional adjudicators were typically retired band directors, college professors, private contractors, symphony musicians, or pre-service college music education majors from local universities. Adjudicators who did not have students participating in auditions were provided with a monetary supplement for their duties. Many studies (Elliott, 1995a; Barnes & McCashin, 2002; Welker, 1997; Wine, 1996) corroborate this method of adjudicator recruitment and appointment to be common practice in selecting panels in all-state auditions.

**Adjudicator Expertise, Assignments, and Training.** Absent from the official OkMEA procedures is a system to assess an adjudicator's proficiency on their listed instrument of strength or their skillset in assessing the auditions of an assigned instrument group. There was no official training, college degree, or licensure required to adjudicate OkMEA auditions, nor is being a member of the organization itself a prerequisite. However, every effort was made by the OkMEA audition committee to match each adjudicator's stated personal musical instrument of strength with the musical instrument audition room to be adjudicated. Though it was common practice to prefer returning judges annually, there was no official metric to define "experience" or procedure in place to measure any significance that "experience" had on audition results. Furthermore, adjudicators of different musical instrument strengths were occasionally asked to fill in at the last minute when there were unpredicted absences of judges.

OkMEA had only informal control of the level of experience, availability, or competency of the judges used for the OASBOA. In all, adjudication and assignments were a good-faith effort based on the professionalism of the teachers, outside adjudicators, and the OkMEA audition committee. The effort was buoyed by baseline expectations that adjudicators possess at least a moderate musical skillset.

**Need for the Study**

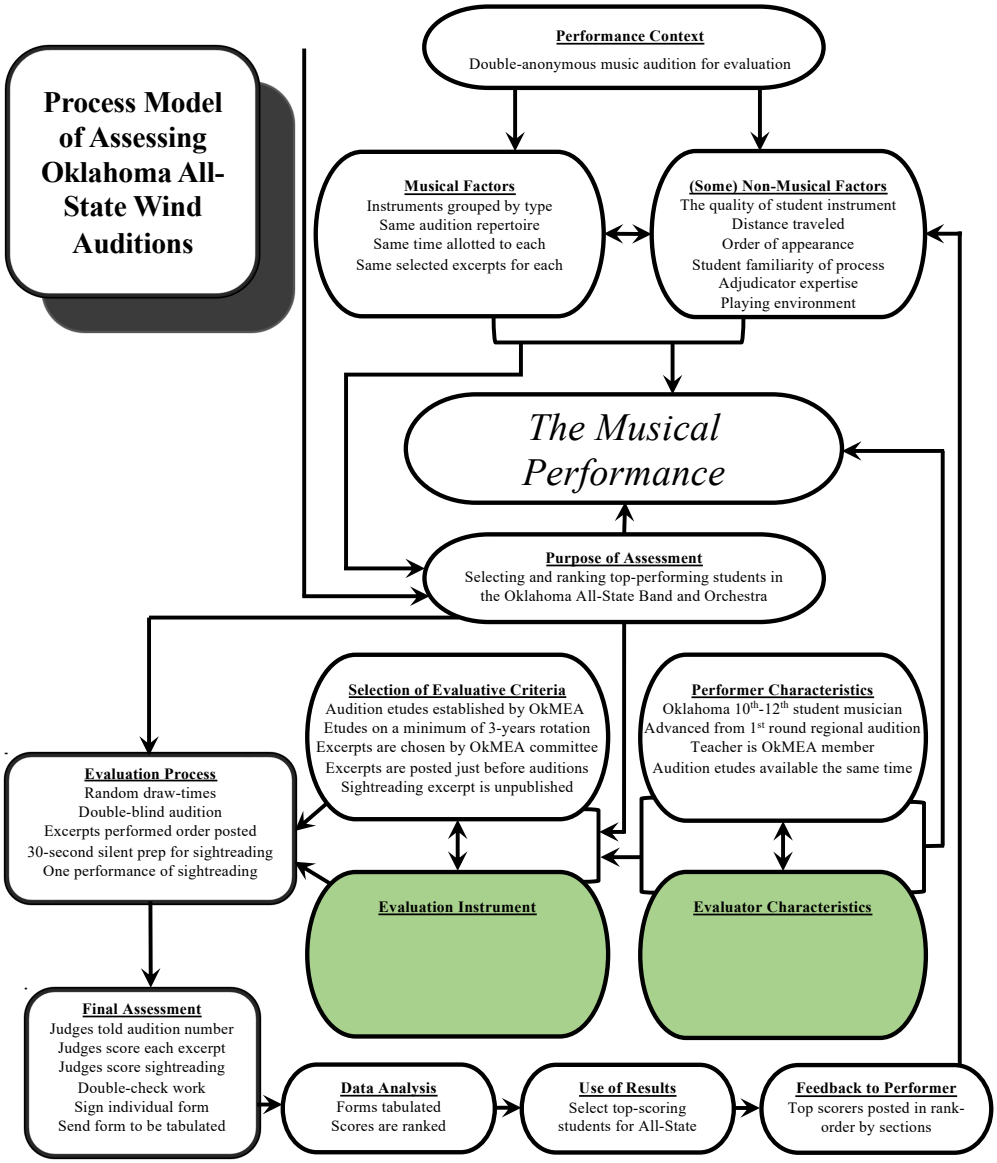
Auditioning for all-state is an intense, high-stakes competition (Hickman, 2015). But even without a universal music performance evaluation model, there are many agreed-upon similarities and a body of research to support development (Williams & Chapman, 2020). Likewise, all stakeholders expect a fair and reliable audition measurement tool, which is vital for the organization's success (Elliot et al., 2000; Glesjer & Heyndels, 2001). Like other all-state honor band auditions, the OASBOA tabulated results are intended to be summative and final. The process aims to rank the best players in the highest chair reliably and to outfit a selective all-state ensemble with the best high school musicians possible. In 2012, the band committee leadership of the OkMEA embarked on researching and submitting possible suggestions to revise the OASBOA scoring process to improve fairness and reliability. As a committee member, I saw the need to thoroughly examine the existing audition procedures and review standing literature on music assessment and evaluation before any new revisions could be brought to the OkMEA band director collective with confidence and fidelity.

Using the McPherson and Thompson (1998) PMAMP as a launchpad, it was possible to define the OASBOA performance context and identify the salient factors framing the audition's musical and nonmusical variables. Once the performance context was established and factors were defined, the information revealed that the existing OASBOA practices controlled for

many known non-interpretive musical variables such as repertoire, instrument groups, and time allotted to perform, as well as many nonmusical variables such as gender, race, deportment, etc., that researchers have shown to affect adjudication. However, subjective musical factors such as artistry and interpretation were scored solely through judgments formed by the cognition and experience of individual judges, who were not trained in adjudicating auditions, accredited, or officially measured for expertise or experience. Thus, the overarching nonmusical factors that influenced OkMEA's previous holistic scoring practice when applying the PMAMP to the OASBOA scenario (Figure 1.3) are 'Evaluator Characteristics" and "Evaluator Instrument."

**Figure 1.3**

*OASBOA within the PMAMP*




*Note*. This figure is adapted from "Assessing music performance: issues and influences," by G. McPherson and W.F. Thompson, W. F.,1998, *Research Studies in Music Education*, *10*(1), p. 13 (https://doi.org/10.1177/1321103X9801000102).

The differentiation of auditioner ability levels cannot be leveled without a system that shows high interrater reliability (Elliot et al., 2000). Even though universal best practices for all-state audition evaluation and adjudicator selection/training do not exist, there is research that could help guide state organizations with their unique audition scenarios and procedures. Like state organizations nationwide, the OkMEA band division had concerns about the reliability of its adjudication and set in motion a committee to suggest constructive revisions or recommend a solution. The body of research and the organization's aspiration for improvement highlight the possibility of a viable and statistically reliable performance measurement tool for Oklahoma All-State band auditions.

**Purpose of the Study**

The purpose of this study was to design, implement, and analyze a research-based adjudication measure, the Barber General Audition Measure (BGAM), for placing and ranking the musicians in the Oklahoma All-State Band and Orchestra final round auditions (OASBOA). Using McPherson and Thompson's (1998) "Process Model of Assessing Musical Performance" as a footing, a measure was designed to function within the existing OkMEA audition procedural framework.

**Research Questions**

1. What was the interjudge reliability of the BGAM within one instrument group?
2. What was the comparison of the BGAM interjudge reliability between different instrument groups?
3. Did the interjudge reliabilities of the BGAM stand up over a five-year period?

**Definitions**

1. The terms *honor band, all-state band*, *all-region band*, and *all-state ensemble* were often used interchangeably to describe similar selective honor groups that are set by various audition methods.
2. The terms *auditioners* and *auditionees* were used interchangeably to refer to participants in auditions for either a summative ranking or formative evaluation.
3. The terms *global* and *holistic* score were used interchangeably to describe the approach of using a single score to evaluate an individual performance or audition.
4. The term *interjudge* or *interrater* reliability is defined as the level of agreement between adjudicators extracting any chance agreement (Zhao et al., 2022).
5. The acronym OkMEA (Oklahoma Music Educators Association) recently replaced the acronym OMEA to address confusion with the Ohio Music Educators Association. To avoid the same confusion, the Oklahoma Music Educators Association was referred to as OkMEA in this research regardless of the timestamp of any specific reference.
6. The term *holistic* rubric is defined by Wesolowski (2012) as an adjudication form, whereas adjudicators provide single scores for a musical performance based on generic and simple descriptors that can be applied to many performance scenarios.
7. Music Educators National Conference (MENC) and the National Association for Music Education (NAfME) are the same organizations. MENC is the nomenclature used to represent the organization from 1934 to 2011.

**Delimitations**

1. Study participants included students limited to Oklahoma tenth through twelfth-grade band students who participated in the final round of auditions and whose teachers were current Oklahoma Music Educators Association members.
2. The analysis was limited to the five years (2015–2019) during which the instrument was implemented in the final round of auditions for the Oklahoma All-State Band.
3. Auditions occur on the first Saturday of December every year at Westmoore H.S. in Moore, Oklahoma. A bad weather day is in place to postpone the auditions one week if needed. This occurred once during this study.

## Chapter 2: Review of Literature

Auditions are ingrained in our musical culture (Silvey, 2009; Waddell & Williamon, 2017). Whether for competition or formative analysis, accurate and useful audition results are paramount to our music education programs. This chapter addresses literature relating to issues surrounding the complexities and variables associated with designing and employing music performance measurements. It highlights studies on musical factors and criterion-based assessment practices as compared to holistic evaluations, as well as nonmusical performance variables that can affect results. Philosophical and historical literature concerning performance measurement is also examined.

### Criterion and Holistic Measurement

#### *Musical Factors and Criterion-Based Assessment*

For the last 50 years, numerous researchers (Abeles, 1973a, 1973b; Bergee, 1987,1995; Cooksey, 1977; Russell, 2010; Smith & Barnes, 2007; Wesolowski, 2017; Zdzinski & Barnes, 2002) experimented with utilizing a criterion-based model to construct music assessment rating scales that exhibited strong reliability and validity. Seeking to improve music assessment measures primarily based on subjective judgment, Abeles (1973b) crafted a measure for objective music assessment by employing Butt and Fiske's (1968) facet-factorial methodology. In this approach, a researcher extracts qualitative facets from commonly agreed-upon performance standards or traits and conventionalizes them into a quantitative measure or rating scale. Guided by Hosmer's (1949) "As the Adjudicator Hears It," Abeles posited that an objective rating scale could be developed through general performance standards that all adjudicators intuitively share. Using factor analysis, Abeles constructed a Clarinet Performance Rating Scale (CPRS) comprised of six performance standards: interpretation, intonation,

rhythm continuity, tempo, articulation, and tone (Abeles, 1973a; Abeles, 1973b). Ultimately, the study reported that a facet-factorial approach to rating scale development showed strong interjudge reliability of .90 and criterion-related coefficient validity above .80 when assessing clarinet solo performances. Abeles' work became the foundation for many researchers (Bergee, 1987, Cooksey, 1977; Russell, 2010; Smith & Barnes, 2007; Zdzinski & Barnes, 2002) who replicated similar approaches for different instruments and ensembles (Zdzinski, 1991).

In 1997, Cooksey (1977) engineered a similar criteria-based rating scale using the facet-factorial approach to assess high school choir performances. In this model, a seventh measurement factor, diction, was added to precision, dynamics, tone control, tempo, balance/blend, and interpretation/musical effect. A choir performance rating scale (also CPRS) was created based on these seven factors and 36 criteria that were extracted and used as subscales. A pool of judges ($N = 50$) was charged with adjudicating 100 high school choirs. The data analysis found high interjudge reliability and high inter-criteria reliability. To test for validity, a second test was employed using a second set of adjudicators ($N = 20$) to assess a small group of the recordings and apply a global score. Using the global scores as the comparison criterion, the model produced high validity with the criteria-based rating scale (generally above .80) and high interjudge reliability (.84). Cooksey concluded that measures built with objective criteria could be used effectively when adjudicating choirs.

DCamp (1980) found similar results to those of Cooksey (1977) when creating a high school band performance rating scale using the same factor analytics approach. Using a test and re-test method, the examiner found good reliability scores from all seventeen judges ($r = .91$). When analyzing the three-judge panels by applying the Spearman-Brown Prophecy formula to investigate interjudge reliability, the researchers found concern with the low coefficient (r =

.64). The Spearman-Brown Prophecy formula, also known as the Spearman-Brown Predictive formula, was used to predict the reliability of comparing a global criterion of ranking the performances as compared to the researcher's rating scale results. The researcher surmised the variance was likely a result of adjudicator inexperience and the researcher's lack of control of the adjudicator pool.

Bergee (1987) also replicated the facet-factorial approach of Abeles (1973a) and Cooksey (1977) and designed a euphonium-tuba performance rating scale (ETPRS) for solo assessments but obtained slightly different results. The researcher's factor analysis extracted an even smaller four-factor structure than the six- and seven-factor scales developed by Abeles (1973b) and Cooksey (1977). The study concluded that qualitative elements of solo euphonium and tuba performances almost always consist of (a) musicality or artistic ability, (b) tone quality, (c) intonation, (d) at least one factor describing musical movement, and e) technique, with 30 items used to outline the subscales. In a separate study differing from Abeles (1973b), Bergee (1995) offered that the judgment of concert band performances comprised a three-factor hierarchy of basic, intermediate, and upper levels. The basic level consisted of well-defined items, whereas the intermediate level comprised interrelated items. The uppermost level consisted of a single predominant higher-order factor. Using a factor analysis of music, the researcher extracted the primary factors of Tone Quality/Intonation, Musicianship/Expressiveness, and Rhythm/Articulation. To test for criterion-related validity, the analysis compared the criterion scores to the commonly used band assessment global scores (I–V) as applied by the adjudicators. Interjudge reliability and validity coefficients ranged consistently from .84 –.99.

Building on the three-factor hierarchal theory, Bergee and Rossin (2019) developed a middle school band assessment model using a two-phased, mixed-method approach of factor analysis to extract items for a rating scale. Other studies (Russell, 2010; Wesolowski, 2017) likewise found similarly high reliabilities by employing the facet-factorial approach to create music performance rating scales in rating guitar performances, jazz improvisation, and jazz rhythm sections, respectively. Departing from the facet-factorial approach, Smith (2009) found that a facet-rational derived rating scale for assessing jazz improvisation (also outlined by Butt & Fiske, 1968) is similarly effective, with Cronbach's alpha values ranging from .81 to .95.

The body of criterion-based music assessment research centers on quantifying musical factors to justify an assessment model or to provide justified feedback to the participants. However, Saunders and Holahan (1997) focused on the accuracy of criteria-specific rating scales in an honor band audition scenario and whether the criteria-specific rating scales could adequately predict total scores; moreover, which measured factors of an audition predicted the bulk of a total score. Strengthening this study, the researchers used the 1994 Connecticut All-State Band woodwind and brass auditions as the sample group. The existing audition procedure of this honor group provided a sample size of $N = 925$ woodwind and brass auditions and 36 qualified elementary, secondary, and collegiate-level teachers/adjudicators in a real-life, real-time scenario. The adjudication form consisting of 11 continuous rating scales and four checklist items was employed. Among the 15 total measurement tools, criteria-specific constructs consisting of tone, intonation, technique/articulation, melodic accuracy, rhythmic accuracy, tempo, interpretation, note accuracy, and musicianship were used. The analysis revealed that areas of a criteria-type assessment model show high reliability and can predict a student's overall score, but validity was not addressed.

### *Other Musical Factors and the Effects on Evaluation*

Commonly understood qualitative musical elements, such as intonation, articulation, tone, rhythm, tempo, and dynamics, have served as anchors for numerous criteria-based reliability studies (Abeles, 1973a; Abeles, 1973b; Fiske, 1977a; Bergee, 1987; Lien & Humphreys, 2001; Russell, 2010; Saunders & Holahan, 1997; Smith, 2009; Stegman, 2009; Wapnick et al., 2005; Zdzinski & Barnes, 2002). However, other musical factors, such as specific audition excerpts used, performance length, metronome markings, and so forth, have also been shown to affect scores. Groulx (2013) investigated adjudicator error detection accuracy while comparing tonal and atonal musical excerpts, finding lower accuracy in atonal excerpts. Pitt (1994) researched pitch, timbre, and their perception by musicians and non-musicians, finding that variations in timbre affected non-musicians' judgments to a greater extent. Wapnick and Ryan et al. (2005) determined that adjudicator rating consistency correlated with excerpt duration and tempo of high-level piano performances. Therefore, although common qualitative musical factors in an audition can be formalized, other less formal musical factors associated with adjudicator experience, musical experience, and other implicit biases that affect adjudicators' evaluations remain complex and ever-present.

### *Criterion-Based Assessment Compared to Holistic Assessment*

The data-driven "age of accountability" brought more widespread use of criteria-based rating scales into the educational arena along with ideological resistance (Mills, 1991). From the onset, many practitioners felt that assessments quantifying specific components of musical performances were forced and impossible to measure objectively (Wrigley, 2013). Using interviews and transcriptions, Stanley et al. (2002) sought to understand the perceptions of performance examiners of a music conservatory made to employ criteria-based rating scales

instead of their former holistic scoring method. The team studied the staff ($n$ = 15) two years after implementing a criterion-based music performance assessment model. The study discovered that while faculty members found using a criteria-focused assessment model provided valuable academic feedback for students, many others found the practice only provided a narrow scope, which hampered the total performance evaluation. It was discovered that most adjudicators using the criteria-based assessment model intuitively adopted a three-pronged approach to scoring:

- Apply an initial "gut reaction" or holistic overall assessment
- Identify performance characteristics to justify the global assessment
- Use the criteria as a checklist to ensure they have listened to all aspects of performance and were not just being "carried away" by a performance (Stanley et al., 2002)

Considering Bergee's (2003) assertion that adjudicators using holistic scoring often cannot support how they arrived at a score, the evidence provided by Stanley et al. (2002) shows an organic process developed of arriving at a score holistically even through segmented criteria ratings were the expectation.

The results when comparing holistic and segmented type scoring have been mixed. For instance, Norris and Borst (2007) examined the effectiveness of two choral festival adjudication forms showing more robust interjudge reliability with a descriptive rubric-based form than a more traditional, less descriptive adjudication form. Fiske (1975, 1977) found holistic scores more reliable than criteria-based score sheets when providing a rank order when adjudicating trumpet solos. However, the musical experience of the rater was statistically significant when providing diagnostic feedback using a segmented score sheet. Fiske (1977) did show that an adjudicator's performance ability and musical knowledge do not significantly affect ranking

evaluations, thus asserting that non-specialists could rank musicians with high interjudge reliability. Mills (1991) analyzed segmented scoring against holistic scoring scales employed by music teachers and student teachers. The researcher concluded that using segmented, non-holistic rating scales recklessly assumed that musical performance could be mathematically measured, and therefore, a performance rating using criteria-based assessment is only equal to the sum of the segmented criteria. Furthermore, without intentional weighting, criteria-based rating only places equal importance on the measured segmented criteria. The researcher debated that holistic assessment has more real-world credibility and reliability because performance is the sum of its parts.

DCamp (1980) raised concerns about interjudge reliability when comparing the criterion rating scale scores of three-judge panels to global scores and independent rankings. In 1985, Levinowitz observed similar variance when comparing the Abeles (1973b) CPRS with a three-part continuous rating scale for clarinet (tone, expression, and rhythm), as well as with a global rating scale. Furthermore, Levinowitz (1985) found that the reliability coefficients for the CPRS were lower than those for the separate rating scale and the global scores (Zdzinski, 1991). Although Saunders and Holahan (1997) reported high reliability in criterion-based rating scales for solo auditions, their study did not measure pairs or groups of judges to assess interjudge reliabilities. Consequently, the researchers concluded that their approach is not to be used for the ordinal placement of participants and recommended further research in this area.

The mixed results in factor-based assessment versus holistic evaluation research are nested in the subjectivity of quantifying specific criteria of musicianship. For example, Smith and Barnes (2007) used the factor-loading approach to develop a seven-factor orchestra rating scale and found strong correlations compared to a segmented criteria-based adjudication form.

Two of the seven factors, “Rhythm” and “Bow,” resulted in less-defined information. The researchers questioned whether the adjudicators applying the rating scale could discern errors as wrong rhythms or bad bowing techniques. The authors implied that adjudicator age, audio versus video sampling, or simply lack of training may have contributed to the ambiguous data on the rating scale. Whatever the cause or combination of causes, the “Rhythm” and “Bow” factors were unclear and subjectively hard to quantify. Both Latimer et al. (2010) and Norris and Borst (2007) found similar problems with the vague term “Rhythm” in their studies. Smith and Barnes (2007) concluded that “Factor analysis is as much an art as a science, and perhaps the overall concept of "competent playing" is sufficiently complex that there is a limit to this form of analysis” (p. 278).

The subjective versus objective judgment debate is not limited to music performance but other arts. In 1939, Eysenck compared a criterion versus a holistic approach to evaluating works of visual art. Bruine de Bruin (2005; 2006) studied human judgment effects in figure skating and dance competitions, respectively. Tam (2018) found low intra-class correlations when researching professors applying the descriptors “Originality and Balanced Views” and “Application of Aesthetic and Contextual Knowledge” while employing a rubric designed to grade students’ written criticism of works of visual art. Outside of the arts, Isaacson and Stacy (2009) debated subjective versus holistic assessment of nursing students, finding trepidation in faculty users and anxiety in students when assessed on specific criteria. Later, Nicholson et al. (2009) compared the interrater reliability of using an analytical rubric versus a holistic score for evaluating the competency of operating room nurses. Like Fiske (1977), the researchers found the holistic scores of the raters to be more consistent, whereas the analytic rubric provided valuable diagnostic feedback.

All in all, criteria-based rating scales have shown strong reliability and usefulness in educational settings where strong and precise performer feedback is essential to the educational process (Abeles, 1973a; Bergee, 1987; Saunders et al., 1997). However, statistical, realistic, and ideological warning signs of employing criteria-based scoring for rankings exist, as the subjective versus objective debate is omnipresent when measuring human behavior.

**Nonmusical Factors Effects on Assessment**

Nonmusical factors and their effects have been a leading concern in music performance evaluation research. Coimbra et al. (2001) argue that unique and memorable music performance contributions are centered on the complexities of a musician's individuality. Their study emphasizes the importance of personal expression and the unique characteristics of each performer in creating impactful musical experiences. Studies have supported that elements such as order of audition, gender, physical appearance, race, and body movement (Elliott, 1995b; Goldin & Rouse, 2000; McPherson & Thompson, 1998; Peynircioğlu et al., 2018; Shih, 2018; Waddell & Williamon, 2017; Wallace, 2020) can generate implicit biases that positively or negatively affect a performance evaluation, thus making accurate evaluation difficult and possibly erroneous. Other studies (Bermingham, 2000; Bergee, 2007; Harrington, 2016; Howard, 2012; Lien & Humphreys, 2001; Kroger & Margolis, 2017; Napoles, 2009; Radocy, 1976; Silvey, 2009; Tsay, 2014; Thompson, 2006; Thompson et al., 2007; Waddell & Williamon, 2017; Wapnick et al., 1997; Wapnick et al., 1998; Wapnick et al., 2013) investigated how nonmusical factors and their effects on scores and rankings. Examples of the investigated nonmusical factors include:

- anonymous vs. open auditions
- facial expressions and other external characteristics of the performer

- the sequence of auditions or positioning
- attire, posture, and physical attractiveness
- distance traveled to the audition site
- enrollment/size of the school
- instrument types/quality
- adjudicator access to musical score/excerpt and adjudicator familiarity with performance music
- authority figures and bias affects
- implied factors, labels, and other "priming" variables
- stage entrance and deportment
- time allotments for judgments

For this review, I have categorized the nonmusical factors under the headings; (a) visual information effect, (b) procedure and order effect, (c) privilege effect, and (d) adjudicator effect.

***Visual Information Effect***

Live music performances provide untold visual stimuli that can introduce implicit bias into the evaluation process. Research by Waddell and Williamon (2017) revealed that poor stage entrances and facial expressions before performing could elicit lower scores from adjudicators. Krahé et al. (2013) found that the ratings of an adjudicator's perceived emotions and a performer's visual and auditory expressions mirrored each other. Howard (2012) reported that performance attire has a major impact on performance ratings, particularly among less experienced adjudicators. This is further supported in 2017 by Waddell and Williamon's secondary finding that adjudicators with varying levels of experience had harsh initial reactions

to a poor stage entrance; however, using a continuous digital measurement interface, the researchers discovered that initial scores from more experienced adjudicators improved once the performance began. Wapnick et al. (1998) also noted that male musicians rated highly for physical attractiveness, attire, and deportment tended to score better in music performance. Additionally, they found that the scoring range between “unattractive” men and “attractive” men was much broader than that of their female counterparts with similar attributes.

In 2014, Tsay enlisted over 1062 novice and expert musicians to participate in a study exploring how visual cues affect the perception of music quality during a performance. The six-part experiment utilized audiovisual, audio-only, and visual-only recordings from a live international chamber group and symphony orchestra competition. The researcher found that participants could accurately select the top-ranked orchestras through video-only excerpts but could not do so with audio-only and audiovisual recordings. Tsay’s findings suggest that visual cues regarding “group dynamics” and “leadership” drive much of one’s judgment when ranking musical performances. In a similar study, Wallace (2020) found that spontaneous hand gestures used by singers in vocal performances affected participants' scoring of “expressivity” and “technique” when distinguishing between audio-only and audiovisual recordings. This supports the Coimbra et al. (2001) finding that second and third-year university vocal students gradually become more aware of their performance body language as they advance through their studies. They discovered that as students gained more experience, they intentionally and methodically worked on their physical presence alongside their musical and mental states to communicate their musical intent more effectively.

Other studies of variables such as race and gender regarding music assessment have provided mixed results and highlighted the complexities of investigating them. Glesjer and

Heyndels (2001) found that historically, women received lower markings than men in the prestigious Queen Elisabeth Music Contest but also found evidence of bias complexities in scoring based on order effect, nationality, and repertoire. Harrington (2016) found that college-aged adjudicators showed explicit biases toward both the gender and age of the performer. Researching "Asian bias" in music performance, Peynircioğlu et al. (2018) found that a bias does exist; however, they concluded the Asian bias was deceptive and was not grounded in awareness of a performer's race but generated by less body movement by Asian performers than Caucasian performers, thus raising the question if the variable is a racial effect or an ethnicity effect.

Goldin & Rouse (2000) studied the impact of anonymous auditions regarding gender discrimination in orchestras by combing the historical audition data of five major symphonies. Interestingly, the empirical analysis provided estimates with significant standard errors, and some coefficients did not pass standard tests of statistical significance. Thus, the data provided no conclusive evidence that anonymous auditions impacted audition scores. However, the overarching finding from the study is that the use of screened auditions had ushered in a 30% increase in female new hires over thirty years, showing improved impartiality in the audition process.

The body of research shows the powerful effect of visual stimuli on performance assessment; the entanglement of complex subtleties among judges and performers makes distinguishing specific biases difficult (Bermingham, 2000).

### *Procedural and Order Effects*

Nonmusical factors can sometimes influence audition scores due to the audition procedures. Thompson et al. (2007) examined timestamps and discovered that the order of

presentation or the length of listening time could affect the perceived fairness of evaluations. Sheldon (2004) found that allowing adjudicators to listen to performances multiple times significantly improved the accuracy and consistency of their evaluations, suggesting that repeated listening can enhance the reliability of performance assessments. Similarly, Napoles (2009) found that adjudicators who were provided with and adhered to reference sheet music during auditions achieved greater consistency in ratings. These findings underscore the importance of allowing adjudicators to review performances more than once, as well as ensuring they are well-informed and consistent.

Order effects, defined by the *APA Dictionary of Psychology* as the "influence of the order in which treatments are administered, such as the effect of being the first administered treatment (rather than the second, third, and so forth)," are found to be a common phenomenon that is difficult to control (APA, n.d.; Bruine de Bruin & Keren, 2003; Bruine de Bruin, 2005). In the case of anonymous music auditions, the treatment consists of the scores assigned to performances by an adjudicator and their relationship to the performance order. According to Bobbitt (2020), types of order effects include:

- practice effects: an adjudicator becomes better at the measurement task as they become more comfortable.
- fatigue effects: an adjudicator performs tasks progressively worse as they become fatigued by doing the same task repetitively.
- boredom effects: an adjudicator performs worse near the end of auditions out of boredom.
- carryover effects: an adjudicator responds to each audition differently based on what was experienced in the preceding audition(s).

Order effects, or serial effects, refer to the phenomenon where the sequence in which stimuli are presented influences the evaluation (Plessner, 1999; Bruine de Bruin, 2005). This effect has been observed across various domains, including evaluations of musical performances (Hollingshead, 1996; Glesjer & Heyndels, 2001), gymnastics (Plessner, 1999), and dance and figure skating (Bruine de Bruin, 2005, 2006). For instance, Flôres and Ginsburgh (1996) and Glesjer and Heyndels (2001) noted higher scores in the rankings of participants who performed in later timeslots during the highly competitive international Queen Elisabeth Music Competition. Wind and Wesolowski (2018) examined how adjudicator fatigue exacerbated order effects over the course of several days among a panel judging instrumental solos. Bruine de Bruin (2005) posits that order effects depend considerably on each rater's unique cognitive tasks in specific situations. Elliott et al. (2000) asserted the importance of recognizing order effects in performance assessments as vital to fair evaluations in competitive settings.

### *Privilege Effect*

Another nonmusical effect correlating with music performance scores is the level of privilege. Barnes and McCashin (2002) explored disparities in access to instruction and resources between small and large schools. Their research involved analyzing educational resources and their impact on student outcomes, revealing that small schools often face challenges in accessing resources, which can affect the quality of education and ultimately be reflected in audition results. Bergee and Platt (2003) analyzed ratings from 7,355 small solo and ensemble events at two Midwestern state festivals. They focused on the effects of time of day, event type, and school size and identified significant differences in their analysis. Delving deeper, Bergee and McWhirter (2005) extended this research by using logistic regression to

model the influence of extramusical variables, including geographical location and district expenditure, on festival ratings. Their studies underscored that these factors influenced the evaluation process, emphasizing the importance of considering contextual variables to ensure fair and equitable assessments in music education. Reinforcing both studies, Bergee and Westfall (2005) extended the analysis by employing a multinomial logistic regression model to allow for more nuanced results. The research concurred that the significance of extramusical privilege factors can predict adjudication results.

Race, gender, and socioeconomic factors also correlate with performance ratings. Perry (2023) analyzed the demographics of performers in Alabama's all-state music programs, revealing disparities based on race and socioeconomic status. By utilizing demographic data, the study identified a significant underrepresentation of certain groups, indicating the need for initiatives that enhance diversity and inclusion in Alabama's music education programs. Harrington (2016) investigated the impact of implied gender and age on music performance evaluations, examining evaluative biases related to these cues. The results showed that implied gender and age can influence evaluators' judgments, potentially leading to biases. Overall, this body of research underscores the necessity for awareness and action to address the effects of privilege in evaluative biases, ensuring fair assessments of music performance.

### *Adjudicator Effect*

**Adjudicator Experience.** Winter (1993) writes, "The problem in music performance assessment, even when appropriate criteria have been established, is that many factors may affect the assessment process. In the examination context, these include the training of the examiners, the experience of the examiners, and the preparation of the students" (p. 34).

Research in adjudicator aptitude, education level, practical experience, and the effects on performance measurement and evaluation have been investigated thoroughly.

Fiske (1977) is a foundational study for understanding the impact of domain-specific expertise on music performance evaluations. Fiske's research showed no relationship between performance ability and judge reliability, but it demonstrated that brass experts provided more consistent and accurate assessments than non-brass experts. Wapnick and Darrow (2013) also highlighted the role of domain-specific expertise among piano majors in offering more reliable evaluations. This built upon Fiske's findings by exploring various dimensions of music performance evaluation. Byo (1989) found that music majors, much like Fiske's brass experts, provided more consistent and reliable ratings than non-music majors, suggesting that musical training enhanced evaluative skills. Kroger and Margulis (2017) examined the influence of nonmusical factors, such as priming and order position, on the evaluations of non-musician adjudicators. Their findings revealed that these nonmusical factors significantly impacted the ratings, indicating that non-musician adjudicators are more susceptible to biases. Together, these studies emphasize the importance of musical training and awareness of potential biases in achieving accurate performance evaluations.

Expanding on the role of adjudicator expertise, Pitt (1994) demonstrated that musicians show greater accuracy and consistency in distinguishing pitch and timbre compared to non-musicians. Kinney (2009) emphasized the importance of excerpt familiarity, revealing that music majors exhibited enhanced internal consistency in their evaluations, particularly for familiar excerpts. These findings align with Fiske's (1977) emphasis on domain-specific expertise and suggest that training and familiarity contribute to reliable outcomes. Silvey (2009) explored the influence of experience on using evaluative labels during music

performance assessments. The study involved both experienced and inexperienced adjudicators using labels to evaluate performances. The results indicated that experienced adjudicators were more consistent and accurate in using evaluative labels than their less experienced counterparts. This highlights the importance of experience in effectively using evaluative tools, suggesting that training and expertise are crucial for achieving reliable and consistent performance evaluations.

Other studies (Radocy, 1976; Brand & Burnsed, 1981; Hewitt & Smith, 2004; Hewitt, 2007; Hedden & Johnson, 2008; Shih, 2018; Parasotto, 2023) collectively emphasize the importance of experience and expertise in music performance evaluations. Hedden and Johnson (2008) found that experienced teachers exhibited greater accuracy in pitch detection, while Brand and Burnsed (1981) demonstrated that experience and ability improved error detection skills. Research by Radocy (1976), Shih (2018), and Parasotto (2023) further supported the importance of experience in achieving reliable evaluations. These findings aligned with similarities observed in previous studies regarding the influence of experience, consistency, and accuracy in evaluations and the role of training and familiarity in enhancing evaluative skills. Collectively, these studies emphasize the essential role of experience, training, and expertise in attaining accurate and consistent music performance evaluations.

**Adjudicator Training.** Among the limited studies focusing on all-state band audition procedures, there is little evidence of a consistent practice of adjudicator training, systems for adjudicator training, or even universal best practices for adjudicating honor band auditions. Welker (1997) found that 35 states lacked adjudicator training, while 14 states had it. Researchers have suggested the need for training on evaluation measurement tools (Stegman, 2009). Both Eyesnick (1972) and Doerksen (1999) emphasized the importance of specialized

training in improving evaluative skills, though Eysenck focused on technical teaching while Doerksen highlighted diagnostic abilities. Benkert (1998) and Sheldon (2004) further support that perceptual and evaluative skills can be developed through targeted training and repeated exposure, reinforcing the importance of experience and familiarity in achieving reliable music performance evaluations. However, Smith and Barnes (2007) posit that "while it may be an admirable goal to develop a detailed and comprehensive inventory of aspects of performance, such a scale could likely be burdensome (requiring a considerable time to complete an evaluation) and necessitate judge training" (p. 269).

**Conclusion**

These studies, reviewed collectively, highlight the multifaceted nature of music evaluations by incorporating factors such as individual expression, standardized criteria, training, experience, and social influences. The potential blending of social, religious, regional, gender, age, expertise, class, and conduct at any moment during the measurement process can create a myriad of complex intricacies that affect real-time adjudication.

# Chapter 3: Methodology

## Purpose of the Study

The purpose of this study was to design, implement, and analyze a research-based adjudication measure for placing the musicians in the Oklahoma All-State Band and Orchestra final round auditions (OASBOA). Using McPherson and Thompson's (1998) "Process Model of Assessing Musical Performance" (PMAMP) as a footing, a measure was designed to function within the existing OkMEA audition procedural framework.

The development of the adjudication measure focused on adhering to the existing double-anonymous audition parameters set by OkMEA, the established adjudicator selection process and pool, while being rooted in the current spectrum of music performance measurement and assessment research. In this chapter, I explored the OASBOA audition process established by the OkMEA, the measurement tool created for the North Central Directors Association (NCDA) all-district band audition pilot study, and the primary study implementation model utilized in the OASBOA for the OkMEA.

## Research Design

In 2011, the Band Division members of OkMEA formed a committee to research, develop, and design a new adjudication rubric for adoption to address reliability and validity concerns. Using the PMAMP as a framework for analysis of the OkMEA audition procedures, coupled with institutional knowledge, I established the following factors and sub-questions to guide the development of the BGAM (Barber General Audition Measure):

1. OkMEA has limited control of the judging pool.
    - Most adjudication positions are filled by band directors who have identified a specific instrument as their strength. This listing operates on the honor system.

OkMEA assumes that a band director is college-trained on this instrument. Since no licensure, training, competency, or experience is formally required to adjudicate the OASBOA, what are the realistic expectations for reliably applying specific descriptors on a formal rubric necessary to pass judgment and rate effectively?

- With limited control of the judging pool, how can a measurement tool equally value the adjudication experience of a 20-year teacher versus a 3-year performer or a college professor of a particular instrument judging for the first time versus a rural school band director who has judged 15 previous years?
- How can a measurement tool account for the adjudicator called to assess an audition room for instruments and music excerpts with which they are unfamiliar?

2. The sole purpose of OkMEA is to extract the top outliers on each instrument.
    - No constructive or critical feedback was provided to the participants.
    - The double-anonymous audition, short excerpts, and one-attempt scenario limit the amount of time an adjudicator has to pass judgment.
    - Adjudicator fatigue and order effects could be factors in adjudicating large instrument groups.
3. OkMEA has control of the audition process and tabulation of scores.
    - The OASBOA uses consistent musical excerpts throughout the audition process.
    - The anonymity of judges and participants is maintained throughout the auditions.

- Physical factors (general appearance, race, gender, body movement, etc.) are not a factor in adjudicating auditions.
- Judges are not allowed to consult with each other when scoring.
- Auditionees and adjudicators get one live performance opportunity per round. There are no recordings, reviews, or re-auditions.
- Time is a factor. All auditions, tabulations, and results postings must be completed on the same calendar day.

**Oklahoma All-State Band and Orchestra Auditions**

The OASBOA takes place annually and includes two rounds of wind and percussion auditions held approximately one month apart. The first and second rounds are scheduled for the first weekends in November and December, respectively, each year. In the first round, each district has some freedom in how it assesses and selects students to advance from its region to the second round of auditions. However, the format for the second round of auditions follows a set agenda aligned with the procedures outlined in the OkMEA Audition manual. Since the audition processes can vary by district, this project concentrates on the second round of auditions that occur every December for consistency. The pilot study was conducted in the NCDA all-district auditions because the NCDA audition process is based on the OkMEA guidelines with only minor differences.

***Audition Music***

Immediately following the completion of the first-round wind and percussion auditions, OkMEA publishes a new list of audition music that will be common for all students statewide who have been promoted to the second round of auditions. The posted music consists of two published standard etudes for the wind instruments that are unique to each instrument. The two

etudes are typically differentiated by being either predominantly lyrical or predominantly technical. Percussion is unique as it focuses on more than one instrument (concert snare, mallet percussion, and timpani); in this case, each instrument is assigned one new etude instead of two. Mallet percussion is adjudicated on both a 2-mallet and a 4-mallet etude, but only a new 2-mallet etude is used for the second round. The 4-mallet etude required for the first round of auditions does not change and is used again for the second round.

Each auditionee is required to perform selected excerpts from the etudes during their scheduled audition time slot. The excerpts chosen by the All-State band committee from the posted etudes are not revealed until the day of the auditions. In addition to the selected excerpts, all wind students must sightread a brief, unpublished excerpt with only 30 seconds of silent preparation time at the end of their audition. Because the percussion field includes multiple instruments, percussionists must complete two sight-reading excerpts: one on the concert snare drum and one on a mallet instrument (e.g., marimba, xylophone).

Due to the high number of auditioners in the second round, larger instrument groups such as clarinet, flute, trumpet, and percussion are divided into preliminary and final audition processes. Each of these large instrument groups utilizes multiple rooms and adjudication teams to narrow the field of auditioners down to a single final round later in the day. In the preliminary rounds, the sight-reading component is not utilized. After the preliminary audition round, the highest-scoring auditioners from each room are posted as progressing to the final round. At this stage, the process begins again for the students selected to participate in the final round. For the small instrument groups, there is only one final round.

On the morning of the second round of auditions, students are assigned an audition number by a randomizing computer program, along with their designated audition room and a

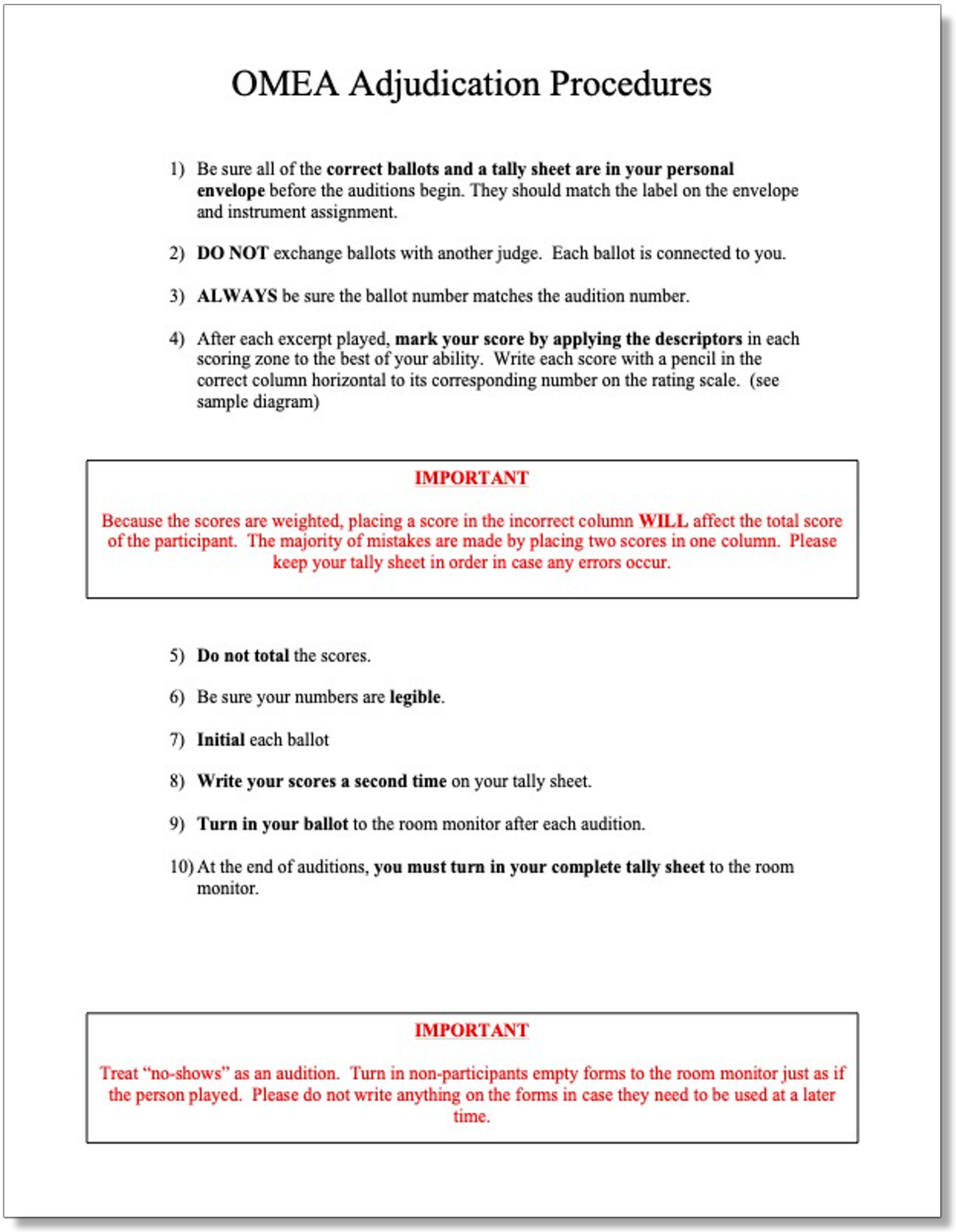
randomly assigned audition time or order. Once the random numbers and audition assignments are posted, the students sign in at the audition location with an adult supervisor, known as an outside room monitor according to OkMEA. The outside room monitor's role is to ensure the efficiency and procedures of the audition process while protecting the identities of the students and adjudicators.

### *Audition Process*

Adjudicators' Meeting. Before the auditions begin, each adjudicator receives a number from one to five (Judge 1, Judge 2, etc.). This number remains consistent throughout the audition round and is used to preserve the anonymity of the adjudicators' identities and scores from the participants, room monitors, and tabulators. The audition contest manager and the judges hold a brief meeting just before the auditions to review procedures and expectations and familiarize themselves with the scoring forms (see Figures 3.1 and 3.2).

**Figure 3.1**

*Sample of Adjudicator Procedures Provided Prior to Auditions*



# OMEA Adjudication Procedures

1) Be sure all of the **correct ballots and a tally sheet are in your personal envelope** before the auditions begin. They should match the label on the envelope and instrument assignment.

2) **DO NOT** exchange ballots with another judge. Each ballot is connected to you.

3) **ALWAYS** be sure the ballot number matches the audition number.

4) After each excerpt played, **mark your score by applying the descriptors** in each scoring zone to the best of your ability. Write each score with a pencil in the correct column horizontal to its corresponding number on the rating scale. (see sample diagram)

**IMPORTANT**

Because the scores are weighted, placing a score in the incorrect column **WILL** affect the total score of the participant. The majority of mistakes are made by placing two scores in one column. Please keep your tally sheet in order in case any errors occur.

5) **Do not total** the scores.

6) Be sure your numbers are **legible**.

7) **Initial** each ballot

8) **Write your scores a second time** on your tally sheet.

9) **Turn in your ballot** to the room monitor after each audition.

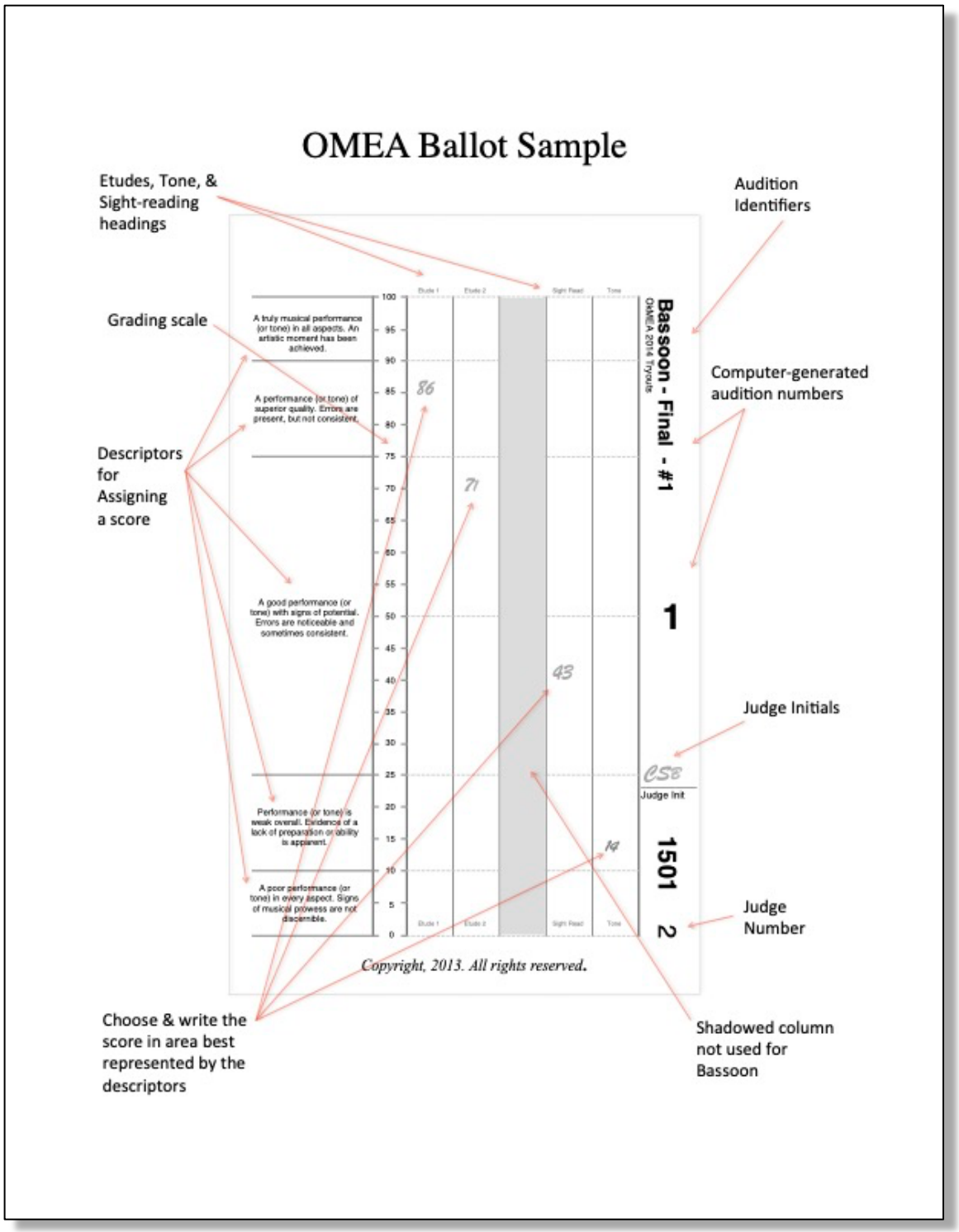
10) At the end of auditions, **you must turn in your complete tally sheet** to the room monitor.

**IMPORTANT**

Treat "no-shows" as an audition. Turn in non-participants empty forms to the room monitor just as if the person played. Please do not write anything on the forms in case they need to be used at a later time.

**Figure 3.2**

*Sample of Ballot Instructions Provided to the Adjudicators Prior to Auditions*



The adjudicators are advised not to communicate with one another or the participants during an audition. They are also instructed not to reach a consensus with the other adjudicators about the participants between auditions. Scoring is meant to reflect the personal judgment of each auditionee by each individual judge. During this meeting, each judge receives separate BGAM (see Figure 3.4) assessment ballots for each player, marked explicitly with the assigned numbers of both the judge and the auditioner. Each ballot is a replica of the BGAM designed for the instrument group being adjudicated.

**Inside the Audition Room.** The adjudication of participants involves a double-anonymous audition scenario. Each audition room features five judges behind a screen to maintain anonymity. This level of concealment reduces the impact of external, nonmusical characteristics on the assessment (Bermingham, 2000). Each audition room includes an adult supervisor (referred to by OkMEA as an inside room monitor or inside monitor). The inside monitor's responsibilities encompass informing the adjudicators of the auditionee's assigned performance number, overseeing the sight-reading process at the end of the audition, organizing and managing completed scoring sheets, and serving as a liaison between the adjudicators and auditionees when necessary. The room monitor is the only participant permitted to communicate with both parties and is allowed to move freely between the screen and the audition area. See Appendix A for a complete list of monitor duties.

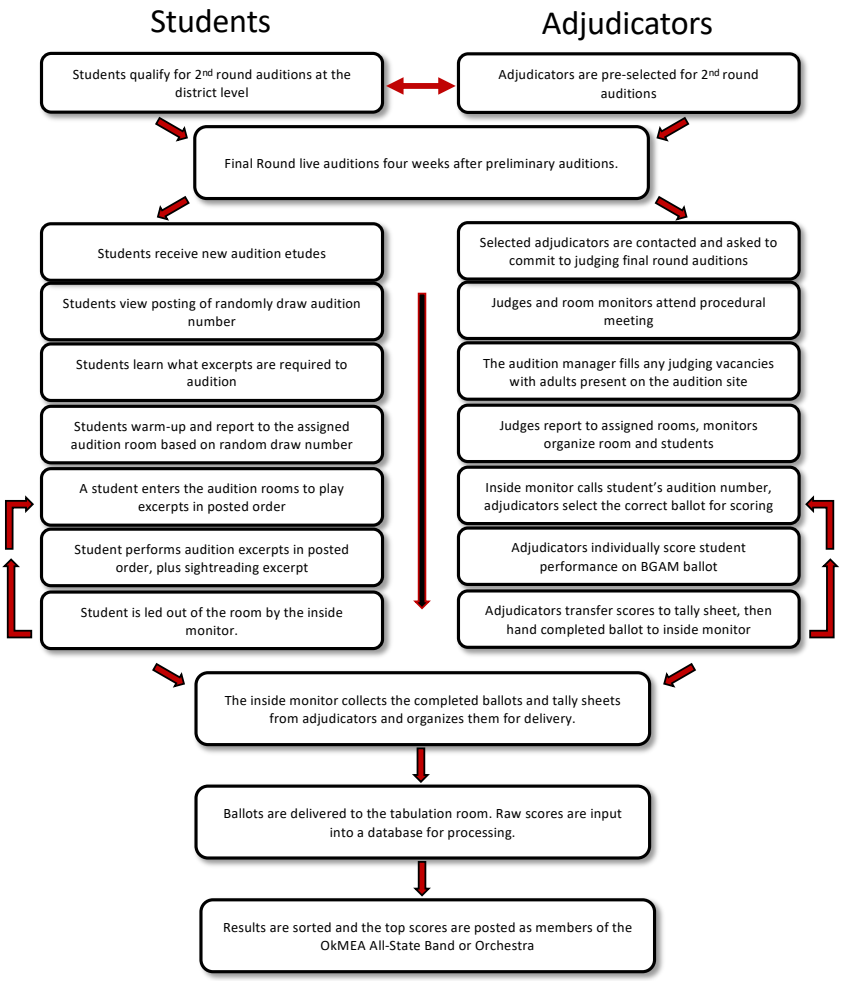
The room monitor facilitates the audition process for every student from start to finish. Initially, the room monitor vocally announces each participant's assigned code number as they enter the room, allowing adjudicators time to ensure that the audition number corresponds with the correct adjudication ballot. The audition begins promptly with a performance of the pre-selected excerpts from the etudes, followed by sight-reading. Adjudicators are expected to score

each component as it is completed. During the sight-reading portion of the audition, the inside monitor places the pre-selected sight-reading music on the music stand and informs the auditionee that they have 30 seconds to study the excerpt. The monitor then starts a 30-second timer and instructs the auditionee to play the excerpt once the time is up. After the participant performs the sight-reading music, the inside room monitor collects the excerpt, escorts the auditionee to the door, gathers the assessment ballots, and welcomes the next contestant. This process continues until all participants have auditioned.

During the audition, adjudicators score each component of the audition on the BGAM by hand, then record their scores on a separate tabulation sheet (used only as a reference in case of a tabulation error; see Appendix C) and submit the completed individual ballots to the inside monitor. The inside monitor organizes the ballot sheets by adjudicator number and student number. Once organized, the inside monitor hands the sorted ballots to an OkMEA representative. At the end of the auditions, adjudicators sign their separate tabulation sheets and submit them to the OkMEA audition committee in a sealed envelope. This process assists in referencing or correcting any possible mistakes affecting tabulation (e.g., scores written in the wrong column, illegible scores) after the adjudicators leave the premises. Figure 3.3 displays the complete audition process.

**Figure 3.3**

*OkMEA Audition Process*




**Audition Room Results.** Both preliminary and final audition rounds follow similar processes. The differences between the preliminary and final auditions are (a) the preliminary auditions do not include the sight-reading component, (b) the musical excerpts selected from the etudes for the final round change, and (c) a new adjudication panel is chosen for the final round. At the end of the final round of auditions, individual scoring ballots are sent by an

OkMEA representative to an audition tabulation room for processing. The consensus top scorers are sorted and displayed as members of the Oklahoma All-State Bands or the wind and percussion section of the Oklahoma All-State Orchestra. The number of students selected for each instrument varies across the ensembles and is based on a fixed allotment outlined in the OkMEA procedure manual. These allotments can be adjusted to meet the needs of the literature and the preferences of the guest conductor.

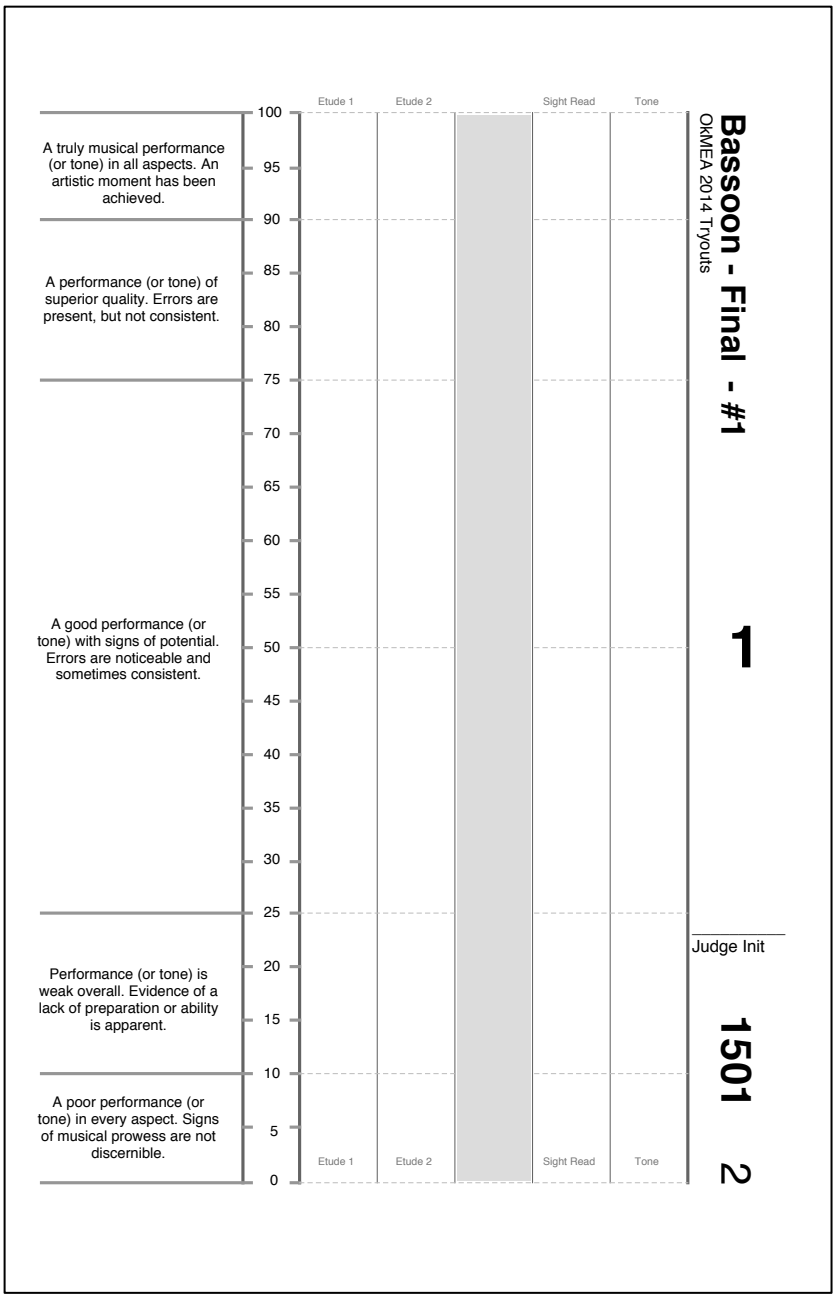
***The BGAM Ballot Design***

The assessment tool developed for this project emphasizes ease of use, anticipating that a diverse panel of adjudicators will utilize it. The tool consists of a half-sheet paper ballot for each audition. Each ballot (see Figure 3.4) includes scoring columns, an auditioner number, an instrument name, judging numbers, a space for adjudicators to verify their ratings with initials, and a computer code that refers to the instrument and the auditioner number.

The ballot features five rubric-type point zones applied to each event in an audition. These point zones are categorized based on percentages that reflect a normal distribution but are rounded to the nearest five for convenience. A column worth 100 points is utilized to score each audition portion. (Each column score is later weighted according to a district's point specifications. For example, for final rankings, as per the OkMEA procedures, etudes are valued at 50 percent of the total score, while sight-reading accounts for 15 percent of the total score.) Columns not necessary for specific instruments are shaded to indicate to adjudicators that they are not required.

**Figure 3.4**

*Barber General Audition Measure*




The BGAM is a hybrid rating scale/rubric featuring intentionally vague, artistry-based assessment descriptors applicable to all musical instruments during every stage of the audition.

It also allows each adjudicator to make subjective judgments based on their expertise. The rating scales are divided into point blocks corresponding to each descriptor and based on percentages representing the normal distribution. The rubric descriptors are:

- 90–100 A truly musical performance (or tone) in all aspects. An artistic moment has been achieved.
- 75–89 A performance (or tone) of superior quality. Errors are present but not consistent.
- 25–74 A good performance (or tone) with signs of potential. Errors are noticeable and sometimes consistent.
- 10–24 Performance (or tone) is weak overall. Evidence of a lack of preparation or ability is apparent.
- 0–10 A poor performance (or tone) in every aspect. Signs of musical prowess are not discernible.

The purpose of employing a normal distribution-like point range for the descriptor sections was to encourage adjudicators to use the entire point spectrum, potentially assisting in identifying outliers. For the all-state band, the highest-scoring outliers are the preferred candidates for placement.

***Pilot Testing***

On November 4, 2012, the BGAM v1 (see Appendix D) was applied to the NCDA HS All-Region Band auditions. The NCDA utilized the same audition process as outlined in the OkMEA procedure manual, except (a) there was only one final round for all instrument groups, and (b) only two adjudicators per room were used instead of five. Like OkMEA, participating

band directors were used as adjudicators and outside judges were employed to fill in the gaps in certain instruments and fulfill the two-adjudicator requirement.

At the audition, the adjudicators were briefed on using the new adjudication ballot, addressing any questions, and instructed to complete a short survey via Survey Monkey about the usability of the new ballot (see Appendix D) at the end of the audition day. The survey link was emailed to all adjudicators. After the auditions, I received a flash drive from the contest manager containing all the scores from every judge for each student audition ($N = 347$) and a key sheet with codes assigned to each instrument group.

All the received data were de-identified and in CSV format. Using the code sheet, I descriptively analyzed each subgroup and loaded the audition data for each instrument group and each judge into SPSS for analysis. I utilized Cronbach's alpha for the study, and all subscale scores were reliable, exceeding the conventional acceptability level of .70 (see Table 3.1).

**Table 3.1**

*Reliability Estimates for NCDA Auditions*

| Instrument | *n* | Cronbach's alpha (a) |
|---|---|---|
| Piccolo | 4 | 1.0 |
| Flute | 27 | .98 |
| Oboe | 13 | .99 |
| Bassoon | 13 | .99 |
| Eb clarinet | 2 | .92 |
| Bass clarinet | 9 | 1.0 |
| Contra clarinet | 5 | .95 |
| Alto saxophone | 23 | .86 |
| Tenor saxophone | 11 | .98 |
| Bari saxophone | 15 | .90 |
| Trumpet | 39 | .97 |
| Horn | 22 | .94 |
| Trombone | 23 | .96 |
| Euphonium | 13 | .92 |
| Tuba | 16 | .98 |
| Percussion 1 (snare drum) | 25 | .97 |
| Percussion 2 (keyboard) | 22 | .99 |
| Percussion 3 (timpani) | 12 | .95 |

*Note.* $N = 347$.

## Procedures

### *Pre-Data Collection*

The preliminary document was sent to dissertation committee members on January 8, 2024. After the dissertation committee approved the research proposal, I submitted the required documentation online to OkMEA and OU-IRB for approval and acquired the existing audition data needed for this study.

### *Data Collection and Analysis*

After approval from OU-IRB and OkMEA, I obtained the de-identified weighted scores from the second-round Oklahoma All-State Band and Orchestra Auditions (OASBOA) wind and percussion auditions from 2015–2019 from the contest database manager. This raw data was provided in CSV format on a single flash drive.

Once received, the data were entered into SPSS, labeled by audition years and instrument group. I used exploratory analysis procedures to descriptively analyze the data (Morgan et al., 2013). I then conducted reliability estimates (internal consistency) for interjudge reliability by instrument and year. Specifically, I utilized Cronbach's alpha ratios and intraclass coefficients (ICC) for this analysis. This analysis method is preferred when there are three or more raters (Shrout & Fleiss, 1979; Portnoy & Watkins, 2000) rather than the two adjudicators in the pilot study. In the case of the main study, each event consistently consists of 5 raters or adjudicators ($k = 5$). I identified the proper ICC form for reliability analysis by employing the Koo and Li (2016) flowchart for ICC form selection and verified it with the McGraw and Wong (1996) flowchart for selecting an ICC. Both charts arrived at employing a mean-rating ($k = 5$), consistency, 2-way mixed-effects model for reporting each event's ICC and 95% confidence

interval (CI). Finally, I compared the measure's reliability longitudinally across the entire data set.

## Chapter 4: Results

The purpose of this study was to design, implement, and analyze a research-based adjudication measure for placing the musicians in the Oklahoma All-State Band and Orchestra final round auditions (OASBOA). Using McPherson and Thompson's (1998) "Process Model of Assessing Musical Performance" as a footing, a measure was designed to function within the existing Oklahoma Music Educators Association (OkMEA) audition procedural framework. Specifically, I aimed to create and implement the Barber General Audition Measure (BGAM), find the interjudge reliability within each separate instrument group using the measure, compare the results to the other instrument groups, and check the consistency of the results over five years. The development of the BGAM focused on operating within the existing double-anonymous audition parameters set by OkMEA and the current adjudicator selection process, and it was rooted in the ongoing spectrum of music performance measurement and assessment research.

From 2015 to 2019, the BGAM was implemented in the OKMEA's final rounds of the OASBOA. The audition manager provided data from the participants ($N = 2699$) from all five years in CSV format, which I uploaded and analyzed using SPSS version 28.0. I employed exploratory data analysis procedures to descriptively analyze and report the data (Morgan et al., 2013). I then examined each instrument group's interjudge reliability analysis, with results displayed separately for each year over the five years. The display presents the number of auditioners ($N$), followed by the Cronbach's alpha ($\alpha$), the intraclass coefficients (ICC), and then the display of confidence intervals (CI). The ICC estimates and their 95% confidence intervals were based on a mean-rating ($k = 5$), consistency, 2-way mixed-effects model.

**Descriptive Analysis**

Each instrument group has a varied range of auditioners (*n*). Groups with consistent numbers of auditionees from year to year (e.g., alto saxophone) use an additional preliminary audition round and send a specific number of participants to the final round. All participants are qualified high school musicians who have earned the opportunity to audition based on results from one or two preliminary audition rounds before the final round. Every event, every year, featured a panel of five adjudicators for each instrument group, with each independent adjudicator applying the BGAM for each participant. Due to the strict anonymity practiced, it is unknown yet possible that some adjudicators may have participated in more than one instrument group in any given year's dataset. However, an adjudicator cannot appear twice in the same instrument group in the same year. Considering that most of the adjudicator pool participates in the auditions each succeeding year, it is highly likely that many judges are represented in the same adjudicator group in each separate year of the data set. However, I did not evaluate individual judges longitudinally in this study, as it does not fall under the scope of the current research.

In the following tables, I present the results from each instrument group from 2015 to 2019, using the total weighted scores. The OASBOA contest manager provided me with the total weighted scores of each adjudicator used to place students in the final rank order of each instrument group. The display presents the number of auditioners (*n*), followed by the Cronbach alpha ($\alpha$), the intraclass correlation coefficients (ICC), and then the 95% Confidence Intervals (CIs) of each year for each instrument. Five adjudicators or raters ($k = 5$) were consistently used for auditions each year.

Since $k > 3$ in each tested group, I applied the accepted ICC reliability thresholds suggested by Portney & Watkins (2000) for the analysis: values less than .50 indicate poor reliability, values from .50 to .75 indicate moderate reliability, values from .75 to .90 indicate good reliability and values of .90 and above indicate excellent reliability. All Cronbach's alpha values at or above the .70 threshold were considered reliable, with values above .80 considered the standard for basic research (Nunnally & Bernstein, 1994, as cited in Lance et al., 2006).

***Instrument Group Results by Year of Audition***

For the piccolo group, the number of participants varied from $n = 17$ to $n = 30$ each year between 2015 and 2019 (see Table 4.1). The Cronbach's alpha indicated excellent internal reliability, ranging from .95 to .97. The ICC values were also mainly above the .90 threshold, indicating excellent reliability values. The notable ICC drop to .79 with a 95% CI [.415, .923] in 2019 indicates good reliability values for that specific year.

**Table 4.1**

*Reliability Results for Piccolo Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 17 | .97 | .93 | [.780, .925] |
| 2016 | 19 | .97 | .93 | [.788, .972] |
| 2017 | 27 | .97 | .95 | [.899, .977] |
| 2018 | 30 | .95 | .93 | [.849, .971] |
| 2019 | 21 | .95 | .79 | [.415, .923] |

Like the piccolo group, the flute group's Cronbach's alpha values also indicate high internal consistency, ranging from .87 to .94 over the five years (see Table 4.2). The ICC values fluctuated from .90 in 2017 to .69 in 2019, suggesting potential issues with the rater agreement. The 95% CIs provided greater insight into the precision of the ICC values, with narrower intervals [.876, .955] in 2018, while the broader intervals in 2016 [.272, .865] and 2019 [.304, .851] mirrored the alpha levels in indicating increased variability. Like piccolo, the flute audition results exhibited lower ICCs in the 2019 auditions, although still above the accepted .50 value for moderate consistency.

**Table 4.2**

*Reliability Results for Flute Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 42 | .87 | .80 | [.624, .891] |
| 2016 | 42 | .94 | .70 | [.272, .865] |
| 2017 | 42 | .93 | .90 | [.801, .946] |
| 2018 | 42 | .93 | .92 | [.876, 955] |
| 2019 | 42 | .89 | .69 | [.304, .851] |

For the oboe group, Cronbach's alpha values were consistently high, ranging from .96 to .98, indicating excellent internal consistency in the BGAM (see Table 4.3). The ICC values were also consistently high, with an ICC above .95 for all years, indicating excellent reliability. The 95% CIs suggested "good" to "excellent" internal consistency, with slight uncertainty noted in 2015 CI [.803, .967]. Overall, the data demonstrate high reliability and consistency in oboe audition scoring, with excellent rater agreement across all metrics and years.

**Table 4.3**

*Reliability Results for Oboe Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 27 | .96 | .96 | [.803, .967] |
| 2016 | 24 | .97 | .96 | [.920, .982] |
| 2017 | 26 | .98 | .96 | [.899, .984] |
| 2018 | 27 | .98 | .98 | [.963, .989] |
| 2019 | 26 | .97 | .95 | [.894, .980] |

The bassoon audition Cronbach's alpha values from 2015 to 2019 were consistently high ($\alpha$ = .93–.97), indicating strong internal consistency (see Table 4.4). The ICC values varied, peaking at .94 in 2016 and dipping to .85 in 2019, yet all solidly above the .75 accepted threshold for good internal consistency. The 95% CIs were tightest in 2016 [.873, .974] and the widest in 2018 [.628, .952], reflecting varying consistency. The data revealed high reliability and consistency in scoring, with only minor variability in rater agreement across the years.

**Table 4.4**

*Reliability Results for Bassoon Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's ($\alpha$) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 31 | .93 | .87 | [.694, .942] |
| 2016 | 23 | .96 | .94 | [.873, .974] |
| 2017 | 35 | .96 | .93 | [.849, .965] |
| 2018 | 36 | .97 | .88 | [.628, .952] |
| 2019 | 33 | .93 | .85 | [.656, .932] |

The Eb clarinet audition adjudication exhibited high reliability and consistency in applying the BGAM over five years (see Table 4.5). However, there was more substantial variability in rater agreement and consistency compared to the oboe and clarinet auditions. The Cronbach's alpha values remained high, ranging from .88 to .98, indicating strong internal consistency. The ICC values fluctuated, peaking at .98 in 2018 before dropping to .83 in 2019. The 95% CIs were narrower in 2018 [.932, .995], indicating higher accuracy, while those in 2019 [0.519, 0.949] indicated increased uncertainty.

**Table 4.5**

*Reliability Results for Eb Clarinet Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 6 | .98 | .97 | [.887, .995] |
| 2016 | 8 | .94 | .94 | [.833, .986] |
| 2017 | 8 | .88 | .87 | [.653, .970] |
| 2018 | 8 | .97 | .98 | [.932, .995] |
| 2019 | 11 | .92 | .83 | [.519, .949] |

In the Bb clarinet auditions from 2015 to 2019, Cronbach's alpha values consistently remained high, ranging from .94 to .97, indicating excellent internal consistency (see Table 4.6). The ICC values varied slightly, with the highest at .97 in 2017 and the lowest at .87 in 2018, yet all years were well above the .75 threshold, indicating "good" to "excellent" reliability. The 95% CIs showed narrower ranges in 2017 [.949, .980], reflecting more precise estimates, while the intervals in 2018 [.675, .940] were wider, indicating greater uncertainty. The BGAM data demonstrated high reliability and consistency in the Bb clarinet audition scores, with minor variability in rater agreement across the years.

**Table 4.6**

*Reliability Results for Bb Clarinet Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 42 | .94 | .89 | [.767, .946] |
| 2016 | 48 | .96 | .94 | [.876, .966] |
| 2017 | 52 | .97 | .97 | [.949, .980] |
| 2018 | 52 | .95 | .87 | [.675, .940] |
| 2019 | 52 | .94 | .91 | [.832, .953] |

The bass clarinet auditions in 2015, with 40 participants ($n = 40$), showed high reliability and consistency, evidenced by a Cronbach's alpha of .96 and an ICC of .96, with a narrow 95% CI [.935, .976] (see Table 4.7). The following year, 2016, saw a substantial increase in participants ($n = 54$), with a slight dip in Cronbach's alpha to .94 and an ICC of .93, maintaining a relatively narrow 95% CI [.891, .956]. In 2017, with 36 participants ($n = 36$), the

reliability metrics returned to the high levels seen in 2015, $\alpha = .96$, ICC = .96, and 95% CI [.930, 0.976]. However, 2018 marked an increased drop in rater agreement, with the ICC falling to .68 but still within the accepted threshold of .50–.75, indicating moderate reliability. This year also recorded the lowest Cronbach's alpha at .91 and the most expansive 95% CI [.272, 0.853], suggesting increased uncertainty and variability in the ratings. In 2019, the reliability metrics returned to the patterns established from 2015 to 2017 with $\alpha = .96$ and ICC = .94 with a 95% CI [.876, .969], reflecting excellent reliability.

**Table 4.7**

*Reliability Results for Bass Clarinet Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's ($\alpha$) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 40 | .96 | .96 | [.935, .976] |
| 2016 | 54 | .94 | .93 | [.891, .956] |
| 2017 | 36 | .96 | .96 | [.930, .976] |
| 2018 | 42 | .91 | .68 | [.272, 853] |
| 2019 | 42 | .96 | .94 | [.876, .969] |

Auditionee numbers (*n*) ranged from 13 to 21 in the contra clarinet audition results (see Table 4.8). Cronbach's alpha values were consistently high, ranging from .93 to .98, indicating strong internal consistency. The ICC values were generally high, peaking at .96, but dropped to .86 in 2016. Except for the broader interval in 2016, which indicated some uncertainty, the 95% CIs were narrow in most years' adjudications using the BGAM.

**Table 4.8**

*Reliability Results for Contra Clarinet Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 19 | .97 | .96 | [.925, .985] |
| 2016 | 21 | .95 | .86 | [.603, .948] |
| 2017 | 13 | .96 | .96 | [.910, .986] |
| 2018 | 14 | .93 | .92 | [.829, .971] |
| 2019 | 17 | .98 | .96 | [.901, .987] |

In 2015, the alto saxophone auditions exhibited moderate reliability, with a Cronbach's alpha of .83 and a moderate ICC of .61, accompanied by a wide 95% CI [.231, .821], indicating considerable uncertainty (see Table 4.9). In 2016, the reliability improved substantially, with a Cronbach's alpha of .94 and an ICC of .93 with a narrow CI [.861, .965], reflecting precise estimates. The data from the 2017 auditions show a slight decrease in reliability, with a Cronbach's alpha of .85 and an ICC of .77 with a CI [.547, .895]. In 2018, the reliability metrics improved again, featuring a Cronbach's alpha of .93, an ICC of .93, and a narrow CI [.869, .965], maintaining a high level in 2019. Overall, the alto saxophone auditions

demonstrated high reliability and consistency in most years, with some variability in rater agreement and precision, particularly in 2015 and 2017.

**Table 4.9**

*Reliability Results for Alto Saxophone Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 24 | .83 | .61 | [.231, .821] |
| 2016 | 24 | .94 | .93 | [.861, .965] |
| 2017 | 24 | .85 | .77 | [.547, 895] |
| 2018 | 24 | .93 | .93 | [.869, .965] |
| 2019 | 24 | .92 | .91 | [.842, .958] |

The analysis of total weighted scores of tenor saxophone auditions from 2015 to 2019 indicated the number of auditioners (*n*) varied each year, ranging from 27 to 38. The reliability of the scores, as indicated by Cronbach's alpha, was consistently high, with values between .96 and .98. The ICC also demonstrated excellent reliability, ranging from .92 to .98. The 95% Confidence Intervals for the ICC values were consistently narrow, indicating confident estimates of reliability, with the widest CI occurring in 2018 [.778, .965]. The results can be found in Table 4.10.

**Table 4.10**

*Reliability Results for Tenor Saxophone Auditions Using Total Weighted Scores*

| Year | n | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 33 | .97 | .95 | [.895, .977] |
| 2016 | 28 | .98 | .98 | [.949, .988] |
| 2017 | 27 | .96 | .96 | [.926, .978] |
| 2018 | 36 | .97 | .92 | [.778, .965] |
| 2019 | 38 | .98 | .97 | [.937, .983] |

Like the tenor saxophone group, baritone saxophone participants ($n$ = 17–29) fluctuated over the five years (see Table 4.11). The reliability of the adjudications, as indicated by Cronbach's alpha, was consistently excellent, with values ranging from .96 to .99. The ICC also exhibited strong reliability, ranging from .94 to .98. The 95% CI for ICC values was narrow, indicating precise reliability estimates, with the lowest interval in 2016 [.834, .976] and the highest in 2019 [.968, .992]. Overall, the audition results indicate excellent and consistent reliability.

**Table 4.11**

*Reliability Results for Baritone Saxophone Auditions Using Total Weighted Scores*

| Year | n | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 23 | .97 | .95 | [.878, .979] |
| 2016 | 27 | .97 | .94 | [.834, .976] |
| 2017 | 28 | .96 | .95 | [.900, .973] |
| 2018 | 17 | .98 | .98 | [.944, .990] |
| 2019 | 29 | .99 | .98 | [.968, .992] |

While the reliability of the trumpet auditions remained relatively high over the five years, there was a noticeable decline in both Cronbach's alpha and ICC values from 2015 to 2018, followed by a slight improvement in 2019 (see Table 4.12). The Cronbach's alpha for the trumpet audition scores ranged from .89 to .95, indicating only a minor decrease in reliability over the years. The ICC range (.73 to .92) varied more considerably, with a 95% CI for ICC values being wider in some years, suggesting a slight pattern of less consistent estimates of reliability, particularly in 2016, 2017, and 2018 [.450, .898], [.437, .881] and [.388, .875], respectively.

**Table 4.12**

*Reliability Results for Trumpet Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 36 | .95 | .92 | [.843, .960] |
| 2016 | 36 | .91 | .77 | [.450, .898] |
| 2017 | 36 | .89 | .75 | [.437, .881] |
| 2018 | 36 | .89 | .73 | [.388, .875] |
| 2019 | 34 | .92 | .82 | [.576, .918] |

The horn audition data analysis indicates excellent reliability across the entire dataset (see Table 4.13). The Cronbach's alpha consistently showed excellent results, ranging from .93 to .95, while the ICC values varied from .81 to .91, indicating reliability levels ranging from "good" to "excellent." However, the 95% CI for the ICC values were relatively wide in some years, suggesting varying confidence in the reliability estimates, with the lowest interval recorded in 2017 at [.400, .897].

**Table 4.13**

*Reliability Results for Horn Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 32 | .94 | .91 | [.803, .967] |
| 2016 | 32 | .93 | .81 | [.825, .958] |
| 2017 | 36 | .94 | .88 | [.400, .897] |
| 2018 | 36 | .95 | .82 | [.618, .925] |
| 2019 | 36 | .95 | .87 | [.490, .907] |

Like the horn audition scores, the ICC values of the trombone dataset varied substantially more, ranging from .76 to .92. Likewise, the 95% CI for the ICC values were relatively broad, indicating varying confidence in the reliability values, with the lowest interval being in 2017 [.400, .897] and the highest interval in 2015 [.803, .967]. The Cronbach's alpha ranged from .91 to .96, indicating a slight decline in reliability over the years, but still well above the accepted .70 threshold. Results can be found in Table 4.14.

**Table 4.14**

*Reliability Results for Trombone Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 28 | .96 | .92 | [.803, .967] |
| 2016 | 28 | .94 | .91 | [.825, .958] |
| 2017 | 28 | .92 | .76 | [.400, .897] |
| 2018 | 28 | .91 | .83 | [.618, .925] |
| 2019 | 28 | .91 | .79 | [.490, .907] |

In analyzing the Euphonium results, the number of auditioners (*n*) varied each year, with a notable drop in 2019 to just 16 due to the introduction of a preliminary round that year. The Cronbach's alpha remained high, ranging from .94 to .99, while the ICC values indicated strong reliability, ranging from .76 to .97. The 95% confidence interval (CI) for ICC values was generally narrow, except for 2019, which had a wider interval of [.336, .919]. This decline in reliability and the smaller sample size in 2019 suggest a potential issue that may affect the consistency of the scores for that year.

**Table 4.15**

*Reliability Results for Euphonium Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 46 | .99 | .97 | [.940, .987] |
| 2016 | 48 | .96 | .94 | [.880, .966] |
| 2017 | 47 | .96 | .95 | [.919, .971] |
| 2018 | 52 | .94 | .92 | [.846, .954] |
| 2019 | 16 | .95 | .76 | [.336, .919] |

Similar to the euphonium dataset, the number of tuba participants (*n*) varied each year, experiencing a significant drop to 23 participants in 2019 due to the addition of a preliminary audition round for this instrument group. Over the years, Cronbach's alpha consistently remained high, ranging from .96 to .98. The ICC values also demonstrated strong reliability, ranging from .86 to .98. Like the euphonium, the 95% CI for ICC values were generally narrow, indicating reliable confidence estimates, except for 2019, which showed a broader interval of [.609, .945].

**Table 4.16**

*Reliability Results for Tuba Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 56 | .96 | .93 | [.867, .963] |
| 2016 | 49 | .97 | .95 | [.905, .976] |
| 2017 | 62 | .98 | .98 | [.973, .988] |
| 2018 | 57 | .98 | .96 | [.917, .981] |
| 2019 | 23 | .96 | .86 | [.609, .945] |

Nonetheless, the sharp decline in the number of participants and the lower ICC value in 2019 for tuba auditions correspond with the trend observed in euphonium auditions for that same year. This suggests that the reduced number of participants in 2019 posed challenges that affected the reliability of scores across other instruments, highlighting a suggestive deviation that warrants further discussion. In summary, although the tuba auditions demonstrated high reliability overall, the 2019 results highlight a noteworthy divergence, consistent with patterns seen in auditions for other instrument groups with fewer participants.

The analysis of the percussion group dataset indicated consistently high reliability, with Cronbach's alpha values ranging from .91 to .96 (see Table 4.17). The ICC values varied between .84 and .95, and the 95% CIs generally narrow. In summary, the percussion auditions demonstrated strong and consistent reliability over the years.

**Table 4.17**

*Reliability Results for Percussion Auditions Using Total Weighted Scores*

| Year | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| 2015 | 30 | .91 | .87 | [.738, .934] |
| 2016 | 30 | .96 | .95 | [.903, .973] |
| 2017 | 30 | .96 | .94 | [.879, .971] |
| 2018 | 30 | .93 | .89 | [.759, .948] |
| 2019 | 30 | .91 | .84 | [.653, .925] |

### *Analysis Comparing Instrument Groups by Audition Year*

In the final round, all adjudicators across all instrument groups utilized the same multidimensional BGAM to express their preferences and qualify their audition results. The following results were analyzed for all auditioned instrument groups for each year separately to assess the internal consistency of the BGAM as a whole for that year and to compare instrument groups to each other.

My analysis of the 2015 dataset for all instrument groups revealed some initial insights into the reliability of the BGAM use as a whole. The number of participants (*n*) ranged from 6 to 56, averaging 31.29. The Cronbach's alpha ($\alpha$) averaged .95, indicating excellent reliability across most adjudicator groups. The intraclass correlation coefficient (ICC) averages .90, also reflecting excellent reliability. Notably, the euphonium data exhibited the highest reliability, with a ($\alpha$) = .99, ICC = .97, and 95% CI [.940, .987]. Also of note, the alto saxophone group showed varying reliability, with a Cronbach's alpha of .83, ICC of .61 and 95% CI [.231, .821]. Instruments groups such as the bass clarinet, contrabass clarinet, and tenor saxophone consistently demonstrate high reliability, with both Cronbach's alpha and ICC values around .96 to .97, respectively. Overall, I concluded that the 2015 results exhibited highly consistent interrater reliability and internal consistency, with a few exceptions. The results can be found in Table 4.18.

**Table 4.18**

*2015 Reliability Results for All Instrument Groups*

| Instrument | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| Piccolo | 17 | .97 | .93 | [780. .925] |
| Flute | 42 | .87 | .80 | [.624, .891] |
| Oboe | 27 | .96 | .96 | [.803, .967] |
| Bassoon | 31 | .93 | .87 | [.694, .942] |
| Eb clarinet | 6 | .98 | .97 | [.887, .995] |
| Bb clarinet | 42 | .94 | .89 | [.767, .946] |
| Bass clarinet | 40 | .96 | .96 | [.935, .976] |
| Contra clarinet | 19 | .97 | .96 | [.925, .985] |
| Alto sax | 24 | .83 | .61 | [.231, .821] |
| Tenor sax | 33 | .97 | .95 | [.895, .977] |
| Bari sax | 23 | .97 | .95 | [.878, .979] |
| Trumpet | 36 | .95 | .92 | [.843, .960] |
| Horn | 32 | .94 | .91 | [.806, .955] |
| Trombone | 28 | .96 | .92 | [.803, .967] |
| Euphonium | 46 | .99 | .97 | [.940, .987] |
| Tuba | 56 | .96 | .93 | [.867, .963] |
| Percussion | 30 | .91 | .87 | [.738, .934] |

In my analysis of the 2016 instrument groups, I found slight improvements in reliability compared to 2015 (see Table 4.19). The average number of participants (*n*) remained similar to 2015, with an average of 31.82. The Cronbach's alpha remained static, with an average of .95 in 2016, maintaining excellent reliability. The average ICC also remained stable at .91 compared to the 2015 average of .90. I also found that the 2016 tenor saxophone group exhibited the highest reliability and internal consistency across all metrics, $\alpha = .98$, ICC = .98, and 95% CI [.949, .988].

On the other hand, the trumpet group exhibited the lowest ICC of .77, indicating lower reliability than other instruments, although still above the accepted .75 ICC threshold for good internal consistency. I also found that the flute's Cronbach's alpha improved from .87 in 2015 to .94 in 2016, though its ICC decreased from .80 to .70 with a 95% CI [.272, .865], revealing mixed reliability results, yet still above the .70 accepted moderate ICC threshold. The oboe, bass clarinet, and euphonium groups displayed high reliability in both years, with consistently elevated Cronbach's alpha and ICC values. Overall, with my analysis, I concluded that there was a slight enhancement in reliability from 2015 to 2016, with all instrument groups maintaining excellent reliability in adjudication with the BGAM.

**Table 4.19**

*2016 Reliability Results for All Instrument Groups*

| Instrument | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| Piccolo | 19 | .97 | .93 | [.788, .972] |
| Flute | 42 | .94 | .70 | [.272, .865] |
| Oboe | 24 | .97 | .96 | [.920, .982] |
| Bassoon | 23 | .96 | .94 | [.873, .974] |
| Eb clarinet | 8 | .94 | .94 | [.833, .986] |
| Bb clarinet | 48 | .96 | .94 | [.876, .966] |
| Bass clarinet | 54 | .94 | .93 | [.891, .956] |
| Contra clarinet | 21 | .95 | .86 | [.603, .948] |
| Alto sax | 24 | .94 | .93 | [.861, .965] |
| Tenor sax | 28 | .98 | .98 | [.949, .988] |
| Bari sax | 27 | .97 | .94 | [.834, .976] |
| Trumpet | 36 | .91 | .77 | [.450, .898] |
| Horn | 32 | .93 | .81 | [.511, .918] |
| Trombone | 28 | .94 | .91 | [.825, .958] |
| Euphonium | 48 | .96 | .94 | [.880, .966] |
| Tuba | 49 | .97 | .95 | [.905, .976] |
| Percussion | 30 | .96 | .95 | [.903, .973] |

In 2017, the number of instruments tested ranged from 8 to 62, averaging 33.24. All instrument groups had a Cronbach's alpha average of .94, while ICC averaged .90, indicating

high overall reliability for the year. I observed that the Tuba group exhibited the highest reliability in 2017, achieving a Cronbach's alpha of .98 and an ICC of .98 with a 95% CI [.973, .988], reflecting exceptional internal consistency. The analysis of the oboe and Bb clarinet groups also demonstrated high reliability with similarly strong metrics. In contrast, the trumpet and alto saxophone groups showed the lowest reliability among the instruments, with ICCs of .75 and .77, respectively.

Additionally, instrument groups such as the flute, horn, bass clarinet, tenor sax, and baritone saxophone consistently showed high reliability scores. Conversely, the trombone group exhibited good reliability, with an ICC value of .76. Although the tuba displayed the highest reliability, the alto saxophone and trumpet groups indicated areas for potential improvement. Overall, the reliability metrics remained relatively stable for the entire year, with only minor fluctuations in Cronbach's alpha and ICC values. Results found in Table 4.20.

**Table 4.20**

*2017 Reliability Results for All Instrument Groups*

| Instrument | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| Piccolo | 27 | .97 | .95 | [.899, .977] |
| Flute | 42 | .93 | .90 | [.801, .946] |
| Oboe | 26 | .98 | .96 | [.899, .984] |
| Bassoon | 35 | .96 | .93 | [.849, .965] |
| Eb clarinet | 8 | .88 | .87 | [.653 .970] |
| Bb clarinet | 52 | .97 | .97 | [.949, .980] |
| Bass clarinet | 36 | .96 | .96 | [.930, .976] |
| Contra clarinet | 13 | .96 | .96 | [.910, .986] |
| Alto sax | 24 | .85 | .77 | [.547, .895] |
| Tenor sax | 27 | .96 | .96 | [.926, .978] |
| Bari sax | 28 | .96 | .95 | [.900, .973] |
| Trumpet | 36 | .89 | .75 | [.437, .881] |
| Horn | 36 | .94 | .88 | [.723, .945] |
| Trombone | 28 | .92 | .76 | [.400, .897] |
| Euphonium | 47 | .96 | .95 | [.919, .971] |
| Tuba | 62 | .98 | .98 | [.973, .988] |
| Percussion | 30 | .86 | .94 | [.879, .971] |

In the 2018 data analysis for all instrument groups, I found that the total participants (*n*) tested ranged from 8 to 57, with an average of 34.12, slightly increased compared to 31.29 in

2015 and 31.82 in 2016. The Cronbach's alpha averaged .94 in 2018, marginally lower than the 2015 and 2016 averages ($\alpha$ = .95). Likewise, the ICC averaged .90 in 2018, the same as in 2015 and 2016. Also, in the 2018 dataset, I was able to establish that the oboe and baritone saxopohone groups exhibited the highest reliability, with a Cronbach's alpha of .98 and an ICC of .98, while the bass clarinet groups had the lowest ICC at .68 with a 95% CI [.272, .853]. This is the third occurrence of ICC values found to be below the accepted threshold of .75 for good reliability yet still above the accepted .50 for moderate reliability. Instrument groups such as the Bb clarinet, tenor sax, and tuba had high reliability across all four year's analyses. The trumpet group had mixed results in 2018, with a Cronbach's Alpha of .89 and an ICC value of .73 with a wide 95% CI [.388, .875], indicating the possibility of lower reliability than in previous years. Results found in Table 4.21

**Table 4.21**

*2018 Reliability Results for All Instrument Groups*

| Instrument | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| Piccolo | 30 | .95 | .93 | [.849, .971] |
| Flute | 42 | .93 | .93 | [.876, .955] |
| Oboe | 27 | .98 | .98 | [.963, .989] |
| Bassoon | 36 | .97 | .88 | [.628, .952] |
| Eb clarinet | 8 | .97 | .98 | [.932, .995] |
| Bb clarinet | 52 | .95 | .87 | [.675, .940] |
| Bass clarinet | 42 | .91 | .68 | [.272, .853] |
| Contra clarinet | 14 | .93 | .92 | [.829, .971] |
| Alto sax | 24 | .93 | .93 | [.869, .965] |
| Tenor sax | 36 | .97 | .92 | [.778, .965] |
| Bari sax | 17 | .98 | .98 | [.944, .990] |
| Trumpet | 36 | .89 | .73 | [.388, .875] |
| Horn | 36 | .95 | .82 | [.464, .924] |
| Trombone | 28 | .91 | .83 | [.618, .925] |
| Euphonium | 52 | .94 | .92 | [.846, .954] |
| Tuba | 57 | .98 | .96 | [.917, .981] |
| Percussion | 30 | .93 | .89 | [.759, .948] |

Overall, the reliability metrics remained relatively stable for 2018 compared to the last three years, with only minor fluctuations in Cronbach's Alpha and ICC values. The highest reliability

averages observed in 2016 were slightly lower in 2018, yet the general reliability of all instrument groups remained commendably high.

In my analysis of the 2019 dataset, the baritone saxophone group exhibited the highest reliability with a Cronbach's alpha of .99 and an ICC of .98 with a 95% CI [.944, .990]. This continued the exceptional internal consistency and reliability trend for that instrument group. In contrast, the flute group recorded the lowest ICC at .69 with 95% CI [.304, .851] yet still demonstrated moderate overall reliability. Instrument group value ranges such as the Bb clarinet, bass clarinet, and tenor saxophone showed excellent reliability, with Cronbach's alpha ($\alpha$ = .94–.97) and ICC values (.91–97). The euphonium group yielded mixed results, $\alpha$ = .95, ICC = .76, and 95% CI [.336, .919], indicating slightly lower reliability than previous years. Overall, I found the reliability for 2019 to be stable, with only minor fluctuations in Cronbach's alpha and ICC values. Results can be found in Table 4.22.

**Table 4.22**

*2019 Reliability Results for All Instrument Groups*

| Instrument | *n* | Cronbach's (α) | ICC | 95% CI |
|---|---|---|---|---|
| Piccolo | 21 | .95 | .79 | [.415, .923] |
| Flute | 42 | .89 | .69 | [.304, .851] |
| Oboe | 26 | .97 | .95 | [.894, .980] |
| Bassoon | 33 | .93 | .85 | [.656, .932] |
| Eb clarinet | 11 | .92 | .83 | [.519, .949] |
| Bb clarinet | 52 | .94 | .91 | [.832, .953] |
| Bass clarinet | 42 | .96 | .94 | [.876, .969] |
| Contra clarinet | 17 | .98 | .96 | [.901, .987] |
| Alto sax | 24 | .92 | .91 | [.842, .958] |
| Tenor sax | 38 | .98 | .97 | [.937, .983] |
| Bari sax | 29 | .99 | .98 | [.968, .992] |
| Trumpet | 34 | .92 | .82 | [.576, .918] |
| Horn | 36 | .95 | .87 | [.655, .945] |
| Trombone | 28 | .91 | .79 | [.490, .907] |
| Euphonium | 16 | .95 | .76 | [.336, .919] |
| Tuba | 23 | .96 | .86 | [.609, .945] |
| Percussion | 30 | .91 | .84 | [.653, .925] |

***Total Averages by Year***

All adjudicators throughout the dataset used the same measure (BGAM) to evaluate each participant, regardless of the instrument. The table below presents an overview of the reliability estimates for total final round weighted totals by audition year from 2015 to 2019 (see Table 4.23). The number of auditioners (*N*) ranged from 502 to 567 each year. The average Cronbach's alpha values were consistently high, ranging from .94 to .95, indicating strong internal consistency. The average intraclass correlation coefficient values varied slightly, ranging from .87 to .91, reflecting high reliability across the years. This table indicates consistent reliability in the audition scoring process over the five-year period using the BGAM.

**Table 4.23**

*Overall Reliability Estimates of the Complete Final Round Weighted Totals by Audition Year*

| Year | *N* | Average α | Average ICC |
|---|---|---|---|
| 2015 | 532 | .94 | .90 |
| 2016 | 541 | .95 | .90 |
| 2017 | 557 | .94 | .91 |
| 2018 | 567 | .95 | .89 |
| 2019 | 502 | .94 | .87 |

*Note.* The average totals for all years combined are α = .95 and ICC = .89

Complete results of all five years on one chart can be seen in Appendix F.

## Chapter 5: Discussion

Being selected for an all-state performance ensemble is among the highest and most prestigious honors a high school student can achieve. All-state selection not only provides students opportunities to perform music with other advanced peers and established conductors, but also to enhance their applications to universities and for scholarships. Auditions are typically high-stakes, competitive, and stressful events for everyone involved. With each state having an organized all-state ensemble and so much at stake, one might assume that a standardized system exists for adjudicating and selecting the best candidates for all-state recognition. Music performance and measurement research has been extensive and prominent for over 50 years. However, a standardized, globally accepted adjudication system for all-state auditions does not yet exist, requiring state organizations to determine what works best for their unique situations. In 2012, the Oklahoma Music Educators Association (OkMEA) band division set forth on a mission to research and find a solution to some of the reliability issues it was experiencing with its existing audition adjudication measurement tool.

Therefore, the purpose of this study was to design, implement, and analyze a research-based adjudication measure for placing musicians in the Oklahoma All-State Band and Orchestra (OASBOA). Using McPherson and Thompson's (1998) “Process Model of Assessing Musical Performance” (PMAMP) as a guideline, I developed a measure to operate within the existing Oklahoma Music Educators Association (OkMEA) audition procedural framework. Specifically, I created and implemented the Barber General Audition Measure (BGAM), determined the interjudge reliability within each separate instrument group using the measure, compared the results across different instrument groups, and assessed the consistency of the results over five years. My development of the BGAM focused on adhering to the existing

double-anonymous audition parameters set by OkMEA and the current adjudicator selection process while being grounded in the spectrum of music performance measurement and assessment research at the time.

**Summary of Major Findings**

The weighted total scores represented the results that the OkMEA used to rank students for selection to its all-state ensembles. I employed exploratory data analysis procedures to organize the data, then analyzed each instrument group's interjudge reliability, displayed the results separately for each section and each year over the five years. I then presented the total averages for the entire 5-year span. The displays show the number of auditioners (*N*), followed by Cronbach's alpha, the intraclass coefficients (ICC), and the confidence intervals (CI) based on a mean rating ($k = 5$) in a consistency, 2-way mixed-effects model. I employed Cronbach's alpha to measure the internal consistency of the BGAM and identify to what extent the raters were rating the same construct (Tavakol & Dennick, 2011). I employed the ICCs and the 95% CIs to measure the rating agreement across the multiple raters (Koo & Li, 2016). The ICC results present the agreements and the degree of correlations between the measurements.

Considering the musical and nonmusical factors embedded in music performance measurement for this particular set of circumstances, I created, tested, and implemented a general music audition measurement tool, referred to as the BGAM in this study, for all-state wind and percussion auditions. In this study, I questioned whether it is possible to show strong interjudge reliability and internal consistency within each instrument/adjudicator group using the BGAM. Is it possible to have strong interjudge reliability and internal consistency for every instrument and adjudicator group using the BGAM for one year? Is it possible to show strong interjudge reliability and internal consistency of the BGAM every year over a five-year period?

In the following sections, I have posted the completed results for the entire study for reference (Figure 5.1) and provided a summary of each research question. I then conclude by presenting the implications of the BGAM, suggestions to the OkMEA, stating the study limitations, and providing recommendations for future research.

**Figure 5.1 Complete Analysis Results**

Reliability (Cronbach's alpha) and Intra-class Coefficients (ICC)

| | | 2015-16 | | | | 2016-17 | | | | 2017-18 | | | | 2018-19 | | | | 2019-20 | | | | 5-Year Total Av | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Code | Instrument* | n | Cron | ICC | 95%CI | n | Cron | ICC | 95%CI | n | Cron | ICC | 95%CI | n | Cron | ICC | 95%CI | n | Cron | ICC | 95%CI | Cron Av | ICC Av |
| 101 | Piccolo | 17 | 0.97 | 0.93 | .925-.780 | 19 | 0.97 | 0.93 | .788-.972 | 27 | 0.97 | 0.95 | .899-.977 | 30 | 0.95 | 0.93 | .849-.971 | 21 | 0.95 | 0.79 | .415-.923 | 0.962 | 0.91 |
| 102 | Flute | 42 | 0.87 | 0.80 | .624-.891 | 42 | 0.94 | 0.70 | .272-.865 | 42 | 0.93 | 0.90 | .801-.946 | 42 | 0.93 | 0.92 | .876-.955 | 42 | 0.89 | 0.69 | .304-.851 | 0.912 | 0.80 |
| 103 | Oboe | 27 | 0.96 | 0.96 | .803-.967 | 24 | 0.97 | 0.96 | .920-.982 | 26 | 0.98 | 0.96 | .899-.984 | 27 | 0.98 | 0.98 | .963-.989 | 26 | 0.97 | 0.95 | .894-.980 | 0.972 | 0.96 |
| 104 | Bassoon | 31 | 0.93 | 0.87 | .694-.942 | 23 | 0.96 | 0.94 | .873-.974 | 35 | 0.96 | 0.93 | .849-.965 | 36 | 0.97 | 0.88 | .628-.952 | 33 | 0.93 | 0.85 | .656-.932 | 0.95 | 0.89 |
| 105 | Eb clarinet | 6 | 0.98 | 0.97 | .887-.995 | 8 | 0.94 | 0.94 | .833-.986 | 8 | 0.88 | 0.87 | .653-.970 | 8 | 0.97 | 0.98 | .932-.995 | 11 | 0.92 | 0.83 | .519-.949 | 0.938 | 0.92 |
| 106 | Bb clarinet | 42 | 0.94 | 0.89 | .767-.946 | 48 | 0.96 | 0.94 | .876-.966 | 52 | 0.97 | 0.97 | .949-.980 | 52 | 0.95 | 0.87 | .675-.940 | 52 | 0.94 | 0.91 | .832-.953 | 0.952 | 0.92 |
| 108 | Bass clarinet | 40 | 0.96 | 0.96 | .935-.976 | 54 | 0.94 | 0.93 | .891-.956 | 36 | 0.96 | 0.96 | .930-.976 | 42 | 0.91 | 0.68 | .272-.853 | 42 | 0.96 | 0.94 | .876-.969 | 0.946 | 0.89 |
| 109 | Contra clarinet | 19 | 0.97 | 0.96 | .925-.985 | 21 | 0.95 | 0.86 | .603-.948 | 13 | 0.96 | 0.96 | .910-.986 | 14 | 0.93 | 0.92 | .829-.971 | 17 | 0.98 | 0.96 | .901-.987 | 0.958 | 0.93 |
| 110 | Alto sax | 24 | 0.83 | 0.61 | .231-.821 | 24 | 0.94 | 0.93 | .861-.965 | 24 | 0.85 | 0.77 | .547-.895 | 24 | 0.93 | 0.93 | .869-.965 | 24 | 0.92 | 0.91 | .842-.958 | 0.894 | 0.83 |
| 111 | Tenor sax | 33 | 0.97 | 0.95 | .895-.977 | 28 | 0.98 | 0.98 | .949-.988 | 27 | 0.96 | 0.96 | .926-.978 | 36 | 0.97 | 0.92 | .778-.965 | 38 | 0.98 | 0.97 | .937-.983 | 0.972 | 0.96 |
| 112 | Bari sax | 23 | 0.97 | 0.95 | .878-.979 | 27 | 0.97 | 0.94 | .834-.976 | 28 | 0.96 | 0.95 | .900-.973 | 17 | 0.98 | 0.98 | .944-.990 | 29 | 0.99 | 0.98 | .968-.992 | 0.974 | 0.96 |
| 113 | Trumpet | 36 | 0.95 | 0.92 | .843-.960 | 36 | 0.91 | 0.77 | .450-.898 | 36 | 0.89 | 0.75 | .437-.881 | 36 | 0.89 | 0.73 | .388-.875 | 34 | 0.92 | 0.82 | .576-.918 | 0.912 | 0.80 |
| 114 | Horn | 32 | 0.94 | 0.91 | .806-.955 | 32 | 0.93 | 0.81 | .511-.918 | 36 | 0.94 | 0.88 | .723-.945 | 36 | 0.95 | 0.82 | .464-.924 | 36 | 0.95 | 0.87 | .655-.945 | 0.942 | 0.86 |
| 115 | Trombone | 28 | 0.96 | 0.92 | .803-.967 | 28 | 0.94 | 0.91 | .825-.958 | 28 | 0.92 | 0.76 | .400-.897 | 28 | 0.91 | 0.83 | .618-.925 | 28 | 0.91 | 0.79 | .490-.907 | 0.928 | 0.84 |
| 117 | Euphonium** | 46 | 0.99 | 0.97 | .940-.987 | 48 | 0.96 | 0.94 | .880-.966 | 47 | 0.96 | 0.95 | .919-.971 | 52 | 0.94 | 0.92 | .846-.954 | 16 | 0.95 | 0.76 | .336-.919 | 0.96 | 0.91 |
| 118 | Tuba** | 56 | 0.96 | 0.93 | .867-.963 | 49 | 0.97 | 0.95 | .905-.976 | 62 | 0.98 | 0.98 | .973-.988 | 57 | 0.98 | 0.96 | .917-.981 | 23 | 0.96 | 0.86 | .609-.945 | 0.97 | 0.94 |
| 119 | Percussion | 30 | 0.91 | 0.87 | .738-.934 | 30 | 0.96 | 0.95 | .903-.973 | 30 | 0.96 | 0.94 | .879-.971 | 30 | 0.93 | 0.89 | .759-.948 | 30 | 0.91 | 0.84 | .653-.925 | 0.934 | 0.90 |
| | | 532 | 0.94 | 0.90 | | 541 | 0.95 | 0.90 | | 557 | 0.94 | 0.91 | | 567 | 0.95 | 0.89 | | 502 | 0.94 | 0.87 | | 0.95 | 0.89 |
| | | | N = 2699 | | | K = 425 *k = 5 in every instance | | | | | | | | ** split to 2 rounds in 19-20 which is why the n is lower | | | | | | | | | |

Accepted Threshold for Cronbach's Alpha Values (Nunnally & Bernstein, 1994)

<.70 = poor reliability
.70–.80 = reliable
.80–.95 = standard for research
>.95 = acceptable for applied use

Accepted Threshold for ICC Values (Portney & Watkins, 2000)

<.50 = poor reliability
.50–.75 = moderately reliability
.75–.90 = good reliability
>.90 = excellent reliability

## Individual Instrument Group BGAM Results Per Year

The reliability of the BGAM application used for audition scores among the various wind and percussion groups was a critical first step in ensuring a fair and consistent audition process for the OASBOA. Overall, the results for the instrument groups showed no identifiable

pattern of concern over the five years. However, by identifying a few anomalies and variations in these reliability metrics, I highlighted areas where rater consistency was weaker and used this information to suggest improvements that could enhance the overall reliability of the audition process.

For 2019, the piccolo auditions stood out with an ICC value of .79, significantly lower than those of other years, where ICC values ranged from .93 to .95. In flute auditions, the years 2016 and 2019 exhibited notable anomalies, with ICC values of .70 and .69, respectively. These values were considerably lower than those of other years, indicating less consistency in rater agreement during those periods. However, all ICCs remained above the accepted threshold for good reliability. It was also plausible that in 2019, the flute and piccolo adjudicator panels included some or all of the same adjudicators, suggesting the possibility of rater *DRIFT* (differential rater functioning over time) or some particular effects of DRIFT (Wolfe et al., 2001).

Although several types of DRIFT effects exist, Wolfe et al. (2001) focused on the construct of time interactions with primacy/recency effects, differential centrality/extremism, and practice/fatigue. Wind and Wesolowski (2018) found that these particular DRIFT effects have influenced the variability of adjudicator results for solo music performances over time. Similarly, DRIFT effects may have plausibly played a role in the OkMEA auditions, which included similar consorts (piccolo/flute, clarinets, saxophones, euphonium/tuba, etc.), where OkMEA occasionally duplicated raters for more than one audition event. Furthermore, duplicated adjudicators who also served in groups with larger numbers of participants (*n*), such as flutes and Bb clarinets, could have further increased this impact. The results of the 2018 audition year revealed a similar anomaly in the Bb clarinet and bass clarinet auditions. The Bb

clarinet ICC value of .87, notably lower than in other years, alongside the bass clarinet auditions that same year (.68), indicated some inconsistency in rater agreement during that year. Once again, this presents an instance reflecting the possibility of adjudicator fatigue or inexperience affecting overall reliability.

The most perceptible anomaly occurred in the 2015 alto saxophone auditions, which showcased an ICC value of .61, considerably lower than in other years, suggesting substantial variability in rater agreement for that year. With the low participant count ($n$ = 24), it is possible that this anomaly was spawned by an outside factor such as differential/centrality extremism, as defined by Wolfe et al. (2001) as a rater who either only employs the outer extremes of a rating scale, or a rater employing only a central swath of the available point range. Supported by Wind and Wesolowski (2018), time-construct as a factor could provide evidence to raters possibly adjudicating multiple events before the alto saxophone auditions, which resulted in fatigue and ultimately, inconsistency. Or it could also show a lack of experience of one or more judges adjudicating a highly concentrated group of the best finalist players with less variability in their performances. This is supported by Fiske (1977) and Passarotto et al. (2023), who found that experience does provide greater accuracy in higher-level scoring.

Overall, it is essential to note that no patterns in any evaluated instrument group assessed by adjudicators using the BGAM indicated weak reliability. However, the identified anomalies highlight the need for regular analysis of results to inform practices such as rater training and calibration or potentially addressing unethical adjudicator behavior to ensure consistent and reliable assessments (Elliot, 1995a; Dugger, 1997). Analyzing the results annually and addressing these inconsistencies can enhance and secure the overall reliability of the audition process (Elliot, 1995a).

**Total BGAM Results Comparing Instruments by Individual Years**

The second question aimed to examine each year of BGAM use and analyze the results for consistency and patterns of all instrument groups from year to year. Similar to the previous analysis, slight variability in ICC values across different instrument groups and years revealed occasional anomalies; however, the results did not indicate any overarching negative patterns of inconsistency or low reliability on a year-to-year basis.

In 2015, most instrument groups exhibited high reliability, with Cronbach's alpha values above .90 and ICC values indicating strong rater agreement. A notable exception was the alto saxophone, which had an ICC of .61 with a 95% CI of [.231, .821], suggesting uncommon variability in rater agreement. The 2016 results show a general improvement in reliability for most instrument groups, with higher Cronbach's alpha values and ICCs. However, the flute auditions had an ICC of .70, and the contra clarinet had an ICC of .86, both lower than other instruments. The trumpet group in 2016 also showed a lower ICC of .77, suggesting a potential area of interjudge inconsistency for the year. In 2017, the reliability remained high for most instruments, with Cronbach's alpha values consistently above .90. However, the alto saxophone and trumpet showed lower ICC values of .77 and .75, respectively, which indicated some inconsistency in interjudge agreement. The bass clarinet and contra clarinet maintained high reliability with ICC values of .96. The 2018 results also showed high reliability for most instruments. Anomalies included the bass clarinet, which had an ICC of .68, and the trumpet, which had an ICC of .73. In 2019, several anomalies were observed. The piccolo group had an ICC of .79, and the flute had an ICC of .69, both lower than in other years, possibly caused by duplicating adjudicators. The euphonium also showed a lower ICC of .76. The 2019 smaller

sample sizes in euphonium and tuba struggled to define whether an event had less agreement between the judges or lacked variability in the performances, resulting in less reliable ICCs.

Despite these anomalies, most year groups maintained high reliability, with Cronbach's alpha values above .90 yearly. Similar to the individual instrument group data, the results grouped by year also lead to inconsistencies most likely caused by adjudicator inexperience (the need for more training, being provided a benchmark for scoring, or more experience understanding the particular instrument/etudes), adjudicator fatigue, or possibly other isolated DRIFT effects (Fiske, 1977; Passarotto et al., 2023; Wind & Wesolowski, 2018; Wolfe et al., 2001).

In conclusion, the reliability of the audition scores resulting from using the BGAM remained high overall from year to year. The anomalies in the ICC values, as previously reported for instrument groups, were only minor areas of isolated lower consistency and had minimal overall effect on each year. Despite these anomalies, all BGAM values stayed above the threshold of acceptable reliability values from year to year.

**Overall BGAM Results as a Whole from 2015-2019**

The overall reliability estimates for each audition year aligned with the findings from the individual instrument groups analyzed, and likewise, the consistently high annual Cronbach's alpha values mirrored the particular year groups. The overall ICC values observed in the individual year groups were consistently strong, with all but one group showing overall Cronbach's alpha averages above .91, thus reflecting the total average ICC of .89.

In conclusion, the BGAM has demonstrated consistent reliability in evaluating audition scores across the various instrument groups and years of the Oklahoma All-State Band and Orchestra Auditions (OASBOA). The overall Cronbach’s alpha average of .95 for the 2015–

2019 OASBOA met the preferred reliability threshold for applied use of $\alpha = .95$ (Lance et al., 2006; Nunnally & Berstein, 1994). While the overall reliability was high, the anomalies suggested areas for future discussion and improvements regarding adjudicator selection and training.

## Implications of BGAM Reliability

A fair audition system coupled with accurate results is paramount for the student participants, educators, and the governing organization. This study's findings have implications for the OkMEA and other state organizations concerned about reliable audition results.

### *Suggestions for OkMEA*

Although the results of this study demonstrated strong and consistent reliability in the BGAM audition scores, some areas of concern arose from the presence of anomalies. As noted in numerous reliability studies (Fiske,1975; Kinney, 2009; Passarotto et al., 2023), adjudicator experience and training can influence outcomes. While Fiske (1975) found that error detection may not necessarily correlate with experience level, Passarotto et al. (2023) indicated that experience plays a substantial role in making value judgments on the higher end of musical performances. In this particular instance, the OASBOA aimed to identify the outliers, the very best musicians, to represent the state's premier all-state ensembles. At the time of this study, the OkMEA had no official expectations regarding an adjudicator's ability to produce reliable scores, merely a good-faith assumption that an adjudicator's stated experience in music aptitude would suffice. Naturally, this encompassed all the intrinsic and extrinsic biases of each adjudicator. Although OkMEA effectively controls various entry points for biases (e.g., set music, set venue, double-anonymous auditions), it lacked an aural framework for adjudicators to reference the performance levels with its rating scales. Studies by Bergee (2007) and DCamp

(1980) have shown that levels of agreement could be boosted if the judges had a consensus on expected performance levels before the auditions. Consequently, without any training or aural frameworks, OkMEA's past and current procedural approach to rating performances is entirely centered on each adjudicator's personal experiences and biases about what constitutes an exceptional outlier performance. This nuance was considered when I developed the BGAM. However, the anomalies revealed in this study show that added attention to adjudicator training, selection, and experience could still boost the audition results to remarkable levels of consistency.

Secondly, unethical behavior among adjudicators unfortunately does exist. News from the world of sports over the last century has revealed illegal and unethical activities in value-judgment-based Olympic events such as gymnastics and figure skating (Frenger et al., 2019; Zitzewitz, 2012). If such activities have occurred in highly regulated and scrutinized events, one can only assume they have also happened in less protected events like music auditions. Dugger (1997) studied the effectiveness of the OkMEA's previous practice of Olympic-style scoring (i.e., dropping each participant's top and bottom scores) to guard against errant judging at the OASBOAs. This practice is not used anymore and is not included in the scope of this study. However, it does imply that cheating was one of OkMEA's concerns. The OkMEA occasionally inserts adjudicators who are some of the auditionees' teachers and/or parents. Even with a double-anonymous scenario, a level of trust is still needed to achieve fairness, and an adjudicator can bend the rules. Directors, parents, and employers often measure success through the achievements of their students and children under their influence. With this in mind, occasional anomalies in the analysis could provide evidence of possible unethical behavior amongst a judge or panel that should be investigated.

My suggestions for OkMEA resulting from this study and related literature are listed below:

- At the end of each year's auditions, calculate the Cronbach's alpha and intraclass coefficient values for each audition room and examine the reliability results for any inconsistencies or irregularities.
- Keep a formal record of each judge's experience levels and any previous anomalies within the adjudicator pool. Use this information to create balanced adjudicator panels in the future.
- Establish a training system where adjudicators listen to pre-recorded aural examples of etudes corresponding to the BGAM rating scale descriptors. For example, play three recordings of an etude—one of the highest quality, one of the average quality, and one of the lowest quality—before the auditions. This will require annual preparatory work for the organization, but it could empower a less experienced judging panel to apply the BGAM consistently in highly competitive scenarios.

**Limitations**

This study was limited to the final round of all-state auditions conducted by the Oklahoma Music Educators Association, which included its adjudicator pool and student participants from 2015 to 2019; therefore, its generalizability is limited. While it can be assumed that other states have similar audition processes, only the unique factors related to the OASBOA guidelines were used to develop the BGAM. Any other entity restructuring its audit measures should exercise caution when using the BGAM in this study. Factors such as audio recordings, open auditions, random music selections, multiple listenings, and so forth could

yield different results in reliability and consistency (Glesjer & Heyndels, 2001; Harrington, 2016; Shih, 2018).

The data used in this study were entirely de-identified; therefore, it was impossible to draw any correlations between the results and adjudicator experience, instrument choice, education, and so forth. Furthermore, nonmusical privilege factors concerning the auditioners, such as school size, distance traveled, and socioeconomic status (Barnes & McCashin, 2002; also see Bergee & Platt, 2003; Harrington, 2016; Perry, 2023), could not be utilized in the analysis. Overall, demographics outside the participant grade level (10–12) and instrument type were unavailable, limiting a deeper analysis of rater agreement.

This study focused solely on the final results of the weighted total scores from the adjudicators. The weighted totals reflected the sum of the sub-scores for each audition, calculated by applying percentages as detailed in the OkMEA contest manager's procedure manual. The analysis centered on the final scores used to rank the participants. No examination was conducted on the sub-scores or scores from the first and second rounds of auditions. Additionally, due to the complete anonymity of the results, it remained unclear whether adjudicators served on more than one panel, which could have resulted in adjudicator fatigue or other forms of adjudicator DRIFT. Furthermore, it was impossible to ascertain if any of the few anomalies presented in the analysis were merely isolated instances of user error.

This study was limited to a five-year period of auditions from 2015 to 2019. While the BGAM was implemented in 2014, I excluded that year to account for any potential one-time effects from introducing a new measurement tool. Furthermore, although the BGAM is still in use as of this study, I chose to conclude the analysis in 2019 for two reasons: (a) the years 2020 and 2021 included online auditions and other temporary hybrid procedures due to the global

COVID-19 pandemic, and (b) a consistent five-year period more accurately represented the research questions.

Finally, Portney and Watkins (2000) suggest that ICC groups with three or more raters should obtain at least 30 similar samples for reliable interpretation. Since some instrument groups fell below that threshold, caution is warranted when interpreting their lower ICC values.

**Suggestions for Further Research**

This study did not account for the intricacies of the sub-scores from the auditions, nor did it collect any relevant demographic data from adjudicators or participants. Collecting and incorporating demographic information into the analysis could provide a deeper understanding of adjudicator patterns and assist in comparing these effects between nonmusical factors, such as an adjudicator's experience, training, or fatigue (Bergee, 2007). I also suggest comparing this existing dataset with other studies on order effects. It is unclear whether leveraging the BGAM during this period impacted the potential for order effects, as this was outside the scope of this study. However, data indicating a positive or negative impact, or even a neutralization of order effects, would be valuable information for the organization's future auditions. Research similar to Wind and Wesolowski (2018) using the Rasch measurement approach could examine and help understand possible order effects and rater accuracy of the BGAM using time as a construct.

Considering the artistry-based, generic descriptors used in the BGAM rating scale, it can inherently be applied to any musical performance on any instrument, vocal, or ensemble evaluation that does not require formative feedback. In 2019, the OkMEA orchestra division questioned the results of its audition measure. It adopted the BGAM for its portion of the final round of OASBOA string auditions based on the general success of the audition measure and

its ease of use. At the time of this study, the audition data had not been analyzed to determine the reliability of the BGAM's application to string auditions. Further research should explore whether using the BGAM in contexts other than OASBOAs yields the same high reliability and levels of consistency. Additionally, comparing results with adjudicators who pair the BGAM with a facet-factorial based rubric assessment, replicating Saunders & Holahan (1997), could effectively address the scope of validity of the BGAM.

**Conclusion**

Selection into an all-state music ensemble requires trust that the results reliably and accurately represent the thousands of first, second, and final rounds of auditions determining the final selection. This study is unique in that it focused on a specific organization's exclusive audition scenario, utilized information gathered from existing research on music performance measurement, and created a bespoke general measurement tool that ultimately produced statistically high reliability and consistency. The strength of this study's results is supported by its access to a complete population of participant data in an unsimulated real-world application, combined with numerous events replicated over five years, and a professional organization's willingness to utilize existing research to find a solution. Through this project, I conclude that other state organizations engaged in discussions about questionable audition results should also adopt a similar, realistic approach to reviewing their processes and construct a reliable measurement tool to help secure the integrity of their audition results.

## References


Abeles, H. F. (1973a). A facet-factorial approach to the construction of rating scales to measure complex behaviors. *Journal of Educational Measurement*, *10*(2), 145–151. https://doi.org/10.1111/j.1745-3984.1973.tb00792.x

Abeles, H. F. (1973b). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, *21*(3), 246–255. https://doi.org/10.2307/3345094

Alford, E. (1985). A model for all-state percussion auditions. *Music Educators Journal*, *71*(9), 37–38. https://doi.org/10.2307/3396520

American Psychological Association. (n.d.). *APA dictionary of psychology*. Retrieved August 30, 2023, from https://dictionary.apa.org/order-effect

Barnes, G. V., & McCashin, R. (2002). All-state orchestras: A survey of practices and procedures. *Update: Applications of Research in Music Education*, *20*(2), 16–20. https://doi.org/10.1177/875512330202000205

Barry, N. (2009/2010). Evaluating music performance: Politics, pitfalls, and successful practices. *College Music Society*, *49/50*, 246–256. http://www.jstor.org/stable/41225250

Benkert, S. M. (1998). *The effect of training judges on their ability to quantify performances* (Publication No. 9931909) [Doctoral dissertation, The University of Kansas]. ProQuest Dissertations & Theses Global.

Bergee, M. J. (1987). *An investigation of the facet-factorial approach to scale construction in the development of a rating scale for euphonium and tuba music performance* (Publication No. 8813388) [Doctoral dissertation, The University of Kansas]. ProQuest Dissertations & Theses Global.

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. https://www.jstor.org/stable/40318730

Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education*, *51*(2), 137–150. https://doi.org/10.2307/3345847

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., & 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://www.jstor.org/stable/3345517

Bergee, M. J., & Platt, M. C. (2003). Influence of selected variables on solo and small-ensemble festival ratings. *Journal of Research in Music Education*, *51*(4), 342–353. https://doi.org/10.2307/3345660

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

Bergee, M. J., & Westfall, C. R. (2005). Stability of a model explaining selected extramusical influences on solo and small-ensemble festival ratings. *Journal of Research in Music Education*, *53*(4), 358–374. https://www.jstor.org/stable/3648433

Bermingham, G. A. (2000). Effects of performers' external characteristics on performance evaluations. *Applications of Research in Music Education*, *18*(2), 3–7. https://doi.org/10.1177/875512330001800202

Bobbitt, Z. (2020). *What are order effects?*. Statology. https://www.statology.org/order-effects/

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

Bruine de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, *118*(3), 245–60. https://doi.org/10.1016/j.actpsy.2004.08.005

Bruine de Bruin, W. (2006) Save the last dance II: unwanted serial position effects in figure skating judgments. *Acta Psychologica*, *123*(3), 299–311. https://doi.org/10.1016/j.actpsy.2006.01.009

Bruine de Bruin, W., & Keren, G. B. (2003). Order effects in sequentially judged options due to the direction of comparison. *Organizational Behavior and Human Decision Processes*, *92*(1–2), 91–101. https://doi.org/10.1016/S0749-5978(03)00080-3

Burdett, N. D. (1985). *The high school music contest movement in The United States* (Publication No. 8609289). [Doctoral dissertation, Boston University]. ProQuest Dissertations & Theses Global.

Butt, D. C., & Fiske, D. W. (1968). A comparison of strategies in developing scales for dominance. *Psychological Bulletin*, *70*(6), 505–519. https://doi.org/10.1037/h0026830

Byo, J. L., & Crone, L. J. (1989). Adjudication by non-musicians: A comparison of professional and amateur performances. *Missouri Journal of Research in Music Education*, *26*, 60–73.

Campbell, W. P. (1955). *A study of district and all-state music festivals in Pennsylvania with implications for the Pennsylvania Music Educators* (Publication No. 0014898) [Doctoral dissertation, Pennsylvania State University]. ProQuest Dissertations & Theses Global

Coimbra, D., Davidson, J., & Kokotsaki, D. (2001). Investigating the assessment of singers in a music college setting: The students' perspective. *Research Studies in Music Education*, *16*(1), 15–32. https://doi.org/10.1177/1321103X010160010301

Cooksey, J. M. (1977). A facet-factorial approach to rating high school choral music performance. *Journal of Research in Music Education*, *25*(2), 100–114. https://doi.org/10.2307/3345190

DCamp, C. B. (1980). *An application of the facet-factorial approach to scale construction in the development of a rating scale for high school band music performance* (Publication No. 8022012) [Doctoral dissertation, University of Iowa]. ProQuest Dissertations & Theses Global.

DeCarbo, N., Fiese, R., & Boyle, J. D. (1990). A profile of all-state instrumentalists. *Research Perspectives in Music Education*, *1*(1), 32–40.

Doerksen, P.F. (1999). Aural-diagnostic and prescriptive skills of preservice and expert instrumental music teachers. *Journal of Research in Music Education*, *47*(1), 78–88. https://doi.org/10.2307/3345830

Dugger, R. (1997). Inter-judge reliability for the 1994 Oklahoma all-state band auditions based on an Olympic-style judging system. *Journal of Band Research*, *32*(2), 66–75.

Elliott, C. A. (1995a). All-state band practices and procedures: A nation-wide survey. *Journal of Band Research*, *30*(2), 87–94.

Elliott, C. A. (1995b). Race and gender as factors in judgments of musical performance. *Bulletin of the Council for Research in Music Education*, *127*, 50–56. http://www.jstor.org/stable/40318766

Elliott, C. A., Schneider, M. C., & Zembower, C. M. (2000). Influence of the audition hour on selection to an all-state band. *Journal of Band Research*, *35*(2), 20.

Eysenck, H. J. (1939). The validity of judgments as a function of the number of judges. *Journal of Experimental Psychology*, *25*(6), 650–654. https://doi.org/10.1037/h0058754

Eysenck, H. J. (1972). Personal preferences, aesthetic sensitivity and personality in trained and untrained subjects. *Journal of Personality*, *40*(4), 544–557. https://doi.org/10.1111/j.1467-6494.1972.tb00079.x

Fiske, H. E., Jr. (1975). Judge-group differences in the rating of secondary trumpet performances. *Journal of Research in Music Education*, *23*(3), 186–196. https://doi.org/10.2307/3344643

Fiske, H. E., Jr. (1977). Relationship of selected factors in trumpet performance adjudication reliability. *Journal of Research in Music Education*, *25*(4), 256–263. https://doi.org/10.2307/3345266

Flôres, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth Musical Competition: How fair is the final ranking? Journal of the Royal Statistical Society. *Series D (The Statistician)*, *45*(1), 97–104. https://doi.org/10.2307/2348415

Frenger, M., Emrich, E., & Pitsch, W. (2019). Corruption in Olympic sports: Prevalence estimations of match fixing among German squad athletes. *SAGE Open*, *9*(3). https://doi.org/10.1177/2158244019865361

Geringer, J. M., Allen, M. L., MacLeod, R. B., & Scott, L. (2009). Using a prescreening rubric for all-state violin selection: Influences of performance and teaching experience. *Update: Application of Research in Music Education*, *28*(1), 41–46. https://doi.org/10.1177/8755123309344109

Geringer, J. M., & Madsen, C. K. (1995). Focus of attention to elements: Listening patterns of musicians and nonmusicians. *Bulletin of the Council for Research in Music Education*, *127*, 80–87. http://www.jstor.org/stable/40318770

Geringer, J., & Madsen, C. (1998). Musicians' ratings of good versus bad vocal and string performances. *Journal of Research in Music Education*, *46*(4), 522–534. https://doi.org/10.2307/3345348

Glesjer, H., & Heyndels, B. (2001). Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth Music Contest. *Journal of Cultural Economics*, *25*(2), 109–129. https://www.jstor.org/stable/41810752

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *The American Economic Review*, *90*(4), 715–741. https://doi.org/10.1257/aer.90.4.715

Groulx, T. J. (2013). The influence of tonal and atonal contexts on error detection accuracy. *Journal of Research in Music Education*, *61*(2), 233–243. https://doi.org/10.1177/0022429413485602

Haley, K. (1999). *Watkins-Farnum revisited: Application of the Rasch model to measures of musical performance* (Publication No. 9923403) [Doctoral dissertation, Boston University]. ProQuest Dissertations & Theses Global.

Harrington, A. M. (2016). *The effect of implied performer age, implied performer gender, and performance quality level on music majors' evaluations of solo musical performances* (Publication No. 10306929) [Doctoral dissertation, University of Kentucky]. ProQuest Dissertations & Theses Global.

Hedden, D. G., & Johnson, C. (2008). The effect of teaching experience on time and accuracy of assessing young singers' pitch accuracy. *Bulletin of the Council for Research in Music Education*, *178*, 63–72. https://doi.org/10.2307/40319339

Hewitt, M. P., & Smith, B. P. (2004). The influence of teaching-career level and primary performance instrument on the assessment of music performance. *Journal of Research in Music Education*, *52*(4), 314–327. https://www.jstor.org/stable/3345385

Hewitt, M. P. (2007). Influence of primary performance instrument and education level on music performance evaluation. *Journal of Research in Music Education*, *55*(1), 18–30. https://doi.org/10.1177/002242940705500103

Hickman, J. (2015). *Competition in music education: A mixed method study of the Iowa all-state chorus audition process* (Publication No. 3729164) [Doctoral dissertation, University of Florida]. ProQuest Dissertations & Theses Global.

Hollingshead, A. (1996). The rank-order effect in group decision making. *Organizational Behavior and Human Decision Processes*, *68*(3), 181–193. https://doi.org/10.1006/obhd.1996.0098

Hosmer, H. (1949). As the adjudicator hears it. *The Etude*, *67*, 224–262.

Howard, S. A. (2012). The effect of selected nonmusical factors on adjudicators' ratings of high school solo vocal performances. *Journal of Research in Music Education*, *60*(2), 166–185. https://doi.org/10.1177/0022429412444610

Isaacson, J., & Stacy, A. (2009). Rubrics for clinical evaluation: Objectifying the subjective experience. *Nurse Education in Practice*, *9*(2), 134–140. https://doi-org.ezproxy.lib.ou.edu/10.1016/j.nepr.2008.10.015

Johnson, P. (1997). Performance as experience: The problem of assessment criteria. *British Journal of Music Education*, *14*(3), 271–282. https://doi.org/10.1017/S0265051700001248

Kinney, D. W. (2009). Internal consistency of performance evaluations as a function of music expertise and excerpt familiarity. *The Journal of Research in Music Education*, *56*(4), 332–337. http://doi.org/10.1177/0022429408328934

Koo, T. K., & Li, M. Y. A (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155. https://doi.org/10.1016/j.jcm.2016.02.012

Krahé, C., Hahn, U., & Whitney, K. (2013). Is seeing (musical) believing? The eye versus the ear in emotional responses to music. *Psychology of Music*, *43*(1), 140–148. https://doi.org/10.1177/0305735613498920

Kroger, C., & Margulis, E. H. (2017). "But they told me it was professional": Extrinsic factors in the evaluation of musical performance. *Psychology of Music*, *45*(1), 49–64. https://doi.org/10.1177/0305735616642543

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Landy, F., & Farr, J. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72–107. https://doi.org/10.1037/0033-2909.87.1.72

Latimer, 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*, *58*(2), 168–183. https://doi.org/10.1177/0022429410369836

Levinowitz, A. (1985). *A criterion-related validity study of three sets of rating scales used for measuring and evaluating the instrumental achievement of first and second year clarinet students (facet-factorial, Individual-differences, Continuous, Edwin E. Gordon, Additive)* (Publication No. 8521102) [Doctoral dissertation, Temple University]. ProQuest Dissertations & Theses Global.

Lien, J. L., & Humphreys, J. T. (2001). Relationships among selected variables in the South Dakota All-State Band auditions. *Journal of Research in Music Education*, *49*(2), 146–155. https://doi.org/10.2307/3345866

McDow, G. H. (1989). *A history of instrumental music in the public schools of Oklahoma through 1945* (Publication No. 9003672) [Doctoral dissertation, University of Oklahoma]. ProQuest Dissertations & Theses Global.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McPherson, G. E. (1995). The assessment of musical performance: Development and validation of five new measures. *Psychology of Music*, *23*(2), 142–161. https://doi.org/10.1177/0305735695232003

McPherson, G. E., & Thompson, W. F. (1998). Assessing music performance: Issues and influences. *Research Studies in Music Education*, *10*(1), 12–24. https://doi.org/10.1177/1321103X9801000102

Mills, J. (1991). Assessing musical performance musically. *Educational Studies*, *17*(2), 173–181. https://doi.org/10.1080/0305569910170206

Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2013). *IBM SPSS for introductory statistics: Use and interpretation (5th ed.)*. Routledge.

Music for All. (n.d.). Honor Band of America. https://festival.musicforall.org/honor-band/

Napoles, J. (2009). The effects of score use on musicians' ratings of choral performances. Journal of Research in Music Education, 57(3), 267–279. https://doi.org/10.1177/0022429409343423

Neil, R. J. (1944). *The development of the competition-festival in music education* [Unpublished Doctoral dissertation]. George Peabody College for Teachers.

Newman, M. (2008). Judging the judges. *Teaching Music*, *16*(1), 26.

Nicholson, P., Gillis, S., & Dunning, A. (2009). The use of scoring rubrics to determine clinical performance in the operating suite. *Nurse Education Today*, *29*(1), 73–82. https://doi.org/10.1016/j.nedt.2008.06.011

Norris, C. E., & Borst, J. D. (2007). An examination of the reliabilities of two choral festival adjudication forms. *Journal of Research in Music Education*, *55*(3), 237–251. https://doi.org/10.1177/002242940705500305

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric Theory. *Academic Internet Publishers*.

Oldham, H. G. (1966). *The organization and administration of all-state bands and orchestras* (Publication Order No. 6607003) [Doctoral dissertation, University of Nebraska-Lincoln]. ProQuest Dissertations & Theses Global.

Oklahoma Music Educators Association: Band Division. (2011, January 22). *Annual Band Director Division Meeting*. Unpublished manuscript.

Owen, C. D. (1969). *A study of criteria for the evaluation of secondary school instrumentalists when auditioning for festival bands*. [Unpublished doctoral dissertation]. East Texas State University.

Passarotto, E., Altenmüller, E., & Müllensiefen, D. (2023). Music performance assessment: Noise in judgments and reliability of measurements. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000574

Perry, W. E. (2023). *A demographic profile of the schools and communities represented in the Alabama High School All-State Band* (1986—2019) (Publication No. 30567634) [Doctoral dissertation, The University of Alabama]. ProQuest Dissertations & Theses Global.

Peynircioğlu, Z., Bi, W., & Brent, W. (2018). The "Asian bias" illusion in musical performance: Influence of visual information. *The American Journal of Psychology*, *131*(3), 295–305. https://doi.org/10.5406/amerjpsyc.131.3.0295

Pitt, M. A. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 976–986. http://dx.doi.org/10.1037/0096-1523.20.5.976

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport & Exercise Psychology*, *21*(2), 131–144. https://doi.org/10.1123/jsep.21.2.131

Pope, D., & Barnes, G. (2015). Influence of education, performance level, and presentation medium on evaluations of string orchestra performances. *Bulletin of the Council for Research in Music Education*, (203), 7–22. https://doi.org/10.5406/bulcouresmusedu.203.0007

Portney, L.G., & Watkins, M.P. (2000). *Foundations of clinical research: Applications of practice*. Prentice Hall.

Radocy, R. E. (1976). Effects of authority figure biases on changing judgments of musical events. Journal of Research in Music Education, *24*(3), 119–128. https://doi.org/10.2307/3345155

Radocy, R. E. (1986). On quantifying the uncountable in musical behavior. *Bulletin of the Council for Research in Music Education*, *88*, 22–31. http://www.jstor.org/stable/40317998

Raymond, M. R. (2009). [Review of Assessing Performance: *Designing, Scoring, and Validating Performance Tasks*, by R. L. Johnson, J. A. Penny, & B. Gordon]. *Journal of Educational Measurement*, *46*(4), 474–477. http://www.jstor.org/stable/25651529

Russell, B. E. (2010). The development of a guitar performance scale using a facet factorial approach. *Bulletin of the Council for Research in Music Education*, *184*, 21–34. https://doi.org/10.2307/27861480

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

Sheldon, D. A. (2004). Effects of multiple listenings on error-detection acuity in multivoice, multitimbral musical examples. *Journal of Research in Music Education*, *52*(2), 102–115. https://doi.org/10.2307/3345433

Shih, Y. (2018). *Evaluation of music performance: Computerized assessment versus human judges* (Publication No. 10992875) [University of Hawai'i at Manoa]. Available from ProQuest Dissertations & Theses Global.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Silvey, B. A. (2009). The effects of band labels on evaluators' judgments of musical performance. *Applications of Research in Music Education*, *28*(1), 47–52. https://doi.org/10.1177/8755123309344111

Sloboda, J. A. (1991). Music structure and emotional response: Some empirical findings. *Psychology of Music*, *19*(2), 110–120. https://doi.org/10.1177/0305735691192002

Smith, B. R. (2004). Five judge's evaluation of audiotaped string performance in international competition. *Bulletin of the Council for Research in Music Education*, *160*, 61–69. https://doi.org/10.1177/002242940705500103

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. https://doi.org/10.1177/002242940705500307

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

Stanley, M., & Brooker, R., Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Research Studies in Music Education*, *18*(1), 46–56. https://doi.org/10.1177/1321103X020180010601

Stegman, S. (2009). Michigan state adjudicated choral festivals: Revising the adjudication process. *Music Educators Journal*, *95*(4), 62–65. https://doi.org/10.1177/0027432109335852

Tam, C. O. (2018). Evaluating students' performance in responding to art: The development and validation of an art criticism assessment rubric. *International Journal of Art & Design Education*, *37*(3), 519–529. https://doi.org/10.1111/jade.12154

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thompson, S. (2006). Audience responses to a live orchestral concert. *Musicae Scientiae*, *10*(2), 215–244. https://doi.org/10.1177/102986490601000204

Thompson, S., Valentine, E., & Williamon, A. (2007). Time-dependent characteristics of performance evaluation. *Music Perception: An Interdisciplinary Journal*, *25*(1), 12–29. https://doi.org/10.1525/mp.2007.25.1.13

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception: An Interdisciplinary Journal*, *21*(1), 21–41. https://doi.org/10.1525/mp.2003.21.1.21

Tsay, C. (2014). The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes*, *124*(1), 24–33. https://doi.org/10.1016/j.obhdp.2013.10.003

Vasil, T. (1973). *The effects of systematically varying selected factors on music performance adjudication* (Publication No. 7400020) [Doctoral dissertation, University of Connecticut]. ProQuest Dissertations & Theses Global.

Waddell, G., & Williamon, A. (2017). Eye of the beholder: Stage entrance behavior and facial expression affect continuous quality ratings in music performance. *Frontiers in Psychology*, *8*, 513. https://doi.org/10.3389/fpsyg.2017.00513

Wallace, K. (2020). *The role of gesture in perceptions of expressivity and technique in solo vocal performance* (Publication No. 27995396) [Doctoral dissertation, Florida State University]. ProQuest Dissertations & Theses Global.

Wapnick, J., Darrow, A.-A., Kovacs, J., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education*, *45*(3), 470–479. https://doi.org/10.2307/3345540

Wapnick, J., & Darrow, A.-A. (2013). Sectional versus intact evaluation of four versions of a Chopin etude. *Journal of Research in Music Education*, *60*(4), 462–474. https://doi.org/10.1177/0022429412465318

Wapnick, J., Mazza, J. K., & Darrow, A.-A. (1998). Effects of performer attractiveness, stage behavior, and dress on violin performance evaluation. *Journal of Research in Music Education*, *46*(4), 510–521. https://doi.org/10.2307/3345347

Wapnick, J., Ryan, C., Campbell, L., Deek, P., Lemire, R., & Darrow, A.-A. (2005). Effects of excerpt tempo and duration on musicians' ratings of high-level piano performances. *Journal of Research in Music Education*, *53*(2), 162–176. https://doi.org/10.1177/002242940505300206

Welch, G.F. (1994). The assessment of singing. *The Psychology of Music*, *22*, 13–19.

Welker, L. G. (1997). *All-state bands in the United States* (Publication No. 9735765) [Doctoral dissertation, University of Alabama]. ProQuest Dissertations & Theses Global.

Wesolowski, B. C. (2012). Understanding and developing rubrics for music performance assessment. *Music Educators Journal*, *98*(3), 36–42. https://doi.org/10.1177/0027432111432524

Wesolowski, B. C. (2017). A facet-factorial approach towards the development and validation of a jazz rhythm section performance rating scale. *International Journal of Music Education*, *35*(1), 17–30. https://doi.org/10.1177/0255761415590524

Williams, L., & Chapman, D. (2020). Wind band adjudication forms in the United States: Values for listening. *Journal of Band Research*, *56*(1), 17–29.

Wind, S. A., & Wesolowski, B. C. (2018). Evaluating differential rater accuracy over time in solo music performance assessment. *Bulletin of the Council for Research in Music Education*, *215*, 33–55. https://doi.org/10.5406/bulcouresmusedu.215.0033

Wine, T. (1996). All-State Choruses: A Survey of Practices, Procedures, and Perceptions. *The Choral Journal*, *36*(8), 21–27.

Winter, N. (1993). Music performance assessment: A study of the effects of training and experience on the criteria used by music examiners. *International Journal of Music Education*, *os–22*(1), 34–39. https://doi.org/10.1177/025576149302200106

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurument*, *2*, 256–280.

Wrigley, W. J. (2005). *Improving music performance assessment* [Unpublished doctoral dissertation]. Griffith University. ProQuest Dissertations & Theses Global.

Wrigley, W. J., & Emerson, S. B. (2013). Ecological development and validation of a music performance rating scale for five instrument families. *Psychology of Music*, *41*(1), 97–118. https://doi.org/10.1177/0305735611418552

Zdzinski, S. F. (1991). Measurement of solo instrumental music performance: A review of literature. *Bulletin of the Council for Research in Music Education*, *109*, 47–58. www.jstor.org/stable/40318448

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. https://doi.org/10.2307/3345801

Zhao, X., Feng, G. C., Ao, S. H., & Lui, P. L. (2022). Interrater reliability estimators tested against true interrater reliabilities. *BMC Medical Research Methodology*, *22*(1), 1–19. https://doi.org/10.1186/s12874-022-01707-5

Zitzewitz, E. (2012). Does transparency reduce favoritism and corruption? Evidence from the reform of figure skating judging. *Journal of Sports Economics*, *15*(1), 3–30. https://doi.org/10.1177/1527002512441479

# Appendix A: Institution Review Board Documentation

The UNIVERSITY of OKLAHOMA

**Institutional Review Board for the Protection of Human Subjects**

**Approval of Initial Submission – Exempt from IRB Review – AP01**

**Date:** October 26, 2012 **IRB#:** 1438

**Principal Investigator:** Christopher Scott Barber **Approval Date:** 10/25/2012

**Exempt Category:** 2 - Observation public behavior

**Study Title:** The Creation and Validation of a Multi-Dimensional Rating Scale for High School Instrumental Honor Band Auditions

On behalf of the Institutional Review Board (IRB), I have reviewed the above-referenced research study and determined that it meets the criteria for exemption from IRB review. To view the documents approved for this submission, open this study from the *My Studies* option, go to *Submission History*, go to *Completed Submissions* tab and then click the *Details* icon.

As principal investigator of this research study, you are responsible to:

- Conduct the research study in a manner consistent with the requirements of the IRB and federal regulations 45 CFR 46.
- Request approval from the IRB prior to implementing any/all modifications as changes could affect the exempt status determination.
- Maintain accurate and complete study records for evaluation by the HRPP Quality Improvement Program and, if applicable, inspection by regulatory agencies and/or the study sponsor.
- Notify the IRB at the completion of the project.

If you have questions about this notification or using iRIS, contact the IRB @ 405-325-8110 or irb@ou.edu.

Cordially,

E. Laurette Taylor, Ph.D.

Chair, Institutional Review Board

The UNIVERSITY of OKLAHOMA

**Institutional Review Board for the Protection of Human Subjects**

**Human Research Determination Review Outcome**

**Date:** February 02, 2024

**Principal Investigator:** Christopher Scott Barber

**Study Title:** Longitudinal Analysis of the Oklahoma All-State Band Audition Evaluation Measure

**Review Date: 2/2/2024**

I have reviewed your submission of the Initial Submission Packet for the above-referenced study. I have determined this research does not meet the criteria for human subject's research. The proposed activity involves de-identified data from the Oklahoma All-State Band and Orchestra final round auditions for analysis. No geographic or demographic data included. No interaction with individuals. Therefore, IRB approval is not necessary so you may proceed with your project.

If you have questions about this notification or using iRIS, contact the HRPP office at (405) 325-8110 or irb@ou.edu. Thank you.

Cordially,

Lara Mayeux

Lara Mayeux, Ph.D.
Chair, Institutional Review Board

## Appendix B: OkMEA Room Monitor Procedures

# OMEA Inside Room Monitor Procedures

**Prior to Auditions:**

- Be sure the room is set up so judges and participants remain anonymous.
- Be sure to know the order of performance.
- Keep the music provided by OMEA on hand in case a participant needs a copy.
- After 3 or 4 auditions, collect all of the no-show audition cards from the outside monitor.

**The Music Etude Audition:**

1) Welcome the next participant into the room and take his/her audition card
2) Tell the participant the order of performance
3) Introduce participant by the **audition number only**
    a. **Inform the judges if participant is auditioning out of numeric order**
4) Tell the participant that he/she may begin the audition
5) During the music etude audition:
    a. arrange the previous participant's completed ballots by judge number
    b. staple the audition card to the top of the arranged ballot pack
    **c. check to make sure scores are legible and in each column**
    d. **mistakes** - ask the judge to check his/her tally sheet for any corrections
    e. stack audition ballot pack by audition number onto a single finalized pile
6) **Treat no-shows as auditions** by stapling corresponding audition cards and judging ballots together and place on finalized stack by audition order

**Sight Reading Audition:**

7) As soon as the music etude portion of the audition is complete, place the sight-reading music on the participant's music stand
8) Inform participant that or he/she has 30 seconds to review the music w/out playing
9) Clock 30 seconds. Use a timer or a watch with a second hand for accuracy.
10) After 30 seconds elapse, clearly say, "You may begin."
11) After the audition, collect the sight-reading music and open the door (if needed).
12) Gather the completed participant's ballots from the judges during transition.
13) Begin process again until all auditions are completed
14) Turn in "finalized ballot pile" only to official office runners.

**End of Auditions**

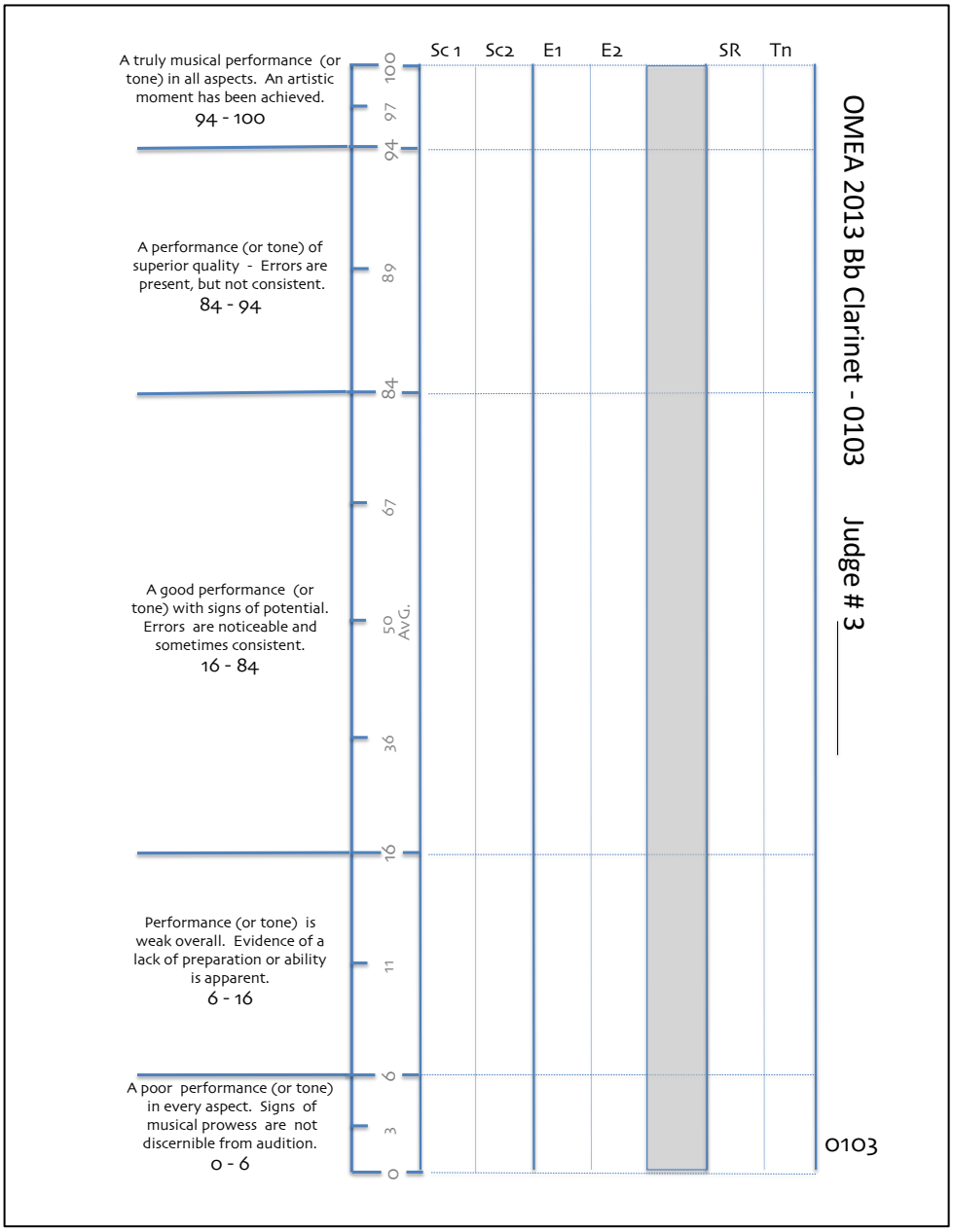
15) **Collect tally-sheets from each judge**. Make sure each one is signed and judging number is correct. Do not let them leave until this is completed.
16) Collect envelopes, music, and any leftover supplies.
17) Return the tally sheets and any remain completed ballots promptly to the contest office.

## Appendix C: Adjudicator Tally Sheet

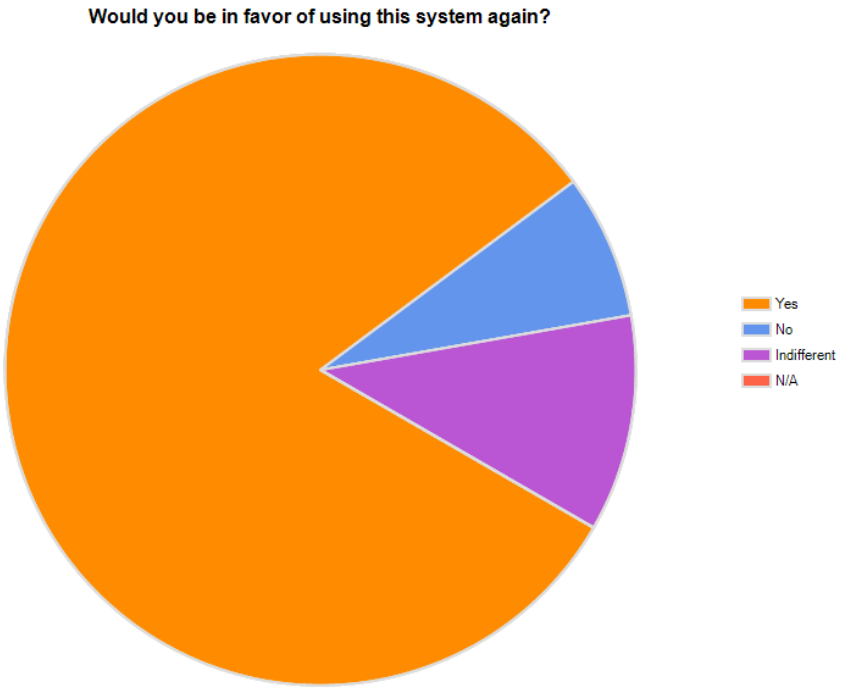
Judge Name_____ Judge Number_____ Event Judged_____

| Audition # | Column1 | Column2 | Column3 | Column4 | Column5 | Column6 | Column7 |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

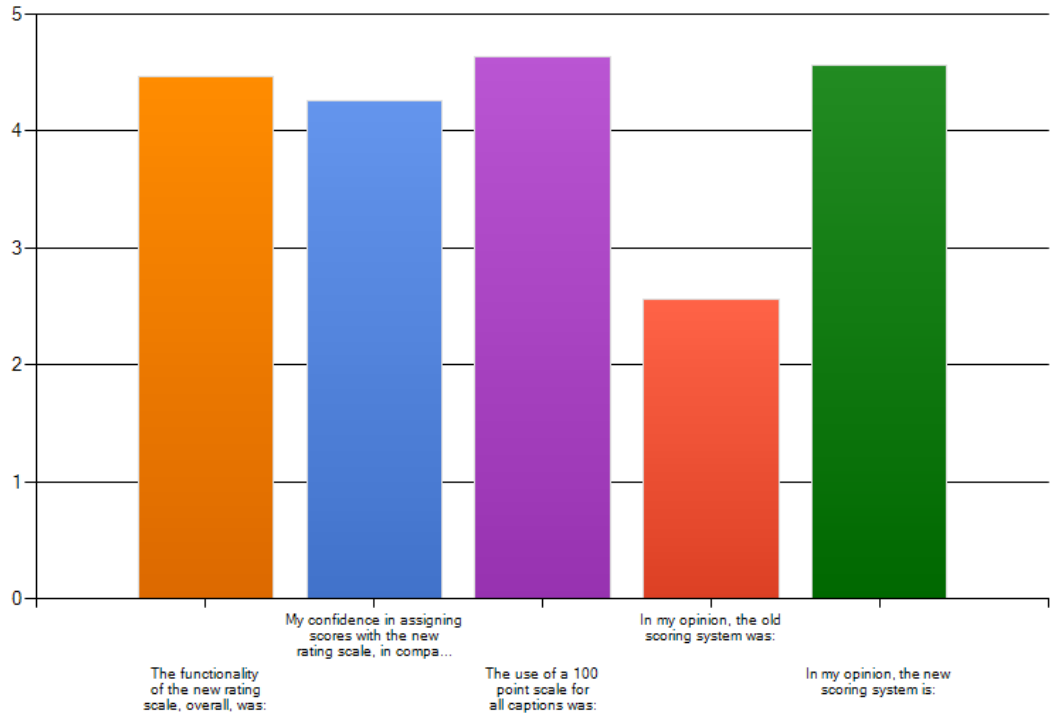
# Appendix D: BGAMv1

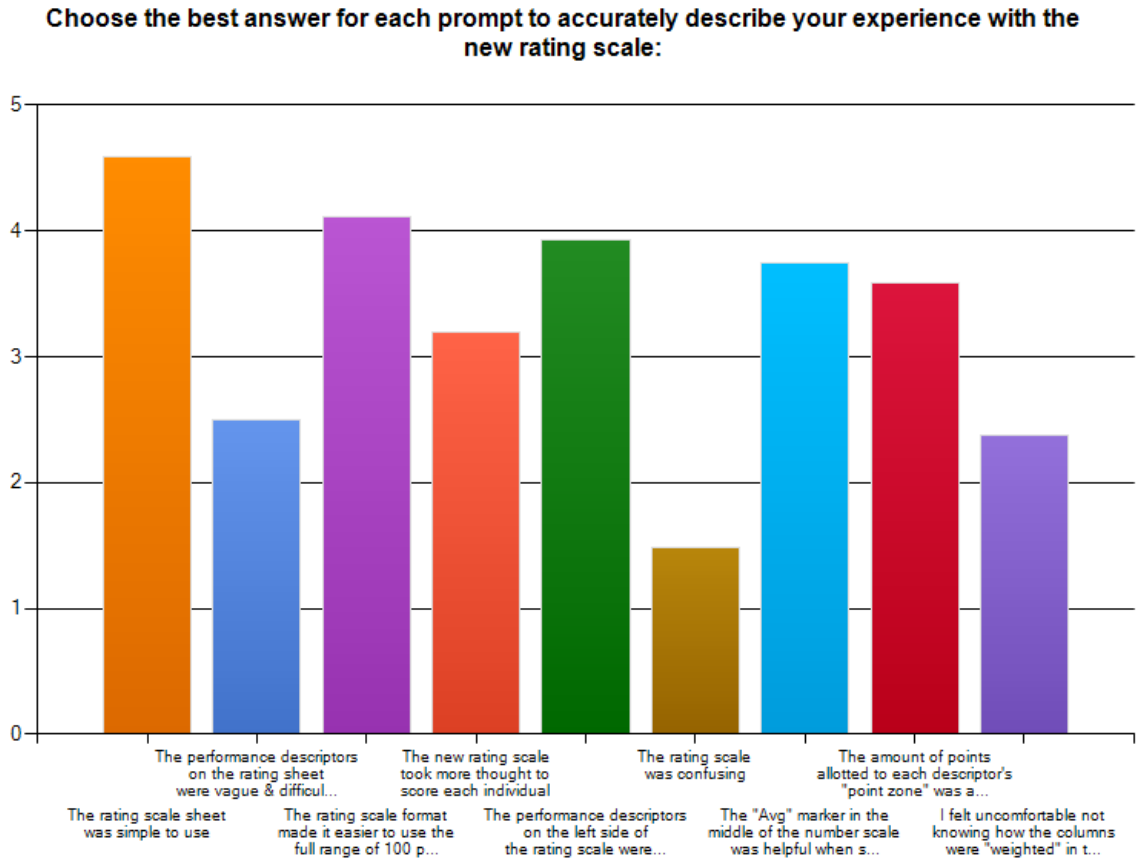


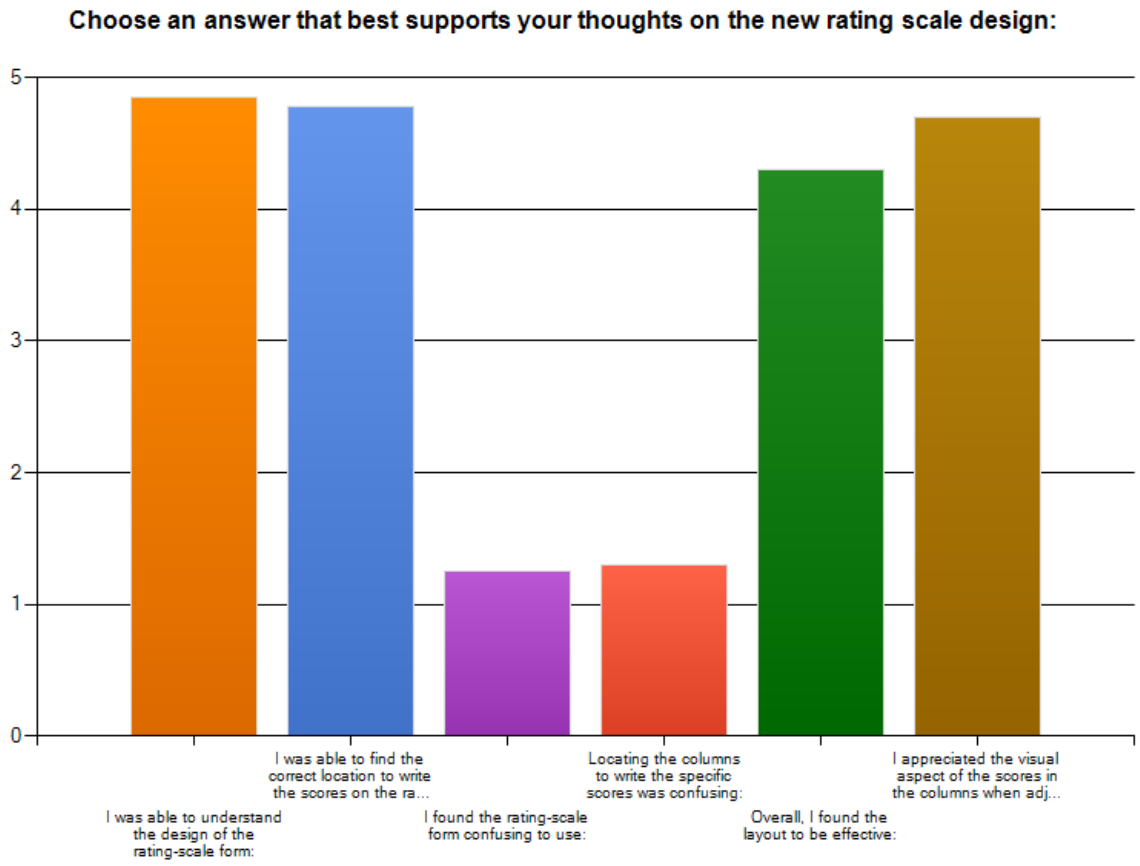
# Appendix E: Pilot Study Survey



**Choose an answer to best describe your opinion when comparing this rating scale to the old whole-number rating scale used by NCDA. If you have not used the old model, select N/A.**



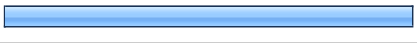

Choose the best answer for each prompt to accurately describe your experience with the new rating scale:
5
4
3
2
1
0
The rating scale sheet was simple to use
The performance descriptors on the rating sheet were vague & difficul...
The rating scale format made it easier to use the full range of 100 p...
The new rating scale took more thought to score each individual
The performance descriptors on the left side of the rating scale were...
The rating scale was confusing
The "Avg" marker in the middle of the number scale was helpful when s...
The amount of points allotted to each descriptor's "point zone" was a...
I felt uncomfortable not knowing how the columns were "weighted" in t...


Choose an answer that best supports your thoughts on the new rating scale design:
5
4
3
2
1
0
I was able to understand the design of the rating-scale form:
I was able to find the correct location to write the scores on the ra...
I found the rating-scale form confusing to use:
Locating the columns to write the specific scores was confusing:
Overall, I found the layout to be effective:
I appreciated the visual aspect of the scores in the columns when adj...

## NCDA Rating Scale Survey

SurveyMonkey

### Would you be in favor of using this system again?

| | | Response Percent | Response Count |
|---|---|---|---|
| **Yes** |  | **81.5%** | **22** |
| No |  | 7.4% | 2 |
| Indifferent |  | 11.1% | 3 |
| N/A | | 0.0% | 0 |
| | | **answered question** | **27** |
| | | **skipped question** | **0** |

## NCDA Rating Scale Survey

SurveyMonkey

**Choose an answer that best represents how you felt about not having access to the true total score of each individual:**

| | | Response Percent | Response Count |
|---|---|---|---|
| Positively affected |  | 7.4% | 2 |
| Negatively negatively |  | 3.7% | 1 |
| **Indifferent - not affected at all** |  | **44.4%** | **12** |
| Negatively affected at first, but it subsided as the auditions progressed |  | 11.1% | 3 |
| Positively at first, but it subsided as the auditions progressed | | 0.0% | 0 |
| I felt free to judge without having to manage numbers |  | 33.3% | 9 |
| | | Other (please specify) | 4 |
| | | **answered question** | **27** |
| | | **skipped question** | **0** |

**If you judged more than one age level (i.e, Jr. High & High School), do you feel the rating scale descriptors transferred to all age levels?**

| | Response Percent | Response Count |
|---|---|---|
| **Yes** | **59.3%** | **16** |
| No | 3.7% | 1 |
| N/A | 37.0% | 10 |
| **answered question** | | **27** |
| **skipped question** | | **0** |

**If you judged more than one type of wind instrument (i.e tenor sax & bassoon), do you feel the rating scale descriptors transferred to all of the instruments?**

| | | Response Percent | Response Count |
|---|---|---|---|
| **yes** |  | **55.6%** | **15** |
| no | | 0.0% | 0 |
| N/A |  | 44.4% | 12 |
| | | **answered question** | **27** |
| | | **skipped question** | **0** |

## NCDA Rating Scale Survey

SurveyMonkey

**Choose an answer to best describe your opinion when comparing this rating scale to the old whole-number rating scale used by NCDA. If you have not used the old model, select N/A.**

| | Much worse | Somewhat worse | About the same | Somewhat better | Much better | N/A | Rating Average | Response Count |
|---|---|---|---|---|---|---|---|---|
| The functionality of the new rating scale, overall, was: | 3.8% (1) | 3.8% (1) | 11.5% (3) | 15.4% (4) | **53.8% (14)** | 11.5% (3) | 4.46 | 26 |
| My confidence in assigning scores with the new rating scale, in comparison to the old system, was: | 3.7% (1) | 3.7% (1) | 22.2% (6) | 14.8% (4) | **44.4% (12)** | 11.1% (3) | 4.26 | 27 |
| The use of a 100 point scale for all captions was: | 3.7% (1) | 0.0% (0) | 11.1% (3) | 14.8% (4) | **55.6% (15)** | 14.8% (4) | 4.63 | 27 |
| In my opinion, the old scoring system was: | 18.5% (5) | **55.6% (15)** | 3.7% (1) | 7.4% (2) | 3.7% (1) | 11.1% (3) | 2.56 | 27 |
| In my opinion, the new scoring system is: | 0.0% (0) | 7.4% (2) | 7.4% (2) | 18.5% (5) | **55.6% (15)** | 11.1% (3) | 4.56 | 27 |
| | | | | | | **answered question** | | **27** |
| | | | | | | **skipped question** | | **0** |

## NCDA Rating Scale Survey

SurveyMonkey

**Choose the best answer for each prompt to accurately describe your experience with the new rating scale:**

| | Strongly disagree | Moderately disagree | Neither agree or disagree | Moderately agree | Strongly agree | Rating Average | Response Count |
|---|---|---|---|---|---|---|---|
| The rating scale sheet was simple to use | 0.0% (0) | 7.4% (2) | 3.7% (1) | 11.1% (3) | **77.8% (21)** | 4.59 | 27 |
| The performance descriptors on the rating sheet were vague & difficult to apply | **30.8% (8)** | 26.9% (7) | 19.2% (5) | 7.7% (2) | 15.4% (4) | 2.50 | 26 |
| The rating scale format made it easier to use the full range of 100 points | 7.4% (2) | 3.7% (1) | 11.1% (3) | 25.9% (7) | **51.9% (14)** | 4.11 | 27 |
| The new rating scale took more thought to score each individual | 11.1% (3) | 11.1% (3) | **37.0% (10)** | 29.6% (8) | 11.1% (3) | 3.19 | 27 |
| The performance descriptors on the left side of the rating scale were effective guideposts for scoring | 7.4% (2) | 3.7% (1) | 14.8% (4) | **37.0% (10)** | **37.0% (10)** | 3.93 | 27 |
| The rating scale was confusing | **70.4% (19)** | 18.5% (5) | 7.4% (2) | 0.0% (0) | 3.7% (1) | 1.48 | 27 |
| The "Avg" marker in the middle of the number scale was helpful when scoring | 11.1% (3) | 11.1% (3) | 7.4% (2) | 33.3% (9) | **37.0% (10)** | 3.74 | 27 |
| The amount of points allotted to each descriptor's "point zone" was adequate | 11.1% (3) | 11.1% (3) | 14.8% (4) | **33.3% (9)** | 29.6% (8) | 3.59 | 27 |
| I felt uncomfortable not knowing how the columns were "weighted" in the total score. | **44.4% (12)** | 11.1% (3) | 14.8% (4) | 22.2% (6) | 7.4% (2) | 2.37 | 27 |
| | | | | | **answered question** | | **27** |
| | | | | | **skipped question** | | **0** |

## NCDA Rating Scale Survey

SurveyMonkey

**Choose an answer that best supports your thoughts on the new rating scale design:**

| | None of the time | Some of the time | Neutral | Most of the time | All of the time | Rating Average | Response Count |
|---|---|---|---|---|---|---|---|
| I was able to understand the design of the rating-scale form: | 0.0% (0) | 0.0% (0) | 0.0% (0) | 14.8% (4) | **85.2% (23)** | 4.85 | 27 |
| I was able to find the correct location to write the scores on the rating-scale form: | 0.0% (0) | 0.0% (0) | 3.7% (1) | 14.8% (4) | **81.5% (22)** | 4.78 | 27 |
| I found the rating-scale form confusing to use: | **81.5% (22)** | 14.8% (4) | 0.0% (0) | 3.7% (1) | 0.0% (0) | 1.26 | 27 |
| Locating the columns to write the specific scores was confusing: | **77.8% (21)** | 14.8% (4) | 7.4% (2) | 0.0% (0) | 0.0% (0) | 1.30 | 27 |
| Overall, I found the layout to be effective: | 0.0% (0) | 7.4% (2) | 11.1% (3) | 25.9% (7) | **55.6% (15)** | 4.30 | 27 |
| I appreciated the visual aspect of the scores in the columns when adjudicating: | 0.0% (0) | 0.0% (0) | 7.4% (2) | 14.8% (4) | **77.8% (21)** | 4.70 | 27 |
| | | | | | **answered question** | | **27** |
| | | | | | **skipped question** | | **0** |

## Appendix F: Complete Data Chart

**Reliability (Cronbach's alpha) and Intra-class Coefficients (ICC)**

| | | 2015-16 | | | | 2016-17 | | | | 2017-18 | | | | 2018-19 | | | | 2019-20 | | | | 5-Year Total Av | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Code | Instrument* | n | Cron | ICC | 95%CI | n | Cron | ICC | 95%CI | n | Cron | ICC | 95%CI | n | Cron | ICC | 95%CI | n | Cron | ICC | 95%CI | Cron Av | ICC Av |
| 101 | Piccolo | 17 | 0.97 | 0.93 | .925-.780 | 19 | 0.97 | 0.93 | .788-.972 | 27 | 0.97 | 0.95 | .899-.977 | 30 | 0.95 | 0.93 | .849-.971 | 21 | 0.95 | 0.79 | .415-.923 | 0.962 | 0.91 |
| 102 | Flute | 42 | 0.87 | 0.80 | .624.-891 | 42 | 0.94 | 0.70 | .272-.865 | 42 | 0.93 | 0.90 | .801-.946 | 42 | 0.93 | 0.92 | .876-.955 | 42 | 0.89 | 0.69 | .304-.851 | 0.912 | 0.80 |
| 103 | Oboe | 27 | 0.96 | 0.96 | .803-.967 | 24 | 0.97 | 0.96 | .920-.982 | 26 | 0.98 | 0.96 | .899-.984 | 27 | 0.98 | 0.98 | .963-.989 | 26 | 0.97 | 0.95 | .894-.980 | 0.972 | 0.96 |
| 104 | Bassoon | 31 | 0.93 | 0.87 | .694-.942 | 23 | 0.96 | 0.94 | .873-.974 | 35 | 0.96 | 0.93 | .849-.965 | 36 | 0.97 | 0.88 | .628-.952 | 33 | 0.93 | 0.85 | .656-.932 | 0.95 | 0.89 |
| 105 | Eb clarinet | 6 | 0.98 | 0.97 | .887-.995 | 8 | 0.94 | 0.94 | .833-.986 | 8 | 0.88 | 0.87 | .653-.970 | 8 | 0.97 | 0.98 | .932-.995 | 11 | 0.92 | 0.83 | .519-.949 | 0.938 | 0.92 |
| 106 | Bb clarinet | 42 | 0.94 | 0.89 | .767-.946 | 48 | 0.96 | 0.94 | .876-.966 | 52 | 0.97 | 0.97 | .949-.980 | 52 | 0.95 | 0.87 | .675-.940 | 52 | 0.94 | 0.91 | .832-.953 | 0.952 | 0.92 |
| 108 | Bass clarinet | 40 | 0.96 | 0.96 | .935-.976 | 54 | 0.94 | 0.93 | .891-.956 | 36 | 0.96 | 0.96 | .930-.976 | 42 | 0.91 | 0.68 | .272-.853 | 42 | 0.96 | 0.94 | .876-.969 | 0.946 | 0.89 |
| 109 | Contra clarinet | 19 | 0.97 | 0.96 | .925-.985 | 21 | 0.95 | 0.86 | .603-.948 | 13 | 0.96 | 0.96 | .910-.986 | 14 | 0.93 | 0.92 | .829-.971 | 17 | 0.98 | 0.96 | .901-.987 | 0.958 | 0.93 |
| 110 | Alto sax | 24 | 0.83 | 0.61 | .231-.821 | 24 | 0.94 | 0.93 | .861-.965 | 24 | 0.85 | 0.77 | .547-.895 | 24 | 0.93 | 0.93 | .869-.965 | 24 | 0.92 | 0.91 | .842-.958 | 0.894 | 0.83 |
| 111 | Tenor sax | 33 | 0.97 | 0.95 | .895-.977 | 28 | 0.98 | 0.98 | .949-.988 | 27 | 0.96 | 0.96 | .926-.978 | 36 | 0.97 | 0.92 | .778-.965 | 38 | 0.98 | 0.97 | .937-.983 | 0.972 | 0.96 |
| 112 | Bari sax | 23 | 0.97 | 0.95 | .878-.979 | 27 | 0.97 | 0.94 | .834-.976 | 28 | 0.96 | 0.95 | .900-.973 | 17 | 0.98 | 0.98 | .944-.990 | 29 | 0.99 | 0.98 | .968-.992 | 0.974 | 0.96 |
| 113 | Trumpet | 36 | 0.95 | 0.92 | .843-.960 | 36 | 0.91 | 0.77 | .450-.898 | 36 | 0.89 | 0.75 | .437-.881 | 36 | 0.89 | 0.73 | .388-.875 | 34 | 0.92 | 0.82 | .576-.918 | 0.912 | 0.80 |
| 114 | Horn | 32 | 0.94 | 0.91 | .806-.955 | 32 | 0.93 | 0.81 | .511-.918 | 36 | 0.94 | 0.88 | .723-.945 | 36 | 0.95 | 0.82 | .464-.924 | 36 | 0.95 | 0.87 | .655-.945 | 0.942 | 0.86 |
| 115 | Trombone | 28 | 0.96 | 0.92 | .803-.967 | 28 | 0.94 | 0.91 | .825-.958 | 28 | 0.92 | 0.76 | .400-.897 | 28 | 0.91 | 0.83 | .618-.925 | 28 | 0.91 | 0.79 | .490-.907 | 0.928 | 0.84 |
| 117 | Euphonium** | 46 | 0.99 | 0.97 | .940-.987 | 48 | 0.96 | 0.94 | .880-.966 | 47 | 0.96 | 0.95 | .919-.971 | 52 | 0.94 | 0.92 | .846-.954 | 16 | 0.95 | 0.76 | .336-.919 | 0.96 | 0.91 |
| 118 | Tuba** | 56 | 0.96 | 0.93 | .867-.963 | 49 | 0.97 | 0.95 | .905-.976 | 62 | 0.98 | 0.98 | .973-.988 | 57 | 0.98 | 0.96 | .917-.981 | 23 | 0.96 | 0.86 | .609-.945 | 0.97 | 0.94 |
| 119 | Percussion | 30 | 0.91 | 0.87 | .738-.934 | 30 | 0.96 | 0.95 | .903-.973 | 30 | 0.96 | 0.94 | .879-.971 | 30 | 0.93 | 0.89 | .759-.948 | 30 | 0.91 | 0.84 | .653-.925 | 0.934 | 0.90 |
| | | 532 | 0.94 | 0.90 | | 541 | 0.95 | 0.90 | | 557 | 0.94 | 0.91 | | 567 | 0.95 | 0.89 | | 502 | 0.94 | 0.87 | | 0.95 | 0.89 |
| | | | N = 2699 | | | K = 425 *k = 5 in every instance | | | | | | | | ** split to 2 rounds in 19-20 which is why the n is lower | | | | | | | | | |

Accepted Threshold for Cronbach's Alpha Values (Nunnally & Bernstein, 1994)

<.70 = poor reliability
.70–.80 = reliable
.80–.95 = standard for research
>.95 = acceptable for applied use

Accepted Threshold for ICC Values (Portney & Watkins, 2000)

<.50 = poor reliability
.50–.75 = moderately reliability
.75–.90 = good reliability
>.90 = excellent reliability