The Pennsylvania State University

The Graduate School

School of Music

EXPLORING THE EFFECTS OF MULTIPLE ASSESSMENT TOOLS ON TEACHERS' EVALUATIONS OF STUDENTS' MUSICAL PERFORMANCES

A Master's Paper

by

David J. Bridgewater

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Music Education

May 2009

David J. Bridgewater

The Master’s Paper of David J. Bridgewater was reviewed and approved* by the following:

Robert Gardner, Ph.D.
Assistant Professor of Music Education
Master’s Paper Advisor

Linda Thornton, Ph.D.
Assistant Professor of Music Education
Second Reader

*Signatures are on file in the School of Music

Abstract

In the field of music education, teachers often wonder how effective they are in assessing students' musical performance abilities. The purpose of this study is to explore the effects of multiple assessment tools on teachers' evaluations of students' musical performances. Three research studies (Abeles, 1973; Rothlisberger, 1992; Saunders and Holahan, 1997) indicate the reliability of assessment tools utilized for assessing general musical performances of individual band students. The three assessment tools utilized in this study include the Clarinet Performance Rating Scale, Audition Performance Rating Scale, and the Woodwind/Brass Solo Evaluation Form. A sample of 19 adolescent band musicians and five band directors with experience in evaluating individual student's musical performances were acquired for this research study. Interjudge and intrajudge reliability were calculated to measure the consistency of the judges' evaluations. All findings for interjudge reliability were statistically significant. Articulation had the highest reliability of the six categories explored among the three assessment tools. The Clarinet Performance Rating Scale had the highest reliability for the overall scores, however the Woodwind/Brass Solo Evaluation Form and the Audition Performance Rating Scale did have similar reliability. The calculations for intrajudge reliability revealed that one of the five judges had non-significant findings for the categories of tone, rhythm, tempo, and interpretation. Band directors must be careful when selecting or designing an assessment tool for evaluating students' musical performances. Articulation may have been the most reliable category because teachers compare what they hear from a student's performance to what they see on music written by the composer. Recommendations for future educational research include acquiring larger samples of adolescent band musicians and band directors. The experience of the band directors should be determined by a certain number of years with teaching band in the public schools.

Chapter 1

Rationale for the Study

In the field of music education, teachers often wonder about how effective they are in assessing students' musical performance abilities. There are many assessment tools that teachers utilize to measure the abilities of their students. Cope (1996) explained that "assessing student learning in a fair and accurate manner is no simple task" (p. 39). On an assessment tool, there are different elements to consider for scoring the overall rating. The evaluation can help the teacher and student identify elements that need improvement. When measuring students' musical performances, there are different aspects to consider, including the grade level of the students, what performance benchmarks need to be assessed, and how educators evaluate the information (Radocy, 1989).

Public school teachers in the United States are expected to have a structured system for evaluating student progress, based on national and state standards for achievement in all subject areas. According to Adam and Mary Bell (2003), teachers need to be more effective in how they assess student learning to help them determine if they are meeting these standards that are recommended at the national, state, and local levels. One of the biggest challenges in assigning grades to students is providing tangible evidence that reflects accountability for the students' learning based on the teachers' instructional expectations. Previous research has shown that "when high school directors rely heavily on attendance to determine grades, they may be 'teaching as they were taught'" (McCoy, 1991, p. 189).

Many teachers demonstrate how they address the standards in their curricula through their assessment of student learning. One particular type of assessment used in the music classroom is the performance test. The challenging part about being a music teacher is determining what elements should be included on an assessment rubric. Burnsed, Hinkle, and King (1985) explain that “the evaluation of a musical performance is constantly being criticized” (p. 28). The authors found some elements on an assessment tool are highly related to each other and could possibly not represent separate entities of an instrumental musical performance.

It is important to include elements that have been clearly communicated to the students through the teacher’s instruction when developing a rubric to assess a musical performance because students should understand the expectations prior to performing the test. Results of two studies (Asmus, 1999; Cope, 1996) indicate when determining the effective elements on a rubric, multiple ratings help a teacher identify certain aspects of the student’s performance that need improvement. In addition, the authors suggest that the assessment can also indicate whether or not the teacher is utilizing appropriate instructional strategies. Cope (1996) explains that “conscientious and deliberate use of regular assessment can strengthen any program and provide valuable assistance to the student developing the skills needed to form a lifelong involvement with music” (p. 40). Music teachers can positively influence their students’ development by using appropriate assessment measures.

Music educators should strive to be as objective as possible when assessing the concepts that students have learned, and how they apply those concepts to their musical performance (Haag et al., 1988). The authors suggest that parents and administrators will be supportive of a fair grading system for students. Teachers can provide an explanation about concepts on the assessment tool taught to their students and the activities utilized in order to enhance their learning.

There are many different criteria to consider when assessing students’ instrumental musical performances. Previous researchers have identified several musical elements commonly found on assessment tools, including tone quality, intonation, rhythm, technique, musicianship, articulation, expression, phrasing, balance, and interpretation (Burnsed et al., 1985; Goolsby, 1999). Regardless of the level of performance is being evaluated, tone quality is an aspect that should always be included in an assessment (Goolsby, 1999).

When assessing musical performances, there are generally two types of scales that teachers have used to measure student achievement. Traditional rating scales, which include a numbering system (1) low to (5) high or lettering system (F) low to (A) high, assign a number to each category that are added up for a total score. These scales have often been used to measure different musical categories of a performance (Robinson, 1995). The other type of scale often used in measuring music performance is the criteria-specific rating scale. According to Robinson (1995), criteria-specific rating scales are defined as “tools that are intended to help educators come up with somewhat more objective evaluations of performance-based activities” (p. 29). Music educators have different views on what types of scales are effective in measuring students’ musical performances.

Researchers have investigated the effectiveness of criteria-specific rating scales, particularly with performances using woodwind and brass instruments. T. Clark Saunders and John Holahan (1997) suggested that “criteria-specific rating scales can be used to evaluate student woodwind and brass performance with substantial reliability” (p. 270). The authors recommended that future researchers should investigate interjudge reliability to measure the level of consistency in their evaluations. Criteria-specific scales have been shown to be effective, and they deliver an objective approach to measuring students’ musical performances.

One of the problems in assessing students’ performances is the reliability between the perceptions of multiple adjudicators. Hewitt and Smith (2004) recommended that investigators who conduct future research studies should “determine the role of the rubric or rating scale in the consistency and accuracy of evaluations made by different groups of judges” (p. 325). The amount of training and experience music teachers have in using assessment tools to evaluate students’ musical performances also influences their effectiveness.

The purpose of this study is to explore the effects of multiple assessment tools on teachers’ evaluations of students’ musical performances. Research in assessment is an ongoing procedure for improvement in music education. In order to fulfill this study, the following research questions were investigated:

1. Are teachers’ evaluations of instrumental students’ musical performances consistent between multiple judges when using a specific assessment tool?
2. Is a teacher’s evaluation of an instrumental student’s musical performance consistent when using different assessment tools?

Chapter 2

Literature Review

The purpose of this study was to explore the effects of multiple assessment tools on teachers' evaluations of students' musical performances. The researcher examined studies that investigated authentic assessment, the development of instrumental performance assessment tools, the influence of training and experience on teachers' evaluations of students' musical performances, and the reliability of instrumental performance assessment tools. Findings in this chapter support the need for conducting this research study.

Zdzinski (1991) did a review of literature on types of assessment tools utilized for measuring instrumental performance. Assessment tools that have been researched include the Watkins-Farnum Performance Scale, Facet-Factorial Rating Scales such as the Clarinet Performance Rating Scale and the Euphonium Tuba Performance Rating Scale, and Computer Measurement in Music Performance. The author expresses concern that "there is a need for further development and refinement of performance evaluation measures for solo instrumentalists" (p. 56).

Authentic Assessment

Ames and Archer (1988) investigated the relationship between specific motivational processes and the salience of mastery and performance goals in the classroom. The authors defined the mastery goals as students' progress improvement, effort with learning, challenging work, learning processes, and ability to learn new material. Performance goals were defined as students' high grades, high normative performance, and performance abilities in relation to other students. Factors of mastery and performance dimensions included and evaluated were goal orientation, student questionnaires, modeled learning strategies, identified task challenges, attitude toward class, and each student’s perception of their own abilities in relation to their learning.

The authors examined the relationship in the final analysis between each student’s perception or interpretation of the classroom and the individual student variables (mastery structure, performance structure, learning strategies, task challenges, attitude toward class, self-perception of competence, and attributions for success and failure by ability, effort strategy, task, and luck). Findings indicated that “mastery and performance goals provide a meaningful way of differentiating students’ perceptions of the classroom learning environment” (p.264). The authors suggested that modifying students’ experiences in the classroom may provide a viable way of redirecting their achievement goal orientation.

Authentic assessments with constructivist frameworks for their own learning are critical for instilling responsibility and motivation in each student. Herrington and Oliver (2000) studied situated learning environments in three ways: they identified critical characteristics, operationalized the critical characteristics, and investigated students’ perceptions of their experiences. All aspects of their investigation were done using a multimedia program on assessment.

The authors contended that there has always been disconnect between what is known about quality education and what is actually done in the classroom. Investigations of high quality educational programs revealed that they focused on providing authentic contexts which reflect how knowledge will be used in real life, utilized expert performances and provided modeling of the processes, and allow students to consider multiple roles and perspectives. In addition, these programs promoted collaborative construction of knowledge, reflection to enable abstractions to be formed, and articulation to enable tacit knowledge to be made explicit. They also provided coaching and scaffolding by the teacher at critical times and authentic assessment of learning within the tasks. Herrington and Oliver’s (2000) study supports other research that shows that situated learning is effective for accelerating the acquisition of knowledge in a meaningful and motivating environment.

The purpose of a study by Moon, Brighton, Callahan, and Robinson (2005) was to investigate differentiated authentic assessment designed to promote meaningful learning in middle school classrooms aligned with state academic standards. The guidelines for authentic assessment measures included being focused on essential concepts, in-depth which leads to further questions, feasible and safe for the school environment, the ability to produce a quality product, the development and display of student strengths, and clearly understood criteria. In addition, these guidelines provided multiple ways for students to demonstrate their skills and knowledge and required scoring that is reflective of the authentic task.

The authors argue that “collecting reliability and validity evidence on the authentic assessments is only useful to the degree that a teacher would implement the assessments in his or her classrooms” (p.127). In general, the feedback was provided by students and teachers with whom the authentic assessments were used was positive. Students found the clearly written rubrics helpful for planning, completing work, and checking for accuracy and completeness.

McPherson and Thompson (1998) developed a theory about issues and influences with assessing students in the field of music education. In their article they discuss items to utilize on an assessment tool such as musical factors and non musical factors. The authors investigated the “Process Model of Assessing Musical Performance.” When assessing students on a performance the authors suggest that “the most critical determinant of any assessment is the quality of the musical performance” (p. 14). The authors concluded that researchers need to fill certain gaps of assessment based on their literature findings for improvement in music education.

Bell and Bell (2003) designed an action research study to assist elementary music teachers with creating and improving methods of assessment for beginning band students in fifth grade. The authors combined the concepts of multiple intelligences, research-based teaching techniques and performance assessments for the study. The authors used surveys, reflection logs, and anecdotal records to measure the affects of authentic assessment methods for fifth grade band classroom. Multiple types of assessments were used, including student portfolios, student section leaders’ information, compositions, and audiovisual documentation.

The authors recognized that various levels of student experiences and teaching needs, so they had a variety of tools that could be chosen from the appendixes which included parent and student surveys, teacher reflection, two log forms, anecdotal records, composition rubric, performance critique sheet, teacher checklist, and grade report. The authors suggested that because “teachers are being held more accountable for national, state, and local standards, their [the teachers’] focus should be to effectively teach and assess these goals and objectives to a large number of students in the small time provided” (p. 35). Valid authentic assessment practices happen only through careful planning and implementation, taking into account input from all stakeholders.

Development of Instrumental Performance Assessment Tools

Fiske (1975) performed a study in which the purpose was to investigate the relationship between evaluations of musical performances by recent graduate students of a music education program and their performing ability and musical knowledge. Thirty-three judges participated in the study. The author constructed a recording of 20 individual trumpet players performing a solo called “The Hollow Men” by Vincent Persichetti. Fiske then added the same 20 performances to the recording in order to test the reliability of the judges who assumed they listened to 40 different performances. A rating sheet for evaluating the musical performances was developed by the author. The judges were required to rate the musical elements of intonation, rhythm, technique, and phrasing for each musical performance on a scale from five (high) to one (low). There was also an overall rating that included all relevant aspects of the performance rather than an average of the other elements. The author also examined background information of all the judges’ grades from their undergraduate studies in applied music, music history, and music theory.

At the conclusion of this study, intonation, rhythm, phrasing, and technique had moderate correlations with judge stability and the overall score a high correlation with judge stability of (*r* = .71). Fiske (1975) found that interjudge reliability of recent music education graduates had no statistical significance. Although there was no significance with interjudge reliability of recent graduates, the author found that “there is a statistically significant inverse relationship between judge reliability and non-performance music achievement as measured by music history and music theory grades” (p. 29). The author suggested from this finding that evaluation ability may be relevant to certain forms of problem solving strategies through one person’s educational background.

Mills (1987) conducted a study in which the two purposes were collect and analyze written comments from assessments of students’ solo musical performances, and to design and test an assessment tool based on those comments in subsequent evaluations. For the first phrase, evaluators were put into two groups: Group 1 included music teachers and music specialists who were college students studying to be music teachers and Group 2 included non-specialists who were college students not studying music, but had experience in musical performance. The two groups were asked to listen to five performances and write a comment about each performance. Then each evaluator gave each performance a score out of thirty points without any other specific guidelines for the rating.

The author then explained that "phase two was an investigation of the extent to which assessments of performances can be predicted from the 'constructs' elicited in phase one" (p. 121). The assessors completed two assessment tasks for each performance recording. The assessors completed the same task in phase one, but at the conclusion of each performance, they had a list of 12 dichotomous statements with each representing a construct to evaluate the performances. Each statement contained either a positive or negative comment pertaining to the performance and the assessors had to circle the appropriate comment. The only high correlation was rank order between the two assessment groups was similar in placement of the performers.

Stanley, Brooker, and Gilbert (2002) conducted a case study investigating teachers' perceptions on introducing criteria for music performance assessment procedures at a conservatorium of music. The authors interviewed teachers regarding the use of criteria for assessing musical performances. The authors found that "most respondents acknowledged that criteria help them concentrate on relevant aspects of the performance" (p. 51). One participant explained that "there are plenty of things that come into discussions which are not part of the criteria and it is important that the examiners focus their mind on what is assessable" (p. 52).

Most teachers found criteria for evaluating musical performances useful. The authors conclude that "examiners feel criteria provide a useful focus during the assessment process and help articulate desirable performance characteristics in feedback to students" (p. 53). The authors suggest that in order to develop criteria for a performance, procedures should be established for evaluation.

Dressman (1992) performed a study in which the purpose was to develop, administer, and analyze results of a test for evaluating junior high school instrumentalists' selected wind instrument performance competencies. The assessment tool designed by the author was called the Instrumental Performance Competency Test. The executive skills measured by the author's assessment tool included posture, hand/finger position, embouchure, breath support, and tongue movement. The performance skills measured on the tool included tone, dynamics, phrasing, articulation notation, interpretation, and meter. The wind instruments utilized in this study were flute, clarinet, saxophone, trumpet, and trombone. Seventy-five middle school students were selected randomly from seven different middle schools in Florida. Five groups of instruments consisted of flute, clarinet, saxophone, trumpet, and trombone. Fifteen students were placed into each instrument group.

During the course of this study, Dressman (1992) measured interjudge reliability of both executive and performance skills on each instrument. The author also measured intrajudge reliability by having the judges complete a follow up evaluation of twenty randomly selected performances with the same tool. Each assessment tool for each instrument utilized in this study had the same characteristics that included embouchure, posture, hand/finger position, breathing, and tongue movement for executive skills and tone, dynamics, phrasing, articulation, interpretation, and meter for the performance skills. Objectives within the characteristics were related to common mistakes that occur on each wind instrument. The author found that the interjudge reliability between the judges' evaluations of both executive and performance skills and the students' musical performances showed a moderate to high positive correlation. The intrajudge reliability of each judge’s evaluation on executive and performance skills and the students’ musical performances showed a high positive correlation. The author concluded that the Instrumental Performance Competency Test is a useful assessment tool in measuring students’ musical performances pertaining to their own instrument.

Bergee (2003) investigated the relationship between music professors’ evaluations and end of semester applied music performances. The author used different assessment tools for each type of instrument including brass, woodwinds, percussion, voice, piano, and strings. A five point Likert scale was utilized to assess each category on the assessment tool. The assessment tools were used by the professors for students’ applied music juries for the end of the semester. After examiners completed their evaluation, they were required to assign a grade ranging from F to A+.

Results of Bergee’s (2003) study indicated that “all subscale interjudge reliabilities for all groups except percussion were statistically significant, with the exception of the suitability scale in voice” (p. 143). The author also discovered that “all rating scale total score interjudge reliability coefficients were statistically significant, as were all for the global letter grade assessment” (p. 143). Bergee found no significant differences between experienced and inexperienced evaluators.

Abeles (1973) investigated the validity of an assessment tool called the Clarinet Performance Rating Scale (CPRS). The author developed the tool for assessing clarinet performances. The six categories of musical performance evaluated with this assessment tool included interpretation, intonation, rhythm continuity, tempo, articulation, and tone. Each of the categories contained five comments on music performance for evaluation on the CPRS.

In order to measure the validity of the CPRS, thirty-two instrumental music teachers enrolled in graduate music education courses at the University of Maryland were divided into three adjudication groups. The results indicated that interjudge reliability displayed high correlation with group one was r = .993, group two was r = .985, and group three was r = .978. Because most items on the assessment tool were not specifically related to the clarinet, the author suggests that "this six-factor structure for clarinet performance would also seem appropriate for classifying music performance in general" (p. 254).

Influence of Training and Experience on Teachers' Evaluations of Students' Musical Performances

Doerksen (1999) explored the relationship between aural-diagnostic and prescriptive skills of preservice and experienced instrumental music teachers. Preservice teachers were junior and senior level undergraduate music education majors at Ohio State University and the experienced teachers were high school band directors from Ohio whose ensembles received at an overall performance rating of "I" (one) at four different Ohio Music Education Association state band competitions. These two groups listened to four different recordings of bands, which were chosen as examples of four types of performances: an excellent performance of difficult music, an average performance of difficult music, excellent performance of moderate music, an average performance of moderate music. The author created an assessment tool called the Aural Diagnostic and Prescriptive Skills Test that contained nine elements. The examiners were provided a full band score of the pieces as they listened to each recorded performance in order to help them with their evaluation.

Results of Doerksen's (1999) study showed that there was a significant difference for between-group comparisons in the rating of intonation, but no significant differences were reported in the other eight elements. As far as interactions between the two groups, tone quality, intonation, articulation, and dynamics were significant factors after listening to all four groups perform. The results did indicate that expert teachers were more specific in their comments than were the preservice teachers. In regards to ratings, the preservice teachers rated musical elements in the excellent performance band category lower than the expert teachers.

Hewitt and Smith (2004) conducted a causal-comparative study examining relationships between teaching-career level, their primary instrument, and the seven subarea scores on the Wind/Brass Solo Evaluation Form. Teacher-career levels consisted of in-service teachers, upper division undergraduate students in music education, and lower division undergraduate students in music education. The seven subarea score on the assessment tool included tone, intonation, technique/articulation, melody, rhythm, tempo, and interpretation. The subjects evaluated recordings of six junior high trumpet players performing “Gigue” from Robert King’s *French Suite for Trumpet*. Before the students recorded their performance, the students practiced individually on their own time. The assessment tool was completed by the participants for each performer.

After completing this study, Hewitt and Smith (2004) found no significant relationship of a three way interaction between teaching level, primary instrument, and the performer. However, the authors did find that “the only statistically significant two-way interaction was teaching level and performer” (p. 321). Of the seven subareas evaluated in this study, intonation was the only subarea that was significant related to teaching-career level. With few significant findings, the authors indicated that “the results of this study seem to support the finding that the primary performance instrument of the evaluator has no influence on the assessment evaluation of junior high trumpet performances” (p.323).

The purpose of a study conducted by Winter (1993) was to examine the influence of training and experience on music teachers' assessments by observing video recordings of three piano performances. Judges were selected based on four different levels of training and experience to evaluate three piano performances utilizing the Musical Performance Assessment (MPA). The four categories of evaluators consisted of the following: untrained and inexperienced, trained and inexperienced, untrained and experienced, and trained and experienced. The author defined trained participants as those "who attended a short preparation course presented as part of the study" (p. 35). Experienced judges were defined as those "who had previous involvement as examiners in any formal music performance assessment situation." (p. 35).

Thirty-three descriptors on the Musical Performance Assessment were considered in the evaluation. Only five of the descriptors had non-significant differences between the evaluators (Winter, 1993). The author found that judges who had both training and experience produce more consistent and accurate reports of the musical performances than judges who had experience only. This result indicated that trained examiners were considered "more uniform in their approach to the assessment as measured by the variance of responses to the MPA testing instrument, and were more sensitive to the performer's situation than the untrained examiners" (p. 35). The author contended that training in assessment could offer more consistent results when evaluating a musical performance.

Reliability of Instrumental Performance Assessment Tools

The purpose of a correlation study by Stivers (1973) was to investigate the reliability and validity of the Watkins-Farnum Performance Scale. The author explored the effectiveness of the assessment tool in many ways including the equivalent forms reliability, test-retest reliability, intrajudge reliability, and interjudge reliability. In addition, the author explored the relationship between the assessment scores and IQ scores, grade point averages, musical aptitude scores, length of music study, sight reading scores, and practiced performance scores. Practiced performance was defined by the author as "the score obtained on an administration of the Watkins-Farnum to a student after he has had the musical examples in his possession for one week." (p. 3).

After the scores on the Watkins-Farnum Performance Scale were determined, Stivers (1973) explored the overall scores along with the findings of other aspects of students' performance. The author discovered a moderate positive correlation between scores on the Watkins-Farnum Performance Scale and IQ (*r* = .50), academic grades (*r* = .52), and band grade (*r* = .42). There was one high correlation between sight reading performance and practiced performance at (*r* = .95). The findings in this study indicate that when evaluating students' musical performances, the Watkins-Farnum Performance Scale is a reliable assessment tool and can motivate students to learn and be successful in their musical studies.

Burnsed, Hinkle, and King (1985) investigated interjudge reliability of the Concert Band and Orchestra Adjudication Form when used at four different contest festivals. The authors also explored the interjudge reliability of the criteria (tone, intonation, technique, balance, interpretation, and musical effect) included on the rating scales. Judges for each festival were chosen by nomination and popular vote through local ensemble directors. Three judges evaluated ensemble performances at each festival. There were 110 ensembles that participated in the four festivals combined and a total of 330 performance evaluation forms were collected.

The authors found that the judges at the first festival had significant differences on the criteria of tone, intonation, balance, and musical effect. At the three other festivals, the authors discovered disagreement on one or two criteria. The authors found that “three of the four groups of judges disagreed on tone and two of the four disagreed on intonation and balance” (p. 24). Although there were slight disagreements on criteria, the judges’ final ratings were reliable.

The purpose of a study by Bergee (1988) was to investigate interjudge reliability of the Euphonium Tuba Performance Rating Scale (ETPRS) when used to evaluate university brass juries. Five brass university instructors (two trumpetists, one hornist, one trombonist, and one tubist) were asked to evaluate college students’ brass jury performances. Twenty-four students participated in the study. Twenty-seven descriptions related to brass musical performances were evaluated. Each description contained a five point Likert Scale. Judges were strongly cautioned to read each comment before deciding the appropriate score that pertained to each student’s performance.

Bergee (1988) then placed the twenty-seven dimensions into four categories which included interpretation/musical effect, tone quality, technique, and rhythm/tempo. Each of the four categories displayed a high correlation in relation to jury grades ranging from *r* =.74 to .91. The author also found that “the coefficient between ETPRS total scores and jury grades is *r* = .912, suggesting a high degree of positive relationship” (p. 17). A squared multiple regression which indicated that jury grades with a large proportion of variance accounted by ETPRS scores was statistically significant. Based on these findings, Bergee concluded that the Euphonium Tuba Performance Rating Scale is an adequate assessment tool for evaluating brass juries.

Rothlisberger (1992) conducted an experimental study investigating video modeling preparation and its effects on instrumental students’ performance achievement at auditions measured by an assessment tool called the Audition Performance Rating Scale (APRS). The author spent three weeks prior to band festival auditions preparing students for their upcoming audition. The author divided students who participated in the study into three treatment groups: video instruction, script instruction, and regular band instruction. At the conclusion of this experiment, the author found that students who viewed the video tape instruction did not achieve higher scores in contrast to the script instruction and regular band instruction groups.

When students auditioned for acceptance into the band festival, they were adjudicated by two judges using the APRS. Rothlisberger (1992) measured the interjudge reliability between the two judges at the band festival. After all auditions were completed, the author found that “the interjudge reliability for the composite APRS showed a high positive correlation between judges’ evaluation and the students’ musical performances. The author concluded that the APRS is a reliable assessment tool for evaluating students’ musical performances.

Saunders and Holahan (1997) conducted a correlation study of an assessment tool called the Wind/Brass Solo Evaluation Form. Subjects for this study included 926 students who auditioned for the Connecticut All-State Band. Thirty-six judges evaluated students’ musical performances from their auditions. The judges assessed the students’ musical performances of a chromatic scale, one major diatonic scale, one prepared piece, and one sight-reading piece.

Results of the study by Saunders and Holahan (1997) indicated that total scores from the assessment tool used in the auditions displayed a high correlation (r =.82-.95) with each group of instruments. The authors performed stepwise multiple regression and analysis of variance and found that five performance dimensions of the assessment tool (tone, technique/articulation, rhythm accuracy, and interpretation) contributed significantly with the prediction equation. The multiple regression was statistically significant accounting for 92% variance among total scores on the assessment tool. The authors concluded that “ the data in this study provide direct evidence that criteria-specific rating scales can be used to evaluate student woodwind and brass performances with substantial reliability” (p. 270).

Summary

The investigator learned valuable information from research studies that explore authentic assessment, the development of instrumental performance assessment tools, the influence of training and experience on teachers’ evaluations of students’ musical performances, and the reliability of instrumental performance assessment tools. Three reliable assessment tools were used for evaluating students’ musical performances. Music teachers with experience in evaluating students’ musical performances have shown to be effective and will serve as judges. Research on interjudge reliability has shown that evaluating musical performances with one assessment tool can be effective with multiple judges. Researchers have created assessment tools for students’ musical performances based on feedback made by judges. The uses of authentic assessments in the classroom can help to improve students’ musical achievement.

Researchers have investigated one specific assessment tool, however the investigator has not found research on at least two or more assessment tools with specific criteria for evaluating individual musical performances within the same study. The principal investigator explored the effects of three different assessment tools. The procedures conducted for this research study are discussed in the next chapter.

Chapter 3

Methodology

The methodological paradigm and procedures of this study are discussed in this chapter. The purpose of this study was to explore the effects of multiple assessment tools on teachers' evaluations of students' musical performances. This study was conducted by identifying a school district with a sample of students. The investigator consulted with the teachers and chose a piece for all groups of instruments. Three assessment tools were utilized to measure teachers' evaluations of the students' musical performances. A pilot test was administered prior to conducting the main study. Students' musical performances were recorded. Judges received a packet of information for evaluating students' musical performances. Data analysis was measured using a computer software program after the judges completed their evaluations.

The research paradigm chosen for this project was a quantitative study. A quantitative paradigm was chosen for this study because the findings could be generalized to a large population of students and judges. This project was a correlation study because the purpose was to examine the effects of using multiple assessment tools on multiple judges' ratings of student musical performances. Reliability was calculated for the overall scores of each assessment tool, as well as scores from all comparable categories, for all participating judges.

Permission to conduct this study was acquired from the school district through the assistant superintendent. The request addressed the purpose and the importance of conducting the research study as well as reasoning for why the school district was chosen. After the school district approved the study with an official letter to the investigator, permission was acquired from the Institutional Review Board (IRB). Once the IRB approved the project, the principal investigator read a recruitment script to the students and then a letter was sent home to the interested students’ parents, asking permission for their son or daughter to participate in the study.

Two populations consisting of adolescent band musicians and band teachers were represented in this study. Stratified random sampling was used in order to identify a sample of adolescent band musicians for the study. The sample for this study consisted of 19 students in grades seven through eleven who participated in public school band classes. The sample included at least one student from each of the following instruments: flute, clarinet, saxophone, trumpet, horn, trombone, euphonium, and tuba. The students received instruction from their teachers with a preparation period of six weeks.

The band directors represented the judges who evaluated the students’ musical performances. Random sampling was used in order to acquire a sample of band teachers for the study. The sample consisted of five band teachers who had experience in evaluating performance tests. The band teachers evaluated the students’ performances with the provided assessment tools.

The investigator identified a piece for all participating students to perform. The selection, “Etude No. 3” from the *Melodious Etudes* book by Johannes Rochut, was made in consultation with the students’ band teachers, so that the decision was informed by the teachers’ experience and knowledge of the ability level of the students. The students performed a moderate level piece for this study. The moderate level was defined as challenging, so that there would be variance in the quality of performances, but not so difficult that the students would not be able to perform the piece. The selected solo lasted approximately 60 seconds in length. It was desired to keep the piece in the same key, Concert Eb Major, for all instruments with the exception of the horn part in Concert Bb Major to ensure an appropriate level of difficulty for all students. The teachers provided instruction for the students during their regularly scheduled lesson time, once per six-day cycle, over the course of 5 six-day cycles. The amount of instruction time for all students varied during each lesson based on their performance abilities.

The judges used three assessment tools to measure the students' musical performances. The Audition Performance Rating Scale, the Woodwind/Brass Solo Evaluation Form, and the Clarinet Performance Rating Scale were used in this study. These tools were chosen based on previous research measuring their reliability, and because the assessment tools are generally familiar to many music teachers. All of these assessment tools measure music performance in general, although each assessment tool is unique in how it measures various aspects of students' musical performances. The Watkins-Farnum Scale could not be utilized for evaluation in this study because it does not contain Likert Scale measurements which make this assessment tool different for comparison.

The overall score and six categories were utilized for comparison between the three assessment tools. The categories assessed by the judges on all three tools were tone, intonation, rhythm/continuity, tempo, interpretation, and articulation. Each assessment tool also had its own categories or instrument specific descriptions that could not be used for comparison. Individual categories and instrument specific descriptions on one assessment tool only were excluded from evaluation.

The Woodwind/Brass Solo Evaluation Form contains assessments on a prepared solo piece, scales, and sight reading. The assessments within the prepared solo section were the only section utilized for evaluation. The solo evaluation assessment had seven categories and one category which is melodic accuracy was not used to measure the overall score of the students' musical performances. The Audition Performance Rating Scale also contained assessments on two prepared solo pieces, scales, and sight reading. The two prepared pieces had a rubric for a slow, lyric selection and a fast, technical selection. The numbers and the descriptions on the two rubrics for each of the categories were the same. The prepared piece evaluation tool was used to measure the students' musical performances. This assessment tool had seven categories and one category which was note accuracy was not utilized to measure the overall score of the students' musical performances. The Clarinet Performance Rating scale contained 25 descriptions related to the six categories and 20 descriptions were utilized to measure the overall score of the performances because five descriptions related to clarinet performance.

A pilot test was conducted to check the quality of the procedures that would be involved in the proposed study. A small sample of students was acquired for the performance, and a digital recording system was utilized for recording the students' musical performances. Enlisted judges received assessment tools and instructions explaining the process for evaluating each musical performance on the recording. The judges then assessed the students' musical performances. When all evaluations were completed the principal investigator scheduled interviews with each judge individually in order to find out if procedures and use of the assessment tools were understood. The judges returned all materials to the investigator upon completion of their interview.

A digital recording device was used during the study for the students' musical performances. The recording process was scheduled two consecutive days: one day for students in grades nine through eleven and one day for students in grades seven and eight. The order was conducted at random both days. Each student performed his or her solo in a large room while the observer recorded.

Five judges were selected to measure the students' musical performances using three assessment tools. Each student's solo was evaluated three times by each judge, using each of the three assessment tools. The investigator provided the judges with a compact disc with all of the recorded performances, which were randomly ordered in an attempt to avoid judges' recognition of the same performances. Each judge also received a packet with instructions explaining the process for measuring students' musical performances, the assessment tools in the order in which they were to complete the evaluations, and the selected piece with each instrument part for the purpose of comparing what they hear to what they see in terms of articulation and interpretation while assessing the performances. The judges were not notified beforehand that they would be evaluating the same musical performances.

The resulting data from the assessment tools was entered into a single database, utilizing a statistical computer software program. Seven variables were used in the data analyses of this study. These variables will be the scores from each section of the assessment tools (tone, intonation, rhythm/continuity, articulation, tempo and interpretation), as well as the overall standardized score from the assessment tool (sum of all section scores). All statistical analyses were calculated using SPSS computer software.

Interjudge reliability was calculated by comparing all five judges' overall scores based on the students' musical performances. The judges' numbers were analyzed to measure the reliability of their evaluations of the performances for each assessment tool. The same procedure was repeated for calculating the interjudge reliability of each category.

Intrajudge reliability was calculated by comparing each judge's overall score of the students' musical performances from the three different assessment tools. The same procedure was repeated for calculating the intrajudge reliability of each category. Results of the analyses are reported in the next chapter. The investigator drew conclusions from the results and suggested recommendations for future research studies based on the findings.

Chapter 4

Results

The results of interjudge reliability and intrajudge reliability will be discussed in this chapter. All data from the judges' evaluations were calculated using SPSS computer software. The three assessment tools included in this study were the Woodwind/Brass Solo Evaluation Form (WBSEF), the Clarinet Performance Rating Scale (CPRS), and the Audition Performance Rating Scale (APRS).

The Intraclass Correlation Coefficient was utilized in SPSS computer software to calculate interjudge reliability of the overall scores from each assessment tool. Reliability was also calculated six sub-categories of tone, intonation, rhythm, tempo, interpretation, and articulation. Table 1 shows that all relationships between all five judges' evaluations and the six sub-categories and overall scores for each assessment tool were statistically significant.

All three assessment tools had the highest interjudge reliability in two categories: articulation and interpretation. The highest interjudge reliability numbers occurred under the category of articulation. The APRS had the highest reliability (r = .576) followed by the CPRS, (r = .552), and then the WBSEF (r = .504). Interpretation had the second largest interjudge reliability among the tools. The CPRS displayed a moderate correlation (r = .489) and had the highest reliability of the three tools in this category.

The overall scores on all three assessment tools showed similar interjudge reliability. The CPRS had the highest interjudge reliability of the three tools. The CPRS displayed a moderate correlation (r = .473) along with the WBSEF and APRS close to this finding.

Table 1

Interjudge Reliability for Three Assessment Tools of Students' Musical Performances

Category	WBSEF	CPRS	APRS
Judges (n = 5)			
Tone	0.247* F(18) = 8.702	0.261* F(18) = 3.486	0.384* F(18) = 7.698
Intonation	0.249* F(18) = 4.109	0.361* F(18) = 5.440	0.392* F(18) = 7.840
Rhythm	0.423* F(18) = 5.451	0.390* F(18) = 5.806	0.232* F(18) = 2.655
Tempo	0.188* F(18) = 2.639	0.103* F(18) = 1.726	0.102* F(18) = 1.844
Interpretation	0.386* F(18) = 5.557	0.489* F(18) = 9.083	0.388* F(18) = 6.131
Articulation	0.504* F(18) = 7.251	0.552* F(18) = 8.559	0.576* F(18) = 7.627
Overall	0.445* F(18) = 8.702	0.473* F(18) = 8.614	0.468* F(18) = 9.182

*statistically significant at the p = .05 level

Two of the three assessment tools had similar interjudge reliability in the rhythm scores. The WBSEF revealed a moderate correlation (r = .423) and the CPRS had a close reliability. Two assessment tools, the CPRS and APRS, also had close interjudge reliability with intonation.

The APRS shows a moderate correlation (*r* = .392). One assessment tool, the APRS, had a moderate correlation (*r* = .384) with tone.

Tempo had the lowest reliability of all six categories for each assessment tool. Tempo on the APRS reveals the lowest reliability (*r* = .102). The CPRS also had low reliability (*r* = .103). The WBSEF shows the highest number in comparison to the CPRS and APRS with low reliability (*r* = .188).

Correlation coefficients were calculated to explore the intrajudge reliability of the overall scores as well as the six categories: tone, intonation, rhythm, tempo, interpretation, and articulation. Table 2 shows that four of the five judges' evaluations under all six categories and the overall score were statistically significant. Judge 2 had significant findings with intonation, articulation, and the overall score.

The overall scores for four of the five judges had high reliability across all three tools. Judge 4 had the highest reliability (*r* = .697) of all five judges. Judges 1, 3, and 5 had similar reliability. Judge 2 had the lowest reliability (*r* = .328) with the overall scores between the three assessment tools.

Judge 1 showed the highest intrajudge reliability across all three assessment tools when assessing tone and intonation. Intonation was rated the highest (*r* = .650) Tone contained the second highest intrajudge reliability (*r* = .544). Judge 3 achieved the highest intrajudge reliability under the categories interpretation and rhythm when assessing students' musical performances across all three tools. Interpretation shows the highest reliability (*r* = .639) and rhythm was the second highest reliability (*r* = .471).

Judge 4 had the highest intrajudge reliability when assessing interpretation and articulation from all three tools. Interpretation was rated the highest (*r* = .693), and articulation contained the second highest intrajudge reliability (*r* = .591). Judge 5 achieved the highest intrajudge reliability under the categories of interpretation and rhythm. Interpretation shows the highest intrajudge reliability (*r* = .549). Rhythm was rated the second highest (*r* = .478). Judge 2 had low correlations with intonation and articulation.

Table 2

Intrajudge Reliability for Three Assessment Tools of Students' Musical Performances

Category	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
Tone	0.544* F(18) = 4.638	0.061 F(18) = 1.238	0.375* F(18) = 2.762	0.375* F(18) = 2.840	0.286* F(18) = 2.341
Intonation	0.650* F(18) = 7.165	0.330* F(18) = 2.952	0.434* F(18) = 3.369	0.517* F(18) = 4.377	0.362* F(18) = 3.511
Rhythm	0.425* F(18) = 4.222	0.166 F(18) = 1.732	0.471* F(18) = 3.585	0.475* F(18) = 4.063	0.478* F(18) = 3.813
Tempo	0.278* F(18) = 2.505	0.082 F(18) = 1.301	0.397* F(18) = 2.996	0.514* F(18) = 4.065	0.271* F(18) = 2.294
Interpretation	0.366* F(18) = 2.640	0.089 F(18) = 1.307	0.639* F(18) = 6.036	0.693* F(18) = 7.900	0.549* F(18) = 6.478
Articulation	0.421* F(18) = 3.150	0.360* F(18) = 2.724	0.438* F(18) = 3.310	0.591* F(18) = 5.311	0.341* F(18) = 2.610
Overall	0.685* F(18) = 7.303	0.328* F(18) = 2.618	0.640* F(18) = 6.245	0.697* F(18) = 7.713	0.538* F(18) = 4.784

*statistically significant at the *p* = .05 level

The mean of each assessment tool utilized by each judge was calculated with SPSS computer software. The purpose of acquiring the mean from each tool is to show consistency within each judge. Figure 1 shows a graph with the lines representing the mean of each assessment tool for each judge.

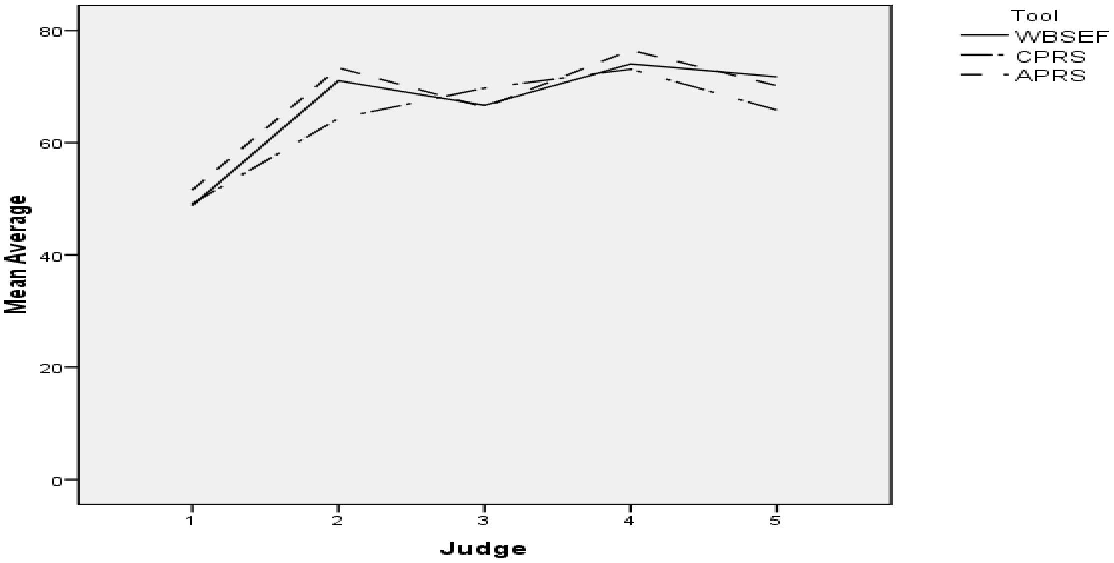


Figure 1. Mean Averages represent consistency for each judge's evaluation between the three assessment tools.

Figure 1 shows that judge 1 was the most consistent with evaluations between the three assessment tools, WBSEF (M = 48.8), CPRS (M = 49.1), and APRS (M = 51.6). Judge 4 was also most consistent between the three assessment tools, WBSEF (M = 74.0), CPRS (M = 73.1), and APRS (M = 76. 5), but with higher averages. Judge 2 was consistent with the WBSEF (M = 71.1) and the APRS (M = 73.3), but rated lower consistency with the APRS (M = 64.3). Judge 5 had similar results with the WBSEF (M = 71.8) and the APRS (M = 70.2), and lower with the APRS (M = 65.8). Judge 3 was also consistent with the WBSEF (M = 66.7) and the APRS (M = 66.3), but slightly higher with the CPRS (M = 69.7).

The standard deviation of each assessment tool utilized by each judge was calculated with SPSS computer software. The purpose of acquiring the standard deviation from each tool is to show variability within each judge. Figure 2 shows a graph with the lines representing the standard deviation of each assessment tool for each judge.

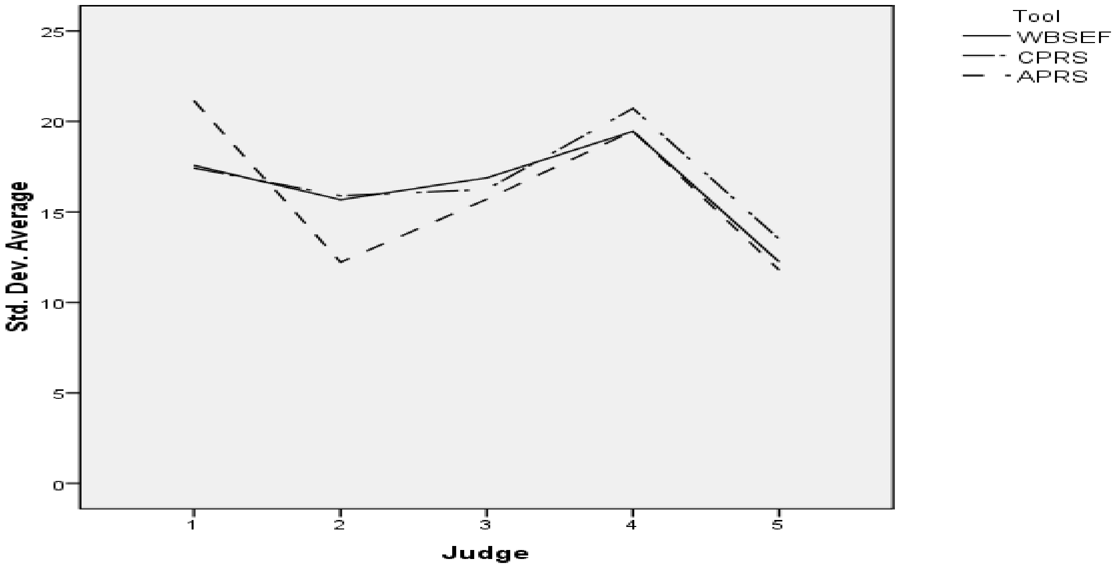


Figure 2. Standard Deviation averages represent the variability of judges' evaluations between the three assessment tools.

Figure 2 shows that judge 1 had greater variability with the APRS (SD = 21.15). Judge 2 had lower variability with the APRS (SD = 12. 22). Judge 3 and judge 4 had little variability between all three assessment tools. Judge 5 had the lowest variability between all three assessment tools as well as the lowest variability in comparison to the other judges. Implications from the results will be discussed in the next chapter.

Chapter 5

Discussion

Implications from the results in the previous chapter and future recommendations for future educational research will be discussed in this chapter. Each of these assessment tools (Woodwind/Brass Solo Evaluation Form (WBSEF), the Clarinet Performance Rating Scale (CPRS), and the Audition Performance Rating Scale (APRS)) compared the overall scores and six categories: tone, intonation, rhythm, tempo, interpretation, and articulation. The purpose of this study is to explore the effects of multiple assessment tools on teachers' evaluations of students' musical performances.

All correlations for interjudge reliability were found to be statistically significant, however reliabilities were not high in this research study. The overall scores were similar between the three assessment tools from all five judges' ratings of the students' musical performances. This finding indicates that judges were reliable in determining the overall results for the students' musical performances. The CPRS was found to be the most reliable tool in this study, however the reliability scores were close enough that the WBSEF or APRS could have been more reliable.

Interjudge reliability was calculated for each assessment tool utilized in this study. The WBSEF displayed a moderate correlation for the category of rhythm. The CPRS also displayed moderate correlation for rhythm between the five judges. The APRS had a moderate correlation for intonation between the five judges' evaluations of the students' musical performances. The results indicate that the judges were not consistent with all six categories on each assessment tool. Each assessment tool also had different levels of reliability with four of the six categories.

The APRS had the highest reliability for articulation of the three tools evaluated in this study. Two categories, articulation and interpretation, had similar reliability from all five judges among all three assessment tools. Articulation showed the highest reliability of all six categories evaluated from all three tools. It is possible the judges had similar agreements with articulation because they could view the chosen piece of music for this study while listening to each student's performance. The judges could determine whether tonguing, slurs, accents, and length of notes were appropriate for each performance.

Interpretation had the second highest reliability between all three tools. This category can be somewhat difficult to evaluate consistently because each teacher has an opinion about appropriate interpretation. There was a moderate correlation between the three assessment tools under this category. The CPRS had the highest reliability with interpretation compared to the other two tools utilized in this study. The reliability from the results page indicates that the comments for articulation were most appropriate on the CPRS for this study. This might not be the case if this study were replicated with larger samples of students and judges.

Intrajudge reliability was calculated between all three tools for each judge's evaluations of the overall scores and category scores. Judge 2 had non-significant findings under the categories of tone, rhythm, tempo, and interpretation. All other findings were statistically significant and the intrajudge reliabilities were higher compared to the interjudge reliabilities.

The overall scores on all three assessment tools displayed high correlations with the exception of judge 2 who revealed a low correlation. Judge 4 had the highest reliability of all five judges. Judges 1, 3, and 5 had similar reliabilities. The results from the overall scores indicate that most judges are reliable in assessing individual student's musical performances.

Judge 1 had high correlation when assessing the category of tone. Judges 3, 4, and 5 displayed a moderate correlation. These results indicate that the judges have different perceptions on assessing appropriate tone quality among the three assessment tools. Judge 1 was the most consistent evaluator in comparison to the other five judges.

Two judges had a high correlation with the category of intonation. Judge 1 was the most reliable of the five judges for evaluating this category and judge 4 had the second highest reliability. Judges 2, 3, and 5 had moderate correlations among their evaluations with the three assessment tools. Each of the five judges had different perceptions on appropriate intonation of students' musical performances.

Four of the five judges had similar reliability scores when assessing rhythm. Although the judges had similar reliability scores, none of the judges had a high correlation with this category. Judges 1, 3, 4, and 5 had moderate correlations from evaluating rhythm. The results from this finding indicate that four of the five judges have similar views and are reliable when assessing rhythm among the three tools.

Tempo is another category that only one judge had a high correlation. Three judges had moderate correlations with this category. Judge 4 had the highest reliability and was the most consistent evaluator among the three assessment tools. This is another category where each of the five judges had different perceptions with determining an appropriate musical performance.

Three of the five judges had high correlations when assessing interpretation across all three assessment tools. Judges 3, 4, and 5 had the best reliability with this category. Judge 4 was the most consistent with evaluations among these three judges. Judge 1 displayed a moderate correlation with interpretation. The results from this finding indicate that judges could be most reliable when assessing this particular category among the three tools.

Four of the five judges revealed a moderate correlation and one judge had a high correlation in the category of articulation. Judge 4 was the most consistent evaluator among the five judges. Judges 1, 2, 3, and 5 had similar reliability scores even though they were moderate. All five judges appear to have different perceptions for evaluating appropriate articulations of students’ musical performances among the three tools.

The results from this study indicate that all judges do not have the same levels of consistency when evaluating students’ musical performances. Since this is the case, it is important for band directors to be careful when selecting or designing an assessment tool for evaluating individual student’s performances. The descriptions on each tool may affect how teachers assess their students’ musical performances.

The judges were most reliable with evaluating articulation for the students’ musical performances. Articulation is a category where teachers compare what they hear in a student’s performance to the music written by the composer on a sheet of paper. Teachers evaluate whether students are tonguing or slurring notes in the appropriate sections of written music. Teachers also listen for note values held for the proper amount of time as well as where accents occur within the music. Those factors under this category are important when choosing or designing an assessment tool for evaluating general music performance.

The second most reliable category between the five judges was interpretation. This is another category where teachers compare what they hear in a student’s performance to dynamic markings written in the music. Teachers also evaluate the interpretation by how a student phrases and provides expression with the music. One of the biggest goals that most music teachers have for their students is to give instruction on how to be expressive when performing music on an instrument. Teachers could be very reliable with evaluating the category of interpretation.

Results from this study were preliminary in nature due to the small sample sizes of the adolescent band musicians and band directors. More research on this topic is needed to generalize findings to the two populations studied. Recommendations for future educational research should include a larger sample size of adolescent band musicians and judges for assessing musical performances. The same assessment tools should be utilized in order to explore whether or not the CPRS continues to be the most reliable assessment tool. Researchers could also consider whether judges' evaluations are consistent among groups of instruments. These recommendations would make greater contribution to the population of all adolescent band musicians and band directors.

Abeles (1973) explained that the CPRS, originally designed for evaluating clarinet performance, could be utilized for assessing general music performance. The CPRS continues to be a reliable tool for evaluating students' musical performances. The overall score was most reliable on this tool in comparison to the WBSEF and APRS.

Researchers should investigate the amount of experience that band directors have with assessing individual student performances. Experience could be defined as a certain number of years or more with teaching band in the public schools. One recommendation may also include comparing band teachers' years of experience with evaluations of students' musical performances. Researchers could explore band teachers with 10 to 15 years of experience and make a comparison with band teachers who have 20-25 years of experience. In addition to the experience, researchers should recruit band teachers who utilize individual performance tests as part of their assessment for students. This may help improve the reliability of using assessment tools in the public schools for music education.

One last recommendation for future research studies is to explore the effects of three assessment tools that were not used in this study. The tools should evaluate general music performance. Researchers could further investigate other assessment tools that have shown reliability from previous research and utilize them for comparison with overall scores and similar categories. Assessment tools could also be designed by researchers with different wording under each category and be used in an educational research study for making comparisons.

References

Abeles, H. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education, 21*(3), 246-255.

Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80*(3), 260-267.

Asmus, E. (1999). Special focus assessment in music education: Music assessment concepts. *Music Educators Journal, 86*(2), 19-24.

Bell, A., & Bell, M. (2003). *Developing authentic assessment methods from a multiple intelligences perspective.* (Report No. SO-035-233). Chicago, IL: University and SkyLight Professional Development Field-Based Master's Program. (ERIC Document Reproduction Service No. ED479391)

Bergee, M. (1988). Use of an objectively constructed rating scale for the evaluation of brass juries: A criterion-related study. *Missouri Journal of Research in Music Education, 5*(5), 6-25.

Bergee, M. (2003). Faculty interjudge reliability of music performance evaluations. *Journal of Research in Music Education, 51*(2), 137-150.

Burnsed, V., Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert band festivals. *Journal of Band Research, 21,* 22-29.

Cope, C. (1996). Steps toward effective assessment. *Music Educators Journal. 83*(1), 39-42.

Dressman, M. (1990). The development and validation of a test to evaluate selected wind instrument performance competencies of middle school/junior high school instrumentalists (Doctoral Dissertation, University of Miami, 1990). *Dissertation Abstracts International, 51* (12), 4053. (UMI No. 9114795)

Doerksen, P. (1999). Aural-diagnostic and prescriptive skills of preservice and expert instrumental music teachers. *Journal of Research in Music Education, 47*(1), 78-88.

Fiske, H. (1979). Musical performance evaluation ability: Toward a model of specificity. *Bulletin of the Council for Research in Music Education, 59,* 27-31.

Goolsby, T. (1999). Special focus assessment in music education: assessment in instrumental music. *Music Educators Journal, 86*(1), 31-35+50.

Haag, R., Russo, M., Fusco, L., Richmond, F., Branum, K., & Marshall, C. (1988). Idea bank: Evaluating music students. *Music Educators Journal, 75*(2), 38-41.

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational Technology, Research and Development, 48*(3), 23-48.

Hewitt, M., & Smith, B. (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.

McCoy, C. (1991). Grading students in performing groups: A comparison of principals' recommendations with directors' practices. *Journal of Research in Music Education, 39*(3), 181-190.

McPherson, G., & Thompson, W. (1998). Assessing music performance: Issues and influences. *Research Studies in Music Education, 10,* 12-24.

Mills, J. (1987). Assessment of solo performance: A preliminary study. *Bulletin of the Council for Research in Music Education, 91,* 119-125.

Moon, T., Brighton, C., Callahan, C., & Robinson, A. (2005). Development of authentic assessments for the middle school classroom. *The Journal of Secondary Gifted Education, 16*(2/3), 119-133.

Radocy, R. (1989). Special focus in music education: Evaluating student achievement. *Music Educators Journal*, *76*(4), 30-33.

Robinson, M. (1995). Alternative assessment techniques for teachers. *Music Educators Journal, 81*(5), 28-34.

Rothlisberger, D. (1992). Effects of video modeling preparation on student instrumental audition performance achievement and performance anxiety (Doctoral Dissertation, University of Maryland, 1992). *Dissertation Abstracts International*, *53*(7), 2287. (UMI No. 9234645)

Saunders, T., & Holahan, J. (1997). Criteria-specific rating scales in the evaluation of high school instrumental performance. *Journal of Research in Music Education*, *45*(2), 259-272.

Stanley, M., Brooker, R., & Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Research Studies in Music Education*, *18,* 46-56.

Stivers, J. (1973). A reliability and validity study of the Watkins-Farnum Scale (Doctoral Dissertation, University of Illinois at Urbana-Champaign, 1972). *Dissertation Abstracts International, 34*(2), 815. (UMI No. 7317440)

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*, *22,* 34-39.

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

Appendix A

Wind/Brass Solo Evaluation Form

TONE The student’s tone (Circle ONE number only):
5 is full, rich, and characteristic of the tone quality of the instrument in all ranges and registers.
4 is of a characteristic tone quality in most ranges, but distorts occasionally in some passages.
3 exhibits some flaws in production (i.e., a slightly thin or unfocused sound, somewhat forced, breath not always used efficiently, etc.).
2 has several major flaws in basic production (i.e., consistently thin/unfocused sound, forced, breath not used efficiently)
1 is not a tone quality characteristic of the instrument.

INTONATION The student’s intonation (Circle ONE number only):
5 is accurate throughout, in all ranges and registers.
4 is accurate, but student fails to adjust on isolated pitches, yet demonstrates minimal intonation difficulties.
3 is mostly accurate, but includes out-of-tune notes. The student does not adjust problem pitches to an acceptable standard of intonation.
2 exhibits a basic sense of intonation, yet has significant problems, student makes no apparent attempt at adjustment of problem pitches.
1 is not accurate. Students’ performance is continuously out of tune.

RHYTHMIC ACCURACY The student performs (Circle ONE number only):
5 accurate rhythms throughout.
4 nearly accurate rhythms, but lacks precise interpretation of some rhythm patterns.
3 many rhythmic patterns accurately, but some lack precision (approximation of rhythm patterns used).
2 many rhythmic patterns incorrectly or inconsistently.
1 most rhythmic patterns incorrectly.

TEMPO The student’s tempo (Circle ONE number only):
5 is accurate and consistent with the printed tempo markings.
4 approaches the printed tempo markings, yet the performed tempo does not detract significantly from the performance.
3 is different from the printed tempo marking(s), resulting in inappropriate tempo(s) for the selection, yet remains consistent.
2 is inconsistent (i.e., rushing, dragging, inaccurate tempo changes).
1 is not accurate or consistent.

INTERPRETATION The student demonstrates (Circle ONE number only):
5 the highest level of musicality including well-shaped phrases and dynamics.
4 a high level of musicality, but has some phrases or dynamic that are not consistent with the overall level of expression.
3 a moderate level of musicality and musical understanding.
2 only a limited amount of musicality and music understanding.
1 a lack of musical understanding.

TECHNIQUE/ARTICULATION The student demonstrates (Circle ALL that APPLY):
1 appropriate and accurate tonguing
1 appropriate slurs as marked
1 appropriate accents as marked
1 appropriate ornamentation as marked
1 appropriate length of notes as marked (i.e, legato, staccato).

Appendix B

Clarinet Performance Rating Scale

1: highly disagree that the statement is descriptive
2: slightly disagree that the statement is descriptive
3: neither disagree nor agree that the statement is descriptive
4: slightly agree that the statement is descriptive
5: highly agree that the statement is descriptive

TONE

1 2 3 4 5 1. Thin tone quality.

1 2 3 4 5 2. There was a lack of tonal color.

1 2 3 4 5 3. The quality of the tone was rich.

1 2 3 4 5 4. Sounded Shallow.

INTONATION

1 2 3 4 5 5. Played out of tune.

1 2 3 4 5 6. The intonation was good.

RHYTHM/CONTINUITY

1 2 3 4 5 7. Uneven rhythm.

1 2 3 4 5 8. Smoothness in execution.

1 2 3 4 5 9. Insecure technique.

1 2 3 4 5 10. The rhythm was distorted.

TEMPO

1 2 3 4 5 11. Played too fast.

1 2 3 4 5 12. Seemed to drag.

1 2 3 4 5 13. Rushed

INTERPRETATION

1 2 3 4 5 14. Effective musical communication.

1 2 3 4 5 15. The interpretation was musical.

1 2 3 4 5 16. Played with musical understanding.

1 2 3 4 5 17. Played with traditional interpretation

ARTICULATION

1 2 3 4 5 18. Attacks and releases were clean.

1 2 3 4 5 19. Free from tonguing noise.

1 2 3 4 5 20. Accents were played as indicated.

Appendix C

Audition Performance Rating Scale

TONE QUALITY

The student's tone quality as demonstrated in the prepared selection (Circle ONE number only):

5 was full, rich, and characteristic of the tone quality of the instrument in all ranges and registers.

4 was of a characteristic tone quality in most ranges, but distorts in occasional passages (i.e., loud or soft, high or low tessitura, etc.).

3 has some flaws in basic tone production (i.e., thin sound, spread tone, unfocused tone, forced tone, air in sound not contributing to the tone).

2 has several major flaws in basic tone production.

1 is not a tone quality characteristic of the instrument.

INTONATION

The intonation of the student's performance of the prepared selection (Circle ONE number only):

5 is accurate throughout, in all ranges and registers.

4 is accurate, but student fails to adjust on isolated pitches, or demonstrated minimal intonation problems.

3 is mostly accurate, but has notes that are played out-of-tune including some significant problems, and did not adjust on some out-of-tune pitches.

2 has a basic sense of intonation, but did not adjust pitches to an acceptable standard of intonation and had significant intonation problems.

1 is not accurate. Students performance is continuously out-of-tune.

RHYTHM

The student's performance of the rhythm patterns of the prepared selection (Circle ONE number only):

5 was accurate throughout.

4 was nearly accurate but lacked precise interpretation of some rhythm patterns.

3 approximated the notated rhythms, but lacked accuracy in performance of some rhythm patterns.

2 demonstrated an inconsistent performance of most rhythm patterns.

1 was not accurate.

TEMPO

The tempo of the student's performance of the prepared selection (Circle ONE number only):

5 was accurate and consistent with the printed tempo markings.

4 approached the printed tempo marking, yet the performed tempo did not detract significantly from the selection.

3 was at a different tempo than the printed tempo resulting in an inappropriate tempo for the selection.

2 was performed with inconsistencies, (i.e., rushing, dragging, inaccurate tempo changes).

1 was not accurate or consistent, with major flaws including rushing, dragging, etc.

MUSICALITY

The student's performance on the prepared selection (Circle ONE number only):

5 demonstrated the highest level of musicality including well-shaped phrases and dynamics. The performance was one which demonstrated musical understanding.

4 demonstrated a high level of musicality, but had some phrases or dynamics that were not consistent with the over-all level of musicality or musical understanding.

3 demonstrated a moderate level of musicality and musical understanding.

2 demonstrated only a limited amount of musicality and musical understanding.

1 demonstrated a lack of musical understanding.

STYLE OF ARTICULATION

The student's performance of the prepared selection (Circle ALL that APPLY):

1 demonstrated appropriate tonguing accuracy.

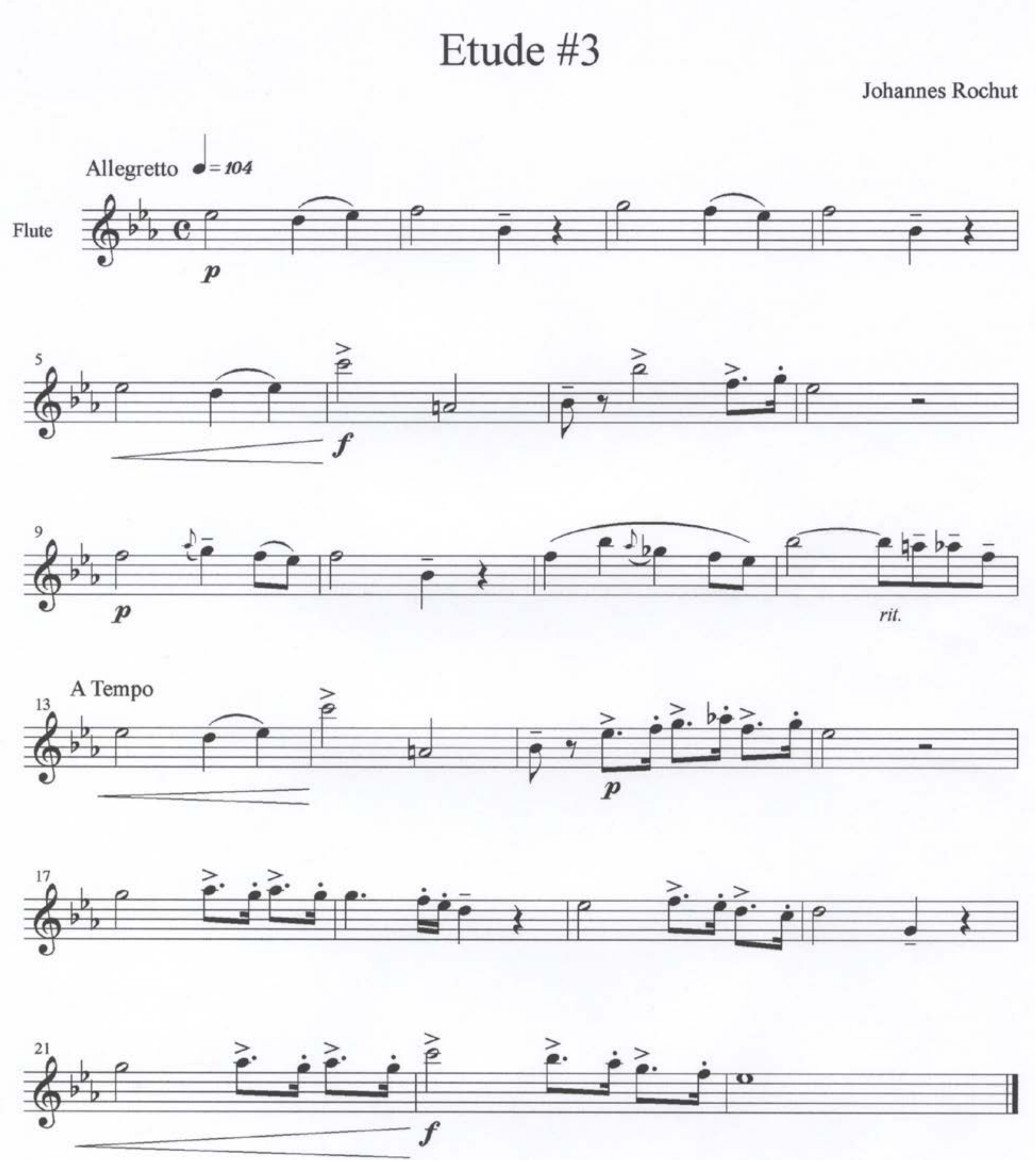
1 demonstrated appropriate slurs as marked.

1 demonstrated appropriate accents as marked.

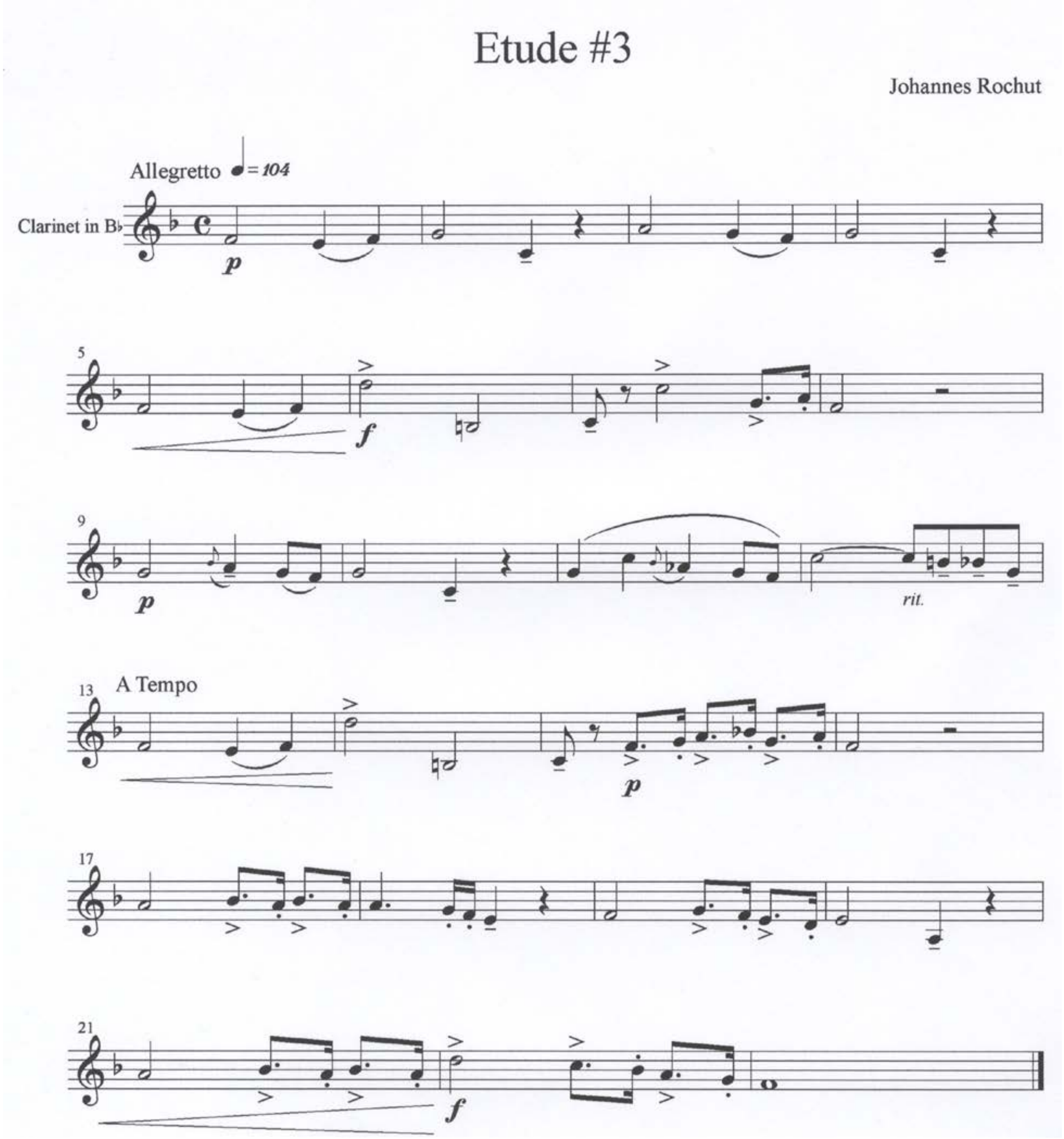
1 demonstrated appropriate ornamentation as marked.

1 demonstrated appropriate length of notes as marked

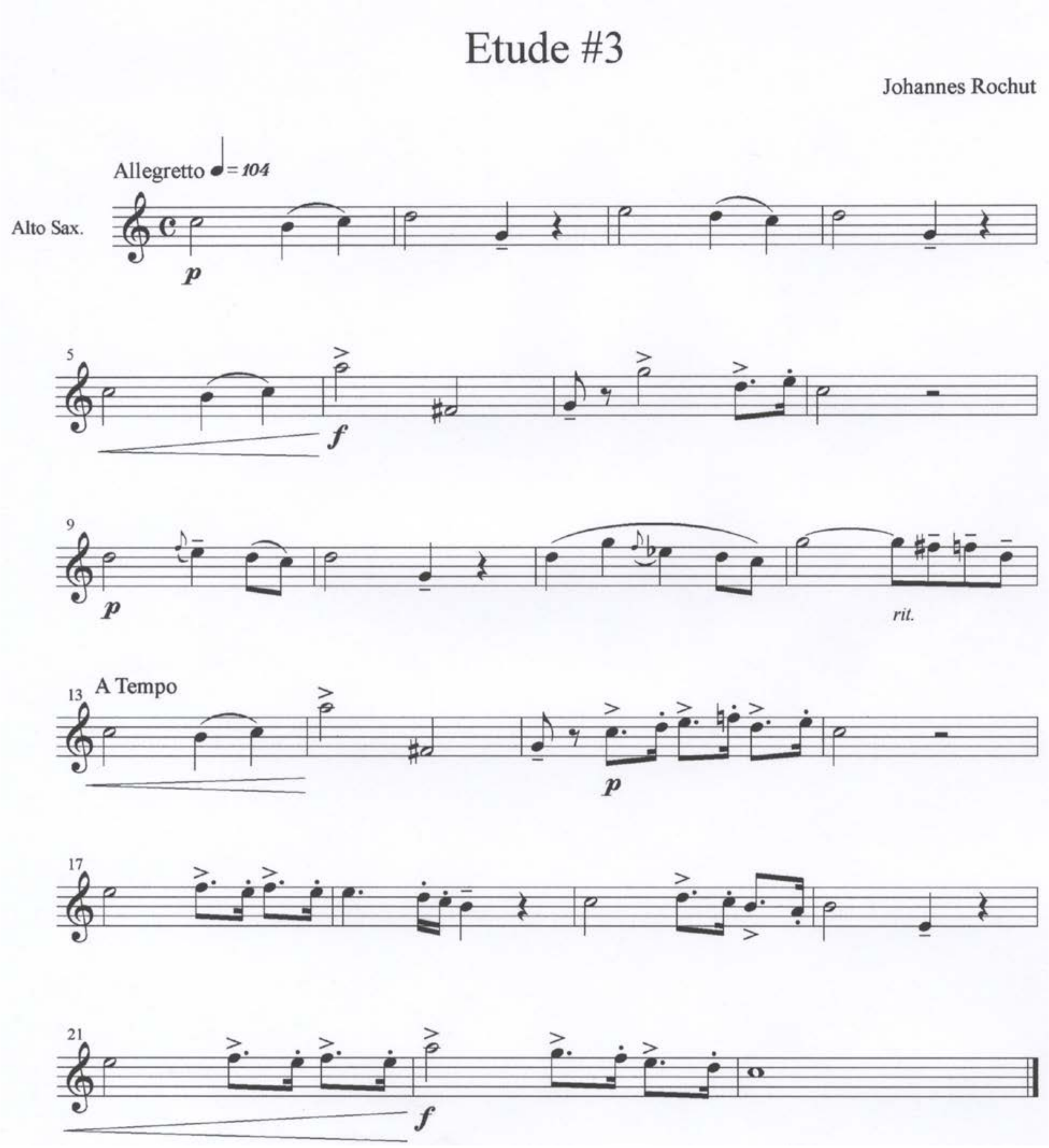
Appendix D


Etude #3
Johannes Rochut
Allegretto ♩= 104
Flute
p
5
f
9
p
rit.
A Tempo
13
p
17
21
f

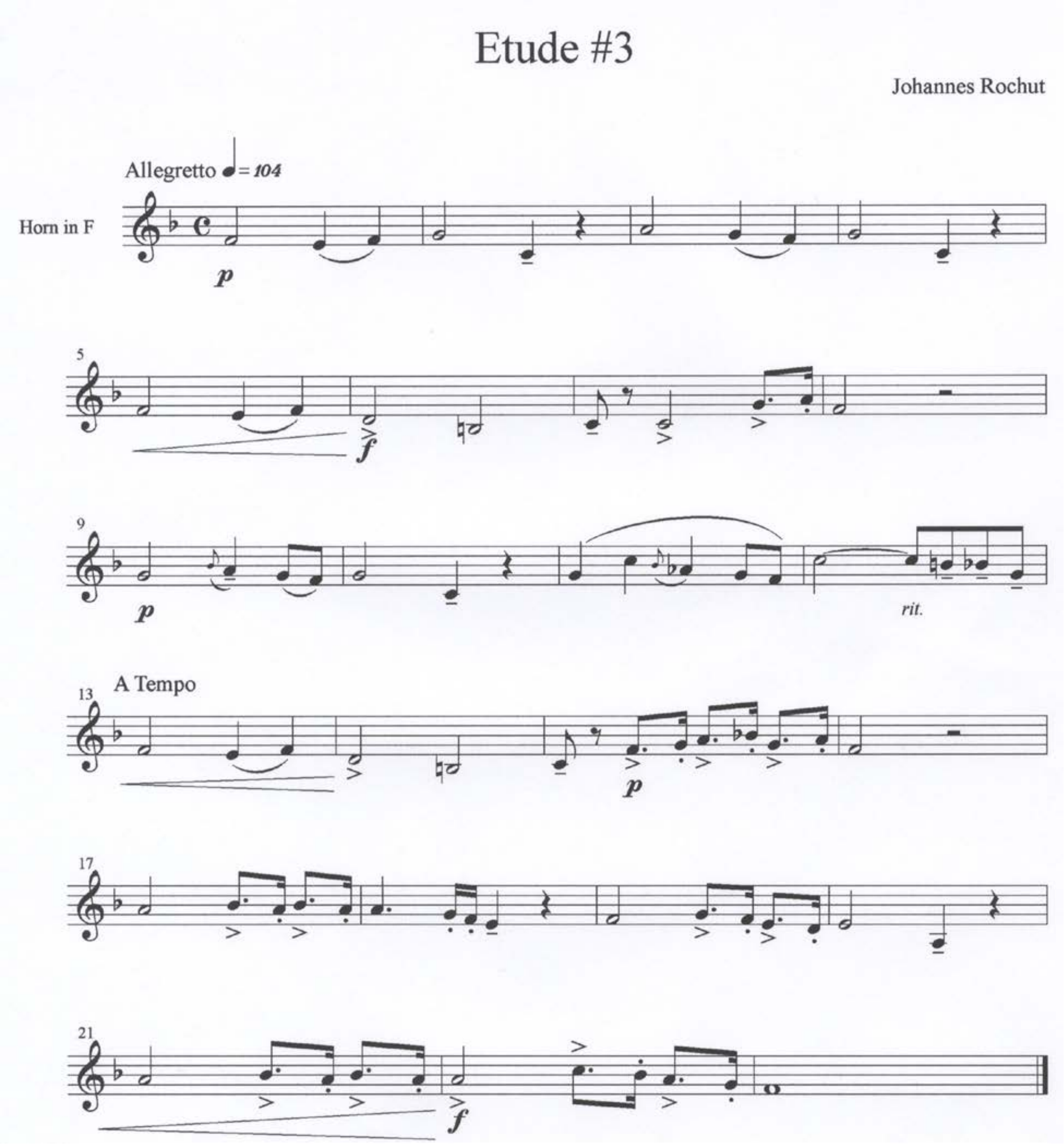
Appendix E


Etude #3
Johannes Rochut
Allegretto ♩= 104
Clarinet in B♭
p
5
f
9
p
rit.
13
A Tempo
p
17
21
f

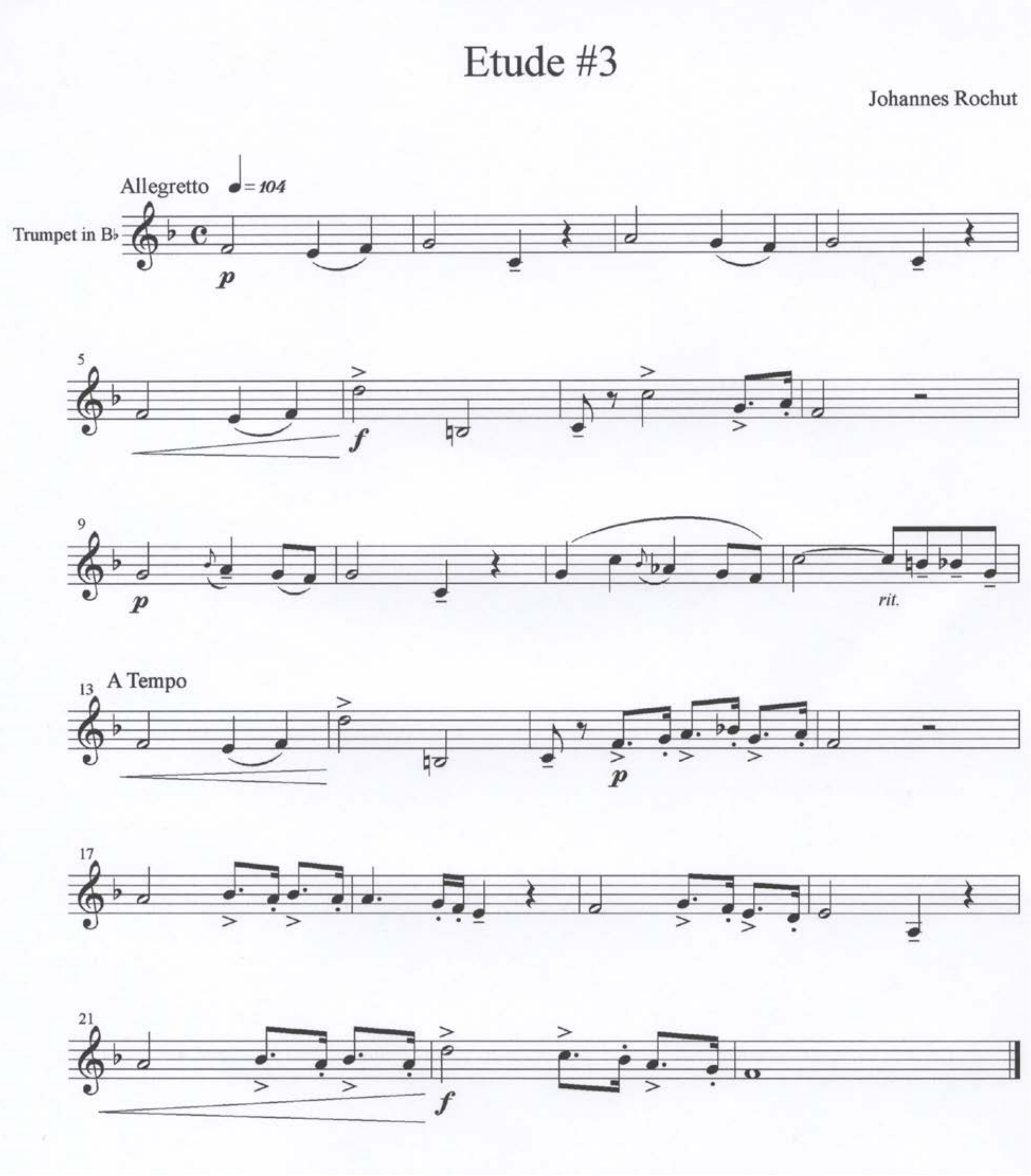
Appendix F


Etude #3
Johannes Rochut
Allegretto ♩= 104
Alto Sax.
p
5
f
9
p
rit.
13
A Tempo
p
17
21
f

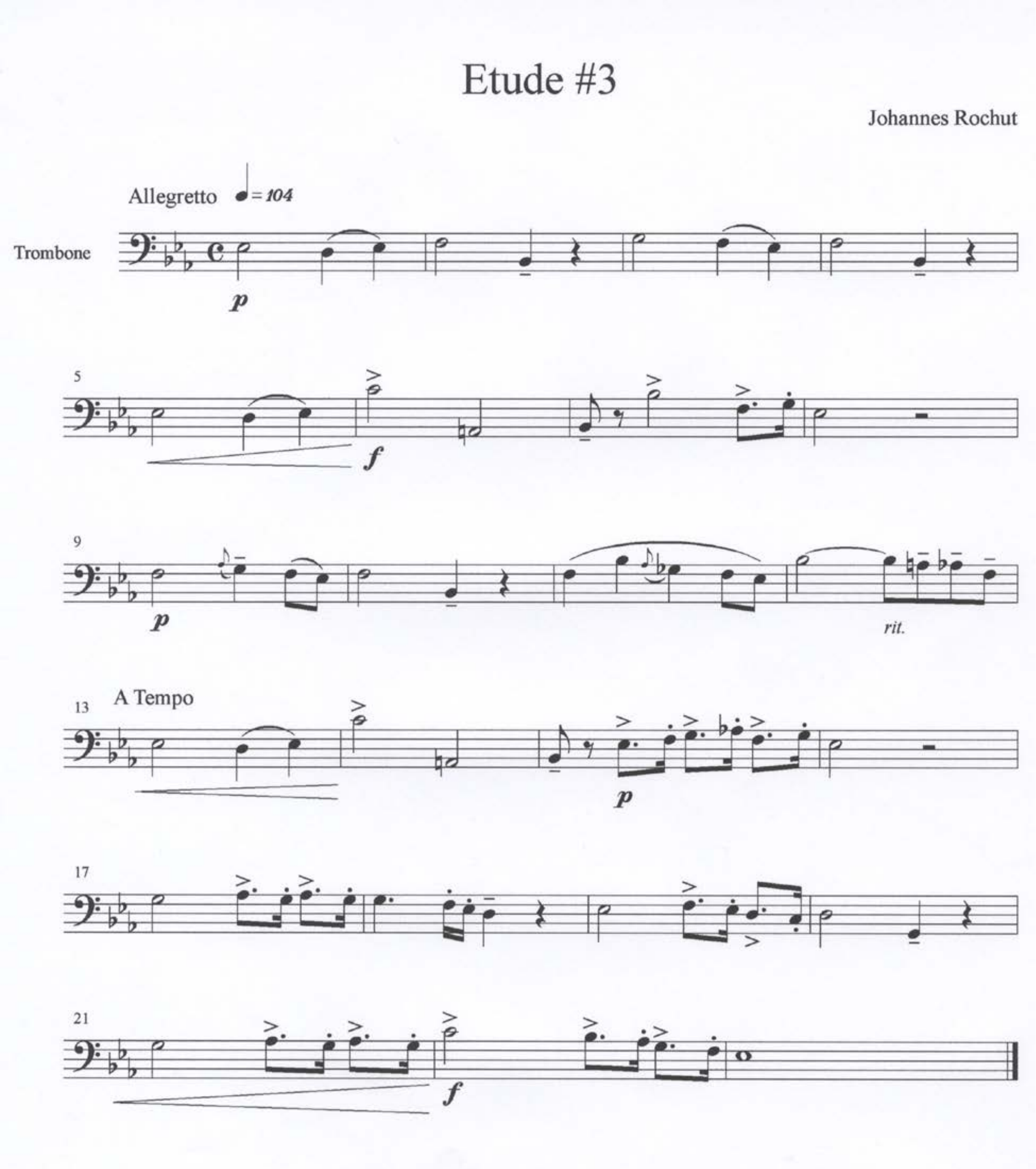
Appendix G


Etude #3
Johannes Rochut
Allegretto ♩= 104
Horn in F
p
5
f
9
p
rit.
13
A Tempo
p
17
21
f

Appendix H


Etude #3
Johannes Rochut
Allegretto ♩= 104
Trumpet in B♭
p
5
f
9
p
rit.
13
A Tempo
p
17
21
f

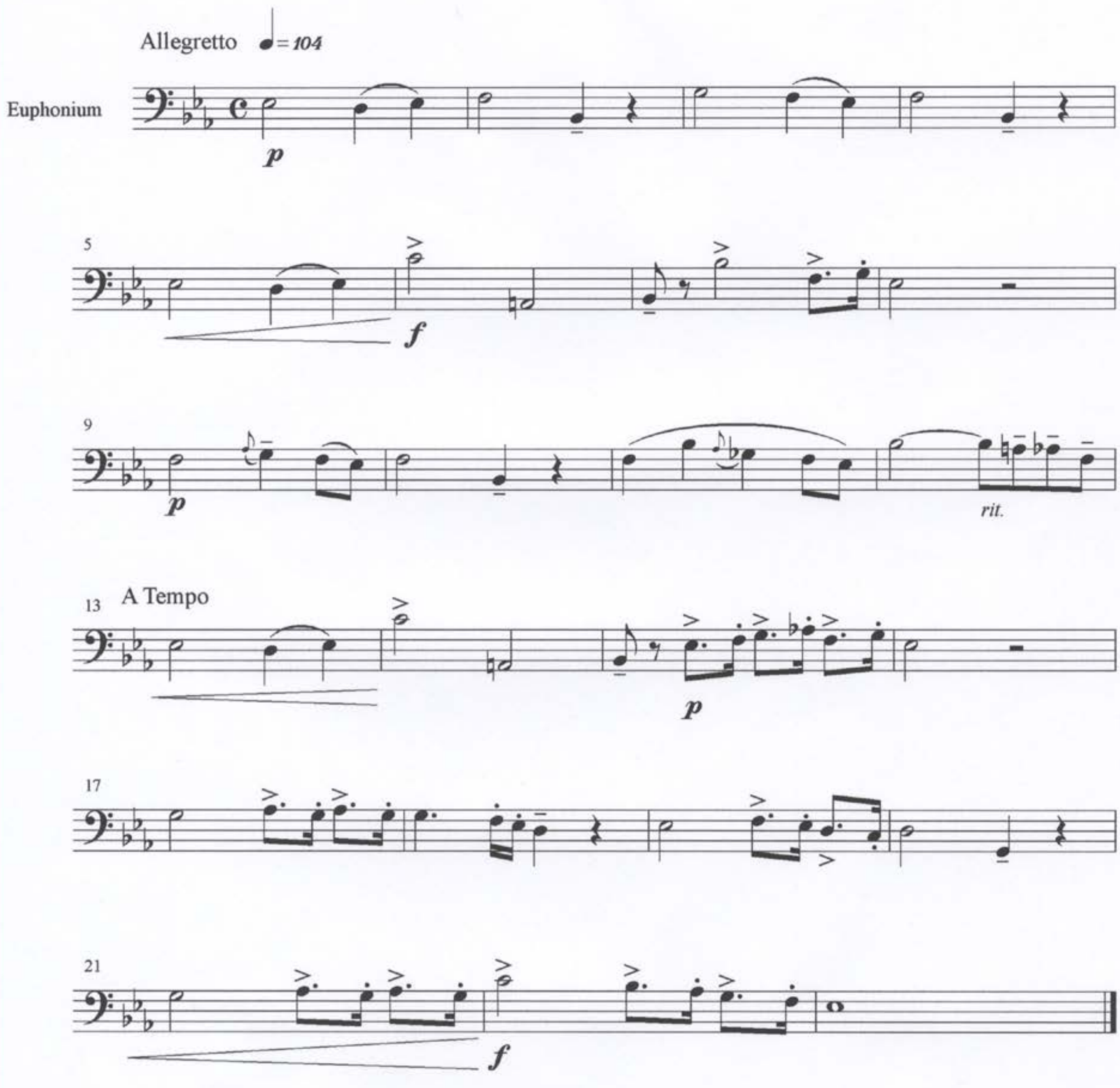
Appendix I


Etude #3
Johannes Rochut
Allegretto ♩= 104
Trombone
p
5
f
9
p
rit.
13
A Tempo
p
17
21
f

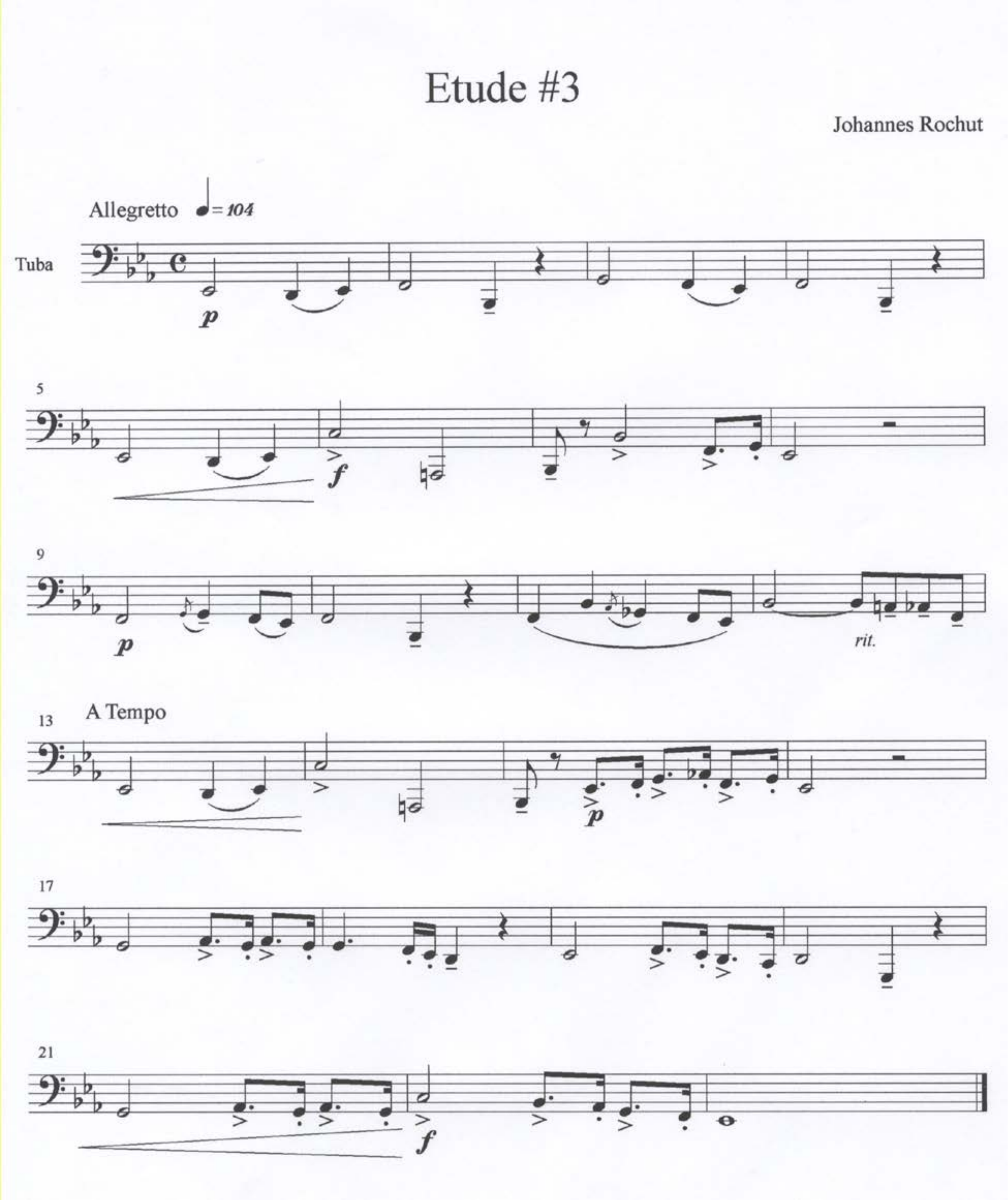
Appendix J

Etude #3

Johannes Rochut



Appendix K


Etude #3
Johannes Rochut
Allegretto
= 104
Tuba
p
5
f
9
p
rit.
13
A Tempo
p
17
21
f