EVALUATION OF MUSIC PERFORMANCE:

COMPUTERIZED ASSESSMENT VERSUS HUMAN JUDGES

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

MUSIC

MAY 2018

By

YI-JU SHIH

Dissertation Committee:

Gabriel Arnold, Chairperson
Ronald Heck
Jeffrey Boeckman
Chet-Yeng Loong
Miguel Felipe

Keywords: music performance, human judges, Computerized Assessment, *SmartMusic*

ACKNOWLEDGMENTS

Sincere appreciation is extended to the dissertation chair, Dr. Barbara McLain (retired), and Dr. Gabriel Arnold, as well as the committee members Dr. Jeffrey Boeckman, Dr. Chet-Yeng Loong, Dr. Miguel Felipe, and Dr. Ronald Heck for their guidance and helpful suggestions throughout the doctoral program.

My eternal gratitude is expressed to my parents, Hui-Hsiung Shih and Ya-Hsiu Shih-Hsieh, as well as my family for their unconditional love, patience, and support from Taiwan as I continued my education.

A heartfelt gratefulness is expressed to my husband, Kameron Slaten, who knows the challenges and endeavor to accomplish a doctoral degree. Without his love, inspiration, and tremendous encouragement, I would not be able to complete my dissertation.

Lastly, I would like to thank all my proofreaders who have helped me to improve my scholarly work, as well as the study participants and human judges for sharing their experiences with me and making valuable contributions to this study.

ABSTRACT

A major concern in the literature on music performance evaluation has been the reliability of assessment. The challenges in fairly and accurately evaluating music performance are often identified as subjective matters, non-musical factors, as well as the methods or tools with which music teachers assess. Subjective matters, such as biases and personal preference, could lead to unfair assessment. Non-musical factors, such as student attitude, effort, and participation, have been given greater weight than musical factors in calculating music grades. Computerized assessments designed for evaluating music performance could improve objective measurement of music assessment. The *SmartMusic* assessment is a technological program commonly used for evaluating musical performance. Although researchers have studied the effectiveness as well as the reliability of the *SmartMusic* assessment, very few quantitative studies have shown evidence of the comparison of computerized assessments with human examiners in assessing music performance. This quantitative experiment compared the evaluations of a set of human judges and the *SmartMusic* assessment. Statistically significant differences between the human judge panel and the *SmartMusic* assessment were found in the variability and the reliability of the ratings. The dependability of the computerized assessment was below acceptable levels of reliability in evaluating student performance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Assessment is a critical part of the educational process for music teachers, students, and parents. Music teachers are called to evaluate the effectiveness of their instruction. Students and parents regularly receive feedback concerning progress in music literacy and performance. Assessment in musical performance however, has limitations due to factors that affect the quality of the evaluation (McPherson & Thompson, 1998; McPherson & Schubert, 2004). Music assessment is often subject to bias based on the opinions and perspectives of the evaluator, such as stereotyping, first impressions, performer attractiveness, stage behavior, dress, social factors, and the evaluators' characteristics. Evaluation procedures, the rating scales utilized in evaluations, and the types of testing may affect the accuracy and fairness of music performance assessments (McPherson & Thompson, 1998, p.12). These non-musical factors have been found to make assessment difficult and inaccurate (McPherson & Thompson, 1998, p.13; McPherson & Schubert, 2004, p.62).

Technology may provide an alternative method for the assessment of music performances, in order to diminish non-musical biases and to provide a more reliable measure of student progress and instructional effectiveness. As Atkins et al. (2010) stated, "technology-based learning and assessment systems will be pivotal in improving student learning and generating data that can be used to continuously improve the education system at all levels" (p. v). This chapter will serve to: (a) describe the validity and reliability of music assessment, (b) provide an overview the paradigms of assessment, (c) discuss the use of technology in general educational assessments and summarize the history of computer-assisted assessments, (d) describe music performance evaluation and problems associated

with music assessment, (e) present the current status of the use of technology in music assessment, and (f) set out the problem statement and the need for this study.

Validity and Reliability of Music Assessments

Validity and reliability of music assessments are important issues for music education. Assessment validity is an indicator of how much meaning can be placed upon a set of test results. Assessment validity is crucial in psychological and educational testing, where the importance and accuracy of tests is paramount. Reliability refers to the consistency of a measure. An assessment is considered reliable if the same results are elicited repeatedly. Stanley, Brooker, and Gilbert (2002) defined test reliability as the following:

> A fundamental concern is that music performance assessments be valid and reliable. Essentially, validity relates to how faithfully assessments measure characteristics they purport to measure, while reliability refers to consistency in the assessment process. There is widespread debate in related literature about the role of subjectivity and objectivity in assessment and the impact those perspectives have on the validity and reliability of assessments. (p. 47)

Boyle (1992) believed that objectivity can be improved by "establishing clear criteria against which the performance will be evaluated (and using) some type of rating scale to indicate the extent to which each performer meets the evaluative criteria" (p. 258). McPherson and Thompson (1998) stated "music performance assessment is the process by which one individual attempts to make a qualitative description. While we might like to think otherwise, the assessment of music assessors is sometimes low and significant biases often influence the results" (p.12). Researchers have investigated inter-judge and intra-judge reliabilities. Fiske (1977) conducted a study on the relationship between selected factors in trumpet performance adjudication reliability including judge reliability, judge performing

ability (as measured by applied music grades), and judge music knowledge (as measured by music history and theory grades) in the sample. No relationship was found between judge performing ability and judge reliability as well as judge performing ability and nonperforming music achievement. However, a "statistically significant inverse relationship" was found between judge reliability and nonperformance music achievement (e.g., music history and theory grades), $p < .05$ (pp. 261- 262).

Thompson and Williamon (2003) argued that although performance assessment has been considered highly subjective and unreliable, empirical research suggests that is not always true, and, in fact, the level of reliability can be dependent on the nature of the assessment plan being utilized (pp. 25-26). They provided an example in which three experienced evaluators rated performances given by students at the Royal College of Music, London, and graded them according to marking rules of the Associated Board of the Royal Schools of Music. Inter-judge correlation coefficients (Spearman's *rho*) were computed across 15 categories, such as overall quality, instrumental competence, technical security, rhythmic accuracy, and so on, $\rho = .50$, range = .33– .65, $p < .05$ (p. 29). The authors pointed out that although the data indicated a positive correlation between judges' marks, the correlations were merely moderate with some evidence of bias based on each evaluator's own instrumental experience.

Bergee (2003) investigated inter-judge reliability of music performance evaluation. The adjudicators consisted of brass ($n = 4$), percussion ($n = 2$), woodwind ($n = 5$), voice ($n = 5$), piano ($n = 3$), and string ($n = 5$) faculty evaluators for end-of-semester applied music juries at a large university. Each judge completed two rating scales including a criterion-specific rating scale for each performer and a global letter grade for each performance. Begree utilized Kendall's coefficient of concordance (W) to measure the agreement among raters. If $W = 0$, then there is no agreement among the raters. The result indicated that except

percussion, all full-group coefficients of concordance based on total score were between W = .70-.80. Accordingly, there were positive agreements among the judges. Inter-judge reliability was consistent, with panel size ranging from an n = 2 for percussion to an n =5 for woodwind, voice, and strings. On the other hand, except percussion and piano, all full-group coefficients of concordance based on the letter grade were between W = .58-.69, which indicated less agreement among the judges when utilizing a global grade than when using a criterion specific rating scale. Bergee suggested that "evaluators might consider using criterion-specific rating scales in lieu of the more common regimen of writing comments, because rating scale total score reliabilities in this study tended to be higher than letter-grade reliabilities, especially among the larger panels" (pp.144-147).

As aforementioned, many factors could affect the validity and reliability of music performance evaluations. Assessment in music performance may vary from teacher to teacher and may be subject to the tools and designs of testing. In order to provide accurate and beneficial assessment to students, there is a need to determine the most valid and reliable evaluation practices.

Assessment Paradigms

Assessment not only provides information regarding the learning process and outcome for students, but also reflects on a teacher's instructional efficacy. The National Association for Music Education (NAfME) stated:

> Some form of regular assessment of music programs should be adopted. The assessment should measure student learning across a range of standards representative of quality, balanced music curriculum, including not only responding to music but also creating and performing music. This assessment should serve the goal of educational accountability by providing data that can be included in the school- or

district-level "report card" disseminated to the public as required by law.

("Assessment in Music Education," n.d., para. 7)

Assessment involves more than simply examining tests. Assessment guides the re-evaluation and re-definition of goals to empower teachers and their students.

Although a few similar terms, such as "evaluation," "measurement," and "assessment," have been sometimes used interchangeably in the literature, each of these might be slightly different in use and meaning. Colwell (2002) explained that *measurement* has been regarded as a single test, the smallest unit in assessment; *evaluation* is differentiated by the making of judgments based on the data collected from measurements and other processes; *assessment* refers to a considerable body of data that has the potential to diagnose the process and provide feedback after data analysis (p. 1129).

Asmus (1999) defined *measurement* as "the use of systematic methodology to observe musical behaviors in order to represent the magnitude of performance capability, task completion and concept attainment," *assessment* as "the collection, analysis, interpretation, and application of information about student performance or program effectiveness in order to make educational decisions," and *evaluation* as "the collection and use of information to make informed educational decisions" (p. 21). According to Asmus' (1999) and Colwell's (2002) definitions, the terms *evaluation* and *assessment* are very similar, while *measurement* is not synonymous with *evaluation* and *assessment*. In this study, *assessment* and *evaluation* will be employed interchangeably.

Straka (2004) asserted that *measurement* and *evaluation* fall under the umbrella of assessment (p. 263). Measurement consists of objective testing, which is a test that has right or wrong answers and so can be marked objectively and should be reliable and valid, while evaluation consists of subjective observations, which is "the ongoing process of making

judgments and decisions based on the interpretation of evidence gathered through assessment" (Nebraska Department of Education, n.d., p. 490).

Several paradigms of assessments include objective versus subjective, formative versus summative, informal versus formal, authentic versus non-authentic, norm-referenced versus teacher-created, and quizzes versus exams. These paradigms are described in the following section.

Objective Versus Subjective Assessments

Assessment is commonly classified as being either objective or subjective. An objective assessment may consist of a selected response format that involves choosing the answer from a set of alternatives. Paper and pencil tests are examples of objective assessments, and the types of questions may include multiple-choice, short answer, true/false, fill-in-the-blank with a word bank provided, and matching (Barley, 2006, p.15). Subjective assessments may also be based on a questionnaire form, but may have more than one correct answer or more than one way of expressing the correct answer. Examples of subjective assessments include short constructed response (providing a brief answer in writing or by drawing a diagram or picture), extended written response (writing the answer), or performance assessment (doing, creating, or performing the answer). Subjective assessments are open to bias as an assessor or judge may interject personal opinions and preferences during the grading process.

Suskie (2004) stated that "an objective assessment is one that needs no professional judgment to score correctly. Subjective assessments, on the other hand, yield many possible answers of varying quality and require professional judgment to score" (p. 99). Objective assessments draw upon quantitative scales to grade student work or performance, while subjective assessments require the instructor's professional skills and developed awareness of quality in academic fairness in order to grade student performance. Similarly, Gronlund

(1976) considered objective assessments to be reliable and fair, whereas subjective assessments have the tendency to be unstable or biased because the scores are influenced by the view or judgment of the adjudicator (p. 22). Computerized or online assessments fit well with the format of objective assessments because a computer does not have emotions or biases.

Formal Versus Informal Assessments

Evaluating music performances vary in form, ranging from informal and spontaneous procedures to formal procedures in highly structured or systemized settings (Barry, 2009, p. 246). According to the New York State Education Department (2009), formal assessments are conducted at the state and local district levels. They are administered to students in targeted grade levels and involve data collection and controlled procedures (p. 23). These types of assessments systematically evaluate student performances on essential knowledge and skills centered around subjects and are administered at set intervals (e.g. weekly, bimonthly, quarterly). The learning outcomes are demonstrated and scores are compared for each student.

Informal assessments involve the facts that teachers collect to evaluate how well students have mastered skills and content covered in class (The New York State Education Department, 2009, p. 23). This type of assessment tends to be less structured and may not be validated or tested for reliability. Examples of informal assessments are observations, interviews, record reviews, and performance reviews. Student scores are not compared with other students, and each student score is compared to past performance. However, informal assessments can give students immediate help and correct their errors for a better learning process.

Summative Versus Formative Assessment

The term *formative evaluation* was coined by Michael Scriven (1967) in reference to curriculum development. Scriven is a British-born academic philosopher who is best known for his contributions to the theory and practice of evaluation. Subsequently, summative and formative assessments were clearly differentiated in the literature by Bloom, Hastings, and Madaus's (1971) seminal writing. The authors regarded summative assessment as the collection of data for providing evaluation and feedback after instruction has occurred. Formative assessment is defined as "a systematic evaluation in the process of curriculum construction, teaching, and learning for the purposes of improving any of these three processes" (Bloom, et al., 1971, p. 117). Thus, formative assessments may be quantitative and qualitative, while summative assessments are only quantitative in nature.

From a temporal perspective, scholars have described that the data collection of formative assessment occurs before instruction, while summative assessment occurs after instruction (Burns, 2008, pp. 2-3). The goal of formative assessment is to assist teachers and students with better achieving their learning goals and processes over time. The goal of summative assessment is to summarize student achievement at set intervals in time, such as at the end of a class, unit, or semester (Bauer, 2014, p. 133). Accordingly, timing is one of the key factors in differentiating summative from formative assessments.

Formative and summative assessments can be further characterized within an "improvement paradigm" versus an "accountability paradigm" (Ewell, 2009, p.8). Formative assessment is considered to be an improvement paradigm derived from the "institution-centered" approach of the mid-1980s, whereas summative assessment is considered to be an accountability paradigm derived from early state mandates (Ewell, 2009, p.8). Formative assessments allow teachers to provide feedback in order to help students throughout the teaching and learning process.

As the assessments mentioned above, formative assessments are associated with informal assessments to facilitate improvement over time for a more affective learning process. Summative assessments are associated with formal assessments, which aim to examine students' overall comprehension at set intervals, and determine whether students have mastered competencies required to move to the next unit or next level of education.

Teacher-Made Tests Versus Standardized Assessments

Teacher-made tests are designed by teachers specifically for their students. They do not include standardized tests created by test companies. Teacher-made tests can be referred to as informal and authentic assessments. Although teacher-made tests can be vital components of the instructional and learning process, Wiggins (1989) stated that "course-specific tests also have glaring weaknesses, not only because they are often too low level and content heavy, they are rarely designed to be authentic tests of intellectual ability; as with standardized tests, teacher-designed finals are usually intended to be quickly read and scored" (p.123).

Standardized tests are classified into two categories: aptitude tests and achievement tests (Popham, 1999, p.8). In general, a standardized aptitude test forecasts the potential ability of students to achieve in a subsequent educational setting. For example, the Scholastic Aptitude Test attempts to predict how well secondary school learners will perform in their continuing education. A standardized achievement test attempts to evaluate skills and knowledge students learned through planned instruction in a given grade level or a certain period of time. Some examples of standardized achievement tests are "Iowa Tests of Basic Skills, Metropolitan Achievement Tests, California Achievement Tests, and Stanford Achievement Tests" (Popham, 1999, p.8).

Although a standardized test may be norm-referenced and measure the relative skills and knowledge of the similar age or grade-level learners using a national-wide test (Popham, 1999), it may not measure educational quality. Standardized tests are intended for large-scale examinations, rather than customized for individual students from specified school districts.

Authentic Assessment

Authentic assessment is a performance-based evaluation, rather than selected-response questions (Bauer, 2014, p. 133). Music is a performing art, and music assessments often involve performance-based evaluations of musical skills and knowledge, as well as of processes (e.g.., improvisation and the creative process in music) and products (e.g., performance). Authentic assessment provides direct feedback and evidence of student learning, progress, and knowledge. It also refers to assessment tasks that are teacher-made in the real world and in school (Frey, Schmitt, & Allen, 2012, p. 1). An example of an authentic assessment in music is having a student clap a rhythmic pattern to ascertain whether or not the student can perform that rhythm pattern correctly. The aim of authentic assessment is to assess many various kinds of abilities in settings that closely resemble actual situations in which those abilities are used.

The Use of Technology in General Educational Assessment

Technology has the potential to empower teachers in instruction and foster student learning due to its customizable quality. Teachers can utilize technology as an alternative assessment tool to evaluate individual student strengths and needs. The terms *computer-assisted assessment*, *computer-based assessment*, *computerized assessment*, or *computer-aided assessment*, refer to the use of computers to assess student progress (Chalmers & McAusland, 2002, p. 2). These terms are interchangeable and commonly used by researchers and educators to describe utilizing computers for evaluation.

The goals of computer-assisted assessments may also vary. CAA may be diagnostic to verify student knowledge prior to starting a course, which enables modifications to be made to a course design. Similar to conventional assessments, CAA may be formative or summative. Features of computerized assessments would allow teachers to conduct both summative and formative assessments. For example, the *SmartMusic* computer program provides immediate feedback and generates data that allows teachers to conduct both summative and formative assessments.

The History of Computer-Assisted Assessment

The history of computer-assisted assessments can be traced back nearly as far as computing courses (Winters & Payne, 2006, p.1). The earliest documented reference of the use of computers to assist grading identified in this study was in 1959 at the Rochester Polytechnic Institute by Jack Hollingsworth. Hollingsworth (1960) had 120 students in a full-semester programming course and argued that the use of a grader was necessary to accommodate such numbers. The computer program of the institute was utilized to test the behavior of students' machine-language submissions. The major concern of Hollingsworth was security because malevolent students' submissions might affect the grading itself (Hollingsworth, 1960, pp. 528-529). A shortcoming of the new grading system was that it limited student creativity. This was due to the fact that computer assessment was based on precise functional grading, where only a complete match was marked correct.

Later, Bunderson, Inouye, and Olsen (1988) postulated that the introduction of computer-based testing (CBT) would change the field of educational assessments. Starting with the first generation of simple transpositions of written tests to computer-based tests, they predicted the four successive generations of computerized tests (p. 26). The fourth generation of intricate integrated assessment systems was named *intelligent measurement,* or "the application of knowledge-based computing to any of the sub-processes of educational

measurement" (p.116). Below is a summary of the four generations of computerized education measurement:

1. The 1st Generation: Computerized testing (CT): conducting conventional tests via computer.

2. The 2nd Generation: Computerized adaptive testing (CAT): modifying the difficulty or level of contents of the next task, or the pace of the next item based on examinees' responses.

3. The 3rd Generation: Continuous measurement (CM); using adjusted measures set in a curriculum to estimate on-going changes in the student's achievement and profile as a learner.

4. The 4th Generation: Intelligent measurement (IM): making intelligent scoring, analysis of individual profiles, and feedback to students and teachers, with knowledge bases and inferencing procedures. (p. 4)

Accordingly, computer-based tests utilize various development and quality assurance processes not available in paper-and-pencil test instruments that benefit educational assessment.

Greiff & Martin (2014) believed that new computer technology not only generated new learning styles and environments, but also created new settings for the design and administration of assessments (p.1). Traditional paper and pencil tests tended to be static with limited interactivity, relied on intricate logistics, and required administration procedures with trained test administrators. The interaction and data processing capacities offered by the computer helped to standardize and automate administration and grading processes. The authors believed that the advantages of computerized testing included recording capability and application of behavioral data, or the possibility of new test administration procedures such as adaptive testing. Additionally, computers allowed large-scale administration for

educational instruction and assessments that efficiently reduced laborious work and improved the accuracy and consistency of educational evaluations and measurements.

Based on the U.S. Department of Education, since the No Child Left Behind Act of 2001 (NCLB; Pub. L. 107-110) was signed, the importance of an accountability system in public school settings has been a central focus of education. Under this law, accountability results from evaluating academic instructional performance on the basis of student performance measures. The components of accountability include: (a) statements of goals, (b) instructional objectives, and (c) an extensive assessment. These provide criteria to determine whether or not students have accomplished the stated objective (Abeles, Hoffer & Klotman, 1994, p. 250). As the result of the NCLB Act, the focus of education was shifted to assessments. That exponentially increased the exam load for K-12 institutions and assessment administrative personnel. Several states, such as Mississippi, Virginia, and Texas, investigated utilizing computer-based-tests for the purposes of rapid score reporting and re-examination of required graduation tests (Lissitz & Jiao, 2012, p.2). The growing reliance on technology and the advantages of improving assessments have been a main focus in the educational field. This requires further study and attention in order to maintain high-quality education.

The U.S. Department of Education (2010) declared the importance of assessing and measuring what matters most in a statement: "Our education system at all levels will leverage the power of technology to measure what matters and use assessment data for continuous improvement" (para. 1). Later, the NCLB Act evolved and was renamed the Every Student Success Act in 2015 (ESSA; 2015, Pub. L. 114–95). The ESSA was passed to ensure equal opportunity for all students and to provide a well-rounded education including music (S.1177, § 8002, p. 298). The ESSA emphasized "innovative assessment" and "accountability demonstration" and defined them as below:

> An innovative assessment system may include competency-based assessments, instructionally embedded assessments, interim assessments, cumulative yearend assessments, or performance-based assessments that combine into an annual summative determination for a student, which may be administered through computer adaptive assessments; and assessments that validate when students are ready to demonstrate mastery or proficiency and allow for differentiated student support based on individual learning needs. (p.129)

The U.S. Department of Education (2016) further published a summary of the ESSA Assessment Regulations, "Creating Better, Smarter, Fairer Tests." This regulation specifically indicated "Leveraging technology to improve assessments" as below:

> States may develop computer-adaptive tests, which could provide a more precise estimate of a students' ability with fewer questions than a traditional test; and require that such assessments report assessment results against grade-level academic achievement standards (or against the appropriate achievement standards if the computer-adaptive test is for students with the most significant cognitive disabilities), to ensure all students are held to the same standards. (p. 2)

Accordingly, technological assistance has been emphasized to improve educational instruction and assessment. That not only benefits students with various levels of learning ability but also educators for teaching and assessing students in an efficient way.

Music Performance Evaluation

Assessing music performance is both subjective and objective. Subjectivity in evaluating music performance refers to qualitative judgments made by adjudicators, while objectivity refers to estimations of solid components, for example, the accuracy of pitch and rhythms (Long, 2011, p. 4). Boyle and Radocy (1987) stated that effective assessment in

music must be precise, functional, comprehensive, and applicable to instructional material displayed (p. 7).

Studies and researchers have found that the means by which music teachers assess student performance need to be reviewed and improved. Four national music assessments were administered in 1971, 1979, 1997, and 2008 by the *National Assessment of Educational Progress* (NAEP) (Zuar 2006, p. 2; Fisher, 2008, paras. 2-4). However, due to limited funding, after the first music assessment in 1971, the rest of the tests were mostly comprised of multiple-choice exams that limited the measurements. One of the reasons for using this type of test was that the multiple-choice format provided a "cost-efficient" means for assessing learning. This type of test, however, could not evaluate many essential artistic behaviors, such as the ability to create or perform musical pieces (Shuler & Connealy, 1998, p.12). Many researchers have claimed that those types of tests cannot measure student achievement in music (Shuler & Connealy, 1998, pp. 16-17; Frankel, 2002, pp. 6-7). Frankel (2002) pointed out the limitations of New Jersey's music assessment as below:

> How can one properly assess a performing art with a traditional multiple-choice examination? All of the questions on the Visual and Performing Arts ESPA were multiple-choice. No audio recordings of music were played. No performances were recorded.... Although grading a multiple-choice examination is certainly easier than grading a student portfolio, the multiple-choice examination cannot properly assess skills such as singing and playing an instrument. The only way to properly assess a musical performance is by listening to it, either in person or using a recording. (Frankel, pp. 6-7)

Kotora (2001) conducted a study to examine assessment strategies and grading practices that were utilized by Ohio high school choral music educators and instructed by Ohio college choral methods faculties. Kotora (2001) classified 12 assessment strategies

from past and current literature regarding assessment in general education, music education, and choral music education. The author created two surveys to collect data, including a high school Choral Music Teacher Survey and a college Choral Methods Teacher Survey (p. xi). There were two research questions: (a) What type of assessment strategies and grading practices are currently being used by Ohio high school choral music teachers to assess individual student learning in the choral music performance classroom? and (b) what type of assessment strategies and grading practices are currently being taught in choral music methods classes at Ohio colleges and universities to pre-service choral music teachers? (pp. 15-16)

Regarding the high school Choral Music Teacher Survey, a total of 608 surveys were mailed out to the Ohio high school districts. The subjects ($N = 246$) completed and returned surveys, resulting in a 43% return rate for the study. For the college Choral Methods Teacher Survey, a total of 38 Ohio college choral methods teachers were identified and sent surveys. The final subjects ($N = 20$) completed and returned surveys. That indicated a 53% return rate (p. xi).

The frequency of use of the 12 assessment strategies are listed respectively from most-used to least-used below:

1. Concert performances.
2. Student participation.
3. Student attendance.
4. Singing tests.
5. Written tests.
6. Student attitude.
7. Audiotaped recordings.
8. Individual performances.

9. Videotaped recordings.

10. Independent study/written projects.

11. Check sheets/rating scales/rubrics.

12. Student portfolios.

Accordingly, non-musical strategies such as student participation, student attitude, and student attendance still occupied a large portion of assessment. This occurrence was found to be slightly reduced with college choral methods teachers during the last two years (p. xii).

Both high school choral music teachers and college choral methods teachers' verbal comments claimed numerous difficulties in assessing students in the choral music performance classroom including: (a) lack of time to assess due to short class periods with full class schedules and large class sizes; (b) maintaining accurate student records; (c) assessing individuals in large choir classes with good class management; (d) making students and parents understand the importance of assessment in choir courses; (e) insufficient administrative support and understanding; (f) insufficient teacher training; and (g) insufficient guidelines to maintain, develop and implement student assessment. Both high school and college choral instructors preferred to apply assessments based on personal choice, rather than influences from local, state, or national guidelines or standards (p. xii).

In Kotora's (2001) study, although the music educators measured student performance, non-musical criteria and written exams were still used frequently in music assessments. The problematic issues listed in the aforementioned paragraph need to be resolved in order to improve music measurement and evaluation. Additionally, efficient and time saving assessments for music performances need to be explored and examined. Computer-assisted music assessments are likely to enhance music assessments as well as provide alternative ways for music educators to assess music performance.

Zuar (2006) conducted a study to examine New York State's music assessments. The study compared the 2002 New York field test with the 1997 national assessment, which mainly utilized multiple-choice and paper-pencil tests. The subjects ($N = 447$) were students from 20 school districts. They were tested on both performance skills and music knowledge based on the New York State Standards. The relationship between the test items, outcome of student performance, and three perspectives including content, curriculum, and demographics were evaluated. The results indicated that students scored higher in the curriculum area of performance. On Standard 4 (understanding cultural dimensions) and Standard 1 (creating and performing), data illustrated the greatest student achievement (Zuar, 2006, p. 117). The findings implied that the 2002 New York field test applied more authentic strategies of assessment and alternative techniques than the 1997 national assessment (Zuar, 2006, p. 117). This study highlighted a positive case of evaluating student performances and encouraged music educators to assess student performances.

Music educators were encouraged to follow the *National Standards for Music Education* (1994) provided by MENC (National Association for Music Education, formerly known as the Music Educators National Conference) specifying what all students should know and be able to do in music with specific guidelines for grades K-4, 5-8, and 9-12. Later in 1996, MENC published the *Performance Standards for Music: PreK-12* to help music teachers with strategies and benchmarks for assessing student progress towards the national standards.

Barkely (2006) conducted a descriptive study to investigate current music teachers' attitudes and practices regarding the assessment of the National Standards for Music Education (1994) in the elementary school general music classroom. The variables included the strategies and the frequency of assessments used by elementary school general music teachers, as well as factors that influenced teachers' frequency of assessments and assessment

practices. The factors included the number of buildings the teacher worked at, the number of students taught per week, class sizes, teacher experience, teacher training, school resources, teacher opinions of the importance of the assessments, the report card grading systems used by the participating teachers' school districts, and the availability of time for assessments.

The researcher developed a survey to examine the research problems, which posed the following questions:

1. What assessment strategies do elementary school general music teachers use to assess the National Standards for Music Education?
2. How frequently do elementary general school music teachers evaluate their students on each of the National Standards for Music Education?
3. What factors influence elementary school general music teachers' abilities to assess the National Standards for Music Education?
4. What are the attitudes of elementary school general music teachers towards assessment of the National Standards for Music Education? (p. 29)

A total of 619 surveys were mailed out to Michigan school districts. Seventy-nine of the respondents indicated having no general music classes for Grades kindergarten through 4. From the remaining 540 surveys, the subjects *(N* = 255) were completed and returned, indicating a 47% return rate. The study findings identified factors that affected music teachers having insufficient time and data at the end of a marking period to assess student achievement. These factors were: (a) the music class often is only scheduled once or twice a week for 30 to 45 minutes, and (b) elementary school classroom teachers usually teach one class and may have only 25 to 30 students to assess, whereas music teachers normally teach a whole grade level or entire school and often have hundreds of students to grade. Insufficient time seemed to be a problem for music teachers to assess students (p. 2). Other difficulties

for music teachers included difficulty implementing and assessing the National Standards for Music Education.

Based on the results of the study (from the perspective of applying National Standards for Music Education into assessment), the surveyed teachers assessed Standard 1 (singing) and Standard 5 (reading and notating music) most frequently, whereas they assessed Standards 4 (composition) and 3 (improvisation) least frequently (p.47). As for the strategies of assessment, observation was the most used assessment strategy across all nine content standards. Observation has been regarded as an important part of classroom assessment; however, this type of evaluating strategy has a tendency to be subjective, and it has to be systematically documented in order to be credible.

Another controversial issue in music assessment is whether or not non-musical criteria, such as student effort and attitude, should be included in assigning music grades. Barkely (2006) believed, "student achievement and student effort are not the same, and that student effort should be separate for objective assignment of grades (p. 6). Similarly, Lehman (1968) stated that in arithmetic or history, attitude could not be an adequate substitute for achievement, and neither in music (p 81). Students' musical grades should be based on musicality. Boyle and Radocy (1987) stated music assessment in the past was too dependent on non-musical/subjective criteria like attitude as a basic source for evaluation (pp. 12-13). Music grades based on non-musical factors did not provide an accurate report of students' musical development or achievement; instead, the grades usually reflect non-musical criteria.

McQuarrie and Sherwin (2013) conducted a study regarding assessment in music education on relationships between classroom practice and professional publication topics. The goals were to:

1. Identify current assessment techniques utilized by elementary school music

teachers.

2. Identify types of assessment techniques included in the current music teaching literature.
3. Identify any relationships between assessment techniques that were most frequently utilized by teachers and those that were most frequently included in teacher-focused music education publications (para.1).

The researchers applied two approaches to conduct this study: (a) investigate the Washington Music Assessment Participant Survey (WMAPS) utilizing a survey by McQuarrie in 2008; and (b) conduct an analysis of literature review. The survey was designed to identify the assessment practices of elementary school general music teachers. The subjects ($N = 100$) were elementary school general music educators from the Northwestern United States. The analysis focused on the topic of classroom music assessment and the researchers examined 10 years (1999 – 2009) of the national publications *Teaching Music* and *Music Educators Journal* regarding this subject. The variables were the frequency of use and frequency of inclusion in the classroom and literature. The results were then examined in order to identify possible relationships.

The results illustrated the most "frequently used" strategies: "grading based upon participation (about 80%), grading based upon effort (about 79%), assessing individual performances using informal observation (70%), large group performances (61%), and grading based upon behavior (59%)." Conversely, the results showed that participants did not use: "(a) standardized music achievement tests (73%), (b) music assessment software (about 72%), (c) formative assessment strategies (72.16%), (d) portfolios (about 68%), and (e) music aptitude tests (about 56%)" (McQuarrie & Sherwin, 2013, para. 12). Additionally, the findings indicated a possible disconnection between the assessment strategies used by music educators in classroom settings, and those referenced in music education publications.

Although intrinsic aspects of assessments such as performance were graded, they were mostly done so using informal assessments or large group performances that might not reflect individual student capacities. The majority of the grading values from music educators was based on non-musical aspects, such as participation, effort, and behavior, rather than musical goals included in the National Standards for Music.

The Status of the Use of Technology in Music Assessments

Technology can be helpful and useful for music assessment, including the examination of listening skills, musical creativity (e.g., composition) and knowledge, and performance techniques. In *Vision 2020: The Housewright Symposium on the Future of Music Education*, Yarbrough stated:

> The rise of computer technology, distance education, telecommunications and television will impact the speed and accuracy of the delivery of information to everyone involved in the educational process. Computers will increase the ability of musicians and non-musicians to self-educate in virtually every aspect of music. (p. 196)

Technological tools may be applied to: (a) develop traditional assessment implementation, (b) transform traditional implementation in new ways, and (c) enable new approaches with regard to the assessment of student learning (Bauer, 2014, p. 134). Two main aspects of computerized music assessments are: (a) assessing musical knowledge and understanding, and (b) assessing musical performances.

Many software and computer programs can be used for the assessment of student knowledge and understanding. For instance, the document sharing function of Google Docs can be beneficial for multiple teachers to edit and administer tests. Notation software such as *Finale* can be useful for creating notation-based examinations and to design musical audio examinations or audio-visual test items. Numerous applications are available for creating

online quizzes or electronic surveys. Many of these applications allow for a variety of questioning types, including multiple-choice, true/false, short answers, and matching. Teachers can utilize these tools to collect data automatically (Bauer, 2014, p.135), which can help them save time and become better organized.

Teachers may apply technology when assessing student performance skills and technique. There are several types of applications which can be utilized, including: (a) audio and video recordings for teacher, peer, and self-assessment, and (b) audio-visual assessment tools for displaying music scores and content, recording the performance, and offering feedback. Such technologies can enable teachers to assess several parameters of music performance (Bauer, 2014, p.140). A computerized assessment serves as both summative and formative evaluations and could provide digital scores as well as immediate feedback to improve student learning.

Additionally, computer-based assessment might aid in reducing human errors due to its unique strengths, such as no personal or visual bias and no physical fatigue issues. Dowsing, Long, and Craven (2000) conducted a study regarding the differences between human examiners and computerized assessment on IT skills examinations. The results indicated that the computerized assessment system had enhanced stability and accuracy compared to human examiners (p.12). The utilization of technology may not only help to stabilize the consistency of formative and summative assessments, but also to provide an impartial distribution of scores for increased accountability among students, teachers, administrators, local districts, and state and federal departments of education (Macri, 2015, p. 2; Ravitz, 2002, p. 1).

In order to provide an advantageous assessment, a systematic and well-designed evaluation is paramount. Numerous instrumental music directors use a variety of technology to assist them evaluate to their students, and student use of technology enables teachers to

save class time on musical enrichment, allowing for more customized assistance and assessment (Russell, 2014, p. 11). Russell and Austin (2010) conducted research on the assessment practices of secondary music educators. Data showed that 32% of secondary teachers employed audiotapes to assess student performance (p .46). Similarly, LaCognata (2010) stated that 33% of instructors assigned students to record their performance as an evaluating strategy. LaCognata also pointed out that some band directors utilized specific music software, such as the *SmartMusic* program (13.1%) (p. 82).

The *SmartMusic* software can assess pitch and rhythmic accuracy via a computer. The *SmartMusic* program contains screen capture software and can be utilized for teaching, practice, and assessment tools. The software allows audio and visual content to be demonstrated on a screen as well as capture a recorded performance to an audio or a video file. A built-in microphone or external microphone connected to the computer can record sounds instantaneously. This function is developed by MakeMusic to not only allow teachers to give students feedback but also provide auto feedback and corrections to students.

Based on the characteristics of the *SmartMusic* program, it not only provides immediate feedback to students on their progress, but it also can be utilized for the purpose of a standardized test or a term examination as objective, formal, formative, summative, and authentic assessments. This program allows teachers to evaluate student learning outcomes and reflect upon their teaching materials, strategies, and methods, as well as support teachers in coping with the task of assessing each student within a large ensemble.

However, there are caveats regarding computer-based assessments. Based on the information from the *SmartMusic* program, there are limitations regarding its use: “Software can’t go beyond the intersection of microphones and math. That’s where you come in.” For example, the *SmartMusic* program can respond to the accuracy of performers’ rhythm and

pitch, but it cannot measure tone, phrasing, and precise intonation. These are beyond the computer's current capacity at this time.

Problem Statement

Music is primarily an aural and performing art (Shih, 2012, p.1), and music technique and skills cannot be measured using written tests such as multiple-choice or true/false exams. The problems of evaluating music performance are commonly associated with non-musical factors and the way in which music teachers assess. Non-musical factors, such as student attitude, effort, and participation, appeared to be taken into more consideration and rated more often than musical factors towards music grades (Barkley, 2006, p. 6; Keddy, 2013, p. iv). Many music teachers have been using the methods of written exams, such as multiple-choice or worksheets (Shuler & Connealy, 1998, pp. 16-17; Frankel, 2002, pp. 6-7; Zuar, 2006, p. 3), or observation to give students music grades (McQuarrie & Sherwin, 2013, para. 12). Written exams likely measure the music knowledge of students, but they do not help to measure performance skills and musical technique (Zuar, 2006, p. 3).

The subjective concern of musical performance assessments in the educational field necessitates the need for as much objectivity as possible (Bergee, 2003 pp. 137-138). A statement from the National Association of Schools of Music, National Association of Schools of Art and Design, National Association of Schools of Theatre, and National Association of Schools of Dance (1997) explained, "…evaluation of works of art, even by professionals, is highly subjective, especially with respect to contemporary work" (p. 7). The traditional human judge-based assessment has some shortcomings with regard to maintaining fair judgments, the avoidance of personal opinions, and prevention of unstable psychological incidents and situations, such as judges' preferences, both examiner- and test-taker fatigue, and the effect of high pressure during testing.

According to aforementioned phenomena, there is a need for music educators to be able to evaluate students' performing ability using reliable, efficient, and effective approaches. It is essential to use proper tools to evaluate musical technique and skills of students efficiently. Computer-based assessments have the potential to be a practical and useful evaluation tool to help music teachers. It is hoped that computerized assessments will improve the measurement of students' performance skills.

Need for This Study

Although a body of studies exists regarding the *SmartMusic* program as a music practice and instructional tool, fewer than 10 studies have investigated the efficacy of the *SmartMusic* assessment. A computerized assessment can be an alternative peripheral for music teachers and play a significant role in improving music performance assessments. Ruszkowski (2006) described,

> Innovation does not automatically lead to improvements. As new music technology is developed, it must be tested to determine its suitability as an addition to, or replacement for, current equipment. Products such as computer peripherals, electronic instruments, and recording devices, need testing to determine their practicality, validity, and reliability for use in music education and performance. Experimental testing will provide evidence of the effectiveness of a musical product, and whether that product performs to the specifications claimed by the manufacturer. (p. 1)

Unfortunately, not much quantitative research, such as Karas (2005) and Lee (2007), was found comparing computerized assessments with human examiners in evaluating music performance. Thus, there is a need to investigate whether or not computer-assisted assessments are comparable to traditional/human assessments and if computer-assisted assessments will help to improve the quality and efficiency of music assessments.

Null Hypothesis

There will be no significant difference between the music performance assessments of music adjudicators and the *SmartMusic* software.

Summary

This dissertation examines human judges and computer assisted assessments with regard to music performance. In education, a computer-assisted assessment is defined as the use of computerized hardware and/or software to measure and evaluate students' achievement and to help teachers to understand student learning. Since the inception of the launch of music assessment software a couple of decades ago, it has frequently been applied to evaluate music performances. Computer assisted assessment on music performance may be comparable to human judges, and it is foreseeable that it may one day replace human judges in rating orchestra positions, music competitions, and academic use.

The first part of this chapter focuses on the expansion of the importance of assessment, types of assessment, assessment in general education and music education, the use of technology in assessment, current problems in music assessment, and the need for the increased use of technology in music assessments. The specific problems are: (a) questionable validity and efficacy regarding music teacher assessment practices such as utilization of observations and written exams, in the evaluation of students' musical performance skills; (b) over-emphasis of non-musical factors, such as student attitude, effort, and participation, to provide music grades for students; and (c) the adverse effects of judges' subjective opinions and bias, such as performer appearance, gender, and timbre, on the accuracy of assessments.

A need to find a better alternative assessment is critical to the improvement of the reliability of music performance evaluations and ultimately, in assisting students with establishing their musicianship and skills. However, there is a lack of a significant body of

research that has explored computerized assessments versus human examiners--especially for music performance assessment.

Limitations of this Study

1. The investigator has to monitor the technological devices in the room during the performances to ensure the equipment is working properly.
2. The student participants are not all music majors.

Delimitations of this Study

1. This study of technology in music education focuses on assessment and will not cover the use of technology for instruction or practice.
2. The use of technological software in this study is the *Classic SmartMusic*, not the *New SmartMusic* Assessment System released in August, 2016.
3. The performance of this study will only include brass and woodwind instrumentalists, but other artists, such as string players, percussion players, and vocalists will be excluded.
4. The review of literature will thus only include computer-assisted musical performance assessment. Technology for teaching, learning, and practicing will not be covered in this study.
5. Although this study compares computerized assessment versus human judges of performance, the focus of the experiment is evaluating music performance rather than the details of the mechanism or science of computerized assessment.

Definition of Terms

The following terms will be used throughout this study.

***Assessment*:** From the Latin derivation, the term *assess* is "to sit beside, assist in the office of a judge" (Merriam-Webster, 1991, p. 109).

Computer Assisted Assessment: The Joint Information Systems Committee (JISC) (2010) defined Computer-Assisted Assessment (CAA) "as the application of computers to assessment processes, including delivery of tests, capture of responses and marking by either computer or human marker" (para. 1).

***Criterion-Referenced*:** Asmus (1999) stated criterion-referenced assessment as "determining the value of a student's performance by referring to a requirement that was specified prior to the student's performance of a task" (p. 21).

Evaluation: From the Latin derivation, the term of *evaluate* is "to determine or fix the value of" (Merriam-Webster, 1991, p. 429). Boyle and Radocy (1987) stated that "in education, it usually involves or at least implies of the use of tests and measurements, but in addition involves making some judgments or decision regarding the worth, quality, or value of experiences, procedures, activities, or individual or group performances as they relate to some educational endeavor" (p. 7).

***Measurement*:** Colwell (2002) stated that *measurement* has been regarded as a single test, such as the smallest unit in assessment (p. 1129). Asmus (1999) stated measurement as "the use of systematic methodology to observe musical behaviors in order to represent the magnitude of performance capability, task completion and concept attainment" (p. 21). Small in-class quizzes, homework assignments, worksheets, book reports, and research papers are considered to be examples of measurements.

***Norm-Referenced*:** Asmus (1999) explained that *norm* as "the midpoint in a set of scores taken from a large number of representative individuals where 50 percent of the scores are above the point and 50 percent are below" (p. 21) and *norm-referenced assessment* as "the value of a student's performance determined by referring to a norm established from a large number of representative individuals. This value indicates how a student performed in relation to other individuals' previous performances" (Asmus, 1999, p. 21).

CHAPTER 2

LITERATURE REVIEW

Assessing music performance is a key factor in the development of musicianship and is also a primary challenge for adjudicators and music teachers. Wu et al. (2016) stated:

> Despite its inherently subjective nature, a quantitative overall assessment is often desired, as exemplified by U.S. all-state auditions or other competitions. A model that automatically generates assessments from the audio data would allow for objective assessments and enable musically intelligent computer-assisted practice sessions for students learning an instrument. (p. 99)

In this study, the researcher investigated the use of computer assessment of music performance via the *SmartMusic* software. This chapter consists of three main sections. The first section discusses the challenges of music performance assessment. The second section will review the literature on the evaluation of music performance. The third section will review the research on computer-assessment of music performance including: (a) a review of computer evaluation of music performance; (b) music software for assessing music performance; and (c) studies on the *SmartMusic* assessment.

Challenges of Music Performance Assessment

Music performance, which entails interpreting, structuring, and physically creating music, requires a complex series of actions that include psychological, physical, acoustic, social, and artistic decisions (Palmer, 1997, p. 115-117; Widmer & Goebl, 2004, p. 203). In the assessment process, music adjudication is a synthesis of professional analysis and individual thought, impacted by physical and psychological influences at that moment. It cannot be simplified as an assessment of musical value. In the following section, controversial issues and difficulties regarding music performance assessment, such as

musical and non-musical factors, and reliability and validity of measurements, will be described first, then the related research and studies will be discussed.

A problem inherent in performance evaluation is the subjective nature of the musical performance measurement task (Abele, 1973b, p. 246; Mills, 1991, p. 176; Watkins, 1942, p. 12; Wesolowski, 2012, p. 36; Wu et al., 2016, p. 99; Zdinski & Barnes, 2002, p. 246). Mills (1991) stated that "all assessment is subjective, in the sense that human beings determine how it is done" (p. 176). However, human judges' subjective effects tend to affect the fairness of their evaluations.

Scholars have associated subjective effects with non-musical factors in assessing musical performance. McPherson and Thompson (1998) classified the factors relating to evaluating music performance into musical factors and non-musical factors. Additional musical and non-musical factors in the model affect both the musical performance and its assessment. Those musical factors include: (a) choice of repertoire, (b) form and structure of the music, (c) size of an ensemble, (d) skill of accompanying performers, and (e) type of instrument. Non-musical factors affecting the assessment could include: (a) the order in which players perform, (b) social judgments, (c) evaluator's first impression of the performer (e.g., performer attractiveness), and (d) the evaluator's characteristics (McPherson & Thompson, 1998, p.14).

In a later study, McPherson and Schubert (2004) explained that non-musical factors, such as attractiveness and first impression, are:

> Those associated to *validity* – that is, whether evaluators are actually assessing what they think they are assessing. Because non-musical factors produce unfair biases, it is important that educators, adjudicators, and researchers work toward understanding them. (p.73)

Non-musical factors affect the evaluation of music performance, and they highlight the role of subjective perspectives in the judgment of a performance.

McPherson and Schubert (2004) added an additional category, *extra-musical factors*. These *extra-musical* factors are mainly subjective and dependent upon conditions, which can be divided into three sources of extra-musical assessment enhancements:

1. Those that the player can directly control (including less obvious issues such as self-efficacy and cognitive mediation).
2. Those that depend on the playing context.
3. Those that require research about the adjudicator (p.66).

For example, expressive variation in performance is an extra-musical factor directly associated with interpretation, and it is considered deviation from the expressive norm (p.67).

McPherson and Thompson (1998) proposed a model for assessing musical performances that had been developed by Landy and Farr (1980). The model demonstrated a multifaceted set of associating factors that influenced performance and assessment, such as "context, musical and non-musical factors, evaluation instruments and criteria, performer and evaluator characteristics, and feedback to a performer" (McPherson and Thompson, 1998, p. 12).

Four major issues of performance context may influence assessing musical performance including (a) purpose of the assessment, (b) the type of performance, (c) performance proportions, and (d) performance environment" (McPherson and Thompson, 1998, pp. 12-15). First, the "purpose of the assessment" would shift the way a judge listens to or looks for the attributes of a performance. For example, the desired qualities that a judge listens for may be subject to the specific context of a music competition, end of semester exam or recital, placement exam for a music program, or audition for a position in an ensemble.

Second, the "type of performance" may alter a judge's rating, such as sight-reading, "performing rehearsed repertoire, playing from memory, playing by ear, and improvising" (McPherson & and Thompson, 1998, p. 12). Also, performance on different instruments may be evaluated in a distinct way, due to different technique and skills, as well as different repertoire. Several rating systems were created for assessing different instruments, including the *Watkins-Farnum Performance Scale* for brass and woodwind instruments (Watkins & Farnum, 1954), the *Clarinet Performance Rating Scale* (Abeles, 1973), and the *Brass Performance Rating Scale* (Bergee, 1988, 1989).

Third, the "performance proportions" might alter judgment by affecting the aesthetic goals of performance and restraining extra-musical influences on assessment. Morgan and Burrows (1981) indicated that for choral contests, large groups should confine the amount of physical movement to prevent the performance from appearing too busy and altering the balance of sound (p.47).

Fourth, the "performance environment" would affect both the performance and the assessment, including the size and acoustics of the performance space, and the availability of facility and equipment for the performer. Morgan and Burrows (1981) described that each performer may not be the same distance from a microphone (p.47). Different distances of performances from the judges may affect the sound they hear, such as whether it is loud or soft, and cause unfair and inaccurate rating (McPherson & Thompson, 1998, p.14).

In examining Landy and Farr's (1980) model, McPherson and Schubert (2004) stated that the framework for a new model is the "*Johari Window*, a pattern of awareness of behavior and motivation" (p. 75). The *Johari Window* can be classified into four areas:

1. *Public area*: an individual will be aware of some behavior and motivation that also noticeable to others.

2. *Blind area*: some behaviors and motivations will be inaccessible to the individual but accessible to others (hence, this is like a blind spot for the individual).
3. *Secret area*: the individual will hide certain motivations and behaviors, and therefore, this is the secret part of the model.
4. *Hidden area*: there is a part of behavior and motivation of which neither the individual, nor others are aware. (p. 75)

The term "others" in this application refers to an adjudicator or an audience. McPherson and Schubert (2004) described that measurement error falls into the column of the hidden area signifies where assessment factors are concealed from both adjudicator and performer. The secret and blind areas represent the sections where validity is threatened. The secret area is known to the performer rather than adjudicator; inversely, the blind area is known to adjudicator rather than performer. The public area is the only portion known to both performer and adjudicator (pp. 76-78).

In short, the *Johari Window* illustrates the complexity of evaluating music performance that entails not only musical, non-musical, and extra-musical factors but also invisible effects. While Landry and Farr (1980) only described visible effects in the old model, McPherson and Schubert (2004) depicted the invisible effects. These models explain that measurement errors and validity are strongly affected by numerous controllable and uncontrollable influences (McPherson & Schubert, 2004, p.34).

Barry (2009) stated, "the art and science of evaluation involves two basic concepts: *validity* and *reliability*" (p. 246). The evaluation must be valid in that it measures what it is supposed to measure. For instance, a student performing a series of major and minor scales might be a very valid way of evaluating certain technical skills, but it would not be the most appropriate way to evaluate the student's mastery of Baroque performance practice. In

contrast, reliability relates to the consistency of the evaluation. One aspect of reliability pertains to the consistency of one faculty member's ratings (would the same performance receive the same grade at different times). In settings that involve multiple judges, such as juries and competitions, reliability can also relate to consistency across different judges. Reliability is a ratio of agreement divided by disagreement; thus, the higher the rate of agreement among different judges, the higher the reliability (pp. 246-259).

The *Oxford Dictionary of Statistical Terms* (2003) stated, "*measurement errors* refer to the difference between an estimated value of a quantity and its true value" (para. 1). In music assessment, McPherson and Schubert (2004) denoted that "*measurement errors* mean that judge cannot behave in an idealized or machine-like way because there is always likely to be some kind of unbiased, random fluctuation that cannot be easily controlled" (p. 65). *Measurement error* is inversely related to the concept of *reliability*. While *reliability* refers to the degree of test score consistency over many replications of a test or performance task, *measurement error* reflects the discrepancy from an examinee's score over many replications (McPherson and Schubert, 2004, p.65; Meyer, 2010, p. 4).

Inter-judge reliability is the subject of numerous studies (Abeles, 1973a; Fiske, 1975, 1978, 1983; Bergee, 1978, 2003; Burnsed, Hinkle, & King, 1985; dinski & Barnes, 2002). McPherson and Thompson (1998) stated, "the assessment of music performances by adjudicators and teachers is not without difficulties; reliability among assessors is sometimes low and significant biases often influence the results" (p.12). Abeles (1973a) stated, "the replacement of judges' general impressions by ratings arrived at by more systematic procedures is one method which may improve the evaluation" (p. 145). Scholars have noted that performance evaluation, with the appropriate settings, such as valid rating scales, demonstrates good criterion-related validity and inter-judge reliability (Abeles, 1973a; Bergee, 1978, 2003; Saunders & Holohan, 1997; Zdzinski & Barnes, 2002).

Aside from developing correct assessment tools, Fiske (1983) stated that judge consistency, even among experienced judges, was as low as approximately 25% agreement (pp.7-10). A solution to the problem of inconsistent judging is to incorporate the use of a panel of judges and implement training of judges. However, in his earlier study, Fiske (1978) pointed out that judge training alone did not improve evaluation consistency (as cited in Zdinski & Barnes, 2002, p.245).

Numerous researchers and studies have examined the evaluation of music performance and found that judgments of music performance can be subjective and biased due to: (a) labelling and social prejudice, (b) visual effects (e.g., attractiveness, attire, and gender) and audio-visual context, (c) order effects, and (d) musical effects. Some claim that even when teachers are provided with guidelines for assessment, they still apply subjective opinions and judgments while giving grades (Brookhart, 1993, p. 139; Allen, 2005, p. 221). The following portion will discuss the types of value, judgments, and methods that impact music performance assessment mentioned above.

Labelling and Pre-judgements

Duerksen (1972) investigated bias via authoritative labeling in performance assessment by assigning undergraduate students to evaluate two tape recordings of an identical piano performance. The experimental group included 175 music majors and 264 non-music majors, while the control group included 78 music and non-music majors.

A control group was requested to rate two performances, labeled simply as *performance one* and *performance two*. An experimental group also was requested to rate the two performances, labeled as a *professional* and a *student* performance. Half of the experimental group heard the *professional* performance first; half heard the *student* performance first. Control subjects consistently rated *performance two* better; experimental subjects, biased by the authoritative labeling, consistently rated the *professional* performance

better, in technical as well as musical characteristics. Expectancy was such that a *professional* performance was supposed to be better; so, to the listeners, it was better (pp. 268-272).

Radocy (1976) asked undergraduate music students to assess identical performances of compositions under imposed bias conditions (p.119). Participants were assigned to a bias condition by providing them with labeling about performers or composers before the assessment. Participants in the no-bias condition (control group) were not informed anything about the performers. Participants in moderate bias conditions were told fake information about the performers. For instance, one performer might have been marked former symphonic musician while another might have been marked young graduate assistant. Participants in the serious bias condition were provided fake information about the performers and were also informed explanations regarding the performances by the professional performer were purportedly desired by previous audiences. Radocy (1976) observed a general impact of predisposition, yet found that some types' performances (e.g., piano) were more impressionable to those misleading labels than other types (trumpet, orchestra) (pp.119-128).

Visual effects

Visual effects in musical performance assessment include attractiveness, attire, and stage deportment. A series of studies regarding the effects of attractiveness on assessing music performance were conducted by Wapnick, Darrow, Kovacs, and Dalrymple (1997) and Wapnick, Mazza, and Darrow (1998, 2000).

Wapnick et al. (1997) conducted a study to determine whether physical attractiveness of singers would affect judges' ratings of their vocal performances. The subjects ($N = 82$) consisted of 33 undergraduate music majors, 32 graduate music majors, and 17 university music faculty members. Of these, 41 were male, and 41 were female (p. 472). They were

randomly divided into three groups. The visual group evaluated singers on physical effects only through watching the videotape but without sound. The audio-visual group evaluated musical performance from the videotape. The audio group evaluated musical performance from an audiotape labeled from the videotape (p. 473). Based on visual group ratings, male and female singers were classified as more-alluring and less-alluring groups. Four-way mixed-design analyses of variance (treatments by raters' gender, by performers' gender, by performers' attractiveness) were then computed for each of the seven rating groups on the rating forms (p. 474). Other results indicated that: (a) for both male and female artists, male raters were stricter than were female raters; (b) the audio-visual ratings were higher than sound-alone ratings; and (c) the ratings between college majors versus graduate students and faculty combined were not influenced by vocalists' allure (Wapnick et al., 1997 pp. 473-474).

Wapnick et al. (1998) conducted a study on the effects of performer attractiveness, stage behavior, and dress on violin performance evaluations. The purpose was to determinate whether three selected non-musical attributes would affect judges' ratings on violin performances. Twelve violin performances were videotaped from six females and six male violinists. The subjects (N = 72) were graduate students and university music faculty. They were divided into groups based on adjudication format: visual (n = 20), audio (n = 24), or audio-visual (n = 28) (p. 513). The visual group judged a performance without the sound; the audio-visual group judged with both the sounds and the visuals; and the audio group judged a performance with the sounds only. The assessors of the audio-visual and audio groups only rated non-musical characteristics. Results from the audio-visual and audio groups indicated that there were significant interactions on the test items, such as treatment by dress and treatment by stage behavior. The violinists who showed better stage behavior and wore nicer dress rewarded appreciably from videotape assessment, but violinists who did not have those staging traits were not rated differently on either audiotape or videotape. However, as for

attractiveness, less attractive violinists received lower ratings than more attractive violinists. Accordingly, there was no significant difference between the audiovisual and audio conditions. The finding indicated that more-attractive performers might have better musicianship and higher performing skills than less-attractive performers (Wapnick et al., 1998, p. 510).

In terms of attractiveness, there was no significant interaction. More attractive violinists scored higher musical performance ratings than less attractive violinists under both the audio-visual and audio conditions. This suggested to the researchers that more attractive performers may progress to a higher level in their acquisition of performance skills, than less-attractive performers (Wapnick et al., 1998, p. 518).

Wapnick et al. (2000) followed their previous study on the effects of performer attractiveness, stage behavior, and dress by examining the assessment of children's piano performances. The performances of 20 sixth-grade pianists (10 girls and 10 boys) were videotaped. The subjects (N = 123) were musically trained assessors and were divided into three groups as visual (n = 43), audio (n = 40), or audio-visual (n = 40) (p. 325).

The visual group judged a performance without the sound, the audio-visual group judged with both the sounds and the visuals, and the audio group judged a performance with the sounds only. The visual group's assessors were asked to rate the attractiveness of each performer using a 9-point scale (1 = extremely unattractive; 9 = extremely attractive) (p. 326). After the task, they answered two questions as below:

1. How important is external appearance in the evaluation of musical performance?
2. How successful do you think you would be if you rated the musical quality of a performance consciously disregarding the attractiveness of the performer? (pp. 325-326).

The assessors of both audio and audio-visual groups were asked to judge performance quality and did not rate students on non-musical characteristics. Four musical criteria included rhythmic accuracy, dynamic range, phrasing, and overall performance. The assessors were also asked to rate how talented each child seemed to be using a 9-point scale (1 = not talented at all through 9 = extremely talented) (p. 326).

Results indicated that bias affected assessors' ratings. High-attractiveness pianists were evaluated higher than low-attractiveness pianists within the audio group for all three attributes. There was a significant difference of a performer's rating based on gender ($p <$ .02) -- girls were rated equally highly across categories, but boys were rated significantly higher on dress and behavior (p. 330). Unlike results of earlier studies, videotaped performances were not rated higher than audiotaped performances. Female assessors were found to be more compassionate than male assessors. Male and female pianists were affected differently by non-musical characteristics for about half of the test items.

Min (2001) conducted a study on the effects of visual information on the reliability of evaluation of large instrumental musical ensembles. The purpose of the study is to investigate types of presentations (audio-visual versus audio-only) and its potential possible effect on reliability in evaluations of concert bands. Subjects were experienced music teachers ($N = 32$). The evaluators were asked to evaluate five band performances and were randomly assigned to rate either an audio-visual or audio-only presentation first. They then evaluated the alternative presentation after 3 weeks. The evaluation criteria included tone, intonation, balance, precision, and musical effect, along with written commentary.

The data showed that there was a statistical significance found in music effect scores by presentation type, $t = 1.97$, $p <0.00$ (Min, 2001, p. 55). The composite scores were then analyzed by the order of evaluation. The result showed that performance order had no impact on final performer scores, $p < 0.74$ (Min, 2001, p. 58). The findings indicated that different

presentation types, audio-visual and audio-only, affected the evaluation, while the order effect did not make a difference on the assessment.

Ryan and Costa-Giomi (2004) investigated how attractiveness influences the evaluation of young pianists' performances. The assumption was that both the visual and the audio components of a videotaped musical performance influence the viewer's perception of performance quality. Children, musicians, and non-musicians ($N = 75$) were asked to rate the quality of 10 piano performances from audiotapes (sound only) and from videotapes (sound and image) using the 7-point scale. Additionally, the participants rated the attractiveness of the performers from brief videos of the performers getting ready to play (pp. 141-145).

The results indicated that reliability coefficients for attractiveness rankings were high (ranging from $r = .72$ for the children's rankings of the boys to r = .92 for non-musicians' rankings of the girls), except for musicians' rankings of the girls, which yielded a lower reliability score ($r = .63$). Relatively, the data showed that the judges' musical training affected the audio-visual ratings of performance quality, with non-musicians giving higher average ratings, $M = 5.2$, than the other two groups, $M = 4.7$; $F\ (2, 655) = 7.25$, $p = .001$. However, there was no significant difference in the quality rankings of audio performances for all subjects. The factor of the performer's gender influenced the audio-visual ratings, $F\ (1, 655) = 4.39$, $p = .037$, and interacted with the attractiveness rankings, $F\ (2, 655) = 3.09$, $p = .046$). The performance level and attractiveness significantly affected judges' ratings, $F\ (4, 655) = 2.65$, $p = .032$. While attractiveness was favorable towards the best players, rather than the medium and low-level performers, the most attractive performers were given lower quality ratings than were other performers (Ryan and Costa-Giomi, 2004, p. 149).

Ryan and Costa-Giomi (2004) stated that ratings of audio-visual recordings of musical performances are evaluated more reliably than are audio recordings, but also suggested that evaluations may be affected by an attractiveness factor (p.152). The judges'

bias was found to favor-more attractive pianists among the female performers and among the best players, and less attractive pianists among male performers. It was determined that the decision to use more reliable means of evaluation (videotapes or DVDs) at the expense of favoring a particular group of performers, would have to be taken into consideration depending on the outcomes of the situation (pp. 141-151).

Howard (2012) conducted a study regarding the effect of selected non-musical factors (e.g., performance attire and stage behavior) on judges' ratings of high school solo vocalists. The subjects (N= 282), served as adjudicators, and consisted of high school choral students (n = 153), undergraduate (n = 97), and graduate music majors (n = 32). The judges rated recorded solo vocal performances using audio-only and four audio-visual presentation conditions with differentiated combinations of performance attire and stage deportment (p.166).

The results indicated that the ratings of performance quality were significantly affected by attire when singers wore formal attire, F (l) = 5,723.12, $p < .05$, and when singers utilized formal stage deportment, F (l) = 5,080.52, $p < .05$. While adjudicators' gender showed no significant difference regarding rating performance attire presentation conditions, F (l) = .00, $p > .05$, the adjudicator's academic level significantly impacted performer ratings, F (2) = 7.84, $p < .05$ (p. 174). In addition, the adjudicators gave the highest ratings to performances presented in the audio-only condition (Howard, 2012, p. 175).

Platz and Kopiez (2012) conducted a meta-analysis of audio-visual music performances, which sought to determine how strongly the visual component influences the evaluation of music performances, and to quantify the effect of the presentation mode of a music performance on the audience's evaluation. The study combined 15 existing studies with the subjects (N = 1, 2987). The meta-analysis bore an average weighted effect size of d = 0.51 standard deviations for the influence of the visual factor on music performance

assessment regarding liking, expressiveness, or overall quality of music performance. The results indicated that a random-effects model was statistically significant at the specified, a = .05 level, $z = 11.24$, $p < .00$.

Platz and Kopiez (2012) concluded:

> This meta-analysis exposed a medium effect size in evaluation behavior differences varied by the presentation mode of music performance. Considering the small range of the 95% confidence interval around the point estimator, we observed a highly precise estimation of the population effect. We conclude that the visual component is not a marginal phenomenon in music perception, but an important factor in the communication of meaning. (p. 75)

Accordingly, the study suggests that significant differences exist between music and sound in audio-visual contexts, but also implied that performance format and type affects music performance assessment.

Tsay (2013) conducted a study on the impact of visual cues on expert judgment. He argued that, "social judgments are made on the basis of both visual and auditory information, with consequential implications for our decisions" (p. 383). In this experimental study, participant responses were applied to infer the evaluation processes of the original expert adjudicators and determined what factors, either visual or auditory, were most dominant and significant for their judgments in the real-time results of live music competitions. Seven experiments were assigned utilizing various recording conditions including sound recording only, video recording only, or recordings with both video and sound as described below:

> In Experiment 1, the subjects (N=106) were asked to select one of three presentation types, such as audio, video, and audio plus video recordings, that would help them to pick the winner. Results showed that 58.5% chose sound recordings, 14.2% chose video recordings, and 27.4% chose both sound and video recordings. This indicated

that most participants the most important information to evaluate music is audio information.

In Experiment 2, the subjects (N=106) were novice participants who were provided with both video-only and sound-only presentations of 6-second clips of the top performances from international competitions. Although 83.3% of participants stated that the sound affected their assessment of music performance most, they were much more likely to identify the winners of the performances when they were provided with the visual components only. They were significantly above chance, 52.5%, at identifying the winners, t (105) = 10.90, $p < 0.00$. When participants were provided sound-only recordings, they were significantly below chance, 25.5%, at identifying the winners, t (105) = −5.23, $p < 0.00$.

In Experiment 3, the subjects (N=185) were novice participants who were provided with video-only, sound-only, or video-plus-sound versions of the performance clips included in Experiment 2. Data showed that the chance for the participants to identify the winners were: (a) with sound-only recordings, 28.8%; (b) with video-plus-sound recordings, 35.4%; and (c) with silent video-only recordings, 46.4%.

In Experiment 4, the subjects (N=35) were expert participants who were provided with both video-only and sound-only versions of 6-second clips of the top performances from international competitions. A majority, 96.3%, of expert participants stated that the sound affected more for their evaluations. While the chance, 20.5%, was lower for expert participants to identify the winners for sound-only the recordings, the chance, 46.6 % was higher for expert participants to identify the winners for silent video only recordings, t (34) = 4.05, P < 0.00.

In Experiment 5, the subjects (N=106) were expert participants who were provided with video-only, sound-only, or video-plus-sound versions of the performance clips included in Experiment 4. Most of the professional musicians, 82.3%, agreed sound was the most influential information for judgment, $\chi 2$ (2, n = 96) =103.56, $p < 0.00$. While the chances were lower for expert participants to identify the winners for sound-only the recordings, 25.7%, and video-plus-sound recordings, 29.5%, the chance was higher they would identify the winners for silent video-only recordings, 47.0%. Results indicated that experts were significantly more likely to identify the winners of the performances with video only stimuli, t_1 (61) = 4.48, $p <$ 0.00; *Cohen's d* =1.20.

In Experiment 6, it emphasized on the mechanism of the study that examined whether motion impacts the professional judgment of music performance. The subjects (N=89) were professional musicians. After seeing these 6-second silent clips of the three finalists, participants were significantly better than chance (48.8%) at identifying the winners, t (88) =6.49, $p < 0.00$.

In Experiment 7, another mechanism was examined in which the subjects (N =262) were provided with either video-only or sound-only 6-second recordings of the competition performances. The professionals were assigned to isolate the most confident, creative, engaged, determined, enthusiastic, and outstanding performer in each group of three winners in the contest. (Tsay, 2013, pp. 14581-14583)

The findings indicated that (a) passion played an important role in the professional judgment of quality through silent videos. Those selecting "the most passionate contestant" were significantly higher than chance (59.6%) to identify the actual winners; (b) visual stimulus reaches to the level that it is comparatively overweight to auditory information; and (c) human instinctive, automatic, and unconscious reliance on visual influence tends to play a

significant role in judgments (Tsay, 2013, p. 14582). In Tsay's seven experiments, one remarkable finding revealed that although auditory components are the most significant source in assessing musical performance, both experts and general audiences were affected by striking visual information in evaluating musical performance (p. 14580).

Iusca (2014) conducted a study on the effect of evaluation strategy and music format on score variability of music students' performance assessment. Iusca stated:

> Assessing students' music performance level is a multifaceted activity. Its results depend not only on the student's musical training but on a variety of other extra-musical elements related to assessment context, evaluators' characteristics or performer's personality features and psychological states. (p. 120)

Thus, the influence of evaluation strategies (global versus segmented evaluation of students' music performance) and the performance presentation types (audio versus audio-visual) on the variability of the scores was examined.

The subjects (N=50) were undergraduate music students (either strings or woodwinds) being recorded in standard conditions. Four music university professors (a flute player, a cellist, a composer, and a conductor) evaluated the recorded performance. The adjudicators and the performers were unknown to each other.

The students were asked to perform two self-selected instrumental fragments from the performers' repertory; one for demonstrating technical abilities and another for showing their musical expression. Each recording ranged from 1 to 5 minutes. The audio-video performances were converted into audio-only recordings for audio assessment.

Each expert evaluated the recordings four times. In order to diminish the learning effect, there was a one-day recess between evaluation sessions. In the first two sessions, adjudicators rated the audio recordings first using global evaluation and then a segmented scale. In the last two sessions, they rated the audio-visual recording also with global and

segmented assessment separately. For the segmented evaluation, a rating scale reflecting the factorial model developed by Russell (2010) was employed (Iusca, 2014, p. 121). The scale was created for strings, woodwind, and voice, to measure two factors: (a) technique, including tone, intonation, rhythmic accuracy, articulation, and expression; and (b) expression, including tempo, dynamics, timbre, interpretation (Iusca, 2014, p. 121).

A two-way ANOVA was computed to calculate the effect of two independent variables: measurement type (segmented versus global) and presentation format (audio versus audio-video) on music performance score variability. Both technical and expression level of the music performances were found to differ significantly based on measurement type, F (1,49) = 19.58, p = 0.00, for technique and F (1,49) = 8.93, p = 0.00, for expression. Although the interaction between the presentation format and measurement type showed no significance for the technical level of music performance, F (1,49) = 2.66, p = 0.11, the scores were higher on segmented evaluation in the audio condition. Segmented and global evaluation in Iusca's study are describe further on pages 52 and 53.

Order Effect

Flores and Ginsburgh (1996) reported evidence of bias in the Queen Elisabeth musical competition, an international competition for violin and piano organized in Belgium. They examined whether the order of appearance of a candidate had an influence on the final ranking. The data consisted of all results since the inception of the contest in 1951 (with the exception of the 1993 contest). The subjects (N = 253) included a total of 120 violinists among 10 contests and 132 pianists among 11 contests (p. 4). The results indicated that the final rank was not independent of the day in which the candidate appeared. For piano, the cross coefficient for Group1 x Day resulted with a significant difference from zero (z =. 14, SD= .07), showing that the further the day (Day 5) of performance, the larger the chances to be ranked among the first four rankings. Those who appeared during Day 1 of competition

had a lower chance of being ranked among the highest-ranking group (Group 1), while those who performed during Day 5 had a higher chance. The study indicated that rating may be due to the way the competition is organized, and suggested some changes to avoid those biases (p. 1).

VonWurmb (2013) conducted a study on the associations between conditions of performance and characteristics of performers for New York State solo performance ratings. The study examined patterns of external assessments of observed student music performances and analyzed 1,044 performance evaluations commencing solo adjudication ratings of a large suburban school district over a 4-year period (2008-2011) from the New York State School Music Association Spring Festival. The criteria of analysis in performance ratings comprised: (a) conditions of performance, including time of day of performance, level of music performed, and performance medium; and (b) characteristics of performers, including gender, race and ethnicity, and grade level (p. 39).

The data analysis showed a moderate significance on ratings regarding the time of day of performance, $F = 1.98$, $p = .07$ (VonWurmb, 2013, p. 84). While there are statistically differences regarding level of performance, $F = 44.96$, $p < .00$, racial and ethnic group, $F = 5.26$, $p = .00$, and grade levels, $F = 7.36$, $p < .00$, there are no significant difference on ratings regarding performance medium, $F = 1.59$, $p = .19$, and the gender of performance, $t = -.04$; $p = .46$ (VonWurmb, 2013, pp. 86-91).

Accordingly, three major differences were performance level, the racial and ethnic groups, and grade levels. Participants in the study could choose their performance levels by themselves. Those who chose at the highest levels V to VI of difficulty obtained higher ratings than those who chose at levels I to IV. Also, as for the differences on racial and ethnic group, Hispanic and African groups received 4 points lower than Asian and White groups. Two explanations regarding the difference of performance level included that: (a)

student performers could choose which level they perform, thus only those who believed they can do the most difficult criteria chose to perform at highest levels; and (b) there is a difference in scoring method between Levels I to IV and Levels V to VI. The author concluded that "non-significant" findings could be interpreted as evidence that performance ratings do not vary with circumstances and qualities that do not yield of the performance itself" (VonWurmb, 2013, p.1 23).

Musical Factors

Wapnick et al. (2004) conducted a study emphasizing musical factors when assessing music performance. The purpose was to investigate how evaluations of recorded solo performances would be influenced by excerpt duration (e.g., 20 versus 60 seconds) and tempo (e.g., slow versus fast). Two experimental CDs were produced. Each was made of two practice excerpts followed by 19 identical test excerpts. Evaluators (N=167) were musically trained from two universities with men (n=65) and women (n=102). They were undergraduate or graduate students in music majors (n = 135) and university music faculty members combined (n = 32) (p. 165); who had no prior training in music adjudication. Musicians evaluated recordings using six criteria, including note accuracy (NA), rhythmic accuracy (RA), tone quality (TQ), expressiveness (EX), adherence to style (AS), and overall impression (OI) using a 7-point scale (1 = good or worse through 7 = outstanding). Short pauses (5-15 seconds) enabled evaluators to complete ratings before continuing to the next item. Each experimental session lasted approximately 25 minutes (p. 167).

A five-way, mixed-design analysis of variance (ANOVA) was calculated for each of the six test items and the effects of gender, level, major, tempo, and duration. Between-subjects, variables included gender, level (undergraduate versus combined graduate and faculty), and major (non-piano versus piano). Each score was averaged over the two repeated

measures, tempo (fast versus slow) and duration (20 seconds for short versus 60 seconds for long) (Wapnick et al., 2004, p. 171).

Based on the descriptive statistics for significant interactions, the findings revealed the following:

1. Effects as tempo by gender, men rated fast excerpts lower (*Ms*= 4.66) than they rated slow excerpts (*Ms*= 4.95).
2. Effects as piano majors rated slow excerpts higher (*Ms*= 5.21) than fast excerpts (*Ms*= 4.86), and piano majors also rated slow excerpts higher (*Ms*= 5.21) than non-piano majors rated either slow, *Ms*= 4.91, or fast excerpts, *Ms*= 4.95.
3. Effects as duration by level, for undergraduates rated long excerpts, *Ms*=5.03, slightly higher than they did short excerpts, *Ms*= 4.89, but graduate students and faculty rated long excerpts, *Ms*=5.10, noticeably higher than short excerpts, *Ms*=4.64.
4. Effects as level by major, undergraduate piano majors rated performances lower, *Ms*= 4.85, then did undergraduate non-piano majors, *Ms*=5.12, but graduate piano majors and faculty rated performances higher, *Ms*= 5.21, than did graduate and faculty non-piano majors, *Ms*= 4.64.
5. A high correlation indicated that evaluators differentiated the criteria associated with each other (RA x RA, NA x NA, TQ x TQ, EX x EX, AS x AS, and OI x OI) more highly (r = 1.00) than they did cross items, r > .79 (Wapnick et al., 2004, p. 170).

Accordingly, musical factors such as duration and tempo, as well as educational level and major were found to affect judges' ratings.

In summary, bias related to subjective factors has been found to influence music performance evaluations, such as pre-judgements (Duerksen, 1972; Radocy, 1976) and

attractiveness (Wapnick et al.; 1997, Wapnick et al. 1998, 2000; Min, 2001, Ryan & Costa-Giomi, 2004; Howard, 2012; Platz & Kopiez, 2012; Tsay, 2013). Only one study contradicted these findings (Iusca, 2014) and found that instrumental music performance assessments were not affected by the physical appearance of performers.

Other miscellaneous factors have been found to impact judges' ratings. Those included: (a) musical factors, such as different tempos and various lengths of excerpts (Wapnick et al., 2004) and (b) the order effect (Flores & Ginsburgh, 1996). Although Flores and Ginsburgh (1996) stated the order of performance influenced music performance evaluations, Min (2001) claimed judges' ratings do not hinge on the order effect, and VonWurmb (2013) indicated that the time of day of performance had moderate significance.

Research on the Evaluation of Music Performance

Researchers have found it helpful to categorize music performance assessment in several ways. The following section describes: three types of tests (e.g., norm-referenced tests, criterion-referenced tests, and objective-referenced tests); the evaluation methods (e.g., segmented versus global evaluation) which are commonly used in assessing musical performance; and various rating scales with different criteria (e.g., facet-factorial rating scales and the Watkins-Farnum Performance Scale (WFPS).

In their book *Measurement and Evaluation of Musical Experience*, Boyle and Radocy (1987) identified three types of tests: (a) norm-referenced tests, (b) criterion-referenced tests, and (c) objective-referenced tests. The aim of norm-referenced tests is to discriminate or make relative comparisons among individuals' performances; however, the quality of the performance is subject to many factors (Boyle & Radocy, 1987, p. 75). Norm-referenced evaluations are usually employed in competitions and music festivals where the purpose is to rank the musicians or ensembles from most to least accomplished (McPherson & Schubert, 2004, p. 61)

Criterion-referenced testing developed in the psychometric literature in the 1960s and interest increased in the early 1970s (Boyle & Radocy, 1987, p. 76). Glaser (1963) defined that a criterion-referenced test is to compare a performance with an absolute standard, while a norm-referenced test is to make comparisons with a relative standard (p. 519). Boyle and Radocy (1987) provided an example of criterion-referenced evaluation for a music literature class: "93 % of all items answered correctly for an A, 84 % of all items answered correctly for a B, 72 % of all items answered correctly for a C, and 63 % of all items answered correctly for a D" (p. 76). From this list, individual grades will be based on how well students perform on a test. Their grades will be given based on the relation to the criteria, instead of in comparison with others (norm referenced). Criterion-referenced testing may be particularly suitable in "pass-fail" or "can do-cannot do" conditions, such as a required demonstration of precise skills and mastery technique. An example of such a condition would be whether a pianist can, or cannot play, chromatic scales smoothly. This method is commonly applied in school settings to determine how much progress has been achieved, or to determine the level of proficiency in a placement examination.

Boyle and Radocy (1987) stated that similar to the criterion-referenced test, an objective-referenced test is based on a set of goals specific to a given instructional or research setting (p. 80). In objective-referenced testing, the items assessed are indirectly related to objectives. One use of objective-referenced tests might be to evaluate an instructional program. Scores are reported to show how many subjects could answer a particular question or perform a particular task. Normally, results are registered in terms of how many test takers could perform an assignment or answer a question. Individual scores may, or may not be important (pp. 80-81). Among these three types of tests, norm-reference tests and criterion-reference tests are more commonly used in assessing music performance.

Mills (1991), however, cited a study in which a panel of evaluators was asked to rate a performance using both a global approach, which is also known as a holistic rating scale, meaning an overall rating based on adjudicators' perspectives and professional views, and a 12-category segmented assessment. She found that the segmented scheme accounted for approximately 70% of the variability between holistic structures. Mills concluded that there was no benefit to using a segmented assessment, as it may not adequately reflect the process of arriving at a holistic, overall rating. The holistic scheme was found to be more "musically credible" than the segmented assessment (Mills, 1991, p. 179). Several earlier studies on music performance assessments also indicated that inter-judge reliability was higher on global assessments than segmented evaluations (Fiske, 1975, 1977, 1983; Burnsed et al., 1985).

Saunders and Holahan (1997) investigated the suitability of criteria-specific rating scales in the selection of high school students for participation in an honors ensemble. The subjects (N = 926) were the students participating in the selection to the Connecticut All-State Band. They were judged by 36 assessors using the criteria-specific rating scales. The rating scales yielded substantial variability and moderately high to high alpha reliabilities. A Stepwise Multiple Regression was conducted, and the data indicated that student total scores could be predicted from scores of five individual dimensions (*Multiple R* = .96, $p < .001$) that accounted for 92% of the variance among the total scores on the woodwind/brass solo evaluation form (p. 270). The results indicated that criteria-specific rating scales have superior diagnostic validity.

Zdinski and Barnes (2002) addressed that "global rating scales, such as the MENC adjudication ballot (MENC, 1958), provide overall impressions on the performances, but each judge applies internal/subjective standards to evaluate an individual performance using such a scale" (p. 246). Although the issue of subjective assessment lies on a performance

measured by the judges' and observers' perspectives that make the ratings variable, a valuable benefit is that judges make their comments in each area for each performer.

Stanley, Brooker, and Gilbert (2002) investigated preferences of using global or segmented evaluations among conservatoire staff. An interview of 15 conservatorium faculty was conducted. The subjects described their experiences and views of holistic and criteria-specific approaches. The results revealed that

> Some examiners felt using criteria helped them focus on important assessment issues and that criteria were useful for articulating desirable performance characteristics in feedback to students. Other examiners believed criteria-based assessment represented a narrow view, which tended to interfere with their holistic assessments of music performance (p. 46). A majority of examiners stated they would prefer less assessment criteria for "ticking criteria boxes." Instead, they would prefer to spend more time writing more detailed comments. (p. 54)

Iusca's 2014 study examined the effect of evaluation strategy in music performance. Regarding comparing the means on the technical level of music performance, segmented measurement, $M = 5.38$, showed higher scores than global measurement, $M = 4.99$. Conversely, the expression level of music performance, global measurement, $M = 5.43$, showed higher scores than segmented measurement, $M = 5.12$ (Iusca, 2014, pp. 121-122).

The findings of the study suggested that when examining measurement type, the scores were significantly different for the global and segmented evaluations because the experts had higher expectations for technique and lower expectations concerning expression (Iusca, 2014, p. 122).

In addition to segmented and global evaluations, studies have identified and developed different rating scales for assessing music performance, such as the Watkins-Farnum Performance Scale and the facet-factorial rating scale. The Watkins-Farnum

Performance Scale (WFPS) is a standardized sight-reading assessment for all instruments. It is considered the first systematic research endeavor in solo music performance measurement, devised by John G. Watkins (Zdzinski, 1991, p.47). In his landmark dissertation, *Objective Measurement of Instrumental Performance*, Watkins (1942) developed a valid and reliable scale for the measurement of cornet sight-reading performance, called the Watkins Scale. The purposes of his study were to determine the possibility of objectively measuring achievement on a musical instrument, and to find out the relation of sight performance to practiced performance in a group of performance.

Watkins (1942) pointed out several issues of assessment on music performance as follows:

1. Most studies have been conducted in contriving ability tests than in devising in achievement tests.
2. Ability tests in other subjects have been validated against actual performance, whereas in music merely indirect measures of achievement have been presented that led the validation of the supposed ability measures on a subjective and unreliable basis.
3. Some achievement tests in music have been developed using two types: paper and pencil tests to examine knowledge of musical symbols and individual performance tests.
4. The individual performance tests alone have revealed reliabilities high enough to differentiate individuals. The group paper and pencil tests appeared to be greatly correlated with common mental ability as measured by intelligence tests than with any of the present-day music ability tests.

5. Most of the individual performance tests have been devised to evaluate sight singing, scarce work has been investigated in the problems of instrumental performance (Watkins, p. 3).
6. Watkins (1942) stated that Stelzer (1938) was the only test of instrument performance has been built and validated by modern psychometric methods (Watkins, 1942, p. 3; Stelzer, 1938, pp. 35-43).

Based on these concerns above, Watkins (1942) concluded that, "music educators and research workers in music have long needed objective measures of instrumental achievement" (p. 3).

The Watkins scale was designed to reach both musical and scientific criteria for reliability and validity via the following steps: (a) conducting a survey of cornet methods, (b) submitting the questionnaire to instrumental teachers, and (c) analyzing 23 widely known cornet methods to determine the order of introduction of music symbols based on studying weeks (Watkins, 1942, pp. 21-31). The range of difficulty was devised from a simple piece for students who had played only two weeks, progressing to a challenging piece for students who had played at least five years. The melodies were designed to measure 16 separate levels of achievement. Sixty-four exercises, plus four others found necessary, were conducted to evaluate 105 cornet students for various levels of ability (Watkins, 1942, p.8).

In order to verify the reliability of the scale, two forms of the assessment were developed, Form A and Form B, and were provided to students in instrumental music classes. These two forms were comparable in difficulty throughout the entire range. The validity of the WFPS was determined by applying rank-order correlations. The instructor ranked the students and placed the best in number one position and the others in order of their ability. After this, the students took the examination on the WFPS and received a score. A

correlation was calculated between the ranks made by the instructor and the score that students received on the scale. The testing results were:

> Based on the scores made by the 105 cases in the preliminary testing Forms A and Form B correlated .98 with each other. The internal consistency of both forms of the test was high, correlations between scores on the various exercises and scores on the entire test running between .44 to .93. Over half of these were above .80. The dispersions of the scores on the respective exercises and on the entire test were approximately the same for both Forms A and B. (Watkins, 1942, p.82)

The scale comprises a series of 14 exercises for each instrument, except snare drum, which has only 12 exercises. These sight-reading exercises were designed with increasing complexity. The authors, Watkins and Farnum (1954) attempted to make practical use of the scale for bandmasters as an objective tool for assessing instrumental students, such as selecting chair positions and giving semester grades (Lillya & Britton, 1954, p. 174).

As for the administration of the test, students would be examined simply by playing each exercise in sequence, beginning with the first exercise. A student should be stopped after making a zero score in two successive exercises. The WFPS is scored as number of correct measures performed. Each measure can only score one point no matter how many notes are in the measure, or if there is only one error or more than one error being made (Watkins & Farnum, 1954, p. 6). Measures are either correct or incorrect. No partial scores are recorded.

Although the directions above provide detailed and well-informed guidance, this assessment might greatly consume human-power and time from instructors. Each subject is graded using a scoring sheet where each measure is marked when an error occurs. The scoring is based on subtracting the number of measures marked wrong from a "possible" score given on the scoring sheet. The test ends when a subject's score is "0" on two

exercises. An entire fourteen exercises would require about 35 minutes for each individual administration. The test might also be conducted in about 10 minutes if the students are less proficient or unable to play the exercises; then, they will not be tested for more advanced exercises (Lillya & Britton, 1954, p. 174; Haley, 1998, p. 6). The WFPS is similar to a quasi-adaptive version that utilized the Rasch model, a method analyzing test results of examinees' ability by adding on different difficulty of test items (Haley, 1998, p. 5).

Stivers (1972) conducted a study to examine the reliability and validity of the WFPS. The result showed high equivalent, test-retest, inter-judge, and intra-judge reliabilities, $r > .88$ or higher. To verify the validity of WFPS, two types examinations were conducted: (a) content validity and (b) criterion related. While the correlation of content validity was $r < .89$, the correlation of criterion-related validity between teachers' own overall rankings and using WFPS were moderate, $r < .63$, except for horn ($r < .28$ for junior high school; $r < .18$ for senior high school) and saxophone ($r < .18$ for junior high school), which revealed lower correlation. The researcher stated that although the WFPS might not be a comprehensive test for evaluating musicianship and music efficiency, it provides a quantitative score reflecting music reading proficiency (Stivers, 1972, p.103).

MacKnight (1975) conducted a study on music reading ability of beginning wind instrumentalists after melodic instruction using the WFPS, Form A. The result indicated a high-reliability coefficient of ($r = .93$) for all tests for musically select fourth-grade groups. The construct of validity coefficients was ranging from $r = .64$ to .94, median= .79 for performance groups (p. 28).

Another study by Streckfuss (1983), the author asked judges to record each error committed instead of using WFPS scoring, which only counts one error per measure, no matter how many errors are made. High inter-rater reliability coefficients were reported for WFPS scores, total scores, and pitch errors, ranging from $r = .90$ to .92. However, inter-

reliability coefficients for other type errors were rather lower, $r = .85$ for rhythm, $r = .74$, for articulation and expression, as well as $r = .50$ for change of time (p. 66). These results implied that errors of rhythm, articulation, expression, and change of time are subject to each adjudicator's interpretation (pp. 48-49). Similarly, in the earlier study, Stivers (1972) mentioned that the majority of the sight-reading errors for the WFPS might depend on individual interpretation, and the test result may alter from one adjudicator to another (p. 57).

McPherson (1994) conducted a study on factors and abilities influencing sight-reading skill in music using the WFPS, Form A. The result showed a higher, $r < .98$, or similar inter-judge reliability to those reported by Stivers (1972).

Among the studies above, while high reliabilities of WFPS were found (Stivers, 1972; MacKnight, 1975; Streckfuss, 1983; McPherson, 1994) rather lower or moderate validities of WFPS were also reported (Stivers, 1972; MacKnight, 1975; Streckfuss, 1983). The scale served as a standardized test, demonstrating various levels and criteria of music performing skills with high reliability. The way of computing errors and scores is considered to be the justification for lower validities (Streckfuss, 1983, p. 66).

Abeles (1973a) examined a facet-factorial rating scale, which is a technique for development of performance rating scales for evaluating of clarinet music performance. Abeles described, "The facet-factorial approach consists of conceptualizing the behavior as multi-dimensioned and employing factor analytical procedures to select items for the scales" (p.145). A Facet-Factorial rating scale is commonly employed to measure students' achievement involving complex behaviors, not easily assessed using written tests. This is often accomplished using rating scales to evaluate certain criteria.

The criteria-based approaches are based on specific objectives that judges use to grade a performance, while subjective assessments are based on a judge's overall impression of the performance. The criteria-based approaches use a checklist to provide a score or

comments for each criterion, such as articulation, rhythm, intonation, style, and dynamics (Boyle & Radocy, 1987, p. 172). For example, a facet-factorial rating scale can be considered a criteria-based approach. Facet-factorial rating scales as a type of music assessment have been widely applied in assessing music performance. The features of music performance are complex; therefore, the evaluations of music performance consist of multiple-facets (e.g., interpretation, tone, rhythm, intonation, tempo, and articulation) or factors (e.g., the attacks and releases were clean, effective musical communication, played with a natural tone, flat in the low register, and played too slow).

Existing studies using facet-factorial rating scales include the clarinet performance rating scale (Abeles, 1973a), high school band performance rating scale (DCamp, 1980), snare drum rating scale (Nichols, 1985), euphonium and tuba performance rating scale (Bergee, 1987), string performance (Zdzinski & Barnes, 2002), orchestra performance rating scale (Smith & Barnes, 2007), aural musical performance quality measure (Russell, 2010). Rhythm and intonation are common criteria across all the instrumental performance rating scales. The rating scales and evaluating criteria of the studies mentioned above are listed in Table 1.

Table 1

The Studies Utilizing the Rating Scales

	Studies Using Rating Scales	Criteria
1	The Clarinet Performance Rating Scale (Abeles, 1973a, p. 149)	1. Interpretation 2. Tone 3. Rhythm/Continuity 4. Intonation 5. Tempo 6. Articulation
2	The Band Performance Rating Scale (DCamp, 1980, pp. 26-28)	1. Tone Intonation 2. Balance 3. Musical Interpretation 4. Rhythm 5. Technical Accuracy
3	The Snare Drum Rating Scale (Nichols, 1985, p. 30)	1. Technique-Rhythm 2. Interpretation 3. Tone Quality
4	The Euphonium & Tuba Performance Rating Scale (Bergee, 1987, pp. 95-96)	1. Interpretation/Musical Effect 2. Tone Quality/Intonation 3. Technique 4. Rhythm/Tempo
5	The String Performance Rating Scale (Zdzinski & Barnes, 2002, p. 250)	1. Interpretation /Musical Effect 2. Articulation/Tone 3. Intonation 4. Rhythm/Tempo 5. Vibrato
6	The Orchestra Performance Rating Scale (Smith & Barnes, 2007, pp. 272-273)	1. Ensemble 2. Left Hand 3. Position 4. Rhythm 5. Tempo 6. Presentation 7. Bow
7	The Aural Musical Performance Quality Measure (Russell, 2010, p.92)	1. Tone 2. Intonation 3. Rhythmic Accuracy 4. Articulation 5. Tempo 6. Dynamics 7. Timbre 8. Interpretation 9. Technique 10. Musical Expression 11. Overall Performance Quality

Boyle and Radocy (1987) found that one difficulty with specific rating scales was disagreement between judges regarding the relative importance of criteria associated with particular performance characteristics (p. 172). Boyle and Radocy recommended a balance of subjective and criteria-based decisions where the particular performance aspects function as guidance, but not necessarily as specific categories that must be quantified.

Among the aforementioned rating scales and approaches, all have their advantages and disadvantages. The WFPS is the only standardized performance assessment, while the others are non-standardized performance assessments, which serves as teacher-made scales. The benefit of using the WFPS is to make performance assessment objective rather than subjective. Most of the studies using the WFPS reported high reliability, but moderate or lower validity due to the way errors were calculated (Stivers, 1972; MacKnight, 1975; McPheron, 1994). The moderate validity of WFPS is because no matter how many errors were made, each measure only counts one error and deducts one point. This affects the accuracy of the evaluation and the validity of the test. As for other non-standardized performance assessments, almost studies that applied the facet-factorial approach indicated high reliabilities, yet it seems to require more steps and preparation in order to develop the construction of the rating scale.

Computer-Assessment of Music Performance

The development of technology has transformed educational assessment from traditional paper-based to computer-based and Internet-based assessments. Studies have indicated that the development of computer-assisted assessments was an outgrowth of computer-assisted instruction (Peters, 1974; Fukuda, Ikemiya, Itoyama, & Yoshii, 2015).

Recent rapidly-developed audio signal processing technology has not only enabled students to practice playing music instruments without a teacher's help, but also enabled error detection in a music performance. Wu et al. (2016) stated:

> Music performance analysis is a research field that involves the observation, extraction, and modeling of important parameters in music performances. Early research focused on the analysis of symbolic data collected from external sensors or MIDI devices. More recently, the focus has gradually shifted to the analysis of audio recordings. (p. 99)

The basic notions of the interaction between computers and instrument events are associated with *analog* and *digital* representations of events. Williams and Webster (1996) defined the terms as follows, "*Analog* represents events that are recorded as continuous in nature, as opposed to *digital* events that are represented as discrete steps or numbers" (p. 80). Figure 1 illustrates the signals of analog and digital events.

Analog	**Digital**
	
Continuous events	Discrete events
Smooth	In steps
Measured visually (e.g., a meter)	Measured by numbers (e.g., frequency counter)

Figure 1. Comparisons of Analog and Digital Events. Adapted from "Data Structures for Computers and Networking" by B. W. Williams and P. R. Webster, 1999, *Experiencing music technology: software, data, and hardware,* p. 81. New York: Schirmer Books.

In signal process, the conversion from analog (e.g., a sound wave; a continuous signal) to digital (e.g. a sequence of samples; a discrete-time signal) is the process of *sampling* (See Figure 2). Dubois and Thoben (2014) explained how sound travels to the computer from the outside domain via microphone:

> The acoustic pressure wave of sound is first converted into an electromagnetic wave of sound that is a direct analog of the acoustic wave. This electrical signal is then fed

to a piece of computer hardware called an analog-to-digital converter (ADC or A/D), which then digitizes the sound by sampling the amplitude of the pressure wave at a regular interval and quantifying the pressure readings numerically, passing them upstream in small packets, or vectors, to the main processor, where they can be stored or processed. Similarly, vectors of digital samples can be sent downstream from the computer to a hardware device called a digital-to-analog converter (DAC or D/A), which takes the numeric values and uses them to construct a smoothed-out electromagnetic pressure wave that can then be fed to a speaker or other device for playback. (para. 16)

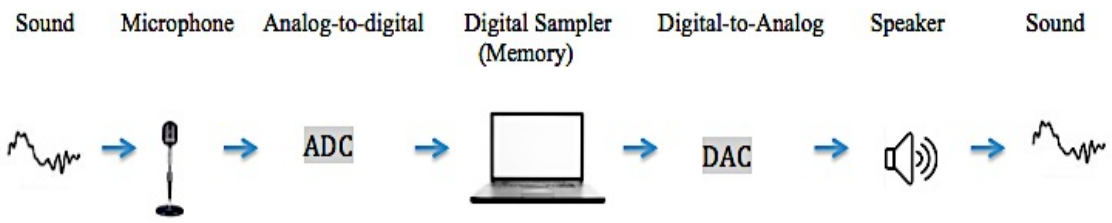


Figure 2. Sound Waves Travel from Players' Performance to Computer Map. Adapted from "Sound," by R. L., DuBois, & W., Thoben (2014). In Reas, C., & Fry, B. (2nd Ed.), *Processing: a programming handbook for visual designers and artists* (No. 6812). MIT Press.

This technology has enabled many music technology companies to develop music learning and evaluation programs. Many commercial music software programs have been designed for evaluating performance skills, including *SmartMusic, iPAS, MusicFirst, and Music Prodigy*.

The following section is divided into two parts. The first part will chronologically review the research for the computer evaluation of music performance. The second part will review existing computerized assessment programs on music performance. The third part will review research on the *SmartMusic* assessment program.

According to Zdzinski (1991, p. 55), the earliest study regarding computer-assisted

evaluation of solo instrumental music performance was authored by Peters (1974). Peters (1974) used an updated PLATO III, a computer program, at the University of Illinois to examine the accuracy of the computer interface in judging trumpet student performance for pitch and rhythm patterns presented by the computer. By applying a pulse-emitting circuit to control the rate of audio inputs from the audio oscillator, the pitch and rhythm responses could be examined (p. 114). The computer interface did not limit the length of each exercise or the number of notes in each frame. The results indicated that the computer was able to judge pitch and rhythmic accuracy (p. 160).

However, Peters (1974) pointed out some issues with the PLATO III system. First, the system was a time-sharing system and the system response was not immediate (p. 155). A time-sharing system utilizes a shared computing resource among many users at the same time. Second, although the system was designed to provide feedback, there was a lack of positive feedback. The author stated:

> The lack of positive feedback was noted early in the administration of the program. To receive any positive reinforcement, the student had to play the entire exercise correctly, melodically and rhythmically. None of the students was able to complete the exercises correctly even for the pitch or melodic judgments. The majority of errors were in intonation, i.e., playing a tone judged out of tune as opposed to a wrong note. (p. 158)

Third, although the interface was technically able to receive and evaluate fast notes, two notes trilled, and lip slurs, legato-tongued notes appeared "a problem when they approached an "interrupted long-tone" state of almost continuous sound" (p. 160). Fourth, the 2% pitch tolerance applied in the study was too extreme for the trumpet students (Peters, 1974, p. 160). The author explained that a tolerance must be allowed because a beginner cannot play a note to the exact frequency value, such as A=220 Hz, $A^{\#}$=232 Hz, B=246 Hz. However, if the

tolerance is set only ± 2% (a level beyond the exactness of the beginner), the note A (220 Hz) would be from 216.6 Hz to 224.4 Hz and A# be from 228.34 Hz to 237.66 Hz (p. 79). The range of pitch tolerance seemed to be too small and precise to obtain good responses from the system. Accordingly, the author suggested that pitch and rhythm tolerances need to have larger tolerance, and provide positive feedback (Peters, 1974, p.160).

Etmektsoglou (1992) conducted a study examining a computer-based evaluation of pitch-matching skills of college freshman students in music. The purpose was to: (a) assess the vocal pitch-matching skills of college freshman music students; and (b) investigate the relationships between these skills and selected performance skills, admission criteria used by the School of Music at the University of Illinois at Urbana-Champaign, former music experiences, and personal characteristics of students. The Computer-Based College Pitch Matching Test (CCPMT) utilized the first 20 items of the Selected Music Skills Test (SMST 20) (Etmektsoglou, 1990a) based on 77 pitches, including single pitches, intervals, and motives/short phrases derived from western tonal music compositions. The subjects (N = 38) were music education freshman students during the fall, 1990 and were individually administered for a 15-minute test.

The hardware and software applied in this study were identical with those utilized for the Computer-Based Music Skills Assessment Project (Peters, 1990b). The hardware included a Dell 286 computer (IBM compatible), Roland MPU-IPC MIDI Card and Cables and Casio MT-240 MIDI keyboard, an AFI 101 pitch- board and cables, and Sony Microphone/Pre-amplifier. The microphone input was transferred to digital MIDI data for student performance assessment. The Dell 286 computer managed the Casio MT-240 MIDI synthesizer, which performed the test items on the preset piano sound. The students were required to repeat every test item using clear articulation and the syllables “tah,” “lah,” or an

alternative syllable of their choice. One percent discrimination of accuracy was applied into data analysis, and "sampling rate of the AFI interface was 1500 samples/second."

The computer was set to detect a sung tone as incorrect when it showed sharp or flat by over a quarter tone or ± 50 cents. The method of scoring was strict. A perfect score would be 77 points, while a potential minimum score was zero. The results showed a wide range of scores from 10 points to 72 points that indicated a high discriminating capacity of CCPMT, M=46.8 and SD = 11.9.

In order to examine the reliability of CCPMT, Etmektsoglou utilized her earlier test (1990a) with the off-line SMST (31) based on 188 pitches to examine 15 participants. They were evaluated by four human judges. Data analysis showed high coefficients for all possible pairs of judges ranged, r = .85 and r = .91 (Etmektsoglou, 1992, p.78). However, while comparing scores given by the music experts and by the computer, a low correlation coefficient, r = .31, ^{r} *ranks* = .18, was found between the judge panel scores and the computer scores (p. 81). The findings of the study indicated that the computer evaluated the response in a very precise approach with much less tolerance than human judges. This was a major distinction of scoring between the computer and the experts.

Fukuda et al. (2015) presented an innovative piano tutoring system that boosts the practice of student pianists by simplifying difficult parts of a musical score based on individual performance skill. By referring to the musical score, the system is theoretically able to detect errors of a performance; identify the difficult parts and then, subsequently, simplify the music. The process involved the following:

1. The audio recording of the user's performance is converted by applying a supervised non-negative matrix factorization (NMF). NMF is a dimension reduction method. The basis spectra of NMF are trained from isolated sounds of the same piano in advance.

2. The audio recording is synchronized with the musical score via the dynamic time warping (DTW), which is an algorithm system for measuring the similarities between two temporal sequences that may differ in speed. The user's errors are then detected by comparing those two kinds of data.
3. Finally, the detected parts are simplified according to three kinds of rules:
    a. Removing some musical notes from a complicated chord.
    b. Thinning out some notes from a fast passage.
    c. Removing octave jumps (Fukuda et al., 2015, p. 1).

The results indicated that the system is capable of transcribing the audio input with high accuracy, and marking the discrepancies between the score and the performance with octave errors. The suggestion recommended by the researchers include conducting more experiments, improving each algorithm, and improving score simplification (Fukuda et al., 2015, p. 4).

Music Software for Assessing Music Performance

This study investigated whether the scorings of computerized assessment and human judges are comparable. Several computerized assessments of music performance have been developed using score input as preceding knowledge to detect performing errors. The following section will describe existing music software for assessing music performance.

Music Prodigy is an interactive practice and assessment software, which is completely digital and cloud-based. *Music Prodigy Core* is a music technological-educational tool that offers immediate feedback for performance accuracy, with polyphonic pitch recognition (multiple notes). The program does not require an external microphone. The aim of this software is to improve student performance and learning outcomes. The platform also provides *Music Prodigy Quiz*, which is a comprehensive evaluation tool for general music students, ensemble students, and university students. This allows music teachers evaluate

students on musical knowledge and performance in one assessment, including questionnaires in various formats, from multiple-choice to audio identification and performance evaluation.

Similarities between *Music Prodigy* and *SmartMusic* include that they both assess students' rhythms and pitches in real-time, give immediate feedback, and offer a rich library where students and teachers have the capability to access a variety of repertories, method books, and exercises for instruments and voices. The major differences are 1) *Music Prodigy* is compatible with either *Finale* or *Sibelius* music notation systems, and 2) *Music Prodigy Quiz* enables teachers to assess not only students' performance but also their music knowledge.

The *iPAS* (Interactive Pyware Assessment Software, n.d.), from Pygraphics, Inc. is an online assessment software. The software provides on-screen music notation and an automatic guidance system to help students complete an assigned exercise outside of school, as well as guide students through a pre-arranged course of practicing. Jacoby (2014) pointed out that as student logs into the program, it will begin with a practice tip, such as "Good posture is important for a good sound" (para, 9), and then students select a chosen exercise or assignment. The practice procedures are the following:

1. Listens to the assigned exercise or musical passage and then plays that assignment into a microphone.
2. The software immediately evaluates the performance with an assigned score.
3. The performance and results may be sent to the student's teacher for grading.

Jacoby (2014) described that the package of *iPAS* includes *Pearson's Standard of Excellence*, a method book. The software has Mac OS and Windows versions available. A teacher's edition of *iPAS* consists of an assignment and gradebook system, as well as a tool for creating custom content (from *Sibelius*, *Finale*, or any MIDI files) for the *iPAS* system (para. 8). The package of *iPAS* includes *Pearson's Standard of Excellence*, a method book. The software

has Mac OS and Windows versions available. A teacher's edition of *iPAS* consists of an assignment and gradebook system, as well as a tool for creating custom content (from *Sibelius*, *Finale*, or any MIDI files) for the *iPAS* system.

On screen, the *iPAS* display of notation is a mixture of a five-line music staff "with the piano-roll view in many MIDI recording programs" (Zanutto, 2007, p.5). However, the display is unusual and may be unfamiliar to users. Also, *iPAS* does not have clear and modern graphic design.

The *SmartMusic* Assessment

SmartMusic, created by MakeMusic, Inc., is an interactive music learning software system, which provides automatic music accompaniment and immediate feedback. As commercial music software, *SmartMusic* can be purchased as an annual subscription. Teachers can create a class or course online, and students can access the course assignments designed by their teachers to receive practice materials tailored to their specific needs. This allows students to receive individualized instruction and assistance outside of school. *SmartMusic* is available to band, string, and vocal students of all ages and skill levels, and is supported on iPad, PC, and Mac. The following sections will provide brief background information about the development of *SmartMusic* technology.

MakeMusic, Inc. built in 1990 develops and sells proprietary music technology solutions under the *Finale* and *SmartMusic* brands (Motiwala, 2011, para. 1). The functions of the software include: (a) *Finale* as a notation software; (b) *SmartMusic* as alternatives to traditional accompaniment, practice, instruction, composition, and assessment tools; (c) *MusicXML* as a standard open format for exchanging digital sheet music and files; and (d) Garritan's virtual instruments.

The software was introduced in 1990, and known as "a hardware-based intelligent-accompaniment product called 'Vivace'," (Rudolph, 2006, p. 10). The early versions were

costly due to the fact that the system consisted of hardware, software, and its library of repertoire.

In April 1998, the company introduced a new and renamed version of the *Vivace* Practice Studio product, *SmartMusic* Studio, (Motiwala, 2011, para. 16). The website was launched and began marketing *SmartMusic* on a licensed subscription basis in the United States in December 2001, with a price scale ranging from $90 per year for the first subscription down to $20 per year for the fourth and more subscriptions (para. 17). The maturity of *SmartMusic* was not fully realized until the hardware was replaced by a web-based version in 2002. Each teacher subscription came with the *SmartMusic Gradebook* to upload assignments to students and download completed assignments from students, along with grading and managing student records.

In 2005-2006, a beta-testing of the software was conducted to verify that the product provided an accessible and friendly service for music instructors, such as electronically sending *SmartMusic* assignments to students and automatically receiving *SmartMusic* assessed grades and recordings of the performances, and managing student grades. During the period 2007-2008, *SmartMusic* 10.0 was released in April 2007, and the *SmartMusic Gradebook* was renamed *SmartMusic Impact* (Motiwala, 2011, para. 19).

The development of applications was completed by the internal team of *SmartMusic* software programmers and testers. MakeMusic (2009) stated that the *SmartMusic* application coordinates includes:

1. Playback of music, either synthesized or audio.
2. Display of music notation on screen with Finale technology.
3. Use of a microphone attachment to record a student's performance.
4. Recognition of notes and rhythms and comparison of a student's performance to what is notated.

5. Communication of errors and correction techniques to students.
6. The support of a growing selection of skill-development features that accelerate student learning. The patented feature, Intelligent Accompaniment, allows student to develop their skills of expression for solo literature (p.5).

MakeMusic explained that the major features of the latest version of *New SmartMusic* (2016) are: (a) it is completely web-based so that students can access the new *SmartMusic* anywhere that has an Internet connection and on any device without installation, (b) when zooming in or out, music is constantly resumed and intelligently active on screen, and (c) it can assess polyphonic performance such as intervals and chords.

Three functions of *SmartMusic* are guided practice, assessment and documentation, and providing a library of repertoire. Lou, Guo, Zhu, Shih, and Szan (2011) identified the program as computer-assisted musical instruction (CAMI) for "interactive, adaptive, learner-controlled, inexhaustible, and unlimited in time, space, and manageability" (cited in Lou et al., 2011, p. 46). Nicole (2014) stated that *SmartMusic* is an interactive music practice system with automatic music accompaniment (pp. 4-5). It also includes a computer-assisted music assessment component that provides instant feedback as well as documented progress (Walls, Erwin, & Kuehne, 2013, p. 9).

The process of the program comprises:

1. Students practice exercises and songs from the repertoire library of *SmartMusic* or upload *Finale* music file.
2. Users can control tempo, key, practice loops, tuner, and more.
3. *Intelligent Accompaniment* follows and responds when students perform.
4. *SmartMusic* assesses student performance and gives immediate feedback on screen.

5. Students can record their performance and submit assignments to instructors or burn to CD.

Aside from downloading the *SmartMusic* software on a PC or Mac, it requires an Internet connection to activate a subscription, receive assignments and submit recording online.

SmartMusic software utilizes several technological programs: (a) IRCAM Real-Time Musical Interactions for automatic accompaniment. The real-time interactive system can transform sound, voice, gestures, memory, and create dialogues between artists and digital media; (b) Audio recordings can be synchronized with the musical score via the dynamic time warping that measures the similarities or differences between two temporal sequences that may differ in speed; and (c) Sound waves allow the computer to detect pitch errors by comparing the original scores with a performance. Through these applications, computers can sense general characteristics such as register, loudness, or density, and can also do score following, which involves moment-by-moment estimations of a performer's tempo.

In August 2016, the new *SmartMusic* program was released and available for purchase. However, the new version is a cloud-based tool for Chromebooks and iPads, which require access to the Internet, and nothing is downloaded to the computer. This study will use the classic version of *SmartMusic*. In this dissertation, use of the term *SmartMusic* will refer to *Classic SmartMusic*.

The Comparison Between the *SmartMusic* and the *iPAS*

Zanutto (2007) compared two online music assessment programs: (a) the *Finale Performance Assessment* (*FPA*) system (now *SmartMusic*); and (b) the *Interactive Pyware Assessment System* (*iPAS*). In this section, *FPA* was used for the *SmartMusic* assessment. The testing results were examined through a brass methods course at California State University, Long Beach, and selected K-12 secondary brass students. The results of field test trials were more accurate and complete with *iPAS*, and inconsistent grading was found with

SmartMusic. Certain conditions may cause unreliable scores, such as poor microphone quality or placement, instrumental tone quality or volume, and surrounding noise. *IPAS* scores for pitch, rhythm, and intonation are illustrated along with a composite total score. *FPA* and *iPAS* both utilized different pitch recognition drivers or software.

The *FPA* "had a significant upgrade that *SmartMusic* now uses, which is the IRCAM pitch recognition engine; reportedly superior to the previous pitch recognition drivers of the FPA program (p.4)." Compared to the previous *FPA* format, the S*martMusic* assessment presented fewer errors, but sensitivity to articulations (i.e. detached vs. legato) remained insignificant. *IPAS* software designers created their own proprietary algorithm for pitch and rhythm recognition with a higher sensitivity to articulations (p.4).

Buck (2008) conducted his dissertation research on *The Efficacy of SmartMusic® Assessment as a Teaching and Learning Tool*. He also pointed out a comparison of the *SmartMusic* assessment with *iPAS* and noted that the *iPAS* assessment feedback includes pitch, rhythm and intonation, while the *SmartMusic* assessment provides pitch and rhythm information but allows severe intonation discrepancies (p. 31).

The Studies Regarding the *SmartMusic* Assessment

Karas (2005) investigated the effects of aural and improvisatory instruction on fifth-grade band students' sight-reading abilities using *SmartMusic* (2004) to measure accuracy of rhythm and pitch. Karas (2005) reported results of testing of the reliability and validity of the measurement. Data showed an acceptable positive correlation, $r = .71$, between the composite score (tonal and rhythm accuracy) for the four measures played a first, then a second time, as scored by *SmartMusic*. The size of this coefficient may have been affected by the fact that *SmartMusic* is not tolerant of dropping or skipping beats when repeats are notated. The internal reliability of *SmartMusic* was estimated to be acceptable for all three analyses, pitch $\alpha=.83$; rhythm $\alpha= .84$; composite $\alpha=.88$ (p. 58). Karas (2005) postulated that

"reliability was established by the technology developers for *SmartMusic*. They recorded examples and played them through the *imic* as if a student was being assessed. The developers reported a high test-retest correlation" (Scheffing, D personal correspondence, April 29, 2005 as cited in Karas, 2005, p. 60).

Karas (2005) used a four-judge panel to determine the validity and reliability of the *SmartMusic* assessment. Inter-judge reliability was measured using the following steps: (a) each printed example of the notation was given to the human judges; (b) the judges were asked to mark errors with one color for incorrect pitch and another color for incorrect rhythm. The correction of inter-judge reliability was found to be high, from $r = .87$ to .96. With inter-judge reliability established, the Spearman Rho method was computed to measure the correlations to *SmartMusic* scores. There was a statistically significant difference between the *SmartMusic* assessment and that of the four-judge panel. A strong concurrent correlation, the ratings of four performance examples ranging from $r = .60$ to $r = .86$, indicated between the grades of *SmartMusic* and judges, shows the *SmartMusic* assessment appeared to be effective in scoring the sight-reading ability of students (p. 59).

Lee (2007) investigated the effects of *SmartMusic* on computer-assisted instruction, previous experience, and time on the performance ability of beginning band students. Cronbach's alpha was computed to estimate validity and reliability. For the reliability analysis, three eighth-graders were assigned to play the test excerpt three times each in order to examine the scoring system of the *SmartMusic* assessment. The coefficient alpha, $a = .91$, indicated that the *SmartMusic* was a reliable testing tool (p. 56). Lee stated, "There was no reason to suspect that the reliability for other exercises in the *SmartMusic* system would show markedly different patterns, therefore, the reliability of the measure as a whole would be deemed the acceptable" (pp. 14-15).

Lee (2007) examined the validity of the *SmartMusic* assessment comparing three local band directors' scoring and the scoring of the *SmartMusic* assessment. The data indicated a high correlation, $r = .93$, between the three-judge panel and the comparisons of the judge's composite scores to the *SmartMusic*, $r = .91$. The validity of the program was considered to be acceptable. Lee stated, "There is no reason to suspect that scoring for this program and other live judges would show markedly different results, therefore, validity of the *SmartMusic* instrument as a whole was deemed acceptable" (p.56).

While Karas (2005) and Lee (2007) supported that the validity of the *SmartMusic* is considered to be acceptable, Long (2011) claimed, "the *SmartMusic* assessment feature is not as comprehensive as a human judge" (p. 42). Long (2011) examined the features of *SmartMusic* to determine the effectiveness of the software for student trombonists. His study was not a quantitative study and did not obtain statistical results. Instead, the trombone students engaged in an evaluation discussion, along with an examination of the essential criteria as follows:

> A trombone etude was performed seventeen times, and one element was changed to compare the original one. Visual criteria in the study included the advantages and disadvantages to having a blind evaluation. Aural criteria included subjective and objective elements in five categories of brass performance evaluation including articulation, rhythm, tone, intonation, and musicianship/style as presented in Wardlaw's (1997) *Performance Rating Scale*. The purpose was to focus upon one component of the evaluation each time and to see how each change affected the assessment feature's assigned grade for each performance. (p. 2)

Prior to examining 17 performances experiments of *SmartMusic* assessment, an original perfect performance was implemented, and the performance was scored 100%. The rest of the 17 performances were divided into several testing sections, and students were assigned to

create specific musical performance errors during each *SmartMusic* assessment session. The details and results are described below:

Section I: The articulation testing

1. Playing imprecise tonguing style similar to a slight glissando: *SmartMusic* graded this performance 100%. This indicated that the *SmartMusic* assessment did not measure and deduct points for imprecise articulation.
2. Playing without using any tongued articulation on any notes throughout the performance: *SmartMusic* again graded this performance 100%, which indicated that it did not measure tonguing or note distinction.
3. Playing using flutter tonguing throughout the entire etude: *SmartMusic* graded this performance 97%. The 3% deduction was a result of the black D-flat that immediately followed another D-flat. This note was the only note in the etude that was the same as note that immediately preceded it. Despite the rapid flutter-tonguing articulation that was inappropriate for this etude, the *SmartMusic* assessment feature did not deem any other notes incorrect throughout this performance.
4. Playing the notes as short as possible with tongue cutoffs to end each note: *SmartMusic* graded this performance 100%. *SmartMusic* did not deduct points for this incorrect technique, nor did the assessment feature deduct points for releasing each note abruptly with the tongue.

Section II: The rhythmic testing

5. Playing every note noticeably late: *SmartMusic* graded this performance 14%. Despite the fact that the subject played all of the correct pitches in tune with clean articulation and appropriate style, consistent playing behind the beat reduced the score by 86%.

6. Playing constant eighth notes on the correct pitches: *SmartMusic* graded this performance 100%. Eighth notes were still played as eighth notes, but quarter notes became two eighth notes, half notes became four eighth notes, and so forth. *SmartMusic* graded this performance 100%. Despite the fact that the subject rearticulated notes that were supposed to be held, the *SmartMusic* assessment feature did not deduct points for adding repeated notes.

Section III: The music style testing

7. Playing swinging the eighth notes in a jazz style throughout the etude: *SmartMusic* graded this performance 83%. The evaluation did not display any red notes for this performance; however, most of the eighth notes on the "and" of the swing rhythm registered as black notes.
8. It did not "hear" these notes played in context of the etude.

Section IV: The timbre and tone quality testing

9. Playing the etude with a poor tone quality: *SmartMusic* graded this performance 100%. An uncharacteristic trombone sound did not disqualify any note.
10. Singing through the microphone rather than playing the trombone: *SmartMusic* graded this performance 93%. The 7% deduction was attributed to intonation flaws in the singing; this deduction was unrelated to timbre.

Section V: The intonation testing

11. Playing every note one partial too high throughout the etude.
12. Playing all of the notes one partial below the correct note: *SmartMusic* graded this performance 0%.
13. Playing the entire etude an octave higher than the indicated notes: *SmartMusic* graded this performance 0%.

14. Starting each note in tune but then quickly bending each note noticeably sharp or flat for the duration of the note: *SmartMusic* graded this performance 97%. The 3% deduction occurred on a G-flat that the subject quickly sharpened.
15. Pulled the tuning slide out as far as possible prior to starting the etude. *SmartMusic* graded this performance 93%. Although the *SmartMusic* assessment feature did not display any red or black notes after this performance, the 7% deduction was most likely due to notes that were so flat that they exceeded the *SmartMuisc* assessment feature's pitch parameters.

Section VI: The dynamic testing

16. Playing the etude very loud instead of the indicated *mezzo piano* dynamic level: *SmartMusic* graded this performance 100%.
17. Playing more expressively than on the other recordings by making noticeable dynamic contrast: *SmartMusic* graded this performance 100% (p. 32-35).

Long (2011) hypothesized that human judges have the capability to evaluate both objective and subjective performance criteria, whereas computerized assessments are limited to objective criteria because a computer collects quantitative rather than qualitative data. Long (2011) claimed, "teachers who promote the *SmartMusic* assessment feature and students who use the *SmartMusic* assessment feature must realize that this feature is not put to proper use when the grading feature is used as a substitute for human evaluation" (p. 42).

Researchers have investigated the effects of using the *SmartMusic* program, and several studies have indicated that the program is a useful and reliable tool for learning music (Karas, 2005; Lee, 2007; Zanuto, 2007; Flanigan, 2008; Astafan, 2011; Nielsen, 2011; Macri, 2015). Among those studies, several have explored the effectiveness of the *SmartMusic* Interactive Practice Software (Flanigan, 2008; Nichols, 2014; Macri, 2015), and some studies have investigated the perspectives of students and teachers using the *SmartMusic* software

(Zanuto, 2007; Macri, 2015). While other researchers have examined the effects of computer-assisted music instruction and practice using the scoring of the *SmartMusic* assessment as dependent variables (Karas, 2005; Lee, 2007; Astafan, 2011), they emphasized the program as a practicing and teaching aid more than an assessment tool. One study was found (Long, 2011) that specifically focused on the usage of the *SmartMusic* assessment. The following sections will provide a review of the reliability and validity of *SmartMusic* assessment and comparisons between the reliability and validity of human judges and the *SmartMusic* assessment.

Based on information from the aforementioned study, the reliability of the *SmartMusic* assessment has been examined by Karas (2005) and Lee (2007), and the results indicated that the *SmartMusic* assessment is a reliable testing tool. However, according to the other studies and information obtained by the publisher, *SmartMusic* does not measure timbre, dynamics, articulation, style, phrasing, or expression (Long, 2011, p. 29-33; Buck, 2008, p. 17). Buck (2008) claimed, "allowances for differing musical styles, i.e. legato, staccato, etc., have not been made in previous studies, though results typically note particular effects on rhythm and pitch." (p.17). In addition, the *SmartMusic* program did not appear to be very useful for assessing higher-level musical skills (Zanutto, 2007, p. 1) because the software can respond to the accuracy of performers' rhythm and pitch, but it cannot measure tone, phrasing, or precise intonation. These are beyond its capabilities.

Summary

Based on the review of literature, numerous factors have been found to influence conventional music performance evaluation including:

1. Labeling by authorities (Duerksen, 1972; Radocy, 1976).
2. Attractiveness (Wapnick et al., 1997; Wapnick et al. 1998, 2000; Min, 2001; Ryan & Costa-Giomi, 2004; Howard, 2012; Platz & Kopiez, 2012; Tsay, 2013;

Iusca; 2014).

3. Order effect (Flores & Ginsburgh,1996; Min, 200; VonWarmb, 2013).
4. Music factors: different tempos and various lengths of excerpts (Wapnick, et al., 2004).

Although more than a half century ago, Watkins (1942) called for a need to improve musical evaluation, arguing that, "music educators and research workers in music have long needed objective measures of instrumental achievement" (p. 3), objectively evaluating musical performance remains a challenge to the present day.

Over the decades, technology for computerized musical performance assessments (e.g., *SmartMusic* and *iPAS*) has developed to improve musical performance assessment. A number of studies have shown positive correlations between the *SmartMusic* assessment and human judges (Karas, 2005; Lee, 2007). However, Long (2011) argued that the *SmartMusic* assessment is not able to assess music performance as effectively as music experts (p. 42).

In Long's study, he examined the efficacy of the *SmartMusi*c assessment using a qualitative research. This study will take Long's (2011) recommendation to conduct a statistical and quantitative study based on the *SmartMusic* assessment versus human judges. It will also investigate the capacity of the program to measure music performance, as well as compare the scorings between the *SmartMusic* program and human experts using rating scales with music criteria including pitch, rhythm, and tempo. Other factors will be taken into consideration based on recommendations of existing studies including the use of a larger sample size (Lee, 2007; Nicole, 2014, p. 29).

CHAPTER 3

METHODOLOGY

The purpose of this study was to investigate the differences between the *SmartMusic* assessments on music performance and assessments by human experts. Permission to conduct this study was endorsed on March 29, 2017 by the Institutional Review Board (IRB) at University of Hawai'i-Mānoa (See Appendix A), and the modification was approved on October 3, 2017 by the IRB (See Appendix B). The modification included: (a) the number of the judge panel may increase due to the risk of absent judges, and (b) the generalizability theory based on an ANOVA calculation will be applied for analytical framework. This chapter provides information concerning the (a) problem underlying the study, (b) study design, (c) procedure, and (d) analytical framework and the pilot study.

Underlying Problem

The *SmartMusic* software has been used as a music performance assessment tool, yet its reliability as a tool to apply to assessment has not been adequately researched. Studies on the *SmartMusic* assessment indicated that pitch and rhythm are the criteria evaluated most accurately (Zanuto, 2007, p.1; Buck, 2008, p.17; Long, 2011, p.29-33). Other musical criteria, such as articulation, timbre, tone quality, intonation, phrasing, and dynamic, are not measured as accurately as expected or not measured at all by the program (Long, 2011, pp. 27-35).

Karas (2005) and Lee (2007) supported the reliability and the validity of *SmartMusic* and considered it to be an acceptable assessment tool. However, Long (2011) claimed that, "the *SmartMusic* assessment feature is not as comprehensive as a human judge" (p.42). While the *SmartMusic* assessment is marketed to detect incorrect notes in terms of pitch and rhythm, human judges often rate a performance using additional criteria, such as articulation, tone, dynamic and musical expression. Music scholars, such as Mills (1991), and several

studies, have argued that bias and subjective judgments are challenges for human judges while assessing music performance (Duerksen; 1972; Radocy, 1976; Wapnick et al., 1997, Wapnick et al., 1998, 2000; Platz & Kopies, 2012).

Study Design

Subjects

A quasi-experimental design was selected to test the null hypothesis. The study utilized undergraduate instrumentalists who enrolled in the woodwind and brass ensembles at the University of Hawai'i-Mānoa in Honolulu. Once the *Invitation to Participate* was accepted and received, each subject was required to sign a consent form before participating in the study (See Appendix C). Thirty-eight of University of Hawai'i-Mānoa (UHM) undergraduate instrumentalists were recruited to participate in the study by performing two sight-reading exercises. Two subjects dropped out, and two recordings were not recorded throughout the entire piece due to playing slower than the indicated tempo. The final subjects of this subject were (N=34) performers.

Materials

The Watkins-Farnum Performance Scale (WFPS) is a standardized sight-reading scale for assessing instrumental performance. The justifications for using the WFPS as the sight-reading exercises included: (a) the scale serves as a standardized music performance, (b) the scale has been used for many studies (Stivers, 1972; MacKnight, 1975; Streckfuss, 1983; and McPherson, 1994), and (c) the music exercises are unfamiliar to the subjects. Two exercises of WFPS Part A, No.5 and No.6, were selected from the WFPS for this study. Exercise No.5 (See Figure 3) was employed as a warm-up exercise and for the use of the pilot study; Exercise No.6 (See Figure 4) was utilized for the real assessment.

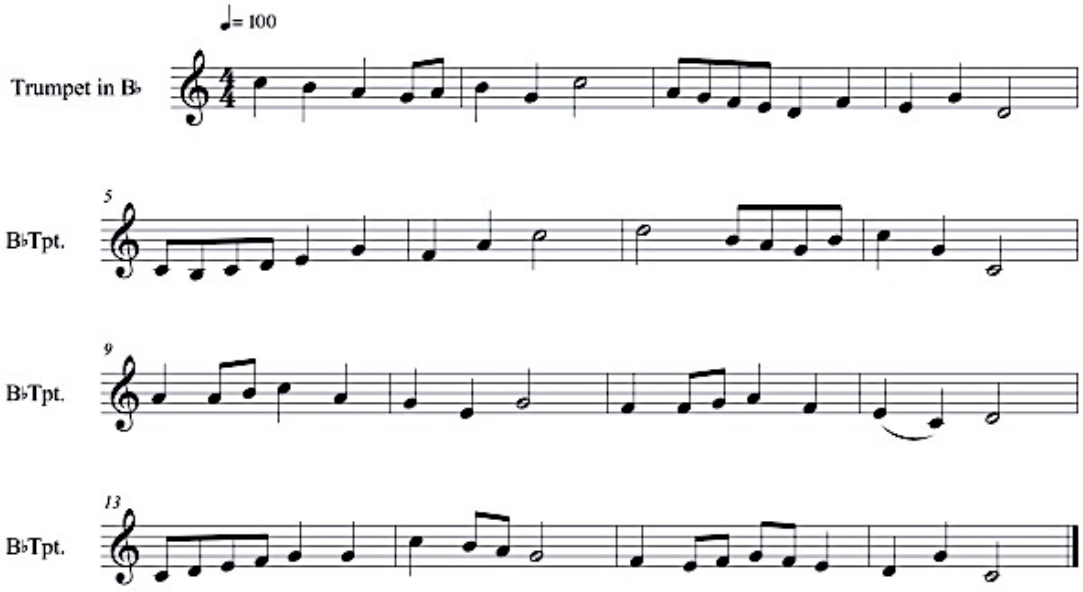


Figure 3. The WFPS Exercise No. 5

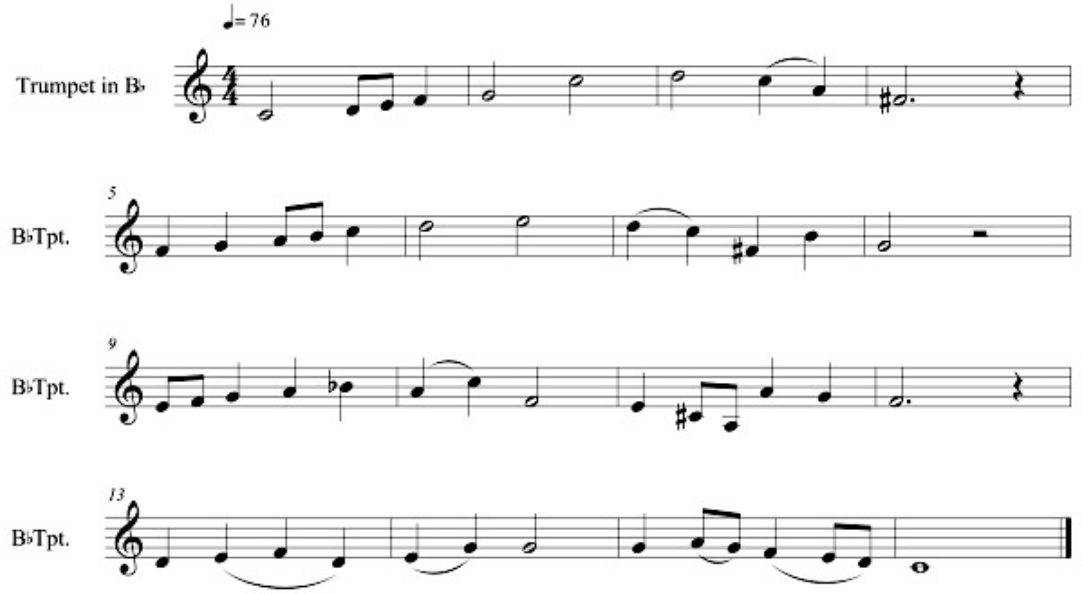


Figure 4. The WFPS Exercise No. 6

The *SmartMusic* program and the *Finale* software

This study employed two software applications including the *Finale* and the *SmartMusic* programs, as well as two types of the *SmartMusic* accounts, including an educator's account and a practice room account. The educator's account was applied for (a) creating a course to assign exercises to each subject and (b) uploading finale-created files/downloading music files from the *SmartMusic* library. In order to make sure each assessment was being conducted through the same procedure, 34 practice room accounts were created for participants by the researcher to obtain the scores of the *SmartMusic* Assessment (i.e., computerized assessment and recordings). The participants did not handle any computer process. The task for participants was to come to a classroom in the music building at the University of Hawaii at Manoa to perform two sight-reading exercises from two hard copy music sheets using the *SmartMusic* program for assessment and recording.

Prior to assessment, each exercise was composed and transposed to the keys for brass and woodwind instruments and adjusted for range thorough the *Finale* software. Each instrument file was then uploaded to the educator's *SmartMusic* library within an educator account. After uploading the finale-created files, the researcher clicked selected files to create assignments for multiple subjects on multiple instruments (e.g. tuba 1, tuba 2… and trumpet 1…).

Figure 5 demonstrates a format of *SmartMusic* assignment creation screen for setting up the assignments. The researcher set the parameters of each exercise as in the example below:

1. Grade style: standard 100 points with recording checked.
2. Tempo: 76.
3. Click: off (4 beats count off, e.g., click-click-click-click).
4. My part: off.

5. Cursor: off.
6. Music on screen: on.
7. Accompaniment: off.
8. Voice count: off.
9. Click Accent Down beats: off.
10. Click Play Subdivisions: off.
11. Range: As published.
12. Assign to students on: (e.g. mm/dd/yy).
13. Due date: (e.g. mm/dd/yy).

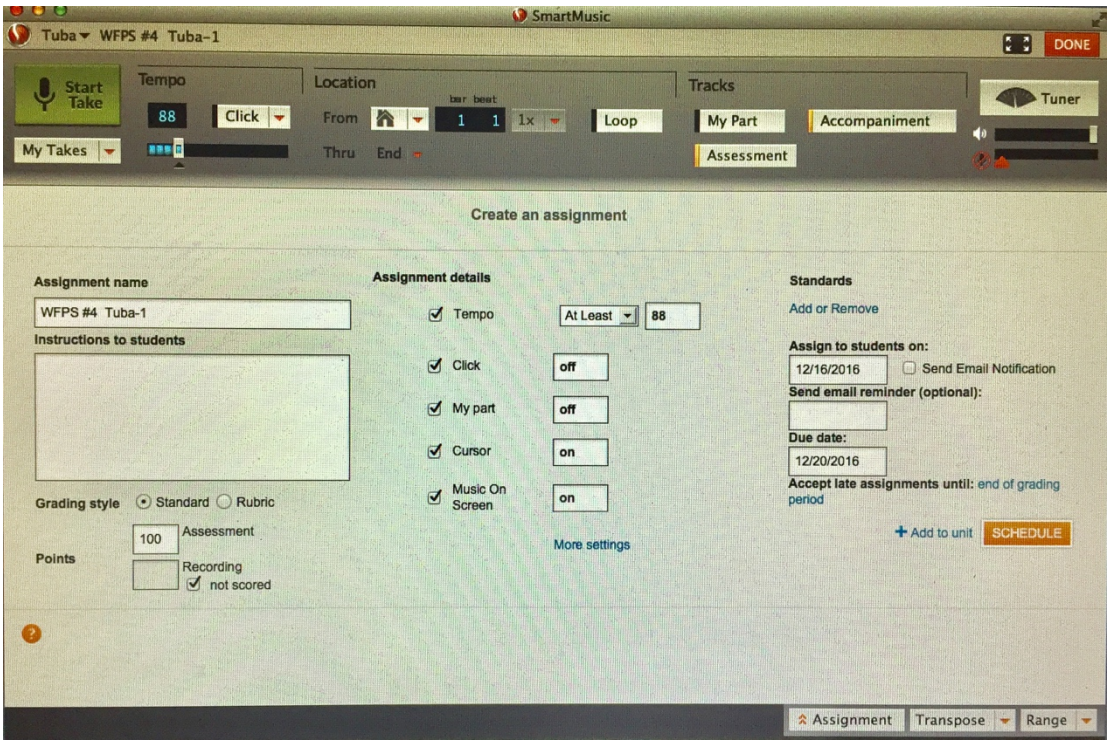


Figure 5. The *SmartMusic* Assignment Creation Screen. The screen was taken under the researcher's educator account.

The *SmartMusic* assessment program evaluated music performance by comparing the original scores with sound waves to detect pitch and rhythmic errors. The program recorded

music performance and gave immediate feedback on the screen. According to the *SmartMusic* program, green note-heads illustrated correct pitch being performed at the correct time, while red note-heads indicated incorrect timing or pitch. Black note-heads meant inaccurate pitch or timing or not being performed. The number of green notes divided by the total number of notes would be the assessment score using a percentage. In this study, the recordings were then provided as audio examples to four music judges who were secondary school band directors on Oahu for evaluation.

Judge Selection

The researcher emailed the band directors from the contact list of the Oahu Band Directors Association. Four human judges responded to participate in the study. The judges were current local secondary school band directors on Oahu. All judges had over five years band or orchestra teaching experience or held an advanced degree in music education. Once the *Invitation to Participate* was accepted and received, each judge was required to sign a consent form before participating in the study (See Appendix D).

Procedure

Part I: the *SmartMusic* Assessment

The researcher set up the assessment equipment including (a) one MacBook Air connected to the *SmartMusic* program, (b) one *SmartMusic*-designed microphone, and (c) one music stand with two music exercises in the music studio. The participants were called to come to the music studio to take the test based on the schedule they were assigned on the scheduling sheet. During the assessment, all performances were evaluated and recorded simultaneously by the *SmartMusic* program.

The participants had a 30-second preparation period prior to each exercise. The procedure was conducted as the following steps:

1. The researcher says, “Please set-up your instrument and have a seat.”

2. The researcher says, “Let’s make sound check,” then press the mic check from the software.
3. The computer says, “*SmartMusic* will help you set your microphone level. Attach your microphone as shown.”
4. The researcher says, “You are going to perform two exercises. You will have 30 seconds to study each score before you play.”
5. The researcher says, “The computer will sound 4 audible clicks at tempo: ♩= 76.
6. The researcher says, “Do you have any questions?” If the participant has no question, then the researcher will put the music on the music stand and the participant will read the music for 30 seconds.
7. After 30 seconds’ study time, the researcher says, “Please find your first note and keep your mouthpiece up while we begin.”
8. The researcher says, “Ready? Listen for the four clicks and begin.”

Part II: The Human Judges.

The judge panel was provided the same recordings from the *SmartMusic* assessment in a randomly chosen order. The judges were asked to listen to the recordings, which were recorded via the *SmartMusic* program and performed by the UHM undergraduate instrumentalist participants. Each recording could be listened to one time only. Each recording was evaluated by these judges using an evaluation form (See Appendix D).

The Human Judges’ Rating Scales

The human judge’s evaluation form was designed based on criteria comprising pitch, rhythm, and tempo, along with the rating rules of the WFPS and the *SmartMusic* electronic assessment feature. The form included two parts: (a) the human judges circle rhythmic, pitch, and tempo errors on the music sheet, and (b) the human judges rate each recording based on

10 standard points. The evaluation form of the judge panel was presented using an A4 size paper (See Appendix D).

In order to compare the judges' assessments with the *SmartMusic* assessment, the scoring criteria need to be comparable. The researcher designed the human judge rating scales using three criteria including pitch, rhythm, and tempo because these are the criteria assessed by the *SmartMusic* assessment.

While creating clear and reliable rules for scoring, the researcher maintained a specific instruction for human judges' scoring. The rules for scoring were adopted from WFPS (Watkins & Farunm, 1954, pp. 6-9), since the music materials were from the Watkins-Farnum Performance Scale Form A. Three types of errors including pitch errors, time errors for rhythm, and change of time errors were utilized for WFPS. The researcher created the human judge rating scales by further utilizing these three types of errors with five rating levels and a 10 standard-points scale (See Figure 6). A perfect score was 10 points. The distribution of the points and the criteria was as follows:

1. A perfect score of pitch was 4 points.
2. A perfect score of rhythm was 4 points.
3. A perfect score of tempo was 2 points.

For pitch and rhythm, there were four rating scales as follows:

1. Rating level I (4 points): All pitch and rhythm played correctly.
2. Rating level II (3 points): 1 to 2 pitch or rhythmic errors.
3. Rating level III (2 points): 3 to 4 pitch or rhythmic errors.
4. Rating level IV (1 point): 5 to 8 pitch or rhythmic errors.
5. Rating level V (0 point): 9 or more pitch or rhythmic errors.

For tempo, the scoring will be:

1. Consistent and correct tempo (2 points).

2. A “change of time” error (1 point).

3. Two or more “change of time” errors (0 points).

Rating / Criterion	V	IV	III	II	I	Points
1. Pitch	9 or more wrong notes (0 points)	5 to 8 wrong notes (1 point)	3 to 4 wrong notes (2 points)	1 to 2 wrong notes (3 points)	All notes played correctly (4 points)	
2. Rhythm	9 or more rhythmic errors (0 points)	5 to 8 rhythmic errors (1 point)	3 to 4 rhythmic errors (2 points)	1 to 2 rhythmic errors (3 points)	All rhythm played correctly (4 points)	
3. Tempo	Correct and consistent tempo (2 points) A “change of time” error (1 point) Two or more “change of time” errors (0 points)					
Total						

Figure 6. The Human Judge Rating Scales. The human judge rating scales was designed by the researcher based on the rating rules of the WFPS. A perfect score of pitch was 4 points. A perfect score of rhythm was 4 points. A perfect score of tempo was 2 points.

Analytic Framework

Generalizability (G) theory was applied to examine the quality of the performance assessments with respect to various types of measurement errors. G theory is based on a calculation of analysis of variance (ANOVA) and represents an extension of Classical Test Theory (CTT) in estimating the reliability of measurements. In CTT, observed scores (X) are assumed to be comprised of an individual’s true score (T) plus an error component:

$$X = T + E \quad (1)$$

observed score true score error

The observed score variance (σ_x^2) is the sum of the true score variance (σ_T^2) and the error variance (σ_E^2):

$$\sigma_x^2 = \sigma_T^2 + \sigma_E^2 \quad (2)$$

Reliability coefficient (ρ_x^2) is statistically defined as the proportion of true score variance (σ_T^2) to the observed score variance (σ_X^2, also denoted as "total score variance"). Once the true score variance and the error variance are available, the reliability coefficient can be calculated as in Equation 3:

$$\rho_x^2 = \frac{\sigma_T^2}{\sigma_x^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \quad (3)$$

The limitation of CTT is that the error cannot be further divided into more specific sources of error, such as error due to various components comprising the test, as well as possible errors due to interaction of components. In contrast, G theory allows for more flexibility in revealing a wide range of measurement conditions to detect as many facets of measurement error as possible. A test score in G theory is considered as a sample from "a universe of admissible observations," which comprises all possible observations (e.g., facets and interactions of facets) on an object of measurement (Shavelson & Webb, 2006, p. 309). Applying G theory within a standard ANOVA framework allows researchers to estimate variance attributable to one or more sources of errors (Lakes, 2003, p. 29).

In this study, a two-facet design consisting of Persons x Raters x Items (*PRI*) was applied. The components included: (1) persons (*P*), which refers to the participants; (2) items (*I*), which refers to pitch, rhythm, and tempo; and (3) raters (*R*), which refers to the human judges and the computerized assessment (see Table 2). Persons are considered the focus of the study, with raters and items (or components comprising the assessment) being possible error facets. The persons are crossed with raters and items to investigate the dependability of the assessments with respect to these two sources (facets) of error.

Table 2

Two-Facet Generalizability Design (PRI)

Source of Variability	Type of Variability	Variance Notation
Persons (P)	Universe-score variance (object of measurement)	σ_p^2
Raters (R)	Constant effect for all persons due to stringency of raters	σ_r^2
Items (I)	Constant effect for all persons due to differences in item difficulty.	σ_i^2
PR	Inconsistencies of raters' evaluations of particular persons' behavior	σ_{pr}^2
PI	Inconsistencies from one item to another for a person's behavior	σ_{pi}^2
RI	Constant effect for all persons due to differences in raters' stringency across items	σ_{ri}^2
PRI (+ e)	Residual consisting of the unique combination of P, R, I; unmeasured facets that affect the measurement, and/or random events	$\sigma_{pri,e}^2$

Note: Adapted from *Generalizability Theory: A Primer* (pp.7-9) by R. J. Shavelson & N. M. Webb, 1991, Newbury Park, CA: Sage Publications.

The person component is the variance in ratings attributable to variance in objects' actual standings. This is referred to as the universe score (denoted as μp) and is analogous to a person's "true score" in CTT (Shavelson & Webb, 2006, pp. 311-312). All other components (e.g., raters and items) usually are regarded as error. In contrast to classical test theory, which separates an individual's observed score into a true score (error-free score) and an error score, a G study facilitates separating observed score variance into variance due to persons (i.e., universe score variance), raters, items, and crossed interactions. This type of analysis results in more specific information about possible sources of bias in the measurement, which can lead to informed choices for improvement in the assessments. The formula of G theory is described in the pilot study in the following section.

The Pilot Study

In order to examine the feasibility of the proposed study, the researcher first conducted a pilot study. It also served as training purpose for human judges. The researcher trained all four judges using five recordings randomly selected from the subject sample. The

samples of recordings were selected from the warm-up exercise for the participants. Each judge was informed of the specific procedure and the scoring rules for using the evaluation forms in advance. A description of the assessment was given, as follows:

> There are 5 recordings including flute, alto saxophone, trumpet, trombone, and euphonium. An evaluation sheet is provided for each recording. Before assessing the recordings, please read the information below and the rules for scoring thoroughly. On the first part of evaluation sheet, exercise No. 5 is presented. Each recording can be listened to one time. You will evaluate each recording based on the criteria: pitch, rhythm, and tempo. Please circle any errors using a pencil on the music score. On the second part of the evaluation sheet, please put the points for each recording.

After the testing, the researcher collected scores that were (a) assessed by the *SmartMusic* software using pitch, rhythm, and tempo as criteria with 100 points and (b) rated by the judge panel using pitch and rhythm as criteria with 10 points. The points of the *SmartMusic* software were divided by 10 and rounded up the decimal to be comparable with the human judges' score, and then the data was computed and analyzed.

The researcher preliminarily looked at the error variance in each of the item components regarding pitch, rhythm, and tempo from the human raters. The largest source of error variance was introduced by the item component (67%). Smaller portions of error were due to the person*item components (13%). The residual error variance is PRI interactions and random error (10%). Importantly, only a small proportion of the overall variability in scores was due to the person effects (10%), or the individuals' performances in the pilot study (See Table 3).

Table 3

Variance Estimates of Each Component on the Human Judges

Component	Estimate	Total Variability (%)
Var (person)	.18	10
Var (item)	1.24	67
Var (rater)	.01	0
Var (person * item)	.23	13
Var (person * rater)	.00[a]	0
Var (rater * item)	.00[a]	0
Var (error)	.19	10
Total Variance	1.85	100

Dependent Variable: score
Method: Restricted Maximum Likelihood Estimation
Note. [a] This estimate is set to zero because it is redundant.

As suggested in Table 3, the total variation in observed scores summed across the components was 1.85. This score can be broken into proportions of variance due to persons and other sources of error. If there was no error associated with the measurement of individuals' performances, we would expect the proportion of variance due to persons to be 100%. As shown in the table, however, in this initial analysis, most of the variance was due to differences in the item components (i.e., pitch, rhythm, tempo).

The absolute error (σ_{Δ}^2) of the design included all sources of error variability which considers error due to items (*i*), raters (*r*), the interaction among persons*items (*pi*), persons*raters (*pr*), raters*items (*ri*), and persons*raters*items (*pri*), plus other sources of error (*e*). The reliability coefficient (ρ_{Δ}^2) was then computed using G theory formula, which is a unified framework of measurement of reliability expressed in Equation 4. When the proportion of variance due to persons was compared with the error components, the dependability (reliability coefficient) of the overall assessment was low, .26 (See Equation 4), against an expected minimum dependability of .70.

$$\sigma_{\Delta}^2 = \frac{\sigma_i^2}{\acute{n}_i} + \frac{\sigma_r^2}{\acute{n}_r} + \frac{\sigma_{pi}^2}{\acute{n}_i} + \frac{\sigma_{pr}^2}{\acute{n}_r} + \frac{\sigma_{ri}^2}{\acute{n}_i \acute{n}_r} + \frac{\sigma_{pri,e}^2}{\acute{n}_i \acute{n}_r} \quad (4)$$

$$1.24/3 + .00/4 + .23/3 + 0 + .00/12 + .19/12$$

$$0.41 + .00 + .08 + 0 + .00 + .02 = .51$$

$$\rho_{\Delta}^{2}=\frac{\sigma_{p}^{2}}{\sigma_{p}^{2}+\sigma_{\Delta}^{2}}=\frac{.18}{.18+.51} = .18/.69 = .26$$

From this result, the conclusion was that examining each component separately would not yield dependable results. Therefore, the researcher decided to examine pitch, rhythm, and tempo as a composite rating for each judge, as opposed to the component scores separately.

The revised analysis is summarized in Table 4. The results suggested that most of the variance in the revised design was due to persons, 2.58, which suggested greater consistency in assessments among the human judges and smaller amounts of error due to raters, raters*persons, and combined residual error. Moreover, the total variability due to person*rater interactions and combined error in this section was 9.8 %. This implied that most of the variability was due to the person components, 90.2 %, rather than the sources of error.

Table 4

Variance Estimates of Total Scores Based on the Human Judges

Component	Estimate	Total Variability (%)
Var (person)	2.58	90.2
Var (rater)	.04	1.4
Vary (Error)	.24	8.4
Total Variance	2.86	100.0

Dependent Variable: score
Method: Restricted Maximum Likelihood Estimation

The absolute error (σ_{Δ}^{2}) of the revised design included all sources of error variability, which considers error due to raters (*r*) and the interaction among persons*raters (*pr*), plus other sources of error (*e*). The correlation coefficient (ρ_{Δ}^{2}) was then computed using G theory formula in Equation 5. The new dependability estimate was significantly improved, .97. This suggested that when the item components were considered more holistically, the scores of the human judges were highly dependable. This also implied the judges were very

consistent in their "total" score ratings of the set of performance assessments. For example, as expected, differences among the persons comprise 90.2% of the total variability in scores. Also, it corresponded with the low error variance by rater and person*rater plus residual error components, 9.8%.

$$\sigma_{\Delta}^2 = \frac{\sigma_r^2}{\acute{n}_r} + \frac{\sigma_{pr,e}^2}{\acute{n}_r} \quad (5)$$

$$= .04/4 + .24/4$$

$$= .01 + .06$$

$$= .07$$

$$\rho_{\Delta}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\Delta}^2} = \frac{2.58}{2.58 + .07} = 2.58/2.65 = .97$$

The last step was to add the computerized assessment in the analysis. The total observed variation based on the human judges and computer was 6.16 (see Table 5). Compared to the human judges alone, adding the computer assessments reduced the variance due to persons considerably (i.e., from 90.2% to 27.4%). Similarly, the error variance introduced by the rater components was much higher (i.e., from 1.4% to 31.7%). In addition, the residual error component increased from 8.4% to nearly 41%). All of these comparisons suggest the *SmartMusic* ratings were considerably different from the human judge ratings.

Table 5

Variance Estimates Based on Total Scores of All Judges with the Computerized Assessment Included

Component	Estimate	Total Variability (%)
Var (person)	1.69	27.4
Var (rater)	1.95	31.7
Var (Error)	2.52	40.9
Total Var	6.16	100.0

Dependent Variable: score
Method: Restricted Maximum Likelihood Estimation

The level of reliability coefficient for the assessments based on the human raters plus the computerized assessment was also considerably lower, .66, as shown below. With the

computer included, the dependability of the assessment fell to slightly below the desired minimal level of .70.

$$\sigma_{\Delta}^2 = \frac{\sigma_r^2}{\acute{n}_r} + \frac{\sigma_{pr,e}^2}{\acute{n}_r} \quad (6)$$

$$= 1.95/5 + 2.52/5$$

$$.39 + .50 = .89$$

$$\rho_{\Delta}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\Delta}^2} = \frac{1.69}{1.69 + .89} = 1.69/2.58 = .66$$

The implication was that the set of performance assessments which included the computerized assessment was not as reliable as the set of the assessments of the human judges alone.

Summary

The researcher utilized G theory in the pilot study and initially examined the variability and reliability of each component. The data indicated that the most variability was due to the three separate item scores (i.e., pitch, rhythm, tempo) within each judge's assessment, and the reliability was quite low, .26. Then, the researcher examined the variability and the reliability of the composite scores based on the human judges. The most variability was due to the person scores, and the reliability was quite high, .97. Lastly, the researcher examined the variability and the reliability of the composite scores based on the human judges with the computerized assessment. The most variability was due to the error component, and the reliability was slightly lower .66, than a common minimal standard set at .70.

The results suggested that the level of reliability was noticeably higher when using a composite rating of pitch, rhythm, and tempo, either with or without computerized assessment. Therefore, the researcher decided to proceed with the composite scores in the

full study. This would facilitate a more focused examination of the consistency of the human judges' assessments versus the computer assessments.

CHAPTER 4

RESULTS

The results of the study were presented in this chapter. The sample ($N = 34$) consisted of six flute recordings, three clarinet recordings, four oboe recordings, two bassoon recordings, three alto saxophone recordings, six trumpet recordings, three trombone recordings, four horn recordings, and three euphonium recordings. The set of performance assessments (N=5) included four human judges and one computerized assessment. Generalizability theory was used to evaluate the quality of the performance assessments. The facets included an individual facet (person), a judge facet (rater), and an error component (persons*items, plus other sources of error).

After examining the variability and the dependability (i.e., reliability) of the assessments using G theory, a series of two-way ANOVA models was estimated to examine the statistical significance of the variability due to persons and raters. The dependent variable in the study was scores (i.e., the participants' performance scores). Other independent variables (the between subjects factors) in the study included persons (i.e., individuals' performances) and raters (i.e., four human judges and one computerized assessment). The researcher first examined the variability of human judges (Raters 1-4), then adding the computerized assessment (Rater 5) to the set of the judges to investigate whether the human judges and computerized assessment are comparable.

Examining of Variability of Human Judges (Raters 1-4)

The raters in the study included the four human judges and the one computerized assessment, which was considered as the fifth rater. In the following section, the dependability of the ratings by the four human judges were examined. This was an important initial step in the analysis in order to investigate the comparative agreement of human raters on the quality of the student performances.

Descriptive statistics were presented for each judge summarizing the average means and variance in ratings across the 34 individuals (See Table 6). The data indicated that judges' mean ratings of the performances were reasonably close, ranging from 7.71 to 8.06. The difference of means among four judges was only .35. The variance of each judges' ratings was also relatively similar, ranging from 3.61 to 7.06. Among the four human raters, the first judge's ratings are more varied, 7.06, while the third judges' ratings were considerably less varied, 3.61.

Table 6

Average Ratings and Variance Across Human Judges for 34 Performances

	N	Range	Mean	Variance
Judge1	34	10.00	7.97	7.06
Judge2	34	9.00	8.06	4.78
Judge3	34	8.00	7.71	3.61
Judge4	34	9.00	7.91	4.33
Valid N (listwise)	34			

Generalizability Theory Results

Generalizability theory can be used to provide an indication of how each potential source of variation in the design of the study contributes to variability in the outcome. In G theory, the variability due to persons (i.e., the individual variability) accounts for the majority of variability in subjects' performance scores. Other potential sources of variability are considered error facets (Shavelson & Webb, 1991, 2006). In Table 7, the results for the four human judges illustrated that the variability, 4.94, was due to differences in the person component, 83.2%, in the study. There was virtually no variance due to the main effect of judges, 0%; however, there was error variance due to the interaction between raters by persons combined with other sources of error, 16.8%. This corresponded that there were more varied on the first judge' ratings and less varied on the third judge' ratings (See table 7).

Table 7

Variance Estimates Across Human Judges

Component	Estimate	% of Var
Var (person)	4.11	83.2
Var (rater)	.00	0.0
Var (Error)	.83	16.8
Total Var	4.94	100.0

Dependent Variable: score
Method: Restricted Maximum Likelihood Estimation

Similar to the pilot study, the dependability of the ratings was compiled by the human judges covering 34 participants in the main study. The total observed variation was 4.94. The absolute error (σ_Δ^2) of the design, that was, all sources of error variability, using G theory formula, which considers error due to raters (r) and the interaction between persons*raters (pr) and other sources of error (e) as follows:

$$\sigma_\Delta^2 = \frac{\sigma_r^2}{\acute{n}_r} + \frac{\sigma_{pr,e}^2}{\acute{n}_r} \quad (7)$$

$$= .00/4 + .83/4$$

$$= .00 + .21$$

$$= .21$$

The dependability of the ratings from the human judges was then estimated the variability due to persons divided by the variability and due to persons plus other sources of error as below:

$$\rho_\Delta^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2} = \frac{4.11}{4.11 + .21} = 4.11/4.32 = .95 \quad (8)$$

The result suggested that the dependability of the ratings provided was statistically high (.95). The variability among the human judges in the following plot (See Figure 7). The human judges' ratings are significantly close in their assessments, with slightly additional variability shown for Individual 7 and Individuals 14 and 16.

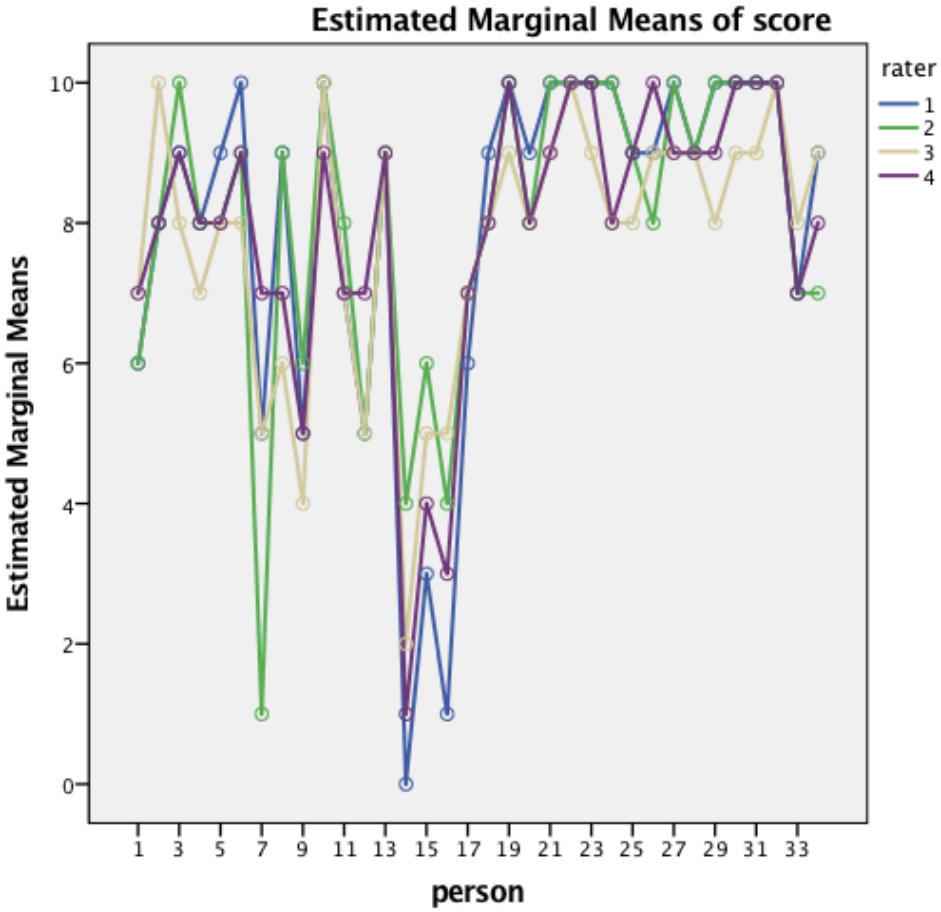


Figure 7. The Variability of Individuals' Scores from Four Human Judges

Examining the Statistical Significance of the Variability

The variability due to raters, persons, and other sources of error was examined using two-way ANOVA. The data indicated that the variability of persons was statistically significant, F (33, 99) = 20.82, $p < .00$. This variability was expected, and it indicated that the person effect explains statistically significant variation in performance across the individuals who participated in the study. The data showed that the variability in performance was due to the individuals in the study and not due to other error facets (e.g., raters, raters x persons, or other error). A crucial indication was that the effect of the human raters is not statistically significant, F (3, 99) = .92, p =.43 (See Table 8).

Table 8

Tests of Between-Subject Effects Across Human Judges

Dependent Variable: score

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	8513.06	1	8513.06	494.36	.00
	Error	551.67	32.04	17.22[a]		
Rater	Hypothesis	2.29	3	.77	.92	.43
	Error	82.21	99	.83[b]		
Person	Hypothesis	570.44	33	17.29	20.82	.00
	Error	82.21	99	.83[b]		

a. MS(rater) + MS(person) - MS(Error)

b. MS(Error)

Adding the Computerized Assessment to the Set of Judges

After adding the computerized assessment as the fifth rater, the data indicated that while the human judges' mean ratings were fairly close, ranging from 7.71 to 8.06, the computer mean was much lower, 2.56. The variance in ratings of the *SmartMusic* assessment, 4.68, was consistent with the human judges, ranging from 3.16 to 7.06. The descriptive statistics were presented for each judge and the computer program summarizing the average means and variance in ratings across the 34 individuals in the study (See Table 9).

Table 9

Average Ratings and Variance Across Human Judges and Computerized Assessment for 34 Performances

	judge1	judge2	judge3	judge4	*SmartMusic*
Mean	7.97	8.06	7.71	7.91	2.56
Variance	7.06	4.78	3.61	4.33	4.68
Range	10.00	9.00	8.00	9.00	8.00

Generalizability Results

In Table10, the results for the five raters (i.e., four human judges and computerized assessment) indicated the total variability was larger (10.59) than the variance when considering only the human judges (4.94, see Table 8). Only 29.8% of the total variance was due to differences in persons component (i.e., individual performances). Most of the variance was due to differences in raters (i.e. the human judges and the computerized assessment) in the study (53.8%). In addition, there was also some error variance due to the interaction between raters assessing individuals combined with other sources of error (16.4%). The evidence from this G study was clear that adding the computer to the set of human judges results in considerable error being introduced due to raters and other sources of error (e.g., raters by persons).

Table 10

Variance Estimates Across the Human Judges and the SmartMusic *Assessment*

Component	Estimate	% of Variability
Var (person)	3.15	29.8
Var (rater)	5.70	53.8
Var (Error)	1.74	16.4
Total Var	10.59	100.0

Dependent Variable: score

Method: Restricted Maximum Likelihood Estimation

After adding the *SmartMusic* assessment to the design, the dependability of the ratings compiled by the total of five judges was estimated. The absolute error (σ_{Δ}^2) of the design was computed, which was all sources of error variability, using the G theory formula (i.e., which considers error due to raters (r) and the interaction between persons*raters (pr), and other sources of error (e) as follows:

$$\sigma_{\Delta}^2 = \frac{\sigma_r^2}{\acute{n}_r} + \frac{\sigma_{pr,e}^2}{\acute{n}_r} \quad (9)$$

$$= 5.70/5 + 1.74/5$$

$$= 1.14 + .35$$

$$= 1.49$$

The dependability of the performance data was then estimated as the variability due to persons divided by the variability due to persons plus other sources of error as shown below:

$$\rho_{\Delta}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\Delta}^2} = \frac{3.15}{3.15+1.49} = 3.15/4.64 = .67 \quad (10)$$

The data indicated that dependability of the ratings provided was dramatically decreased (.63). The reliability of the set of human judges was much higher (.95). The results indicated that the null hypothesis, "there will be no significant difference between the music performance assessments of music adjudicators and the *SmartMusic* assessment," is rejected. For the computer to be used, it was likely that more judges would need to be added to the assessment process to boast the dependability of the assessments up to more acceptable levels (e.g.,75% or 80%).

The plot of the set of judges' ratings is presented in the following figure (See Figure 8). It illustrates that the human judges' scores reflected only slight variability for each performance on the top, except for Individuals 7, 14, and 16, which reflected additional variability. The human judge panel and the *SmartMusic* assessment rated eight out of the 34 recordings (i.e., for Individuals 3, 6, 10, 13, 19, 29, 30, and 32) very differently. Those recordings were rated either a perfect score of 10 points, or with only 1 point deducted by the human judges, whereas the *SmartMusic* assessment rated the same recording only 1 or 2 points. Thus, the computer scores were much lower and very different from the human judges' scores. Only 11.7 % of the ratings (Individuals 17, 18, 21 and 28) were similar between the computerized assessment and the human judges. The discrepancy of the curves

of the ratings between the computerized program and the judge panel suggested that the ratings of the *SmartMusic* assessment were not statistically comparable with those of the set of judges.

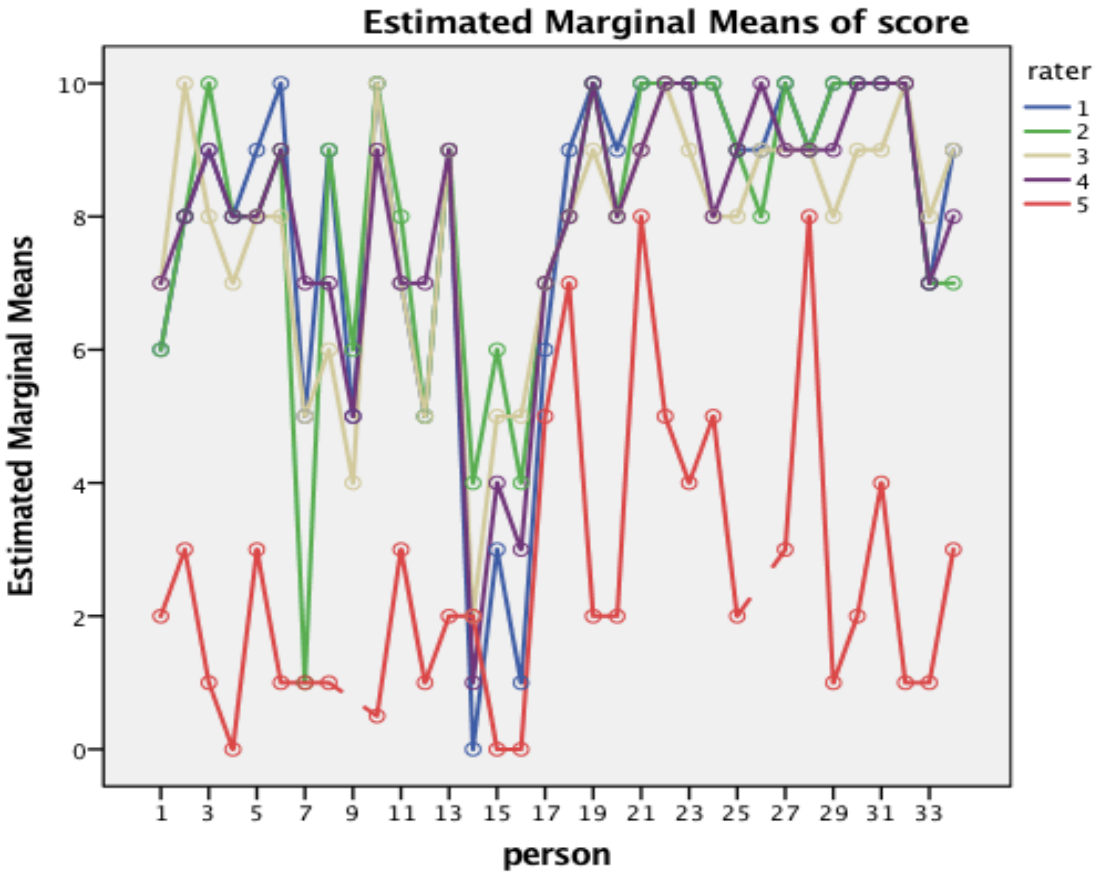


Figure 8. The Means of Individuals' Scores from Four Human Judges.

Examining the Statistical Significance of the Variability

In the two-way ANOVA table below (Table 11), variability due to raters, persons, and other sources of error is examined. The data indicated that person variability was statistically significant, F (33,132) = 10.07, p < .00. This variability was expected, and it indicated the person effect explained statistically significant variation in performance across the individuals who participated in the study. In contrast, however, most of the variability in performance was not due to the individuals in the study but, rather, was due to the error facets

(e.g., raters, raters x subjects, other error). Importantly, the rater effect was statistically significant, F (4, 132) = 112.39, p <.00.

Table 11

Tests of Between-Subjects Effects Across Human Judges and SmartMusic Assessment

Dependent Variable: score

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	7956.29	1	7956.29	37.66	.00
	Error	985.53	4.67	211.18[a]		
rater	Hypothesis	781.68	4	195.42	112.39	.00
	Error	229.53	132	1.74[b]		
person	Hypothesis	577.51	33	17.50	10.07	.00
	Error	229.52	132	1.74[b]		

a. MS(rater) + MS(person) - MS(Error)
b. MS(Error)

Summary

In this chapter, the researcher analyzed the variability and the dependability of the ratings by four human judges, then added computerized assessment to investigate whether the computerized assessment is comparable to the set of human judges. The results indicated that the dependability (i.e., reliability) of the ratings provided was considerably low (.67). The reliability of the set of human judges was .97, which was higher than when the set of ratings included the computerized assessment. This pointed out that the reliability of the assessments did not reach acceptable levels of reliability in evaluating student performance in this study. Thus, the null hypothesis was rejected: there was a significant difference between the music performance assessments of music adjudicators and the *SmartMusic* software. The human judges panel was statistically more reliable than the computerized assessment.

CHAPTER 5

DISCUSSION

The Background and the Purpose of the Study

A major concern in the literature on music performance evaluation has been the validity and the reliability of assessment methods (Begree, 2003; Brophy & Albert, 2008). The challenges in fairly and accurately evaluating music performance are often identified as subjective matters, non-musical and musical factors, as well as the methods or the tools with which music teachers assess performance. The National Association of Schools of Music, National Association of Schools of Art and Design, National Association of Schools of Theatre, and National Association of Schools of Dance (1997) stated, "...evaluation of works of art, even by professionals, is highly subjective, especially with respect to contemporary work" (p. 7). This subjective nature of the art means that any human rater-based assessment of musical performance could have difficulty guaranteeing fair judgments due to non-musical factors.

Researchers found that non-musical factors, such as student attitude, effort, and participation, have been given greater weight than musical factors in calculation of music grades (Barkley, 2006, p.6; Keddy, 2013, p. iv). In addition, written exams (i.e., multiple-choice and short answers) and observation are common methods used to assess students' musicianship and achievement (Shuler & Connealy, 1998, pp. 16-17; Frankel, 2002, pp. 6-7; Zuar, 2006, p. 3; McQuarrie & Sherwin, 2013, para. 12). Music is primarily an aural and performing art (Shih, 2012, p.1), so music technique and skills cannot be measured solely using written exams, such as multiple-choice or true or false questions (Zuar, 2006, p. 3; Shuler & Connealy, 1998, p.12).

Another matter affecting the evaluation of music performance is a deficiency of a systematic or research-based methodology. In today's data-driven educational settings, the

quantity and quality of large-scale assessments have been a provocative issue in music education.

Over the past two decades, computerized assessments, such as the *SmartMusic* assessment, has been designed and used for evaluating music performance. Such software can be an alternative assessing tool to help music teachers and band directors. However, innovation and technology do not automatically improve evaluation of music performance. Experimentation is needed to provide evidence of the efficacy of a music evaluation tool, and whether that product performs to the functions claimed by the manufacturer. Many researchers have studied the *SmartMusic* software as an instructional, practice, and assessment tool (Karas, 2005; Lee, 2007; Zanutto, 2007; Buck, 2008; Flanigan, 2008; Astafan, 2011; Long, 2011; Nichole, 2014). Only two studies, such as Karas (2005) and Lee (2007), have examined the reliability and the validity of the *SmartMusic* assessment. Although Long (2011) conducted a qualitative study on the effectiveness of the *SmartMusic* assessment feature, very few studies, such as (Karas, 2005; Lee, 2007), have shown statistical analyses of the comparison of computerized assessments with human examiners in assessing music performance.

The purpose of this study focused on determining whether there was a difference in reliability between a set of human judges and the *SmartMusic* assessment using a quantitative experiment. Thirty-four (N=34) undergraduate instrumentalists of the University of Hawai‘i performed two sight-reading exercises using the *SmartMusic* program. The performances were assessed and recorded simultaneously via the *SmartMusic* software. The ratings included a set of four human judges and the *SmartMusic* assessment as the fifth rater to determinate whether the *SmartMusic* assessment is comparable to human judges in its ability to evaluate music performance against pitch, rhythm, and tempo criteria. The hypothesis of

this study was: there will be no significant difference between the music performance assessments of music directors and the *SmartMusic* assessment.

Findings and Discussion

According to the analysis of the data, the results indicated that the null hypothesis; there will be no significant difference between the music performance assessments of the human judges and the *SmartMusic* assessment, was rejected. Statistically significant differences between the human judge panel and the *SmartMusic* assessment were found in the variability and the reliability of the ratings.

By comparing the mean ratings, the *SmartMusic* assessment's mean was 2.56, which was much lower than the human judges' mean. The human judges' mean range was significantly higher, 7.71 to 8.06, and the difference of the human's mean range was only .35. Based on the result, the implications included: (a) the human judges had similar views of evaluating music performance based on the criteria of the pitch, rhythm, and tempo; (b) the human judges gave higher scores than the computerized assessment; (c) the computerized assessment detected more errors in terms of pitch, rhythm, and tempo criteria; and (d) while the human judges could be more tolerant for subtle errors, the *SmartMusic* assessment was a much stricter rater due to its more rigid software algorithm.

In previous research, few studies compared or examined the differences between human judges and the *SmartMusic* assessment in evaluating pitch, rhythm, and tempo. Few studies have found that the program is effective in assessing these criteria. Regarding the rhythmic criterion, Karas (2005) claimed that the rhythmic tolerance is under the 16th note of a beat and the pitch tolerance is under 50 cents (p.57). Similarly, Long (2011) also pointed out that the *SmartMusic* assessment was programmed to allow a certain degree of leeway when evaluating the accuracy of pitch and rhythm (p. 39). However, the *SmartMusic* program strictly graded and seriously penalized performers on instances of inconsistent

tempo and incorrect rhythm during the performance, such as late entrance, or playing behind or ahead of the beat throughout the piece. Based on the examples of Long's study, a performer was given a total score of only 14% due to playing each note perceptibly late even though other criteria (i.e., pitch and articulation) were accurate. As for the pitch criterion, Long (2011) described that solely as a result of playing each note partially below or high throughout the piece, the *SmartMusic* rated the performer 0%. This is because that the *SmartMusic* assessment grades each performance based on correct pitches and rhythm played at the exact time (Long, 2011, p.39).

These rating problems in Long's qualitative experiments mentioned above emerged in the present quantitative study as well. Some performances (i.e., Individuals 3, 6, 10, 13, 19, 29, 30, and 32) received high scores like 9 and 10 points from the human judge panel, but the computer rated these performances extremely low, only 1 or 2 points. These performances did not appear to exhibit any obvious error, but they were played slightly slower or faster. The human judges might not notice the subtle change of tempo without a technological device or a metronome. On the other hand, among 34 performances, only four individual performances (i.e., Individuals 17, 18, 21, and 28) were rated similarly by the human judge panel and the *SmartMusic* assessment. For example, all the judges gave Individual 28 a score of 9 points, deducting one pitch point because of two pitch errors, while the *SmartMusic* assessment gave Individual 28 a score of 8 points, deducting for one pitch error, and for just a few notes played not precisely on time as detected by the software. These recordings were rated similarly between the human judges and the *SmartMusic* assessment because they were mostly in tempo. Accordingly, the discrepancy in ratings between human judges and the *SmartMusic* assessment was primarily due to inconsistent or incorrect tempo.

In terms of the correlation and reliability, although Karas (2005) and Lee (2007) found a positive correlation between a human judge panel and the *SmartMusic* assessment,

the reliability of the human judge panel (raters 1-4) in the present study was significantly higher, .95, than that of adding the computerized assessment (raters 1-5), .67. The dependability of this study implied that the computer was below acceptable levels of reliability in evaluating student performance—at least insofar as its ability to evaluate qualities about the performance that human judges responded to more than just the "technical" quality to which the computer attended.

Karas (2005) observed that human judges could identify incorrect pitches but could not notice a certain range of slightly incorrect pitch right before and right after a quarter-step flat or sharp. This, in light of the results of the present study, could explain that the *SmartMusic* assessment deducts points for inaccuracies, possibly because its technology cannot accept subtle inaccuracy that human judges have more tolerance for, or possibly because it can identify subtle inaccuracies that are imperceptible to human judges, which led to the divergence of the reliability between human and computerized assessment.

This quantitative study was based on the recommendation of a qualitative study of Long (2011) to evaluate student performances using both the *SmartMusic* assessment and a panel of human judges. The inter-judge reliability was initially determined with a high dependability. The variability was examined from the set of ratings evaluated by the set of human judges and the *SmartMusic* assessment. The two sets of ratings were then compared to address similarities and differences between human evaluation and the computerized assessment. Statistically significant differences between the human judge panel and the *SmartMusic* assessment were found in both the variability and the reliability of the ratings. Based on the results and implications described above, the *SmartMusic* assessment was statistically not comparable to the set of human judges.

From a subjective versus objective perspective, in Long (2011)'s qualitative study, he claimed that the *SmartMusic* assessment is "incapable of subjective evaluation" and only

assesses performance "with regard to pitch and rhythm on a note-to-note basis" (p. 41). Based on the results of the research, Long's statement was also supported in this quantitative study, in that the human judges appeared to be rational and reasonable in assessing musical performance, yet the *SmartMusic* assessment is absolutely objective without aesthetic perception.

Conclusions

Based on the discussion and findings above, the sole obvious outcome is that the *SmartMusic* assessment is not as reliable as the human judges. However, this statement might be oversimplified. As Long (2011) stated,

> The *SmartMusic* assessment feature is a consistent computer program with precise evaluation parameters that do not change from performance to performance. Consistency is beneficial to any method of performance evaluation, because "if an individual is not able to be consistent in evaluative tasks, it is difficult to place any validity in that individual's assertions about the quality of a music performance." (p.40)

Similarly, Zanutto (2007) stated that "technology is not a panacea. In fact, it is not particularly good at assessing higher level musical skills, but it can provide valuable feedback regarding basic performing skills; note reading, rhythm, pitch, and intonation accuracy" (p. 1).

Instead of rejecting the computerized assessment, better-quality and intelligent music technology needs to be innovated to be comparable with the human judges. However, the current computerized assessment has not been improved adequately over two decades. As Etmektsoglou (1992) stated:

> The computer was programmed to expect a certain pitch response for a specified length of time; if the student sustained that pitch longer than its assigned time, the computer started comparing it with the following expected pitch of the test item, and

judged the second pitch as wrong even when the student had produced the correct pitch response with a small delay. (pp. 80-81)

Accordingly, music technology, such as the *SmartMusic* software, has evolved yet still requires further development to solve the problem of changing tempo or timing to be more comparable to human judges. Adjusting a proper leeway of timing is key to overcome the obstacle of the computerized assessment. For example, performers should not be penalized heavily throughout entire piece due to one instance of changing tempo. A modification to have an adjustable "change time" function would be essential to make a better improvement for assessing music performance.

Further Observations

Aside from the statistical analysis of the study, the researcher had several observations as depicted below:

1. The microphone, which connected with the instrument, at times created some feedback during the assessment, especially for the instruments (e.g., flute and oboe) that produce a higher frequency.
2. For the instruments that require more breathing, such as tuba, the rhythm and tempo may be thrown off due to the subtle late entrance. The *SmartMusic* assessment would continually deduct points for the notes not being played at the assigned time.
3. The SmartMusic program automatically stops recording and assessing a performance at the end of a piece based on its assigned tempo. If a performance is played slower than indicated tempo, not only will it be strictly penalized by the *SmartMusic* assessment due to not being played at the correct time but also the recording will be incomplete.
4. A confounding variable that appeared in the study was that the presence of the

investigator who had to monitor the technological devices in the room during the performances to ensure the equipment was working properly.

Future Recommendations

Further studies need to be conducted to provide more information for music directors and educators on selecting assessment tools for evaluating music performance. Recommendations based on this study are described below.

First, this study did not train the subjects to use the *SmartMusic* assessment before the experiment; thus, a future study is suggested to have performers using the program for a certain amount of time so that they are more familiar with the computerized assessment tool. The familiarity of performers with the *SmartMusic* assessment might increase their scores and statistically improve the reliability of the computerized assessment. In addition, training will enable the participants' ability to use the software themselves to avoid the investigator's presence. A pretest and a posttest then could be performed before and after performers use the *SmartMusic* assessment to determinate whether or not the training improved the variability and the reliability of the computerized assessment.

Second, as for the performers' background in this study, some of them are music majors, and others are not. In a future study, a researcher can further investigate the difference between music majors and non-music majors regarding the use of computerized assessment.

Third, this study used the same sight-reading exercises throughout the experiment for all 34 performers. A future study could select various levels of music examples as well as various levels of performers to examine the differences between human judges and the computerized assessment.

Fourth, this study used a set of human judges with one computerized assessment, the *SmartMusic* program. This researcher suggests that a further quantitative study could utilize

an additional technological tool (e.g., *iPAS* in addition to *SmartMusic*) to form a set of computerized assessments. Such, a study could then compare the reliability of the set of computerized assessment tools to that of a human judge panel.

Lastly, the *SmartMusic* assessment has been calibrated to allow some leeway in pitch and rhythm, but not as much as the judges give (Karas, 2005, p.57; Long, 2011, p.39). An inquiry into mitigating the *SmartMusic* assessment's harsh grading would be worthwhile, especially for assessing instrumental beginners. It would be useful to investigate whether further developments to the software are possible to adjust the amount of leeway the software gives. This may be helpful to reach closer agreements between human judges and the computerized assessment as well as enable the computerized assessment to assist music teachers and band directors in a more practical way.

Implications for Music Education

Although a discrepancy of the ratings was found between the *SmartMusic* assessment and the human judges, the researcher believes that the software is a beneficial and valuable tool for students to practice music. The *SmartMusic* assessment not only provides immediate feedback to assist students to maintain fundamental performing skills, such as the accuracy of pitch rhythm and a consistent tempo, but also it can detect each subtle pitch, rhythmic, and tempo error that human judges may not notice without a metronome or a tuner. Conversely, the traits of human judges, such as more tolerance and forgiveness on subtle errors and changes, are also the reasons for the disagreement between the ratings of the *SmartMusic* assessment and music experts.

Music scholars and researchers have pointed out that subjectivity has been a major issue regarding traditional music assessment/music adjudicators. Aside from comparing the ratings of computerized assessment and human judges, the *SmartMusic* assessment is a reliable tool for assessing music performance (Karas, 2005; Lee, 2007), and the objective

traits of the *SmartMusic* assessment can improve objective measurement in evaluating music assessment. A balanced solution to adjust subjectivity and objectivity for evaluating music performance hence may apply both assessment types, such as incorporating human judges and computerized assessments, as a whole to provide a holistic diagnosis for teachers' instructions and students' music learning.

Closing Remarks

Musical performance criteria, such as pitch, rhythm, and tempo, are the basics and can be measured using objective methods, such as computerized assessments. A good example of musical performance cannot be demonstrated without the accuracy of pitch, rhythm, and tempo. Despite the ratings of the *SmartMusic* assessment was not comparable with the human judges in this study, using computerized assessments to enhance objective measurement for evaluating music performance should still be considered.

The inherence of computerized assessments and music adjudicators is that computerized assessments are objective, while human judges possess both subjective and objective qualities. As Bennett Reimer (1989) stated, "the major function of art is to make objective and therefore accessible the subjective realm of human responsiveness. Art does this by capturing and presenting in its intrinsic qualities the patterns and form of human feeling" (p.153). By combining the qualities of computerized assessment and human judges, performers and adjudicators can enhance the experience and value of each of their roles.

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. Retrieved from http://www.jstor.org/stable/1433910

Abeles, H. F. (1973b). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, *21*(3), 246–255. Retrieved from http://www.jstor.org/stable/3345094

Abeles, H. F., Hoffer, C. R., & Klotman, R. H. (1994). *Foundations of music education* (2nd ed.). New York: Schirmer Books.

Allen, J. D. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *78*(5), 218-223. Retrieved from http://www.jstor.org/stable/30189912

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

Assessment. (1991). In *Merriam-Webster's Dictionary* (9th ed.). Springfield, MA: Merriam-Webster.

Astafan, C. (2011). *SmartMusic: Using technology to assess rhythmic ability within instrumental music in the elementary school classroom.* Retrieved from http://eric.ed.gov/?id=ED518581

Atkins, D. E., Bennett, J., Brown, J. S., Chopra, A., Dede, C., Fishman, B., & Williams, B. (2010). *Transforming American education: Learning powered by technology*. United States: Department of Education. Office of Educational Technology. Retrieved from http://permanent.access.gpo.gov/gpo3612/netp2010.pdf

Barkley, M. (2006). *Assessment of the national standards for music education: A study of elementary general music teacher attitudes and practices* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global (UMI No. 1439697)

Barry, N. H. (2009, January). Evaluating music performance: Politics, pitfalls, and successful practices. In *College Music Symposium* (Vol. 49, pp. 246-256). College Music Society. Retrieved from http://www.jstor.org/stable/41225250

Bauer, W. I. (2014). *Music learning today: Digital pedagogy for creating, performing, and responding to music*. New York, NY: Oxford University Press.

Bergee, M. J. (1987). *An application of the facet-factorial approach to scale construction in the development of a rating scale for euphonium and tuba music performance* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 8813388)

Bergee, M. J. (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. J. (1989). An objectively constructed rating scale for euphonium and tuba music performance. *Dialogue in Instrumental Music Education*, *13*, 65-86.

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

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971*). Handbook on formative and summative evaluation of student learning*. New York: McGaw-Hill.

Brakel, T. D. (2006). Inter-judge reliability of the Indiana State School Music Association High School Instrumental Festival. *Journal of Band Research*, *42*(1), 59-69.

Brophy, T. S., & Albert, K. (Eds.). (2008). *Assessment in music education: Integrating curriculum, theory, and practice*. GIA Publications.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, *30*(2), 123-142. doi: 10.1111/j.1745-3984.1993.tb01070.x

Boyle, J. D. (1992). Evaluation of music ability. In R. Colwell (Ed.), *Handbook of research on music teaching and learning* (pp. 247-265). New York: Schirmer.

Boyle, J. D., & Radocy, R. E. (1987). *Measurement and evaluation of musical experiences*. New York: Schirmer Books.

Buck, M. W. (2008). *The efficacy of SmartMusic® assessment as a teaching and learning tool* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3346520)

Bunderson, C. V., Inouye, D. K., & Olsen, J.B. (1988). The four generations of computerized educational measurement. *ETS Research Report*. Princeton, NJ: Educational Testing Service.

Burns, M. K. (2008). What is formative evaluation. *Minnesota Center for Reading Research*, 1-6.

Burnsed, V., Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert festivals. *Journal of Band Research*, *21*(1), 22-29.

Chalmers, D., & McAusland, W. D. (2002). Computer-assisted assessment. *The Handbook for Economics Lecturers: Assessment, Bristol: Economics LTSN*. Retrieved from http://www.economicsnetwork.ac.uk/handbook/printable/caa_v5.pdf

Civic Impulse. (2017). S. 1177 — 114th Congress: Every Student Succeeds Act. Retrieved from https://www.govtrack.us/congress/bills/114/s1177

Colwell, R. (2002). Assessment's potential in music education. In *The new handbook of research on music teaching and learning: A project of the Music Educators National Conference* (pp. 1128-1158). Oxford University Press, New York.

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* (Doctoral

dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 8022012)

Dodge, Y., Marriott, F. H. C., & International Statistical Institute. (2003). *The Oxford Dictionary of Statistical Terms*. New York: Oxford University Press. Retrieve from https://stats.oecd.org/glossary/detail.asp?ID=6117

Dowsing, R. D., Long, S., & Craven, P. (2000). An Analysis of the difference between traditional and computer-based assessment of IT Skills. In *ASCILITE* (Vol. 2000, pp. 477-486).

DuBois, R. L., & Thoben, W. (2014). "Sound." In Reas, C., & Fry, B. (2nd Ed.), Processing: a programming handbook for visual designers and artists (No. 6812). MIT Press. Retrieved from https://android.processing.org/tutorials/sound/

Duerksen, G. L. (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education*, *20*(2), 268-272.

Etmektsoglou, 1. E. (1990). An evaluation of pitch matching skills at junior high school, and college levels. Unpublished manuscript.

Etmektsoglou, I. E. (1992). *A computer-based evaluation of pitch matching skills of college freshman students in music* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 9305520)

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Ewell, P. T. (2009). *Assessment, accountability, and improvement: Re-visiting the tension*. Champaign, IL: National Institute of Learning Outcomes Assessment.

Figlio, D., & Loeb, S. (2011). School accountability. *Handbook of the Economics of Education*, *3*, 383-421.

Fisher, R. (2008). Debating assessment in music education. *Research and Issues in Music Education, 6*(1), 1-10.

Fiske, H.E. (1975). Judge-group differences in the rating of high school trumpet performances. *Journal of Research in Music Education, 23*(3), 186-189. doi: 10.2307/3344643

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

Fiske, H. E. (1978). The effect of a training procedure in musical performance evaluation on judge reliability (Unpublished manuscript). University of Western Ontario, London, Canada.

Fiske, H. E. (1983). Judging musical performance: Method or madness? Update: *The Applications of Research in Music Education, 1*(3), 7-10.

Flanigan, G. P. (2008). *An investigation of the effects of the use of SmartMusic software by brass players on intonation and rhythmic accuracy* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3401785)

Flôres Jr, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking? *The Statistician*, *45*(1), 97-104.

Frankel, J. T. (2002). The Internet as a means of assessing state and national standards. In Proceedings from *the 2nd National Symposium on Music Instruction Technology*. *Journal of Technology in Music Learning, 1*(2).

Frey, B. B., Schmitt, V. L., & Allen, J. P. (2012). Defining authentic classroom assessment. *Practical Assessment, Research & Evaluation*, *17*(2), 1-18.

Fukuda, T., Ikemiya, Y., Itoyama, K., & Yoshii, K. (2015). A score-informed piano tutoring system with mistake detection and score simplification. In Proceedings from *Sound and Music Computing Conference*. doi: 10.5281/zenodo.851129

Garman, B. R., Boyle, D., & DeCarbo, N. J. (1991). Orchestra festival evaluations: Interjudge agreement and relationships between performance categories and final ratings. *Research Perspectives in Music Education*, *2*(1), 19-24.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*(8), 519-521. doi: 10.1037/10254-024

Graham, G., Parker, S., Wilkins, J. L., Fraser, R., Westfall, S., & Tembo, M. (2002). The effects of high-stakes testing on elementary school art, music, and physical education. *Journal of Physical Education, Recreation & Dance*, *73*(8), 51-54.

Greiff, S., & Martin, R. (2014). Computer-Based Testing. *Oxford Bibliographies* in Education. doi: 10.1093/obo/9780199756810-0031

Gronlund, N. (1976). *Measurement and evaluation in teaching* (3rd ed.). New York, NY: Macmillan Publishing Co., Inc.

Haley, K. (1998). Watkins-Farnum revisited: Application of modern test theory to music performance assessment. American Educational Research Association. Retrieved from http://eric.ed.gov/?id=ED419842

Hattie, J. (2003). Formative and summative interpretations of assessment information. Retrieved from http://www.education.auckland.ac.nz/webdav/site/education/shared/hattie/docs/formative - and - summative - asses sment - %282003%29.pdf

Hollingsworth, J. (1960). Automatic graders for programming classes. *Communications of the ACM*, *3*(10), 528-529. doi: 10.1145/367415.367422

Hollis, E. (2001a). *The Guildhall School's clear performance assessment system: How clear works.* London, England: Guildhall School of Music and Drama Publications.

Hollis, E. (2001b). *The Guildhall School's Clear performance assessment system: Marking schemes for the assessment categories.* London, England: Guildhall School of Music and Drama Publications.

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

Jacoby, M. (2014, July). Technology: Rehearsal software. Retrieved from http://sbomagazine.com/instruments-gear/1430-all/current-issue/technology/4703-technology-rehearsal-software.html

The Joint Information Systems Committee. (2010). *Computer Assisted Assessment.* Nottingham Trent University. Retrieved from https://now.ntu.ac.uk/d2l/lor/viewer/viewFile.d2lfile/111863/54816/CAA%20Revised/page_01.htm

Karas, J. B. (2005). *The effect of aural and improvisatory instruction on fifth-grade band students' sight-reading ability* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global (UMI No. 3199697)

Keddy, M. P. (2013). *Assessment in the secondary school band programs of British Columbia* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global (UMI No. NS28378)

Kotora, E. J. (2001). *Assessment practices in the choral music classroom: A survey of Ohio high school choral music teachers and college choral methods teachers* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global (UMI No. 3036343)

LaCognata, J. P. (2010). *Current student assessment practices of high school band directors* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3436343)

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72-107.

Lee, E. (2007). *A study of the effect of computer assisted instruction, previous music experience, and time on the performance ability of beginning instrumental music students* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3284028)

Lehman, P. R. (1968). *Tests and measurements in music* (Vol. 4). Prentice-Hall.

Lillya, C. P., & Britton, A. P. (1954). Review of the Watkins-Farnum performance scale for all band instruments. *Journal of Research in Music Education*, *2*(2), 173–174.

Lissitz, R. W., & Jiao, H. (2012). *Computers and their impact on state assessments: Recent history and predictions for the future*. Charlotte, NC: Information Age Publishing, Inc.

Long, M. K. (2011). *The effectiveness of the SmartMusic® assessment tool for evaluating trombone student performance* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3457640)

Lou, S-J., Guo, Y-C., Zhu, Y-Z., Shih, R-C., & Dzan, W-Y. (2011). Applying computer-assisted musical instruction to music appreciation course: an example with Chinese musical instruments. *The Turkish Online Journal of Educational Technology*, *10*(1), 45-57.

Lakes, K. D. (2003). *The response to challenge scale: A generalizable study of an observer-rated measure of self-regulation in children* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3101425)

iPas. (n.d.). *Interactive Pyware Assessment System*. Retrieved from: http://www.pyware.com/ipas/

Iusca, D. (2014). The effect of evaluation strategy and music performance presentation format on score variability of music students' performance assessment. *Procedia-Social and Behavioral Sciences*, *127*, 119-123. doi: 10.1016/j.sbspro.2014.03.224

MacKnight, C. B. (1975). Music reading ability of beginning wind instrumentalists after melodic instruction. *Journal of Research in Music Education*, *23*(1), 23-34. doi: 10.2307/3345200

Macrae, R., & Dixon, S. (2010). Accurate real-time windowed time warping. Proceedings from the 11th International Society for Music Information Retrieval Conference. *ISMIR*, 423-428.

Macri, J. I. (2015). *Computer-assisted self-assessment in high school instrumental music: An exploratory case study* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3708842)

MakeMusic (2009). Annual report pursuant to section 13 or 15(d) of the securities exchange act of 1934. Retrieved from https://www.sec.gov/Archives/edgar/data/920707/000095013709001594/c49781e10vk.htm

MakeMusic. (2013). *Computer-assisted practice software*. Retrieved from http://www.smartmusic.com/products/students/

Make Music. (n.d.). *SmartMusic*. Retrieve from: http://www.smartmusic.com

McCreary, T. J. (2001). *Methods and perceptions of assessment in secondary instrumental Music* (Doctoral dissertation). Available from ProQuest Dissertations & Theses (UMI No. 3030187)

McPherson, G. E. (1994). Factors and abilities influencing sight-reading skill in music. *Journal of Research in Music Education*, *42*(3), 217–231. doi: 10.2307/3345701

McPherson, G. E., & Thompson, W. F. (1998). Assessing music performance: Issues and

influences. *Research Studies in Music Education*, *10*(1), 12-24. doi: 10.1177/1321103X9801000102

McPherson, G. E., & Schubert, E. (2004). Measuring performance enhancement in music. *Musical excellence: Strategies and techniques to enhance performance.* In A. Williamon (Ed.), *Enhancing musical performance* (pp. 61-82). Oxford, UK: Oxford University Press. doi: 10.1093/acprof:oso/9780198525356.001.0001

McQuarrie, S. H. (2008). *The influence of statewide music testing on assessment practices of elementary music teachers in Washington state*. (Unpublished doctoral dissertation), Shenandoah Conservatory of Shenandoah University, VA.

McQuarrie, S. H. & Shwewin, R. G. (2013). Assessment in music education: Relationships between classroom practice and professional publication topics. *Research and Issues in Music Education, 11*(1), 14.

Merriam-Webster, Inc. (1991). *Webster's ninth new collegiate dictionary*. Springfield, Mass., U.S.A: Merriam-Webster.

Meyer, P. (2010). *Understanding measurement: Reliability*. Oxford, UK: Oxford University Press.

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

Min, P. E. (2001). *The effects of visual information on the reliability of evaluation of large instrumental musical ensembles* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3029897)

Morgan, J., & Burrows, B. (1981). Sharpen your edge on choral competition. *Music Educators Journal, 67*(8), 44-47.

Motiwala, A. (2011). *MakeMusic*—Niche business with free cash flow and solid balance sheet is music to my ears. *GuruFocus*. Retrieved from

http://www.gurufocus.com/news/120118/makemusic-mmus-niche-business-with-free-cash-flow-and-solid-balance-sheet-is-music-to-my-ears.

Music Prodigy (n.d.). *Music Prodigy*. Retrieved from https://www.musicprodigy.com/about

Music Educators National Conference (1958). *String adjudication ballot*. Reston, VA: MENC: The National Association for Music Education.

Music Educators National Conference. (1994). *National standards for arts education: What every young American should know and be able to do in the arts*. Reston, VA: MENC.

National Association for Music Education (n.d.). *Assessment in music education.* Retrieved from http://www.nafme.org/about/position-statements/assessment-in-music-education-position-statement/assessment-in- music-education/

National Association of Schools of Music, National Association of Schools of Art and Design, National Association of Schools of Theatre, and National Association of Schools of Dance (1997). *A philosophy for accreditation in the arts disciplines*. Reston, VA: National Association of Schools of Music. Retrieved from https://nasm.arts-accredit.org/wp-content/uploads/sites/2/2015/11/Philosophy-for-Accreditation.pdf

Nebraska Department of Education. (n.d.). *The primary program assessment and evaluation*. Retrieved from https://www.education.ne.gov/oec/pubs/pri_pro/Assessment.pdf

New York State Education Department. (1998). Formal and informal assessment of limited English proficient/English language learners. In *The teaching of language arts to limited English proficient/English language learners: A resource guide for all teachers.* Retrieved from https://www.nysut.org/~/media/Files/NYSUT/Resources/1900/January/ESL_RG.pdf

Nichols, J. P. (1985). *A factor-analysis approach to the development of a rating scale for*

snare drum performance (percussion, evaluation) (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 8527988).

Nielsen, L. D. (2011). *A study of K–12 music educators' attitudes toward technology-assisted assessment tools* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3461345)

No Child Left Behind Act of 2001), Pub. L. No. 107-110, 20 U.SC. § 6319 (2002).

Peters, G. D. (1974). *Feasibility of computer-assisted instruction for instrumental music education* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 7414598)

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception: An Interdisciplinary Journal*, *30*(1), 71-83.

Popham, W. J. (1999). Using standards and assessments: Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8-15. Retrieved from https://eric.ed.gov/?id=EJ581564

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. doi: 10.2307/3345155

Ravitz, J. (2002). CILT2000: Using technology to support ongoing formative assessment in the classroom. *Journal of Science Education and Technology*, *11*(3), 293-296. Retrieved from http://www.jstor.org/stable/40186553

Reimer, B. (1989). *A Philosophy of Music Education* (2nd ed). Englewood Cliffs, NJ: Prentice Hall.

Rudolph, T. (2006). The wide world of SmartMusic. *Music Education Technology*, *4*(1), 10-17.

Russell, B. E. (2010). *The empirical testing of a musical performance assessment paradigm* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3411613)

Russell, J. A., & Austin, J. R. (2010). Assessment Practices of Secondary Music Teachers. *Journal of Research in Music Education, 58*(1), 37-54.

Russell, J. A. (2014). Assessment in instrumental music. *Oxford Handbooks Online*. doi: 10.1093/oxfordhb/9780199935321.013.100

Ryan, C., & Costa-Giomi, E. (2004). Attractiveness bias in the evaluation of young pianists' performances. *Journal of Research in Music Education*, *52*(2), 141–154.

Ruszkowski, J. M. (2006). *The effects of the digital music stand on middle school instrumental music sight-reading* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3216083)

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. doi: 10.2307/3345585

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. In Green, J. L., Camilli, G. & Elmore, P. B. (Eds.), *Complementary methods for research in education* (3rd ed.; pp. 309-322). Washington, DC: AERA.

Shih, Y-J. (2012). *Teaching jazz improvisation to middle school recorder learners: A beginning curriculum* (Master's thesis). Available from ProQuest Dissertations & Theses Global. (UMI No. 1522260)

Shuler, S. C., & Connealy, S. (1998). The evolution of state arts assessment: From Sisyphus to stone soup. *Arts Education Policy Review, 100*(1), 12-19. doi: 10.1080/10632919809599445

Makemusic. (n.d.). Assessment. Retrieved from https://usermanuals.smartmusic.com/SmartMusic/content/assessment.htm

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.

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. doi: 10.1177/1321103X020180010601

Stelzer, T. G. (1938). Construction, interpretation, and use of a sight-reading scale in organ music with an analysis of organ playing into fundamental abilities. *The Journal of Experimental Education*, *7*(1), 35-43. doi: 10.1080/00220973.1938.11010113

Stivers, J. E. (1972). *A reliability and validity study of The Watkins-Farnum Performance Scale* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 7317440)

Straka, G. (2004). Measurement and evaluation of competence. In: P. Descy, M. Tessaring (Eds). *The foundations of evaluation and impact research. Third report on vocational training research in Europe: background report* (pp. 263-311). Luxembourg: Office for Official Publications of the European Communities.

Streckfuss, R. J. (1983). *The effect of a sight-reading pacer machine upon the sight-reading ability of college wind instrumentalists* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 8405277)

Suskie, L. (2004). *Assessing student learning: A common sense approach*. Bolton, MA: Anker Publishing Company, Inc.

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

Tsay, C-J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences of the United States of America, 110* (36), 14580-14585. doi: 10.1073/pnas.1221454110

U.S. Department of Education. (2010). *Assessment: Measure what matters.* Office of Educational Technology. Retrieved from http://tech.ed.gov/netp/assessment-measure-what-matters/

U.S. Department of Education. (2016). *Creating better, smarter, fairer tests: Summary of ESSA assessment regulations*. Retrieved from https://www2.ed.gov/policy/elsec/leg/essa/nprmsassessmentfactsheet762016.pdf

Vendlinksi, T., & Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *The Journal of Technology, Learning and Assessment*, *1*(3), 1-20. Available from http://www.jtla.org.

Vispoel, W. P. (1992). Improving the measurement of tonal memory with computerized adaptive tests. *Psychomusicology: A Journal of Research in Music Cognition*, *11*(1), 27-43. doi: 10.1037/h0094134

VonWurmb, E. C. (2013). *A study of the associations between conditions of performance and characteristics of performers and New York State solo performance ratings* (Doctoral

dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3568300)

Walls, K. C., Erwin, P. M., & Kuehne, J. M. (2013). Maintaining efficient ensemble rehearsals without sacrificing individual assessment: *SmartMusic* assessment could leave the director on the podium. *Journal of Technology in Music Learning*, *5*(1), 4-16.

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. Retrieved from http://www.jstor.org/stable/3345540

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. Retrieved from http://www.jstor.org/stable/3345347

Wapnick, J., Mazza, J. K., & Darrow, A. A. (2000). Effects of performer attractiveness, stage behavior, and dress on evaluation of children's piano performances. *Journal of Research in Music Education*, *48*(4), 323–335. Retrieved from http://www.jstor.org/stable/3345367

Wapnick, J., Ryan, C., Lacaille, N., & Darrow, A. A. (2004). Effects of selected variables on musicians' ratings of high-level piano performances. *International Journal of Music Education*, *22*(1), 7-20. doi: 10.1177/0255761404042371

Watkins, J. G. (1942). *Objective measurement of instrumental music.* New York, NY: Teachers College Bureau of Publications, Columbia University.

Watkins, J. G., & Farnum, S. E. (1954). *The Watkins-Farnum Performance Scale: A standardized achievement test for all band instruments: Score sheets*. Winona, MN: Hal Leonard Music, Incorporated.

Wesolowski, B. (2012). Understanding and developing rubrics for music performance

assessment. *Music Educators Journal, 98*(3), 36-42. doi: 10.1177/0027432111432524

Widmer, G., & Goebl, W. (2004). Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, *33*(3), 203-216. doi: 10.1080/0929821042000317804

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership*, *46*(7), 41-47.

Williams, M. J., & Webster, P. R. (1996). Experiencing music technology: software, data, and hardware, New York, NY: Schirmer Books.

Wu, C-W., Gururani, S., Laguna, C., Pati, A., Vidwans, A., & Lerch, A. (2016) Towards the objective assessment of music performances. In Proceedings from *the 14th International Conference on Music Perception and Cognition* (pp. 99-103).

Yarbrough, C. (2000). What should be the relationship between schools and other sources of music learning. In C. K. Madsen (Ed.). *Vision 2020: The Housewright symposium on the future of music education* (pp. 193-208). Reston, VA: The National Association for Music Education (MENC).

Zanutto, D. R. (2007). *Comparison of online music assessment software*. Retrieved from www.csulb.edu/~dzanutto/index.../comparisonofonlinemusicassessmentsoftware.doc

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. Retrieved from http://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. Retrieved from http://www.jstor.org/stable/3345801

Zuar, B. E. (2006). *The New York State music assessment: History, development, and analysis of the data generated by the 2002 field test* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global. (UMI No. 3225209)

APPENDIX A

UNIVERSITY OF HAWAI‘I IRB APPROVAL

UNIVERSITY of HAWAI'I SYSTEM

Office of Research Compliance
Human Studies Program

DATE:	March 08, 2017
TO:	Arnold, Gabriel, University of Hawaii at Manoa, Music
	Shih, Yi-Ju
FROM:	Magno, Norman, Dir, Animal Welfare and Biosafety Prog, Intrn Dir Human Stds Prog, Social&Behav Exempt
PROTOCOL TITLE:	Evaluation of Music performance: Computerized Assessment Versus Human Judges
FUNDING SOURCE:	NONE
PROTOCOL NUMBER:	2017-00073

NOTICE OF APPROVAL FOR HUMAN RESEARCH

This letter is your record of the Human Studies Program approval of this study as exempt.

On March 08, 2017, the University of Hawai‡i (UH) Human Studies Program approved this study as exempt from federal regulations pertaining to the protection of human research participants. The authority for the exemption applicable to your study is documented in the Code of Federal Regulations at 45 CFR 46.101(b) 2.

Exempt studies are subject to the ethical principles articulated in The Belmont Report, found at the OHRP Website www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

Exempt studies do not require regular continuing review by the Human Studies Program. However, if you propose to modify your study, you must receive approval from the Human Studies Program prior to implementing any changes. You can submit your proposed changes via email at uhirb@hawaii.edu. (The subject line should read: Exempt Study Modification.) The Human Studies Program may review the exempt status at that time and request an application for approval as non-exempt research.

In order to protect the confidentiality of research participants, we encourage you to destroy private information which can be linked to the identities of individuals as soon as it is reasonable to do so. Signed consent forms, as applicable to your study, should be maintained for at least the duration of your project.

This approval does not expire. However, please notify the Human Studies Program when your study is complete. Upon notification, we will close our files pertaining to your study.

If you have any questions relating to the protection of human research participants, please contact the Human Studies Program by phone at 956-5007 or email uhirb@hawaii.edu. We wish you success in carrying out your research project.

1960 East-West Road
Biomedical Sciences Building B104
Honolulu, Hawai'i 96822
Telephone: (808) 956-5007
Fax: (808) 956-8683
An Equal Opportunity/Affirmative Action Institution

APPENDIX B

UNIVERSITY OF HAWAI‘I IRB APPROVAL

UNIVERSITY of HAWAI'I SYSTEM

Office of Research Compliance
Human Studies Program

TO:	Arnold, Gabriel, University of Hawaii at Manoa, Music
	Shih, Yi-Ju
FROM:	Rivera, Victoria, Interim Dir, Ofc of Rsch Compliance, Social&Behav Exempt
PROTOCOL TITLE:	Evaluation of Music performance: Computerized Assessment Versus Human Judges
FUNDING SOURCE:	NONE
PROTOCOL NUMBER:	2017-00073

NOTICE OF APPROVAL FOR HUMAN RESEARCH

This letter is your record of the Human Studies Program approval of this study as exempt.

On October 03, 2017, the request for IRB approval of changes to your exempt project noted above has been reviewed and approved. The proposed amendments will be added into your current project file. The proposed changes do not alter the exempt status of your project. The authority for the exemption applicable to your study is documented in the Code of Federal Regulations at 45 CFR 46.101(b) 2.

This approval does not expire. However, please notify the Human Studies Program when your study is complete. Upon notification, we will close our files pertaining to your study.

If you have any questions relating to the protection of human research participants, please contact the Human Studies Program by phone at 956-5007 or email uhirb@hawaii.edu. We wish you success in carrying out your research project.

1960 East-West Road
Biomedical Sciences Building B104
Honolulu, Hawai'i 96822
Telephone: (808) 956-5007
Fax: (808) 956-8683

APPENDIX C

UNIVERSITY OF HAWAI‘I CONSENT TO PARTICIPATE IN RESEARCH

(PERFORMERS)

My name is Yi-Ju Shih. I am a doctorate student at the University of Hawaii (UH). As part of my degree program, I am conducting a research project. The purpose of my project is to investigate whether computerized assessment on music performance is comparable to human judges. I am asking you to participate in this project because you are at least 18 years old, and you are advanced musicians and can perform sight-reading exercises.

Project Description – Activities and Time Commitment: If you decide to take part in this project, you will be asked to play two sight-reading exercises. The study utilizes two types of assessments: (1) *SmartMusic* assessment and (2) human judges. These two assessments are divided into two parts. The 1st part will be conducted on mm/dd/yy. On that day, you will participate in performing two sight-reading exercises using two hard copy music sheets. The researcher will be the administrator to handle the procedure of *SmartMusic* assessment. *SmartMusic* will automatically assess and record your performance while you play. In the 2nd part of assessment, the recordings are presented anonymously to three human judges for rating. I expect 30 or more people will take part in this project.

Benefits and Risks: There will be no direct benefit to you for taking part in this project. The findings from this project may help music/band instructors to improve their assessment of students' music performance. There is low risk to you in participating in this project.

Confidentiality and Privacy: All performance will be anonymous, and the research data will be saved for three years in the researcher's MacBook Air. Afterward, the materials will be destroyed.

Voluntary Participation: You can freely choose to take part or to not take part in this project. There will be no penalty or loss of benefits for either decision. If you do agree to participate, you can stop at any time.

Questions: If you have any questions about this study, please call or email me at 808-321-8292 & yjs@hawaii.edu. You may also contact my advisers, Dr. Gabriel Arnold at garnold8@hawaii.edu. If you have questions about your rights as a research participant, you may contact the UH Human Studies Program at 808.956.5007 or uhirb@hawaii.edu.

Statement of Consent: I have read the above information. I have asked questions and have received answers. I consent to participate in the study.

_____ _____ Email: _____
Signature of Participant Date

_____ _____
Signature of Investigator Date

APPENDIX D

UNIVERSITY OF HAWAI‘I CONSENT TO PARTICIPATE IN RESEARCH

(HUMAN JUDGES)

My name is Yi-Ju Shih. I am a doctorate student at the University of Hawaii (UH). As part of my degree program, I am conducting a research project. The purpose of my project is to investigate whether computerized assessment on music performance is comparable to human judges. I am asking you to be the human judges in this project because (1) you serve as a local secondary school band/orchestra directors on Oahu and (2) you have over five years band/orchestra teaching experience or held a degree in music or an advanced degree in music education.

Project Description – Activities and Time Commitment: If you decide to take part in this project, you will be asked to assess 30 or more instrumental recordings. The study utilizes two types of assessments: (1) *SmartMusic* assessment and (2) human judges. These two assessments are divided into two parts. The 1st part was conducted on Sep 6, 8, 18, and 20, 2017. The subjects performed two sight-reading exercises using two hard copy music sheets. The researcher handled the procedure of *SmartMusic* assessment. *SmartMusic* automatically assess and record each performance. In the 2nd part of assessment, 30 or more recordings will be presented anonymously to you for rating.

Benefits and Risks: There will be no direct benefit to you for taking part in this project. The findings from this project may help music/band instructors to improve their assessment of students' music performance. There is low risk to you in participating in this project.

Confidentiality and Privacy: All ratings will be anonymous, and the research data will be saved for three years in the researcher's MacBook Air. Afterward, the materials will be destroyed.

Voluntary Participation: You can freely choose to take part or to not take part in this project. There will be no penalty or loss of benefits for either decision. If you do agree to participate, you can stop at any time.

Questions: If you have any questions about this study, please call or email me at 808-321-8292 & yjs@hawaii.edu. You may also contact my advisers, Dr. Gabriel Arnold at garnold8@hawaii.edu. If you have questions about your rights as a research participant, you may contact the UH Human Studies Program at 808.956.5007 or uhirb@hawaii.edu.

Statement of Consent: I have read the above information. I have asked questions and have received answers. I consent to participate in the study.

_____ _____ Email: _____

Signature of Participant Date

_____ _____

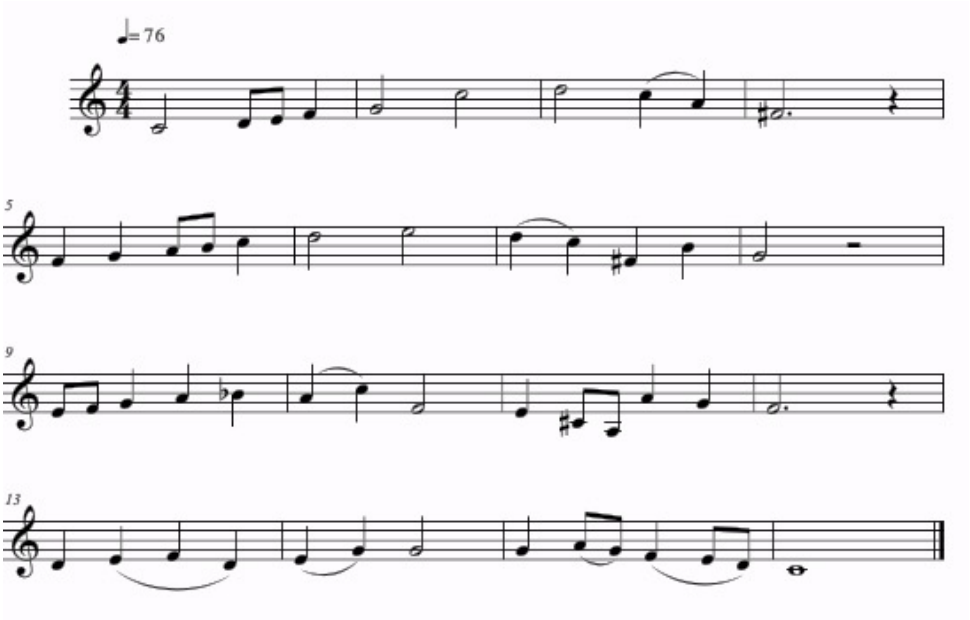
Signature of Investigator Date

APPENDIX E

EVALUATION FORM FOR THE JUDGE PANEL

Trumpet- ____ No.6 Evaluation Form

1. Please circle any errors on the music



2. Please evaluate the recording and give the points on the Rating Scales.

Ratings / Criterion	V	IV	III	II	I	Points
1. Pitch	9 or more wrong notes (0 points)	5 to 8 wrong notes (1 point)	3 to 4 wrong notes (2 points)	1 to 2 wrong notes (3 points)	All notes played correctly (4 points)	
2. Rhythm	9 or more rhythmic errors (0 points)	5 to 8 rhythmic errors (1 point)	3 to 4 rhythmic errors (2 points)	1 to 2 rhythmic errors (3 points)	All rhythm played correctly (4 points)	
3. Tempo	Correct and consistent tempo (2 points) A "change of time" error (1 point) Two or more "change of time" errors (0 points)					
Total						

Notes:

1. A perfect score of pitch is 4 points
2. A perfect score of rhythm is 4 points
3. A perfect score of tempo is 2 points