

Lindenwood University

Digital Commons@Lindenwood University

Dissertations

Theses & Dissertations

Spring 4-2014

The Development and Validation of a Rubric to Enhance Performer Feedback for Undergraduate Vocal Solo Performance

Katherine Herrell

Lindenwood University, kherrell@lindenwood.edu

Follow this and additional works at: <https://digitalcommons.lindenwood.edu/dissertations>



Part of the [Educational Assessment, Evaluation, and Research Commons](#)

Recommended Citation

Herrell, Katherine, "The Development and Validation of a Rubric to Enhance Performer Feedback for Undergraduate Vocal Solo Performance" (2014). *Dissertations*. 384.

<https://digitalcommons.lindenwood.edu/dissertations/384>

This Dissertation is brought to you for free and open access by the Theses & Dissertations at Digital Commons@Lindenwood University. It has been accepted for inclusion in Dissertations by an authorized administrator of Digital Commons@Lindenwood University. For more information, please contact phuffman@lindenwood.edu.

The Development and Validation of a Rubric to
Enhance Performer Feedback for
Undergraduate Vocal Solo Performance

by

Katherine A. Herrell

A Dissertation submitted to the Education Faculty of Lindenwood University

in partial fulfillment of the requirements for the

degree of

Doctor of Education

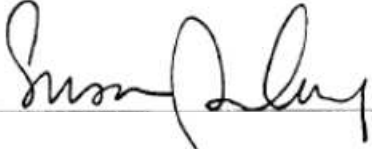
School of Education

The Development and Validation of a Rubric to
Enhance Performer Feedback for
Undergraduate Vocal Solo Performance

by

Katherine A. Herrell

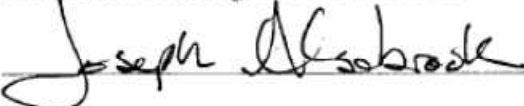
This dissertation has been approved in partial fulfillment of the requirements for the
degree of
Doctor of Education
at Lindenwood University by the School of Education



Dr. Susan Isenberg, Dissertation Chair

4-11-2014

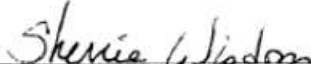
Date



Dr. Joseph Alsobrook, Committee Member

4-11-2014

Date



Dr. Sherrie Wisdom, Committee Member

4-11-2014

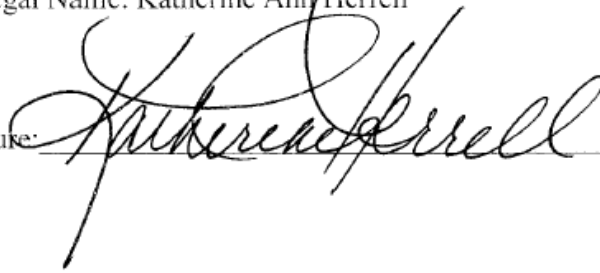
Date

Declaration of Originality

I do hereby declare and attest to the fact that this is an original study based solely upon my own scholarly work here at Lindenwood University and that I have not submitted it for any other college or university course or degree here or elsewhere.

Full Legal Name: Katherine Ann Herrell

Signature: _____

A handwritten signature in cursive script that reads "Katherine Ann Herrell". The signature is written over a horizontal line.

Date: _____

A handwritten date "4/11/17" written in cursive script over a horizontal line.

Acknowledgements

The completion of this dissertation would not have been possible without the guidance and leadership from my dissertation chair, Dr. Susan Isenberg and committee members, Dr. Joseph Alsobrook and Dr. Sherrie Wisdom. I extend my heartfelt gratitude.

I would also like to thank the participating students and faculty at the research institution, as well as the many judges who participated from many universities around the country for the generous gift of their time and talent. The support from the administrators, faculty, alumni, and classmates in the School of Education and the School of Fine and Performing Arts was also instrumental to my success.

I thank my parents, Ron and Theresa Maniscalco, for their unwavering love and support over the years; and finally, I would like to thank my husband, Ken and children, Ben and Tess for their support, encouragement, patience, and sacrifice while I pursued this dream.

Abstract

This is a study of the development and validation of a rubric to enhance performer feedback for undergraduate vocal solo performance. In the literature, assessment of vocal performance is under-represented, and the value of feedback from the assessment of musical performances, from the point of view of the performer, is nonexistent. The research questions guiding this study were 1) What are the appropriate performance criteria, learning outcomes, and meaningful descriptors for various levels of proficiency for undergraduate solo vocal performance? and 2) How do students perceive their use of the feedback from the solo vocal performance rubric to improve future performances? The three groups of stakeholders of the project were voice professors from the research institution who assisted in the development of the rubric; students from the research institution who provided performance excerpts and shared their perceptions about the quality of the feedback; and voice professors from outside the research institution who used the rubric to assess the student performances. Mixed-methods participatory action research was the method used to conduct the study.

Interviews with five experts aided the development of a criteria-specific rubric, which defined performance criteria, learning outcomes, and meaningful descriptors for various levels of proficiency for undergraduate students of singing. The rubric was distributed, along with 20 recordings comprised of 14 students, two professionals, and four repeated student performances, to voice professors who used the rubric to score the performances and provided feedback about the instrument as well as the process. Results of scoring were shared with student performers and interviews conducted about usefulness of the feedback. Seven themes emerged from the research analysis: a) levels

of proficiency, b) performance criteria, c) descriptors, d) numerical scoring, e) comments, f) recording method, and g) song selection relative to the skill level of the singers.

Results of the study determined that the rubric was statistically reliable, and the students received valuable feedback that validated their own self-perceptions and assisted them in long- and short-term goal setting. Practitioners may benefit from further research that explores the validity of the rubric when assigning a grade, assessing live performances, and including additional repertoire.

Table of Contents

Acknowledgements.....	i
Abstract.....	ii
Table of Contents.....	iv
List of Tables.....	xii
List of Figures.....	xiii
Chapter One: Introduction.....	1
Background.....	2
Purpose of the Study.....	3
Rationale.....	3
Research Questions and Hypotheses.....	4
Research Question 1.....	4
Research Question 2.....	4
Hypothesis #1.....	4
Sub Hypothesis #1.....	5
Hypothesis #2.....	5
Sub Hypothesis #2.....	5
Hypothesis #3.....	5
Sub Hypothesis #3.....	5
Hypothesis #4.....	5

Hypothesis #5.....	5
Definition of Terms.....	6
Research Perspective	9
Participatory action research.....	9
Mixed-methods research.....	10
Application of the research methodology.....	11
Limitations	11
Summary.....	12
Chapter Two: The Literature Review	14
Vocal Pedagogy	14
The 16th century.....	16
The 17th and 18th centuries.....	16
The 19th century.....	18
The 20th century.....	19
Characteristics of Good Singing	22
Alignment and breathing.....	22
Tone.....	25
Registration.....	28
Voice classification.....	29
Resonation.....	30

Diction.....	33
Coordination.	34
Expression.....	37
Assessment.....	39
Assessing Musical Performances.....	42
Holistic rating scales.	45
Likert-type scales.	45
Criteria-specific rating scales.....	46
The facet-factorial approach and the development of rating scales.....	48
Rubrics.	55
Developing the rubric.	57
Summary.....	60
Chapter Three: Methodology.....	61
Location	61
Methodology.....	63
Procedures.....	63
Phase I: Rubric design.	63
Phase II: Rubric implementation.	67
Phase III: Student perceptions.	68
Instrumentation	68

Data Analysis	69
Null Hypothesis #1	69
Null Hypothesis #2	69
Null Sub Hypothesis #1	70
Null Sub Hypothesis #2	70
Null Hypothesis #3	70
Null Sub Hypothesis #3	71
Null Hypothesis #4	71
Null Hypothesis #5	71
Participants.....	72
Summary	77
Chapter Four: Results	79
Phase I: Rubric Design.....	79
Interview question #1:.....	80
Interview question #2:.....	81
Interview question #3:.....	82
Interview question #4:.....	84
Comments.	85
Phase II: Rubric Implementation	86
Null Hypothesis #1.	86

Null Sub Hypothesis #1.	87
Null Hypothesis # 2.	89
Null Sub Hypothesis # 2.	89
Null Hypothesis #3.	90
Null Sub Hypothesis #3.	90
Null Hypothesis #4.	91
Null Hypothesis #5.	92
Ratings.	93
Feedback from Judges.....	94
Phase III: Student Perception Results.....	100
Student interview question #1:.....	103
Student interview question #2:.....	106
Student interview question #3:.....	106
Student interview question #4:.....	108
Student interview question #5:.....	109
Emerging Themes	109
Summary.....	110
Chapter Five: Discussion	111
Research Question 1	111
Research Question 2	111

Hypothesis #1.....	111
Sub Hypothesis #1	112
Hypothesis #2.....	112
Sub Hypothesis #2	112
Hypothesis #3.....	112
Sub Hypothesis #3	112
Hypothesis #4.....	112
Hypothesis #5.....	112
Review of Methodology	112
Phase I: Rubric Development	113
Research Question 1.	114
Rubric Organization.....	114
Levels of proficiency.	114
Performance criteria.....	115
Descriptors	117
Breath.	118
Tone.	119
Accuracy.	121
Diction.....	121
Intonation.	122

Vibrato.....	123
Registration.....	123
Agility.....	124
Style.....	125
Expression.....	126
Numerical scoring.....	127
Comments.....	127
Phase II: Rubric Implementation and Validation.....	128
Phase III: Perceived Value of Feedback to the Performers.....	129
Research Question 2:.....	129
Recommendations.....	130
Sliding scale for scoring.....	131
Performance criteria.....	132
Descriptors.....	136
Future Research.....	139
Summary.....	141
References.....	143
Appendix A: Expert Interview Questions.....	149
Appendix B: Research-Based Rubric.....	150
Appendix C: Musical Selection.....	151

Appendix D: Phonetic Transcription of Musical Text.....	155
Appendix E: ANOVA Data for Judges' Scoring.....	156
Appendix F: ANOVA Data for Individual Criteria	157
Appendix G: Student Interview Questions	158
Appendix H: Revised Rubric	159
Vitae.....	160

List of Tables

Table 1. Example of a Likert-Type Scale	46
Table 2. Example of a Criteria-Specific Rating Scale	47
Table 3. Example of a Rubric	57
Table 4. Applied Music Requirements by Major.....	62
Table 5. Explanation of Recording Makeup	67
Table 6. Makeup of Student Participants.....	73
Table 7. Judges' Level of Education.....	76
Table 8. Judges' Years of Experience.....	76
Table 9. Judges' Job Titles.....	77
Table 10. ANOVA Summary for Judges' Scoring	87
Table 11. Results of z-test for Difference in Means of Judges' Scores.....	88
Table 12. ANOVA Summary for Individual Criteria	89
Table 13. Results of z-test for Difference in Means of Individual Criteria	90
Table 14. Intra-Judge Reliability for Ratings of Repeated Items	91
Table 15. Pearson Product-Moment Correlation Matrix for Rubric Categories.....	92
Table 16. Comparison of Calculated Percentages vs. Judges' Holistic Scores	93
Table 17. Average Overall Scores for Student/Professional Performances	94
Table 18. Sample of Comments Presented to Students	102
Table 19. Sliding Scale to Determine Numerical Grades	132

List of Figures

Figure 1. Sample of aggregated data presented to students	101
Figure 2. Rubric organization template	114
Figure 3. Early draft of research-based rubric	117
Figure 4. Rubric descriptors for breath	119
Figure 5. Rubric descriptors for tone	120
Figure 6. Rubric descriptors for accuracy	121
Figure 7. Rubric descriptors for diction	122
Figure 8. Rubric descriptors for intonation	123
Figure 9. Rubric descriptors for vibrato	123
Figure 10. Rubric descriptors for registration	124
Figure 11. Rubric descriptors for agility	125
Figure 12. Rubric descriptors for style	126
Figure 13. Rubric descriptors for expression	127
Figure 14. Comparison of original and revised descriptors for accuracy	135
Figure 15. Comparison of original and revised descriptors for breath	137
Figure 16. Comparison of original and revised descriptors for tone	138
Figure 17. Comparison of original and revised descriptors for diction	139

Chapter One: Introduction

“Performing is considered central to what one must know and be able to do if one is to learn music” (Bergee, 2003, p. 137). For students who are pursuing a baccalaureate degree in music it is recommended that they develop a level of proficiency in technique, have experience performing a variety of repertoire, and possess minimum competency in sight reading (National Association for Schools of Music, 2011-2012). Traditionally, studying applied music at the undergraduate level is conducted in the form of a series of private lessons over the course of a semester and continues over a series of semesters. A “teacher-oriented, master-apprentice relationship” (Bergee, 1993, p. 20) has been and continues to be the norm, at the time of this writing.

Asmus (1999) spoke about the nature of music instruction and the essential factors present in the instructional process. These observations are appropriately applicable to the study of applied music in a private lesson setting.

While music learning may be greatly influenced by the context in which instruction occurs and the entering characteristics of the students who are to receive the instruction, three factors are inherent in all music teaching and learning: (1) the music instruction content and process, (2) the ongoing assessment during instruction, and (3) the outcome of instruction. (p. 20)

The function of the end of semester juried performance has been viewed as an evaluation or summation of the semester's work. The method of evaluating these performances has reflected that paradigm in the use of tools for assigning a grade to the performance. If schools of music and members of the jury panels were to change their

views of the purpose of the end of semester jury to one that is formative in nature; they would need a tool that would not only assign a grade but would also provide meaningful feedback to the students.

Background

Assessment of vocal performance is under-represented in the literature. One of the potential reasons for this lack of research is "vocal research may have been conducted but not reported because sufficient reliability and validity was difficult to obtain" (Wapnick & Eckholm, 1997, p. 429). There are additional considerations present in vocal music that are not present in instrumental performances which can complicate the assessment of vocal performances. For example, "elements such as diction and transmission of the emotional meaning of lyrics have no instrumental counterpart" (p. 429). In addition, the timbral qualities of the vocal instrument are unique to each performer. There are no manufacturing standards for the voice as there are for other instruments. Also, "it is even unclear whether vocal teachers agree on the musical manifestations of certain evaluative adjectives" (p. 429) which can lead to confusion and misunderstanding when describing and assessing vocal performances. In the instrumental and vocal performance assessment research, studies are focused on criteria-specific rating scales, but not on the particular format that is most effective for providing feedback to the performer, the rubric (Saunders & Holohan, 1997; Asmus, 1999; Ciorba & Smith, 2009; Wesolowski, 2012).

Purpose of the Study

The purpose of this study is to develop and validate a research-based rubric with which to assess undergraduate solo vocal performances and which will enhance the feedback provided to students for use in improvement of future performances. The bulk of the research previously published focused on the assessment of instrumental performances, and vocal assessment remains underrepresented. Many researchers (Cooksey, 1977; Levinowitz, 1985; Jones, 1986; Horowitz, 1994; Saunders & Holohan, 1997; Bergee, 2003; Zdzinski & Barnes, 2002) touted the usefulness of the criteria-specific scoring in providing feedback to students, so they could make improvements in future performances. However, I have identified no studies that seek to understand if or how students use the data collected in an evaluation tool to improve future performances.

Rationale

There are many studies that attempted to apply a facet-factorial approach (Butt & Fisk, 1968) to the development of a rating scale evaluating musical performance (Abeles, 1973; Bergee, 1987; Cooksey, 1974; DCamp, 1980; Greene, 2012; Horowitz, 1994; Jones, 1986; Levinowitz, 1985) as well as one study that related the perceptions of the judges using a rating scale (Latimer, Bergee, & Cohen, 2010). The results of these studies demonstrated that the instruments developed were both reliable and valid. All of the studies cited previously considered exclusively the needs of evaluating instrumental performances.

Although vocal performances share characteristics with instrumental performances, there are considerations in evaluating vocal performances that are not present in instrumental performances such as “diction and emotional transmission of the

meaning of the lyrics” (Wapnick & Eckholm, 1997, p. 429). Therefore, it is important to examine the evaluation of vocal performances independently. Related vocal studies include Cooksey (1974) who applied this approach to developing a choral performance rating scale, Jones (1986) who applied this approach to developing a rating scale for high school solo performance, and Wapnick and Eckholm (1997) who tested the validity of a rating scale for undergraduate vocal performance.

Research Questions and Hypotheses

I designed this study to investigate the possibility of developing a tool to assess undergraduate vocal performances that would be both reliable and valid, as well as provide meaningful feedback to the performers. I sought answers to the following research questions:

Research Question 1: What are the appropriate performance criteria, learning outcomes, and meaningful descriptors for various levels of proficiency for undergraduate solo vocal performance?

Research Question 2: How do students perceive their use of the feedback from the solo vocal performance rubric to improve future performances?

The first research question addressed the development of the tool. The second research question addressed measuring the performers' ability to interpret and use the feedback the tool provided. The following hypotheses were designed to test the reliability and validity of the tool:

Hypothesis #1. When scoring performances using the research-based rubric, at least one judge will score differently than the others.

Sub Hypothesis #1. There will be a difference in average mean score on the research-based rubric, when comparing individual judge scoring to the overall group mean score.

Hypothesis #2. When scoring performances using the research-based rubric one category at a time, at least one judge will score differently than the others.

Sub Hypothesis #2. There will be a difference in mean score on the research based-rubric, when comparing individual judge scoring on individual categories to the overall group mean score.

Hypothesis #3. There will be a difference in judges' scoring utilizing the research-based rubric, on repeat performance when compared to the same judges' scoring for the original performance.

Sub Hypothesis #3. The event score when applying the research-based rubric is dependent upon which judge conducted the rating.

Hypothesis #4. There will be a relationship between each of the ratings of characteristics of breath, tone, accuracy, diction, intonation, vibrato, registration, agility, style, and expression, and judges scores utilizing the research-based rubric.

Hypothesis #5. There will be a difference between holistic scores and rubric-based calculated percentage score.

The first hypothesis addressed the inter-judge reliability of the overall instrument. The second hypothesis was designed to determine the criteria-specific validity of the rubric. In other words, the second hypothesis was designed to determine the validity of the instrument on each descriptor. The third hypothesis was designed to determine intra-

judge reliability (the same judge would score a repeated performance the same way twice).

Definition of Terms

Accuracy: Execution of the correct words, pitches, and rhythms (Saunders & Holohan, 1997, p. 264).

Agility: Agility or flexibility is “based on the singer’s ability to negotiate musical challenges nimbly and quickly, including wide pitch intervals, coluratura (fast note) scales and passages, and dynamic variations” (Ware, 2008, p. 97) or simply, “to sing notes rapidly” (Paton, 2006, p. 73).

Assessment: “The collection, analysis, interpretation, and application of information about student performance or program effectiveness in order to make educational decisions” (Asmus, 1999, p. 21).

Assessment for Learning: Type of assessment of which the purpose is “To provide feedback to students to assess the quality of learning and to improve learning behaviors” (Frey & Schmitt, 2007, p. 417). This is only one of the two purposes of Formative Assessment (see definition).

Authentic Assessment: Type of assessment of which the purpose is “to measure ability on tasks which represent real-world problems or tasks” (Frey & Schmitt, 2007, p. 417).

Breath or Breath Support: "The dynamic relationship between the breathing-in muscles and the breathing-out muscles, the purpose of which is to supply adequate breath pressure to the vocal folds for the sustaining of any desired pitch or dynamic level" (McKinney, 1994, p. 53).

Criteria-Specific Performance Scales: These types of scales can be used to conduct performance assessments and "are based on written, objective statements that describe various performance attributes. These objective statements offer more information to the student than assessments using Likert-type scale responses because they offer insight into proficiency levels" (Wesolowski, 2012, p. 37).

Diagnostic Assessment: Type of assessment of which the purpose is to determine "which musical skills a student has already learned" (Hale & Green, 1999, p. 28).

Diction: a "general term that refers to using the prevailing standards of word usage and pronunciation in a comprehensible manner and style" (Ware, 2008, p. 83).

Evaluation: "The collection and use of information to make informed educational decisions" (Asmus, 1999, p. 21).

Expression: The ability to "understand and express not only the sound, but also the meaning of the songs" (Ware, 2008, p. 114).

Formative Assessment: Type of assessment of which the purpose is "to provide feedback to the teacher to assess the quality of instruction or to improve teaching behaviors, or to provide feedback to the student to assess the quality of learning and to improve learning behaviors" (Frey & Schmitt, 2007, p. 417).

Intonation: Refers to the ability to sing in tune, that is, "to reproduce accurate pitches of music scales and modes with a relative degree of accuracy" (Ware, 2008, p. 96).

Performance Assessment: "An assessment that determines a student's ability to perform assigned tasks rather than his or her ability to answer questions" (Asmus, 1999,

p. 21) or an assessment of which the purpose is “to measure a skill or ability” (Frey & Schmitt, 2007, p. 417).

Register: "A particular series of tones, produced in the same manner (by the same vibratory pattern of the vocal folds), and having the same basic quality" (McKinney, 1994, p. 93).

Registration: The ability to sing consistently across and between the different registers of the human voice (Ware, 2008, p. 56).

Rubric: “A set of scoring criteria used to determine the value of a student’s performance on assigned tasks; the criteria are written so students are able to learn what must be done to improve their performances in the future” (Asmus, 1999, p. 21).

Style: Refers to characteristics of a piece of music that could include the type of music, the historical period from which it came, the manner of expression that is used, the expected way of performing it that we associate with a specific composer or school of composers, or the way of performing that belongs to an individual (Paton, 2006, p. 65).

Summative Assessment: “Assessment performed to determine the overall effectiveness of an educational program” (Asmus, 1999, p. 21).

Tone: The sustained phonation that occurs when the vocal mechanism is engaged for singing. Sometimes referred to as tone production, tone quality, vocal sound, timbre, or phonation (Paton, 2006, pp. 16-17).

Vibrato: “a pulsation of pitch, usually accompanied by synchronus pulsations of loudness and timbre, of such extent and rate as to give a pleasing flexibility, tenderness, and richness to the tone” (Seashore, 1938, p. 33).

Research Perspective

I selected a mixed-methods participatory action research methodology. This type of methodology was a good fit for the study because mixed-methods participatory action research is usually practitioner-led in collaboration with key stakeholders to solve a local issue and employs both qualitative and quantitative methods of data gathering (Fraenkel, Wallen, & Hyun, 2012). The action research model was also the best fit to answer the research questions and hypotheses that guided the study.

Participatory action research. Action research is generally conducted by a practitioner to "solve a problem at the local level" (Fraenkel et al., 2012, p. 611) and utilizes generally accepted methods of research "although on a smaller scale" (p. 611). The basic assumptions underlying action research are that "the participants have the authority to make decisions, want to improve their practice, are committed to continual professional development and will engage in systematic inquiry" (Fraenkel et al., 2012, p. 590). I was in a position to recommend and implement changes in instructional and assessment practices in my teaching environment. Participatory action research "attempts to empower participants or bring about social change" (Fraenkel et al., 2012, p. 591) by involving stakeholders in the research process at a level appropriate to their role in the research setting and their expertise. This involvement in the change process fosters buy-in from stakeholders and can facilitate smoother implementation of the resulting changes. I chose to include the administrator, teachers of singing, and the students who were the key stakeholders in my proposed research and resulting proposed changes. Purposive sampling is the common choice for action researchers since they are studying a specific

problem that is local in nature (Fraenkel et al., 2012, p. 595), and it was the type of sampling that I selected for this study.

Typically, action research involves four steps: "identifying the research question or problem, gathering the necessary data, analyzing and interpreting the data and sharing the results with the participants, and developing an action plan" (Fraenkel et al., 2012, pp. 593-595). There are at least five advantages to action research.

It can be done by just about anyone, in any type of school or other institution, to investigate just about any kind of problem or issue. It can help to improve educational practice. It can help education and other professionals to improve their craft. It can help them learn to identify problems systematically. Finally, it can build up a small community of research-oriented individuals at the local level. (Fraenkel et al., 2012, p. 612)

Mixed-methods research. "Mixed-methods research involves the use of both quantitative and qualitative research methods in a single study" (Fraenkel et al., 2012, p. 557). The results of these separate methods are combined to present a more complete picture of the phenomenon under study than either method could produce on its own. In mixed-methods research, the respective strengths of qualitative and quantitative methods are seen as compensating for the respective weaknesses of each method (Fraenkel et al., 2012, p. 558).

"Disadvantages of mixed-methods research involve the time, resources, and expertise necessary to conduct this type of research well" (Fraenkel et al., 2012, p. 583). While quantitative methods are usually associated with positivism and qualitative methods are usually associated with postmodernism, mixed-methods are usually

associated with pragmatism. Pragmatists believe that one should use whatever methods best answer the research question or questions at hand (Fraenkel et al., 2012). This pragmatic view also appears to be a good fit with the action research model.

Application of the research methodology. To apply the mixed-methods participatory action research methodology, I followed a series of steps over three phases. The first phase was the preparation of a research-based rubric, which was achieved via a review of the literature and responses to experts' interviews, and the preparation of recordings which included both student and professional singers. The second phase was the implementation of the research-based rubric where the research-based rubric was used by judges to assess the recorded performances. The third phase was the collection of student feedback which was facilitated through interviews of the student performers in an attempt to determine the information they learned from the completed research-based rubrics used by the judges who assessed their recorded performances.

Limitations

In action research, the results are usually "weak in external validity" (Fraenkel et al., 2012, p. 596) and replication is necessary if the results are to be generalized to other settings. This is true for my study, since it was designed with a particular program, with its own strengths and weaknesses, and set of students in mind, and the data was unique to that particular educational setting. The rubric developed as a part of this research study is appropriate for use with undergraduate vocal soloists only. The levels of development and maturation were defined exclusively to this population, as the rubric was developed. The rubric was considered valid and reliable only with respect to undergraduate vocal soloists from a program with less competitive admissions standards than some

institutions may have employed. Potential threats to validity in an action research study could include

the possibility of collector bias, because the data collector is well aware of the intent of the study. He or she must take care not to overlook results or responses he or she does not want to see. Implementation and attitudinal effects are also a strong possibility, as either implementers or data collectors can, unwittingly, distort the results of a study. (Fraenkel et al., p. 595)

It was important for me as the researcher to commit to reporting all of the data, regardless of whether or not it fit my assumptions. Only audio recordings of vocal solo performances were used to test this rubric. It might be appropriate to consider including visual criteria when adapting this instrument for live performances. Judges were selected from a wide pool of professors at institutions of varying sizes, and selections for critique were performed by students with diverse levels of achievement.

Summary

Vocal performance assessment is under-researched among the music education literature (Wapnick & Eckholm, 1997), and the purpose of this action research was to develop and test a rubric with which to assess undergraduate solo vocal performances, which would also provide constructive feedback to the student. To achieve these purposes, I designed a rubric based on the methodology outlined by Wesolowski (2012). To test the reliability and validity of this rubric, 36 university-level teachers of singing used the rubric to score recorded student performances. After performances were scored, I met with each student performer to collect qualitative data about their perceptions of the quality and usefulness of the feedback provided. The findings from this study contribute

to the literature by a) adding to the limited body of knowledge about valid and reliable vocal performance assessment instruments and b) providing initial insight into student perceptions about the feedback received from a music performance assessment tool.

Chapter Two: The Literature Review

The review of the literature included an historical look at the practice of vocal pedagogy followed by an examination of publications that discussed the characteristics of effective singing. Following the examination of the field of vocal pedagogy, the literature review focused on principles of assessment and narrowed in focus to consider the specific considerations in assessing musical performances. A detailed discussion is provided on several types of assessments that have been used in assessing musical performances, including holistic rating scales, Likert-type scales, criteria-specific rating scales, and rubrics. This examination of the types of assessments used included a discussion of the evolution of the research completed on the application of these types of assessments.

Vocal Pedagogy

The evolution of vocal pedagogy is a long and storied one. It began with the first documented method of singing outlined in a letter by Maffei in 1562 (Sell, 2005, p. 9), and continued through the age of bel canto, or the Italian school, in the 17th and 18th centuries, a verismo style in the 19th century, and the rise of nationalistic singing styles in the 20th century. This entire history is replete with inconsistencies and disagreements about what was the most desirable sound and what was the proper method for acquiring it.

One of the difficulties in finding agreement among vocal experts was the difficulty in defining the word 'pedagogy' itself. "Even dictionaries do not agree. Some say it is 'the art of teaching'; others use the phrase 'the science of teaching'" (Kiesgen, 2005, p. 41), and therein lies the chief problem. It is also important to note that the term

'pedagogy' in the field of applied musical study was inclusive of both children and adults, unlike other fields such as adult learning in which 'pedagogy' is specifically limited to the teaching of children, and 'andragogy' is the term used to refer to the teaching of adults (Henschke, 1998).

Many teachers of singing base their pedagogical knowledge on their own experience rather than an empirical or systematic approach (Himonedes, 2009). Kiesgen (2005) also recognized this general belief that most modern pedagogical practices seemed to be "subjective, reflecting only the personal opinions and taste of the teacher" (p. 41). However, the voice is both a physical and acoustical instrument that operates under the laws of both physics and acoustics (Miller, 2000; Kiesgen, 2005). There were significant advances in the area of voice science such as spectral analysis that are available at the time of this writing, but were not available to teachers of the past (Miller, 2000). The study of voice science was at once, "vital to those who wish to understand the singing voice" (Kiesgen, 2005, p. 44) and an amazing opportunity to have data that were once unattainable (Miller, 2000; Kiesgen, 2005).

Miller (2000) expressed dismay that despite the scientific data available, voice instruction still relied heavily on "the confusing language of imagery" (p. 42), and Himonedes (2009) stated, "there is evidence that teachers of singing customarily use imagery (including kinesthetic and visual imagery) in teaching vocal technique, often allied to a reliance on sensation and the development of aural awareness" (p. 45). Miller (2004) provided the example that "the teacher may well know what 'spin the tone,' 'float the voice' and 'rounder sound' mean, but the terms themselves do not tell the student

how to spin, float, or round the tone. Today's student wants not flowery imagery, but practical assistance" (p. 196).

In addition to the valuable study of voice science, teachers of singing must also recognize that studying the methods of the great teachers of the past is a worthy endeavor since "we find that there is a level of agreement among many of them about the kinds of ideas that work. The fact that so many have found these same ideas to work seems to make them less subjective" (Himonedes, 2009, p. 43). The next section is an examination of the evolution of vocal pedagogy from the sixteenth century through modern times. Throughout this time period, there was an evolution from pure imitation of the teacher to a greater focus on the mechanics of singing which paralleled the advances in scientific knowledge and discovery that occurred during each historical period.

The 16th century. The primary pedagogues of this period were Maffei and Zacconi. According to Sell (2005), the ideal tone quality at the time was *coluratura* and "a light flexible voice that sang softly" (p. 11). Key characteristics of singing emphasized at the time were avoiding nasality, the importance of physical appearance, a "slight and pleasing" (p. 9) tremolo or vibrato, evenness of tone throughout the range, and vocal registration. Much was written at the time about the importance of breathing, but there was little written about how to properly execute breath or about the breathing mechanism. "Most singers and teachers seemed to agree that the best way to learn was by imitating a good teacher, but without suggesting what constitutes a good one" (p. 11).

The 17th and 18th centuries. According to Gerry's 1995 essay (as cited in Austin, 2011), the age of *bel canto*, or the Italian school, was a method of teaching vocal

technique developed in the time period between the late 17th century and the 19th century during which “many of the best musical minds of Italy were occupied in developing the technique of singing and in establishing sound rules and laws for the development of the singer, based entirely upon the empirical approach” (p. 343). During this time period, opera was in its infancy, and singers began to emerge from the shadows of the courts or religious institutions and began to make their mark as individual soloists. "There was new emphasis on vocal display, agility, dramatic ability, and voice production capable of filling not just smaller chambers but large halls and theaters" (Sell, 2005, p. 11).

The beginning of this period was aligned with the Baroque style period. "Four vocal qualities were demanded by the Baroque composers: perfect intonation, good breathing technique, clear diction and meaningful expression of the text" (Sell, 2005, p. 11). There was a focus on "medical research into the singing voice during this time" (Sell, 2005, p. 11), and other characteristics of singing emphasized at the time were vibrato, breathing with some scientific understanding of the breathing mechanism, resonance, clear diction and precise articulation, two areas of registration, the importance of a good ear, raising and lowering of the larynx, legato singing, and the adjustment of the vocal tract for optimal resonation.

One of the primary characteristics of this period of vocal instruction was the uniformity and general agreement among teachers about the process teaching singing (Sell, 2005; Austin, 2011). "The teaching procedure was largely oral. Undoubtedly, at first, it was to a certain extent imitative, for, rest assured, these teachers were also singers . . . It was no unusual thing for singers to work several years before being taught any

repertoire" (Austin 2011, p. 344). Although there was greater understanding at the time about the science of singing, the student was

not required to study the anatomy of the human head, nor was he required to think about or to know the names of muscles over which he really had no conscious control. He was taught only to listen for and to note the sensation of beautiful tone. (p. 344)

According to Sell (2005) many of the developments and traditions of this time period, especially the exercises and vocalises, were handed down over the centuries and still greatly influence modern teachers of singing.

The 19th century. The 19th century was the time of the Romantic period of musical style. The style of singing reflected the changing demands of the style period with the need for vocalists to accommodate larger ensembles with thicker texture and more extreme dynamics. Thus, there was a shift from the Italianate style of singing to a verismo style, which could be described as a more realistic style with a heavier and darker tone quality (Sell, 2005), and a shift from the previous method based mostly on “observation and imitation, to experimentation and more scientifically grounded justifications of pedagogical method” (p. 32).

Garcia II, the son of a pedagogue trained in the Italian school, was an important teacher of singing during this period (Stark, 1999; Sell, 2005). His attempts to "combine science and the art of singing has made Garcia a controversial figure to this day. He was, nonetheless, very highly regarded and is considered by many to have had the greatest ever influence on the art of singing" (Sell, 2005, p. 23). According to Stark (1999), the most important concept introduced during the period was Garcia II's coup de la glotte

which was based on the scientific information about the process of phonation available at the time and was the first introduction of the concept of onset that is a major consideration in modern singing. In addition, Garcia's theory that the vocal tract can and should be adjusted to achieve optimal resonance was a prelude to the singers' formant theory that is prevalent today, and his definition of registers "was a good starting point for a fuller understanding" (p. 90) of modern registration.

Lamperti had long been recognized as the father of *appoggio* which he labeled *la lotte vocale* or vocal struggle (Stark, 1999; Sell, 2005). "This means that when we are singing, the inspiratory muscles labor against the expiratory muscles to retain the breath within the body" (Stark, 1999, p. 24). He believed that "good singing uses surprisingly little breath" (Stark, 1999, p. 24). This was a concept that became a cornerstone of modern singing technique (Miller, 1996).

Another leading pedagogue from this period was Stockhausen. He was a performer and teacher and studied with Garcia II.

Stockhausen was a pioneer in the linguistic approach to vocal pedagogy. He placed great emphasis on the study of vowels as indispensable for beauty of tone, and insisted that vocalizes should be practised on all vowels and was aware of the importance of the tongue, lower jaw movements and laryngeal positioning in vowel formation. (Sell, 2005. p. 26)

This approach is also generally accepted by modern teachers of singing (Miller, 1996).

The 20th century. The early 20th century saw a rise in nationalism across the globe. This socio-political shift was also evident in the singing and pedagogical practices of the period as "traditional methodology began to be converted to supposedly 'national'

styles of singing” (Sell, 2005, p. 32). “Peculiarities of language began to emerge: French nasality; Germanic hard consonants and the Spanish aspirate” (Sell, 2005, p. 32).

Composers began to embrace the history and literature of their homelands as bases for their musical works, and “words were deemed more important than vocal display, and were accorded equal rights with the accompaniment” (Sell, 2005, p. 32). Pedagogues continued to increase their reliance on emerging science to inform their approaches to singing and the teaching of singing.

In Britain, there are at least three tonal ideals still being espoused. Of these one is generally based on the Italianate ideal; the next has traces of German technique, and the third is the very English ‘cathedral’ tone which has its roots in the British liturgical tradition, with its fondness for the seeming ‘purity’ of tones as produced by the choral treble voice. (Sell, 2005, p. 35)

There were three key figures who influenced the vocal pedagogy and the British style of singing. They include Shakespeare who was trained by the Lampertis and advocated the still popular “spreading of the upper back as an alleged assistance for breathing” (Sell, 2005, p. 33) and Greene who also continued the Lamperti tradition, White who was a proponent of the “Sinus Tone Production” theory which held that resonance occurred in the sinuses and was likely to result in the cathedral tone sound. The third is Estill who was an American pedagogue who had a lasting impact on the British. Estill’s method was “a highly organized system which is orientated to ‘feel’” (Sell, 2005, p. 37) and was centered around six voice qualities, “speech, falsetto, cry,

twang, opera and belt” (Sell, 2005, p. 37), the feel of which were each memorized so that they could be replicated consistently.

In Germany, Marienssen-Lohmen was a disciple of Italianate singing and “disliked many German practices, for example, heavy covering of the voice, too much use of head voice, and the low positioned larynx” (Sell, 2005, p. 34). Armin “was noted for the advance of the ‘heroic’ voice of the German school” (Sell, 2005, p. 34) which proved to be a damaging practice.

In France, de Reszke was the most influential teacher. Although trained in the Italianate style, he rejected that style in favor of a more relaxed method of breathing which was characterized by “a collapsed chest with rounded shoulders. He advised the use of the sigh as a means to release the glottis and the tongue; a raised head position . . . and placement of tone in the masque and on the bridge of the nose” (Sell, 2005, p. 35). He ultimately lost his voice, and France did not produce famous singers during this time period. Therefore, his methods were not generally accepted as valid.

Stanley was a British singer and pedagogue whose voice was damaged by the methods by which he was trained. He moved to America to study and embraced a fully scientific approach to the understanding of the vocal process. Another prominent American pedagogue was Vennard. He also embraced a scientific approach to singing, even though he was primarily schooled in the Italian approach. The science of his time (1909-1971) was incomplete, which led Vennard to be a proponent of some of the elements of the German style of singing such as “yawn/sigh, lower abdominal breathing, vocal registration, and vocal tract positions” (Sell, 2005, p. 36).

Miller is an American singer and pedagogue. His book *The Structure of Singing* (1996) became the standard textbook for training American singers and teachers of singing. His belief was, “Artistry cannot be realized without the technical means for its presentation. Systematic vocal technique and artistic expression are inseparable” (Miller, 1996, p. xvi). He devoted his scholarly life to collecting the best scientific information and the most effective and healthy methods of all the national styles, in order to develop an approach to singing regarded as the current standard.

Characteristics of Good Singing

I compared and contrasted five leading books on modern singing technique (McKinney, 1994; Miller, 1996; Paton, 2006; Vennard, 1949; Ware, 2008). I chose these books because they were the leading textbooks for vocal pedagogy and class voice courses at colleges and universities in the United States. The purpose of this comparison was to begin to follow the methodology suggested by Wesolowski (2012), as well as to determine the general performance criteria and desired learning outcomes for solo vocal performance, begin to specify the range and degrees of proficiency for each performance criterion, and develop meaningful descriptors for each criterion performance level. From this comparison emerged eight categories of criteria discussed in the literature: a) alignment and breathing, b) tone, registration, c) voice classification, d) resonance, e) diction, f) coordination, and g) expression. Each of these categories are examined in detail in the discussion that follows.

Alignment and breathing. It is necessary to combine the discussion of alignment and breathing under one heading because their purposes are intertwined, and several of the writers on vocal technique (Vennard, 1949; Miller, 1996; Paton, 2006)

discussed them as a unified part of the breathing process. “Appropriate alignment of the body is extremely important in setting up the right conditions for coordinating the vocal process” (Ware, 2008, p. 41).

Good posture for singing means using the body in such a way that our breathing muscles work easily and there is no interference to the sound we want to produce. Poor posture can mean that your lungs cannot expand fully or that your voice cannot reach all of the notes you should be able to sing because the throat is stretched out of its proper shape. (Paton, 2006, p. 7)

Miller (1996, 2004) referred to the ‘noble’ position which was a concept handed down from the Italian bel canto school of singing, which was an alert and upright, yet free and not stiff posture. McKinney (1994) used the following adjectives to describe proper posture: “alert, balanced, buoyant, erect, expansive, flexible, free-to-move, happy, poised, vibrant” (p. 36). Other descriptors found in Ware (2008) included “vital and balanced”, “anchored to the floor yet buoyant”, “knees flexible and unlocked”, “abdominal area remains relaxed on inhalation and the lower abdominal area remains firm (but not tight) on exhalation”, “chest remains comfortably high, but not pushed and upward”, “shoulders hang loosely”, “neck is held in an erect position, but not rigidly”, and “the head is balanced” (p. 41). Vennard (1949) maintained that the combination of an engaged breathing mechanism and a relaxed body was the most desirable technique (p. 18).

Vennard (1949) believed that breathing was “the most important factor in tone production” (p. 17). He wrote “it may be said that no matter how well a person sings, if

his breathing can be improved his singing can also” (p. 17). He advocated a mixture of rib (costal) breathing and abdominal (diaphragmatic) breathing when at the time they were considered to be separate and exclusive types of breathing.

McKinney (1994) and Ware (2008) described a breathing process that seems to have evolved from the marriage of the former costal model with the diaphragmatic model, which includes four phases: inhalation or inspiration, suspension, controlled exhalation or expiration, and recovery. When inhaling, “the breath seems to move into the body, down to the lungs, and out around the middle of the body” (McKinney, 1994, p. 49). This expansion of the lower abdomen was caused by the displacement of the abdominal organs by the descending diaphragm. The brief suspension period which follows was not present in natural breathing. Its purpose was to prepare the breathing mechanism for the process of phonation. The expiration or controlled exhalation coordinates “with the vocal cords to produce phonation. The length and period of exhalation is determined by the demands of the musical phrase” (McKinney, 1994, p. 51). This process was the exact reverse of the inhalation process. The lungs recoil, the abdominal muscles relax, the diaphragm recoils upward, and air is expelled from the lungs through the trachea (Ware, 2008). Then after the air was expelled, there was a brief period where all of the “muscles associated with breathing relax” (McKinney, 1994, p. 52).

Miller (1996, 2004) advocated the principle of “appoggio” which was sometimes misunderstood to be “narrowly related to the management of airflow during singing” (Miller, 2004, p. 1), but was actually “a complete system of structural support, during which the muscles of exhalation and those of inspiration maintain an antagonistic

balance, inciting a stable but dynamic relationship.” (p. 1). In contrast to diaphragmatic or belly breathing discussed previously,

the appoggio avoids excessive outward distention of the . . . pelvic, lower abdominal regions during both inhalation and phonation. At complete inspiration, the lower torso expands laterally dorsally and frontally. For most of the sung phrase, the large, flat muscles of the abdomen can be trained to remain relatively stable, near the inspiratory position. Breath renewal, silently taken without perceptible chest displacement, re-establishes abdominal expansion . . . Complete abdominal contraction occurs rarely; it is restricted to the termination of exceedingly long phrases. (Miller, 2004, p. 2)

Many teachers of singing advocated the appoggio style of breathing. These teachers spoke of “singing on the gesture of inhalation” (Miller, 2004, p. 13) because singers who used this type of breathing essentially maintained the position of inhalation throughout the breathing cycle. They advised their students to maintain a feeling of fullness throughout the phrase instead of collapsing as the air left their bodies.

Tone. Tone was sometimes referred to as tone production, tone quality, vocal sound, timbre, or phonation. This was the process of the vocal cords activating due to the subglottal air pressure and making sound. McKinney (1994) used the following descriptors when speaking about the characteristics of good tone: “freely produced; pleasant to listen to; loud enough to be heard; rich, ringing and resonant; energy flows smoothly from note to note; consistently produced; vibrant, dynamic, alive; flexibly expressive” (p. 77). Descriptors used by Ware (2008) included: natural sound, freedom

from tension, clear and in tune, elasticity, “ample volume with ringing forward in the mask placement”, chiaroscuro, flexibility and agility (p. 59). Paton (2006) listed the following characteristics in his discussion of good singing: a) audibility, b) resonance, c) clarity, d) intelligibility, e) pure intonation, f) dynamic variety, g) timbre consistency and variety, h) vibrato, i) range, and j) ease of freedom (p. 17). All of the adjectives used by these authors were valuable resources when I set about the process of writing descriptors for the rubric developed as part of this study.

There was lengthy discussion in the literature reviewed about the importance of onset and release to the production of efficient vocal tone. Miller (1996) described this process as “establishing dynamic muscle equilibrium through onset and release” (p. 1) and addressed this as the first and most important consideration in learning or teaching vocal technique. If the onset is lax or aspirate (having too little air pressure) and sounds like a whisper, the tone can be breathy and will likely be under pitch because there is too little subglottal air pressure to cause the vocal cords to fully adduct and too little air pressure to effectively support the pitch. McKinney (1994) referred to this as “hypofunction” (p. 82).

If the onset was pressed or glottal (having too much air pressure), the pressure could build up behind the vocal cords and the onset was like a tiny explosion or grunt. McKinney (1994) referred to this as “hyperfunction” (p. 87). This excess pressure could lead to the vocal cords adducting too much and the tone could be tight and the pitch will likely be sharp. The lax or pressed onset was at times used for dramatic or expressive effect, but the ideal type of onset that produced the most balanced and desirable tone was the balanced onset. A balanced onset produced a balanced tone that would exhibit the

descriptors previously mentioned (McKinney, 1994; Miller, 1996; Paton, 2006; Ware, 2008). Vennard (1949) also discussed the importance of properly executing the “coup de glotte” or “stroke of the glottis” in a balanced manner to initiate a desirable tone and observed that some teachers of singing based their entire methodology on perfecting this technique in the belief that a perfect onset will necessarily lead to good tone (p. 25).

Similarly, the type of onset and tone that a singer produces will continue through the duration of the phrase and affect the also important release at the end of the tone. A lax or aspirate release and phonation was usually followed by a release of the same quality due to a collapsing of the breathing mechanism throughout the phrase and was described by Ware (2008) as lacking “intensity and is often very weak” (p. 59). A pressed or glottal release was

epitomized by the ‘terminal grunt’ one hears when large-voiced opera singers end a loud high note. Although it has its place as a dramatic device in performance situations, this glottal release is out of place in soft to moderately loud dynamic levels in in low-to-medium pitch ranges. (Ware, 2008, p. 59)

Generally, “accomplished singers strive to end most phrases with the same consistent tone quality sustained throughout the phrase. This requires a coordinated or balanced release, with the vocal folds under neither too much tension nor too little” (Ware, 2008, p. 59).

The ability to master the basic skills of breathing and tone production were necessary for students to progress beyond the very beginning stages of singing. Much attention was paid in the beginning of a student's training to these skills. They were necessarily the focus of assessment in the first few semesters of study. Without proper

mastery of both breathing and tone production, it was impossible for students to progress to the more complex abilities that were the hallmarks of advanced, mature singers.

Registration. There was much disagreement among experts when discussing vocal registration, or the different registers of the voice. Registration refers to the idea that there are areas of the vocal range treated differently, technically speaking. One way of viewing the concept of register was to start with the area of the range that was most comfortable for the singer. McKinney (1994) called this the “modal voice,” and Vennard (1949) called this the “full voice.” Vennard (1949) and Miller (2002) also spoke of the “heavy mechanism” when referring to the lower register and the “light mechanism” (Miller, 2002, pp. 152-153) when referring to the upper register. Ware (2008) discussed the popular Three Registers Theory where the lower register was referred to as the “chest register”, the higher register was referred to as the “head register”, and the middle register was referred to as the “mixed or middle register” (p. 54). Vennard (1949), McKinney (1994), Miller (1996), and Ware (2008) discussed extreme, or auxiliary, registers which included the vocal fry or strohbass which was an extremely low part of the male range, falsetto which was a high part of the range outside the male’s modal or full voice range, and the whistle or flageolet register which was an extremely high part of the range outside of the female’s modal or full voice range.

Where the agreement lay was in the idea that “most singers are well aware of unequal tones in their voices” (Paton, 2006, p. 24), and a singer’s voice should be unified throughout the entire range and across all registers. The transition between registers and through the so-called “zona di passaggio” and “pivotal zones” (Ware, 2008, p. 57) between them should be smooth and imperceptible. The ability to skillfully transition

between registers throughout one's range was a hallmark of an advanced or expert singer and was a characteristic to be considered when assessing developing singers such as those studying at the undergraduate level.

Voice classification. Both McKinney (1994) and Ware (2008) spent a portion of their writing discussing voice classification. There were six major classifications of voices, a) soprano, b) mezzo-soprano, c) contralto, d) tenor, e) baritone, and f) bass, into which singers could be categorized according to their a) range, b) tessitura, c) timbre, and d) transition points. Range was the "total compass of a voice part or a singer" (McKinney, 1994, p. 111) or all of the pitches the singer was able to sing. Tessitura "is concerned with that part of the range which is receiving the most use" (McKinney, 1994, p. 111). For female voices, the soprano voice was capable of the highest range and tessitura, the mezzo-soprano voice was capable of a middle or medium range and tessitura. The contralto was capable of the lowest range and tessitura. For male voices, the tenor voice was capable of the highest range and tessitura, the baritone voice is capable of a middle or medium range and tessitura. The bass was capable of the lowest range and tessitura.

Timbre or quality was another consideration in determining voice classification and, therefore, appropriate repertoire. Generally, timbre was described in terms of light or heavy and lyric or dramatic in reference to the "size of voice, kind of tone quality, or style of singing" (McKinney, 1994, p. 112). However, these terms were not mutually exclusive. While there were many singers with light and lyrical voices who had high ranges and tessituras, there were also singers with heavy and dramatic voices who had similarly high ranges and tessituras.

Register transition points were sometimes referred to as lifts or breaks. Despite the disagreement discussed in the previous section about registration, it was generally accepted that most singers have more-or-less clearly defined areas in the voice where there is a 'register' change, a change of quality, or the necessity for some change in technique. It is also generally agreed that the transition points of high, medium, and low voices follow that same sequence, with the higher voices having higher 'lift' notes, etc. The actual pitches on which the transition should occur are not so widely agreed upon, however. (McKinney, 1994, p. 113)

In considering the assessment of solo vocal performances in developing singers, this criteria was important to consider because performers should be singing repertoire that was consistent with their voice classification in order to optimize performance and to avoid injury to the performer.

Resonation. "Resonation is the process by which the basic product of phonation is enhanced in timbre and/or intensity by the air-filled cavities through which it passes on its way to the outside air" (McKinney, 1994, p. 120). Sound produced by the larynx "(the result of airflow and vocal fold approximation) is modified by a mechanical acoustical filter, the vocal tract" (Miller, 1996, p. 48). "The vocal tract resonator tube consists of the pharynx [or throat], the mouth, and at times, the nose. By skillfully combining the resonating cavities, vocal timbre can be controlled" (Miller, 1996, p. 48). "For maximum resonance, the vocal resonators must be optimally enlarged" (Ware, 2008. p. 76).

It is important to know that overtones are involved in resonance and the perception of focus. In the sound spectrum of any instrument there exists

clusters of energy frequencies known as formants that produce specific tonal characteristics. For instance, specific configurations of formants make it possible for us to discern the subtle or not-so-subtle differences between various instruments, voice types, or speech phonemes. (Ware, 2008, p. 78)

"Formant frequencies are peaks that determine the shape of the acoustic spectrum of a vowel" (Miller, 1996, p. 55). In other words, each vowel sound, when properly formed, could be measured with a spectrograph at a specific frequency, which was measured in hertz. The so-called "singer's formant" (Miller, 1996, p. 55) was the optimal frequency at which the singer could access all of the appropriate overtones to create a fully resonant, ringing sound that could penetrate through and above a full symphony orchestra. This frequency varied by voice type, but occurred when the singer was vibrating at around 2800 hertz. This characteristic was only present in the singing voice, not the speaking voice, and in the voices of highly trained singers (Miller, 1996, p. 55).

An open throat was commonly mentioned as a desired characteristic in singing (McKinney, 1994; Miller, 1996; Vennard, 1949; Ware, 2008). Since the pharynx or throat was one of the main resonators of the vocal tract, it must necessarily be as open as possible to provide an appropriate space for the sound to resonate. Miller (1996) advocated a throat position that was approaching a yawning position but "without . . . the muscles tension that must occur in the throat with the yawn posture" (p. 59). McKinney (1994) called this the "beginning-of-a-yawn position" (p. 131) and listed four descriptors of the proper open throat:

1. sufficient size to bring out the low partials, 2. sufficient flexibility to adjust (tune) to different pitches, 3. sufficient softness to absorb undesirable high partials and respond to a broad range of pitches, and 4. sufficient muscle tonus to preserve the character of the tone. (p. 130)

Tone placement was another factor that affected the proper resonance of the vocal instrument. It was a dubious label since singers did not actually place the tone somewhere in their head. Rather, they were being taught to recognize and remember what sensations they felt when they produced a desirable tone.

Vocal pedagogies are not in agreement as to what these sensations should be. "Forward placement" is the aim of some teachers: 'into the masque (mask),' 'into the mouth,' into the upper jaw,' 'out the front,' 'behind the eyes,' 'into the sinuses,' 'at the end of the nose,' 'on the lips,' etc. Other teachers believe the tone should be directed posteriorly: 'down the spine,' at the back of the throat wall,' 'up the back of the throat wall, then over into the forehead,' 'into the body,' 'into the back half of the head,' etc. (Miller, 1996, p. 61)

"Regardless of what theory of 'placement' a teacher may embrace, there is always the peril that the student may not experience the sensation that the teacher's terminology means to elicit" (Miller, 1996, p. 61). Miller (1996) suggested that it was wiser to find a technical way to describe this particular acoustical principal rather than to rely on using subjective or confusing imagery (p. 61), and Vennard (1949) also expressed a similar sentiment.

Because proper placement and effective resonance were characteristics of mature singing, the development of these abilities were important to track in the development of a young singer. These abilities were not developed overnight, but over years of intense study. However, signs of progress toward these goals were evident in the developing singer and should be considered when assessing developing singers.

Diction. Any discussion of diction is necessarily preceded with a discussion of terms since several of these terms were used interchangeably even though each had a unique meaning and each “represents a specific aspect of expressive linguistic communication” (Ware, 2008, p. 82). Articulation refers to the use of the speech organs such as the lips, tongue, jaw, and teeth to form individual speech sounds known as phonemes. Enunciation is the clear production of syllables, words, or sentences. Pronunciation is the ability to pronounce syllable, words, and phrases according to a set of accepted standards. Diction is a “general term that refers to using the prevailing standards of word usage and pronunciation in a comprehensible manner and style” (Ware, 2008, pp. 82-83).

Vowels and consonants are the building blocks of language, but have distinct characteristics that contribute to the production of tone and the communication of ideas. Vowels are “made with a free, unrestricted flow of breath” (Paton, 2006, p. 47) and are the starting point for vocal study. Since much of singing is sustaining vowels on a pitch over time, mastery of proper vowel formation “provides a firm foundation for producing efficient vocal tone” (Ware, 2008, p. 83) and allows access to the appropriate overtones for each particular sound and greatly influences the overall sense of resonance in a singer’s tone. In contrast, consonants are the result of the interruption of air by the

speech organs. Ware (2008) stated that “Consonants carry more ‘information’ than do vowels because they clarify and reveal the meaning and expressive power of languages. Consonants also aid in voice projection by generating positive noise in the acoustic spectrum” (p. 90).

Consonant articulation must be quick and precise because, as interrupters of the tone, there is the possibility that a prolonged consonant could interrupt the tone too much. The articulation of consonants should be exaggerated because they are less sonorous than vowels and do not project as well and the consonants help to “provide the necessary energy for firm phonation” (McKinney, 1994, p. 156). In other words, good consonants are the impetus for good vowels and, thus, good tone.

The evolution of a young singer's ability to master diction from proper vowel formation to proper use of consonants to coordination of vowels and consonants to correct pronunciation in all singing languages including English, Italian, French, and German should be assessed at intervals along the journey of a young singer's development. These individual skills, once mastered and coordinated, provide the foundation for beautifully resonant tone as well as the ability to communicate meaning to the listener.

Coordination. Experts were in agreement that mastery of individual technical skills was only one step toward competent, effective singing. The goal for the developing singer was to develop a mature technique that effortlessly combined the individual elements of singing to produce a pleasing and consistent tone. The proper coordination of the individual elements of singing should result in correct intonation (Ware, 2008), balanced vibrato (McKinney, 1994; Miller, 1996; Vennard, 1949; Ware, 2008), flexibility

and agility (McKinney, 1994; Ware, 2008), sostenuto (Miller, 1996; Ware, 2008), dynamic flexibility and control (Miller, 1996; Ware, 2008), an extended range (McKinney, 1994; Miller, 1996; Ware, 2008), and consistent tone quality with regard to vowel and register alignment (McKinney, 1994).

Intonation referred to the ability to sing in tune. Ware (2008) defined intonation as the ability “to reproduce accurate pitches of music scales and modes with a relative degree of accuracy” (p. 96). He taught that out-of-tune singing was “usually the result of one or more malfunctioning components of the vocal process (respiration, phonation, registration, resonance, and articulation)” (p. 96).

A good vibrato could be defined as “a pulsation of pitch, usually accompanied by synchronus pulsations of loudness and timbre, of such extent and rate as to give a pleasing flexibility, tenderness, and richness to the tone” (Seashore, 1938, p. 33). In addition to being regular, a pleasing vibrato should be free and relaxed, warm and expressive (Paton, 2006, p. 27). An irregular vibrato usually results in an uneven pattern, a too slow pattern or wobble, or a too fast pattern or bleat and is an indication of improper technique or possible injury (McKinney, 1994; Miller, 1996; Ware, 2008).

Agility or flexibility was “based on the singer’s ability to negotiate musical challenges nimbly and quickly, including wide pitch intervals, *coloratura* (fast note) scales and passages, and dynamic variations” (Ware, 2008, p. 97). Paton (2006) simply defined agility as the ability “to sing notes rapidly” (p. 73). Mature singers should be able to meet the demands of sophisticated and intricate music with little evidence of effort.

Sostenuto referred to “the sustaining capabilities of the voice and depends on the coordination of respiration, phonation, resonance, and articulation” (Ware, 2008, p. 98). The challenge with this particular skill lies in the ability to use the appropriate amount of energy while saying free from tension while controlling dynamic levels and accommodating the stylistic demands of a particular musical selection (Ware, 2008, pp. 98-99).

Ware (2008) wrote that range extension would occur “when inhibiting tensions are relieved and vocal efficiency is established.” (p. 100). McKinney (1994) explained that range extension was the coordination of three factors: energy, space and depth. First, as you sing up a scale, each tone requires a little more energy than the one just below it. The total body response is increased; the support mechanism increases its output; more breath pressure is delivered to more resistive vocal cords; and the sound gets louder, for there is a built in crescendo as you sing up the scale. (p. 182)

Second, “as you sing higher, your must use more space” (p. 183). You can choose to increase the opening of the mout or create space at the back of the mouth. Third, “as you sing higher you must use more depth. The natural tendency . . . is to thin out and tighten or whiten at the pitch rises” (p. 183). Depth referred to “actual sensations of depth in the body and vocal mechanism; it also refers to mental concepts of depth as related to tone quality” (p. 183).

As the range extends, more attention must be paid to the consistency of tone quality across the registers. One of the factors that greatly influenced this was vowel formation. It was necessary for singers to adjust the vocal tract away from the so-called

pure vowel formation to negotiate the need for consistency of timbre. This skill of vowel modification or *aggiustamento* “may well be the most subtle of all technical aspects” (Miller, 1996, p. 158) of singing.

The coordination of basic skills to achieve advanced and mature singing abilities was another area that was critical to monitor when assessing developing singers. Therefore, special attention should be paid to these developing abilities when semester assessments are conducted to ensure that progress is being made toward those goals. Any tool that is used to assess and provide feedback to students should include opportunities to remark on these skills of coordination.

Expression. McKinney (1994) and Vennard (1949) limited their discussion to the purely technical aspects of singing without consideration of expression or artistic interpretation. However, Miller (1996) and Ware (2008) spent a good deal of time writing about coordinating the technical skills with the art of communication. Miller (1996) stated, “Technique is of no value except as it makes communication possible” (p. 204) and cautions against spending too little time exploring the “artistry in singing” (p. 197) during lessons in favor of study of the purely technical aspects of singing. Ware (2008) stressed the importance of understanding the music to be performed in the context of the style period which it represents. The performer must have an understanding of style periods (Renaissance, Baroque, Classical, Romantic, and Modern) in the stylistic practices of each period and apply that knowledge to the actual performance.

Ware (2008) also included a discussion of the need for performers to possess “dramatic skills” (p. 106) that is, the ability to communicate to their audience the meaning and emotional context of a song. He suggested that singers become students of

literature and poetry to develop a thorough understanding of “denotation, connotation, imagery, figurative language, allusion, meter, tone, and pattern” (p. 113) to help them to “understand and express not only the sound, but also the meaning of the songs” (p. 114). This internal understanding of the text must be translated through the voice using dynamics, articulation, phrasing and vocal color; through the body using appropriate posture and gesture; and through the face using expression that can register “the emotions called for, honestly and accurately communicating sentiment” (Miller, 1996, p. 202) without taking on the appearance of making “extraneous movements” or “mugging” (p. 202). Ware (2008) mentioned the specific actions and demeanor expected of a performer before, during, and after a performance such as elements of preparation (care of health, positive thinking, and relaxation), performance (entering the stage, acknowledging the audience, presenting the song, acknowledging the accompanist, and exiting the stage), and post-performance (self-assessment of musicianship, technique, diction, stage presence, and dramatic presentation).

It was evident from the literature that understanding of style and the ability to sing expressively were essential skills, although advanced, that each singer must develop. Even though these advanced skills required years of study to develop, it was appropriate to expect that very basic understanding, for example, a simple understanding of the translation of the text, should be expected from the even most inexperienced singers. Therefore, these skills could and should be included in any instrument that attempts to assess the abilities of developing singers.

The comparison of the five books on singing technique (McKinney, 1994; Miller, 1996; Paton, 2006; Vennard, 1949; Ware, 2008) proved invaluable in helping to

determine the general performance criteria and desired learning outcomes for solo vocal performance. The eight categories of criteria identified in the literature: a) alignment and breathing, b) tone, c) registration, d) voice classification, e) resonance, f) diction, g) coordination, and h) expression, were a starting point for developing the criterion performance levels for the new rubric. The adjectives and descriptions collected from the literature were used, in part, to develop the descriptors for each characteristic and the levels and ranges for assessment.

Assessment

“Educational reform and associated accountability issues have made music educators aware of the need to perform assessments that precisely and substantively document what a student has learned” (Asmus, 1999, p. 19). “New mandates and public concerns regarding accountability are additional reasons” (Goolsby, 1999, p. 31) that the need for reliable and valid assessments has increased in prominence.

One of the primary obstacles that presented itself very early in discussion of assessment was the obvious need to clearly define the terms and concepts that were to be used in the discussion. There were many terms used either interchangeable or incorrectly by educators and laypeople. For this reason, the following discussion includes definitions of the terms that are to be used in the context of this paper.

In their seminal work *Understanding by Design*, Wiggins and McTighe (1997) defined assessment as “the act of determining the extent to which the curricular goals are being and have been achieved” by conducting a variety of formal in informal assessments (p. 4). Assessment is not only a means to assign a numerical grade, but it is also a tool the educator can use to identify and design rich learning activities as well as an avenue

for two way communication between the teacher and the learner that will further aid in the student's learning and development (Asmus, 1999, p. 19; Dunbar, 2011, p. 32). Just as assessment had a clearly defined purpose and role in the educational process, evaluation also had its specific function as well. Unfortunately, these two terms were sometimes used interchangeably when, in fact, they have completely different meanings and purposes. Therefore, it was important to examine the role and function of evaluation and its relationship to assessment before continuing the discussion.

Evaluation had a summative role of which the purpose was to "determine the overall effectiveness of an educational program" (Asmus, 1999, p. 21). Evaluations were usually performed at the end of a course of study to determine if the learning objectives had been met and to inform future choices about course content, delivery and instructional methods. The focus of the evaluation was not the student, rather, it was how the students' results indicated the overall effectiveness of the program.

In contrast, assessment's role was more formative. The purpose of formative assessment was "to provide feedback to the teacher to assess the quality of instruction or to improve teaching behaviors, or to provide feedback to the student to assess the quality of learning and to improve learning behaviors" (Frey & Schmitt, 2007, p. 417). The latter purpose was sometimes referred to as "assessment for learning" (Frey & Schmitt, 2007, p. 417). "Formative assessment is not only a powerful measurement tool but also a powerful instructional tool because it allows students to observe their own progress" (Marzano, 2007, p. 24). Hattie (2012) suggested that equal or greater amounts of time should be spent on formative assessment and the feedback students receive should help

them to clearly understand the goal, where they were in relation to the goal, and what they needed to do to close the gap.

There are various types of assessment that may be employed to determine student progress and achievement. Authentic assessment measured a student's ability to apply their skills or knowledge of concepts in a "real world" context (Asmus, 1999; Frey & Schmitt, 2007). In the context of music instruction, this could take the form of students publicly performing music to demonstrate their learning (Asmus, 1999). A juried performance would be an example of a performance assessment.

"Portfolio assessment is a tool for recording both process and product—tangible evidence of a student's learning collected over time. Ideally, an individual's portfolio contains items such as musical programs, teachers' written evaluations, recordings, and written self evaluations" (Asmus, 1999, p. 20). Standards-based assessment was used by educators when they used "local, district, state, or national standards as criteria for student performance" (Asmus, 1999, p. 20). Music educators also employed diagnostic assessments to determine "which musical skills a student has already learned" (Hale & Green, 1999, p. 28). This type of assessment was used either as a pre-test or a placement test. For example, this type of assessment could be used to determine placement in a select ensemble, selection for chair positions (first chair, second chair, etc.).

While the juried performance was usually conducted at the end of a semester of study, it was often confused with an opportunity for an evaluation or summative assessment. I would argue that since the student continues his study over a series of eight semesters, this type of performance assessment should actually be formative in nature. Its goal should be to monitor student progress and provide feedback to the students to

improve their approach to study and ultimately their skill level as they progress through subsequent semesters of study. Therefore, it is important that the method or tool used to assess these performances provides the optimum amount of feedback to the student.

Assessing Musical Performances

“One of the primary goals of music education in general is musical independence” (Goolsby, 1999, p. 35) or the ability to “function autonomously when they leave school” (Hale & Green, 1999, p. 29). Therefore, self-evaluation has been advocated as an important practice in developing students’ independence. Effective use of self-evaluation was described by Goolsby (1999), who stated that students were able to “improve their listening habits and, over a period of years, become rather astute and listening critically to their own performance” (p. 35). They became able to “look beyond their own point of view and to see themselves in relation to the standard” (Hale & Green, 1999, p. 29) they are trying to master. If teachers of singing consistently provided meaningful feedback to their students over the course of their undergraduate training, including the end of semester juried assessment, the students could begin to learn the language and process of self-assessment and prepare to take on the role self-assessor when they leave the protective nest of the undergraduate voice studio.

Many textbooks framed assessment as an objective endeavor for which there was one correct answer for each question posed to the learner. “Music, on the other hand, is a discipline that embraces expressive decisions and divergence of response” (Wesolowski, 2012, p. 36). Researchers agreed that this divergent, subjective nature of the discipline was a chief challenge in assessing musical performances (Abeles, 1973; Bergee, 2003;

Ciorba & Smith, 2009; Dunbar, 2011; Wesolowski, 2012). In fact, Fiske (1983) found that assessors often had no concrete criteria upon which to base their ratings.

Researchers (Abeles, 1973; Cooksey, 1977; Levinowitz, 1985; Jones, 1986; Bergee, 1987, 1989, 2003; Horowitz, 1994; Zdzinski & Barnes, 2002; Ciorba & Smith, 2009; Greene, 2012) demonstrated that “performance assessment under the correct circumstances exhibit good reliability and validity. These investigators succeeded in developing reliable and valid rating scales or rubrics for music performance” (Bergee, 2003, p. 138). Abeles (1973) advocated the use of rating scales, stating that they “improve evaluation because adjudicators must use a common set of evaluative dimensions rather than develop their own subjective criticisms. If the evaluative dimensions adequately sample the content area under investigation, the scale should have satisfactory content validity” (p. 246). Wesolowski (2012) wrote about the advantages of using a rubric instead of a rating scale since “rubrics serve as documentation for student achievement that provides music teachers with a written form of accountability” (p. 36). This practice of documenting student achievement was essential in light of “recent policy initiatives instituted by major accrediting bodies [that] require the implementation of” such measures (Ciorba & Smith, 2009, p. 5).

Wesolowski (2012) also discussed an additional weakness with assessing musical performances through the use of diagnostic and summative assessments. He believed that these types of assessments were unable to provide adequate feedback to students that would enable them to improve. Instead, they conditioned students to learn to avoid making mistakes because errors were the focus of the assessment. He stated, "By implementing more formative methods of assessment, such as the rubric, music educators

can better monitor and improve students' learning as well as shape their instruction . . . in response to what they discover" (Wesolowski, 2012, p. 37).

Another consideration with regard to valid and reliable assessment was the judges themselves. "Researchers have frequently concluded that more than one adjudicator is necessary for good reliability" (Bergee, 2007, p. 345) because a single adjudicator was subject to the effects of "a tight schedule, fatigue, and a myriad of other obstacles" (p. 345). In addition, the judges' level of experience was also an important factor. "Some studies have used student evaluators with acceptable results, but students apparently do not have enough expertise to validly assess high-level performance" (Bergee, 2007, p. 346). Bergee (2007) also noted that interrater consistency was the metric often measured in studies that evaluated judge reliability and validity; however, interrater agreement was an altogether different measurement and recommended that it should be explored. Fiske (1975) and Bergee and Platte (2003) found that area of expertise (e.g., wind experts judging non-wind players) was significant when assessing technique, and that judges should have a "background in the same general family" (Bergee, 2007, p. 356) of the instruments they were judging. Sequence of performances also had an effect on results. Judges tended to rate performances that occurred later in the day more leniently (Bergee, 2007, p. 346). Bergee (2007) also recommended "further research should examine the effect of training evaluators to the assessment protocols they will use. Raters should reach consensus on trial 'anchor' performances before proceeding with the main task" (p. 356).

Holistic rating scales. The first, and most simplistic, approach to rating musical performances was the holistic approach. This approach was researched and advocated by Fiske (1975). He concluded

Judges should be asked to assign only an overall grade for...performances. This trait was shown to be significantly related to all other traits and, therefore, rating other traits and summing or averaging scores for other traits is a needless, time-consuming operation. Judges should give attention to the performance for the purpose of making one decision (and one grade) only rather than making several decisions in a relatively short time. In this way, more time is allowed for making the one decision, greater attention can be given to the performer, and results based on the one score will be subject to no greater error (and probably much less) than would be expected on the basis of several trait ratings. (p. 196)

There were obvious benefits for the judges using this method since it was the least time consuming method because it required the judge to provide only one overall grade. Although this method was deemed reliable, its validity remained questionable. It provided little feedback to the performers about which aspects of their performance were executed well or which aspects needed improvement.

Likert-type scales. The Likert-type scale was usually applied to a collection of criteria (e.g. tone, intonation, etc.) without descriptors to generate a numerical score or rating (Latimer et al., 2010, p. 169). Although these Likert-type scales provide reliable and valid scores, they are lacking in their ability to provide formative feedback to the

performers (Wesolowski, 2012, p. 37). Table 1 provides an illustration of a Likert-type scale.

Table 1.

Example of a Likert-Type Scale

Characteristic	Rating
Tone	1 2 3 4 5 6 7 8 9 10
Intonation	1 2 3 4 5 6 7 8 9 10
Rhythmic accuracy	1 2 3 4 5 6 7 8 9 10

Validation of the Likert-type rating scales as used in the assessment of solo and ensemble music performances was widely found in the literature (Abeles, 1973; Bergee, 1989; Horowitz, 1994; Jones, 1986; Zdzinski & Barnes, 2002) from which one can conclude that these types of scales “often can display a high degree of reliability, but their validity may remain uncertain. Specifically, decisions based on a Likert-type scale reflect an adjudicator’s level of agreement with a general statement concerning a student’s level of performance” (Ciorba & Smith, 2009, p. 6). Similar to the holistic approach to assessing performances, these assessments were far too general to be useful to the performers in the improvement of future performances.

Criteria-specific rating scales. The third approach to evaluating musical performances found in the literature was to use criteria-specific performance scales (Bergee, 2003; Cooksey, 1977; Horowitz, 1994; Jones, 1986; Levinowitz, 1985; Saunders & Holohan, 1997; Zdzinski & Barnes, 2002). “Criteria-specific performance scales are based on written, objective statements that describe various performance attributes. These objective statements offer more information to the student than

assessments using Likert-type scale responses because they offer insight into proficiency levels” (Wesolowski, 2012, p. 37). They were constructed “by a) identifying dimensions central to assessing a specific performance medium . . . b) categorizing the items that best represent each dimension, and c) pairing them with Likert-type, categorical response scales” (Latimer et al., 2010, p. 169). An example of a criteria-specific rating scale is included in Table 2.

Table 2.

Example of a Criteria-Specific Rating Scale

Statement	Level of agreement
Performer plays mechanically	SD D N A SA
Spiritless playing	SD D N A SA
Intonation is inconsistent	SD D N A SA
Plays all registers in tune	SD D N A SA
Performance is clean	SD D N A SA
Poor synchronization of the tongue and fingers	SD D N A SA

Note. From Bergee (2003, p. 10).

Wesolowski (2012) described both the benefits and drawbacks of criteria-specific performance scales saying

The benefits of criteria-specific performance scales are that they are able to assess very specific levels of performance aptitude accurately and reliably. However, adjudicators may find difficulty in judging music performances based on single, generalized objective statements. Also, these scales do not offer any type of quality judgement or convey the level

of achievement. There is only a judgment of “present” or “absent” according to the specified criteria on the checklist. (p. 37)

Consistent with Wesolowski’s (2012) statement about the limitations of criteria-specific performance scales, Saunders and Holohan (1997) argued that “Likert-type scales offer too little information about what causes certain performances to be successful or unsuccessful because they involve responses to a single general statement about a dimension rather than descriptions of various levels of mastery within that dimension” (p. 169).

The facet-factorial approach and the development of rating scales. The facet-factorial approach to rating musical performances has been the primary methodology for developing criteria specific rating scales for musical performances. This approach was based on the methodology developed by Butt and Fiske (1968) to be used as a measurement of dominance in personality.

Strategies for the measurement of dominance were classified as facet vs. trait, and factorial vs. rational, yielding the four approaches compared in the study: rational facet, factorial facet, rational trait, factorial trait . . .

The distinction between trait and facet involves the degree to which the personality variable is conceptually delineated and subdivided before scales are developed to measure it. In the trait strategy, the construct is identified by a label or sentence and is measured at a global level. In contrast, the facet strategy assumes that a trait has several facets, each with several forms or elements. The proposed objective is a homogeneous scale for each specified part of the construct. (Butt & Fiske, 1968, p. 505)

In the context of musical performances, the facet-factorial approach has been widely studied and can be applied in the following manner. The musical performance itself is a construct (complex behavior) which consists of facets (performance components such as tone, intonation, tempo, articulation, etc.) which can further be divided into elements (descriptive statements) for which scales (usually, Likert-type scales) can be developed (Cooksey, 1977).

Abeles (1973) used this facet-factorial approach to develop and validate a clarinet performance adjudication scale. His purpose was to “improve the evaluation of music performance . . . through replacement of judges’ general impressions by ratings arrived at by more systematic procedures” (p. 246). From essays about auditory aspects of junior high school clarinet performances solicited from 17 instrumental music teachers enrolled as graduate students, Abeles was able to identify 54 different descriptive statements, and these statements were categorized by the researcher (p. 246). Then, a list of 40 additional statements were developed from other published literature on clarinet pedagogy and performance. The 94 statements were phrased either positively or negatively and a five point Likert-type scale was developed and used to rate each statement (p. 246).

One hundred recorded performances were assessed by 50 instrumental music teachers using the 94 statements. Each judge was asked to assess two randomly selected performances to which they listened several times. A factor analysis was performed on the results. “The factor analysis solution that best agreed with the a priori structure was a six-factor rotation. The six factors were interpretation, intonation, rhythm continuity, tempo, articulation, and tone” (Abeles, 1973, p. 248). Five items (facets) were selected to represent each of the six factors determined, resulting in a rating scale with 30

statements. The 30 statements were paired with a five point Likert-type scale (Abeles, 1973, p. 248). The rating scale was used by instrumental music teachers enrolled in a graduate program to assess three groups of 10 performances selected from the original 100 (Abeles, 1973, pp. 248-249). Judges heard each performance only once (Abeles, 1973, p. 249). Estimates of interjudge reliability and criterion-related validity were obtained for the “factor scores as well as the total scores” (Abeles, 1973, p. 249) and found that “the six factor structure for clarinet performance seemed essentially the same as the a priori theoretical structure based on the literature search,” (p. 254) and the evaluation instrument was “reliable and valid” (p. 254).

Cooksey (1977) applied the facet-factorial approach to develop a rating scale for high school choral music performance in an effort to develop a “precise, objective measuring instrument,” to mitigate judges’ reliance on “subjective opinions,” and gain some consensus on the criteria for such evaluations (p. 101). To define the criteria, evaluative statements were collected from three major sources: (1) adjudication sheets containing judges’ comments about actual high school choral performances . . . (2) . . . critiques written by choral teachers on recorded performances of high school choruses, and (3) . . . essays written by choral experts on aural aspects of high school choral performances. (Cooksey, 1977, pp. 101-102).

The facet-factorial approach yielded a structure of seven factors of choral performance. The factors were “diction, precision, dynamics, tone control, tempo, balance/blend and interpretation/musical effect. Thirty-six items were selected to form the subscales to measure the seven factors” (Cooksey, 1977, p. 113). These items were

paired with a Likert-type scale, and, like Abeles (1973) the resulting instrument “achieved high inter-judge reliability and high criterion-related validity” (Cooksey, 1977, p. 113).

DCamp (1980) employed the facet-factorial approach to develop a reliable and valid rating scale for high school band music performance. His study was based on the previous works by Abeles (1973) and Cooksey (1977). His study of the facet-factorial approach yielded five factors “central to the evaluation of high school band performance” (DCamp, 1980, p. 41). They were “Tone-Intonation, Balance, Musical Interpretation, Rhythm and Technical Accuracy” (DCamp, 1980, p. 41). He then applied six descriptive statements with the highest loadings to each factor listed, to develop his rating scale, which he then determined through statistical testing to be both reliable and valid. DCamp also determined that there was a need for further research into the type of feedback that such an adjudication could provide to the directors of the bands evaluated (p. 46).

Jones (1986) applied the facet-factorial approach to construct a scale for rating high school vocal solo performance to improve the “precision of measurement, thus providing more structured evaluations” (p. ix). To define the evaluative criteria, Jones (1986) solicited essays from members of the National Association of Teachers of Singing and searched the literature concerning vocal pedagogy. The analysis yielded five factors: a) interpretation/musical effect, b) tone/musicianship, c) technique, d) suitability/ensemble, and e) diction and 32 items for the subscales (pp. ix-x). The rating scale was determined to be reliable and valid. Jones believed that the visual aspects of a vocal performance were critical to the evaluation of such performances and chose to use video recordings for the experts to evaluate; however, his study found that the visual aspects were too

influential and tended to cause the judges to disregard the actual aural characteristics of the performance (Jones, 1986. p. 87).

Bergee (1987) also employed this methodology to develop a rating scale for euphonium and tuba music performance and “sought to determine whether a more homogenous group of performances would affect reliability in a substantial manner” (p. 12) and sought to generalize the rating scale to all brass instruments. Bergee (1987) observed that Abeles (1973) used 10 judges in his study. This number was not typical of an actual end of semester jury evaluation setting. Usually there were far fewer judges present to judge brass performances. He wished to investigate whether reliability could be maintained if he used fewer judges in his sample. His study resulted in a rating scale that was both reliable and valid in the conditions that he sought to investigate.

Horowitz (1994) used the facet-factorial approach to develop a rating scale “designed to measure the ability of a guitarist to perform a jazz improvisation” (p. 13). The scale was designed for “(1) teachers to assist in student evaluation, and (2) students as an aid to self-evaluation and a guide for critical listening” (p. 13). Similar to the previous studies, Horowitz (1994) developed a pool of descriptive statements by analyzing the content of interviews and essays. These statements were then “paired with a five point Likert-type scale and used by 28 judges to evaluate 70 student improvisations” (p. 13). He then performed a factor analysis which “indicated that the scale should consist of three factors: Musicianship, Expression, and Overall Structure . . . [and] ten items were chosen to represent each of the three subscales to form the final 30 item scale” (p. 1). The scale was determined to be reliable and valid.

Like Jones (1986), Wapnick and Eckholm (1997) also attempted to develop a rating scale to rate solo voice performance. They wanted to consider only the aural aspects of vocal performance in constructing a rating scale to assess such performances. They interviewed experts and reviewed the literature on vocal pedagogy which led them to develop a scale based on 12 factors: “appropriate vibrato, color/warmth, diction, dynamic range, efficient breath management, evenness of registration, flexibility, freedom throughout vocal range, intensity, intonation accuracy, legato line, and resonance/ring” (p. 430). They also included a question regarding overall performance, which was to be given independently from the ratings of the twelve factors.

Wapnick and Ekholm (1997) then invited experts to use this new scale to evaluate 19 different performances of the same excerpt and found that “intra-judge reliability was much higher than inter-judge reliability” (p. 435) which supported their belief that there was much disagreement among vocal experts about how to evaluate vocal performances. However, they did find that “evaluations pooled from four or more judges demonstrated considerable inter-judge reliability” (p. 435) suggesting that larger panels of judges were appropriate in evaluating vocal performances in order to ensure reliability of such panel evaluations.

Zdzinski and Barnes (2002) used the facet-factorial approach to develop a valid and reliable rating scale for string performances by middle and high school students. These researchers were able to identify five critical factors in assessing string performances. These factors were “interpretation and musical effect, articulation/tone, intonation, rhythm/tempo, and vibrato” (p. 245), and 28 subscales were identified based on factor loadings. Like the researchers before them, Zdzinski and Barnes (2002) found

this method of developing a rating scale for musical performance to yield high reliability and validity.

Smith and Barnes (2007) used the facet-factorial approach to develop a rating scale for high school orchestra performance. Their analysis identified seven factors: “Ensemble, Left Hand, Position, Rhythm, Tempo, Presentation, and Bow” (p. 268). Several of these factors (left hand, position, and bow) were “unique to string performance, and require visual as well as aural evaluation” (p. 278) which sets this study apart from others in terms of focusing solely on the aural factors of the musical performance. Consistent with other studies, Smith and Barnes (2007) were able to demonstrate that the scale developed using this method was both reliable and valid.

Greene (2012) attempted to develop and validate an instrument to assess high school marching band performance using the facet factorial approach. “Forty-one items were chosen to define subscales for” two separate rating scales (p. v), one of which focused on the musical aspects of the performances, and one of which focused on the visual aspects of the performances. Sixty judges rated nine different high school marching band performances. The underlying factors the study identified for the musical aspects were “1) Communication and Effectiveness, 2) Sound Quality, 3) Program Construction, and 4) Rhythm” (p. v). The factors identified for the visual aspects for the performances were “1) Construction and Performance, 2) Visual Execution, and 3) Quality” (p. v). The results of this study showed high inter-judge reliability with the exception of the third factor of the visual rating, which was quality. Neither the musical nor the visual rating scales yielded an acceptable level of criterion-related validity (Greene, 2012).

Rubrics. Experts agree that rubrics were the most effective means of evaluating musical performances if the goal was to provide feedback to the performers for the purpose of improving future performances (Asmus, 1999; Ciorba & Smith, 2009; Wesolowski, 2012). Rubrics to assess various levels of music achievement in specific performance domains were developed successfully by several researchers (Azzara, 1993; Levinowitz, 1985; Saunders & Holohan, 1997; Norris & Borst, 2007).

These scales provided more information than previously researched scales by including written descriptors of specific levels of performance proficiency. The researchers used these more descriptive assessment tools, or rubrics, to evaluate performances in authentic contexts, often with strong reliability and validity. (Latimer et al., 2010, p. 169)

Wesolowski (2012) praised the use of the rubric in its usefulness in both reliably assessing musical performances and providing valuable feedback that could be used to improve future performances.

The rubric is a form of a criteria-specific performance scale. It is a set of scoring criteria used to determine the achievement level of a student's performance on assigned tasks. A rubric divides a task into constituent parts and offers detailed descriptions of the performance levels for each part. The descriptions are written so students are able to learn what must be done to improve their performances in the future. Because it helps teachers directly assess performance experiences, a rubric is a tool for providing authentic assessment. (p. 37)

“Norris and Borst (2007) compared the reliabilities of a Likert-type rating form with a rubric when adjudicating choral festival performances. The authors reported that the rubric, with its clear performance descriptors, provided a more appropriate format for assessment” (Ciorba & Smith, 2009, p. 7). Further,

According to Asmus (1999), rubrics provide specific advantages when used to assess music performances. First, adjudicators are provided with clear descriptors outlining the graduated levels of performance achievement. Second, performers are provided with (a) specific feedback concerning their performance and (b) useful information needed to improve future performances. (Ciorba & Smith, 2009, p. 7)

When using rubrics, judges

are asked to indicate which of several written criteria most closely describes the perceived level of performance ability. Adjudicators describe what they hear in a performance; they neither indicate whether they like or dislike the performance nor state whether they agree or disagree that the performance meets and indeterminate standard. (Saunders & Holohan, 1997, p. 259)

Table 3.

<i>Example of a Rubric</i>	
The student's intonation	Value
Is accurate throughout, and in all ranges and registers	10
Is accurate, but student fails to adjust on isolated pitches, yet demonstrates minimal intonation difficulties	8
Is mostly accurate, but includes out-of-tune notes. The student does not adjust problem pitches to an acceptable standard of intonation	6
Exhibits a basic sense of intonation, yet has significant problems, student makes no apparent attempt at adjustment of problem pitches	4
Is not accurate. Student's performance is continuously out of tune	2

Note. From Saunders & Holohan (1997, p. 264).

An example of a rubric with graduated and clear descriptors that provide specific feedback about the student's performance is illustrated in Table 3. Clearly, all of the types of assessment tools described in this section including holistic grading, Likert-type scales, criteria-specific rating scales, and rubrics have all been thoroughly researched resulting in evidence of consistent reliability. However, if the goal was to provide the greatest amount of summative feedback to the students to help them monitor their progress, the rubric was the tool that provided the maximum amount of feedback while maintaining reliability and validity. Therefore, the rubric is the tool that I chose to use for my research study.

Developing the rubric. Wesolowski (2012) offered a methodology for developing this type of rubric. First, one must “define the focus, purpose, and objectives of the assessment” (p. 38). The assessor must “include attention to the overall performance structure, the needs of the specific students being assessed, the expectations of what is to be accomplished, and the students’ prior knowledge and skill” (p. 38).

Second, the assessor must “define the performance criteria and learning outcomes” (Wesolowski, 2012, p. 38). “Each criteria . . . should be an important learning outcome for a high-quality performance and understood by the student . . . [and should] reflect your teaching goals” (p. 38).

Third, the assessor must “determine the type of rubric for your assessment” (Wesolowski, 2012, p. 39). “There are two main categories of rubrics: holistic and analytic. Holistic rubrics provide a single score based on an overall assessment of a music performance. The evaluator matches the descriptors of the scale to his or her overall impression of the performance” (Wesolowski, 2012, p. 38).

An analytic rubric contains more than one dimension of evaluative criteria. The multiple criteria are matched with multiple descriptors and the teacher’s feedback, and scoring is based on each of these individual dimensions. Because of the assessment by multiple criteria, the analytic rubric provides more information than does the holistic rubric. . . . A benefit of analytic rubrics is the wealth of specific, individualized assessment information that can be of great value.

(Wesolowski, 2012, p. 38)

Fourth, the assessor must “define the range and degrees of proficiency of performance scale levels” (Wesolowski, 2012, p. 39). For example, the author suggests one set of labels that could be considered amongst many other possibilities. They are “(1) beginning, (2) developing, (3) accomplished, and (4) exemplary” (Wesolowski, 2012, p. 39). Fifth, the assessor must “define appropriate task expectations and meaningful descriptors for each criterion performance level” (Wesolowski, 2012, p. 41). This is the step that sets the rubric apart from other types of criteria-specific rating scales.

Instead of merely assigning a number to each of the criteria, the developer must compose descriptive statements or descriptors for each performance level of each criterion.

The totality of the descriptors provides a comprehensive summary of what is being assessed. The descriptors should be written as clearly and concisely as possible. Avoid any vernacular or terminology that is superfluous in nature. Write descriptors for continuity between levels of performance in each category. The descriptors should define a continuum of the quality throughout each category. Be sure that each descriptor has a clear sense of flow between levels. The descriptors should be detailed enough to limit subjectivity yet concise enough to avoid confusion or ambiguity. (Wesolowski, 2012, p. 41)

Finally, the assessor must “choose an appropriate scoring scale with clearly defined cut points” (Wesolowski, 2012, p. 41). Marzano (2007) suggested that each score on a rubric (which he calls a scale) should describe “specific progress toward a specific learning goal” (p. 24). For example,

a score of 4.0 indicates that the student has gone beyond the information and skill taught by the teacher. A score of 3.0 indicates that the student has learned the target knowledge as articulated by the teacher. A score of 2.0 indicates that the student understands or can perform the simpler information and skills relative to the learning goal but not the more complex information or processes. A score of 1.0 indicates that on his or her own the student does not demonstrate understanding of or skill regarding the learning goal, but with help the student does. Finally a score

of 0.0 indicates that even with help the student does not demonstrate understanding or skill relative to the learning goal. (Marzano, 2007, pp. 24-25)

“Gordon (2002) maintained that the more descriptors included for each dimension, the more reliable the rubric will become, as long as that number does not exceed five” (Latimer et al., 2010, p. 170).

Summary

The examination of the literature in relation to applied music study and vocal pedagogy provided an opportunity to examine the process of vocal instruction and to help determine where in the process this type of formative assessment might fit. The review of leading books on singing provided a rich and extensive list of categories and characteristics of good singing. These categories and the associated descriptors of each were used, in conjunction with feedback from experts, to develop the rubric.

The review of the research about types of tools used in assessing musical performances enabled me to create a hierarchy of methods organized by increasing ability to provide feedback to the students to maximize the feedback loop, improve the instructional process, and provide optimum conditions for students to meet their desired learning outcomes. It was clear the type of tool that would best accomplish these goals was the rubric. Therefore, the review of the literature in developing effective rubrics was instrumental in determining the design of the procedures to follow to create the rubric used for the research study.

Chapter Three: Methodology

Chapter One introduced the purpose of the study and an overview of the research methodology selected. This mixed-methods, participatory action research study was conducted using carefully selected procedures, instrumentation, and data analysis tools. Chapter Three will discuss each step of the methodology and rationale for each of the selected components.

Location

The research institution was a private, four-year liberal arts institution located in a suburban, Midwestern city. The university offered 84 undergraduate and 37 graduate degree programs, with teacher education and business administration representing the largest majors. The institution served approximately 17,000 students, of which approximately 6,000 were traditional full-time day students (Research Site Undergraduate Course Catalog, 2013-2014).

The Music Department, which was housed in the School of Fine and Performing Arts, served approximately 150 music performance, music education, and music business majors. Music performance majors were required to take 16 credit hours of applied music lessons in voice (Undergraduate Course Catalog, 2013-2014). The usual schedule was two credit hours per semester for eight semesters. In addition, the students were required to perform a Junior Recital consisting of 30 minutes of music and a Senior Recital consisting of 60 minutes of music. Music education majors were required to take eight credit hours of applied music lessons in voice (Undergraduate Course Catalog, 2013-2014). The usual schedule was one credit hour per semester for eight semesters. In addition, they were required to perform a Senior Recital consisting of 30 minutes of

music. Music business majors were required to take four credit hours of applied music lessons in voice (Undergraduate Course Catalog, 2013-2014). The usual schedule was one credit hour per semester for four semesters. They were not required to perform a Junior or Senior Recital. All students majoring in music who were taking private lessons were required to present a juried performance at the end of each semester. These performances were assessed by all voice department faculty members using a standard Likert-scale scoring guide. The applied music requirements for each major area of study are summarized in Table 4.

Table 4.

Applied Music Requirements by Major

Major	Credits per semester	No. of semesters	Junior recital	Senior recital	Juried performance
Music performance	2	8	30 minutes	60 minutes	Required
Music education	1	8	Not required	30 minutes	Required
Music business	1	4	Not required	Not required	Required

The Department of Music had no specific admissions criteria beyond the university's admissions policy. Therefore, while there was an audition process for scholarship awards and another audition process for ensemble placement, there was not an exclusionary audition process, and students were not turned away from the program. Instead, they were offered remedial courses when needed, such as Class Voice and Fundamentals of Music, before they were allowed to proceed to private voice lessons and Music Theory I. The net effect of this policy and practice was that most of the entering singers were true beginners.

Methodology

This mixed-methods, participatory action research study was designed to focus on “a specific local problem and . . . [result] in an action plan to address the problem” (Fraenkel et al., 2012, pp. G-1). It met the underlying assumptions underlying action research that “the participants have the authority to make decisions, want to improve their practice, are committed to continual professional development, and will engage in systematic inquiry” (p. 611). The local problem on which the study focused involved the process and mechanism for assessing undergraduate vocal performances at the semester juries. I wished to determine a method to more accurately assess the students as well as to provide them with meaningful and useful feedback that they could convert into action.

Procedures

I followed a series of steps over three phases to complete this research project. The first phase was the preparation of a research-based rubric, which was achieved via a review of the literature and interviews of experts in the field. Recordings were prepared, which included samples contributed by both student and professional singers. The second phase was the implementation of the research-based rubric in use by judges to assess the recorded performances. The third phase was the collection of student feedback, facilitated through interviews of the student performers to determine what information they learned from the completed research-based rubrics the judges used to assess their recorded performances.

Phase I: Rubric design. I compiled lists of criteria from a review of the available literature on vocal pedagogy and the characteristics of successful solo voice performances. Five commonly used books on vocal pedagogy and technique were

examined. They were *Singing: The Mechanism and the Technic* by Vennard (1949), *The Diagnosis and Correction of Vocal Faults* by McKinney, (1994), *The Structure of Singing: System and art in vocal technique* by Miller (1996), *Foundations in Singing: A Guidebook to Vocal Technique and Song Interpretation* by Paton (2006), and *Adventures in Singing: A process for exploring, discovering and developing vocal potential* by Ware (2008).

I interviewed five experts who were university-level voice teachers, to develop a rubric based on Wesolowski's (2012) method. The experts were asked to define performance criteria and desired learning outcomes for solo vocal performance, to specify the range and degrees of proficiency for each performance criterion, to develop meaningful descriptors for each criterion performance level, and to establish appropriate scoring scale and clearly defined cut points for each criterion performance level (Appendix A: Expert Interview Questions). All of these questions were derived from the steps in the article by Wesolowski (2012) and were intentionally aligned with the first research question.

The information collected from the experts was synthesized using the criteria from the review of the literature to form the research-based rubric. The research-based rubric (Appendix B: Research-based Rubric) included space for judges to provide comments about each criteria component, as well as a space for judges to give an independent holistic score from 1 to 100. The judges were asked to give comments about the research-based rubric and to provide information about their levels of education, years of teaching experience, job titles, and a description of the equipment on which they listened to the performances they were assessing.

Jones (1986) found that the “visual dimension of solo performance evaluation” (p. 87) deeply affected ratings of vocal performances. He found that judges tended to disregard the actual “aural clues” (p. 87) in favor of appearance, maturity, and “communicative charisma” (p. 87). Bergee (2003) described several extraneous variables that tended to influence performance assessment including "gender and race, and attractiveness, stage presence, and dress" (p. 343). These findings led me to conclude that an audio recording was the desired means for presenting performances to be adjudicated for this study. An audio recording would eliminate the distraction of visual elements and allow the judges to focus solely on the aural aspects of the performances.

The recording procedure and song selection was modeled after Wapnick and Eckholm’s (1997) study. I, with the assistance of an audio engineer, recorded 14 undergraduate singers of all voice types and abilities, performing an excerpt of the same piece. Like the Wapnick and Eckholm study, students performed an excerpt (mm. 1-27) from Mozart’s art song “Ridente la Calma.” The rationale for choosing this piece was

(a) it was available in both medium-high and medium-low versions, which made it appropriate for most voices; (b) the text was in Italian and fairly easy to pronounce, thus minimizing the possibility that novices would be detected from the pronunciation difficulties alone; (c) the slow, lyrical nature of the song made it suitable for any voice type; (d) it was technically complex enough to reveal strengths and weaknesses in the singer’s vocal production; (e) the range was broad enough to allow evaluation of most of the singer’s range; and (f) it was not too musically complex to be learned in a short period of time. (Wapnick & Eckholm, 1997, p. 431)

Students were able to choose between two transpositions of the piece, one higher and one lower (Appendix C: Musical Selections). They were given the sheet music as well as an International Phonetic Alphabet (IPA) transcription of the text (Appendix D: Phonetic Transcription of Musical Text). Students were given two weeks to prepare the musical selection. Upon request, I provided assistance in learning the pitches and rhythms as well as the pronunciation. I provided no coaching on vocal technique, and did not direct the students to seek coaching from their private lesson instructors.

Two professional singers were recruited to record the same excerpt for the audio recording. These professionals were recorded under the exact conditions as the student recordings, in order to maintain consistency among all of the recordings the judges would hear. Using professionals would serve to further validate the instrument. One would expect the excerpts performed by the professional singers to be rated the highest out of all of the examples.

Four student performances were repeated on the master recording. This served to further validate the results. One would expect these examples to be rated the same as their duplicates. In total, there were 20 recordings: 14 student recordings randomly mixed with two professional recordings and four repeated performances, which were placed at the end of the audio recording. The makeup of the recordings is further illustrated in Table 5.

Table 5.

<i>Explanation of Recording Makeup</i>	
Type of recording	n
Student	14
Professional	2
Repeat of students	4

The recordings were made using a Rode NTK large-diaphragm vacuum tube condenser microphone. The signals were run into a PreSonus DigiMAX D8 microphone preamplifier. The signals were recorded by the Allen & Heath ICE-16 onto a USB flash-drive at a 48-khz sampling rate and a 16-bit depth. The multi-track recordings were transferred to Pro Tools 10 and edited minimally. No processing of any kind was performed other than normalization to -0.5 dbFS (A. Donohue, personal communication, August 8, 2013).

Phase II: Rubric implementation. I distributed copies of the recorded performances and the newly designed rating scale via Survey Monkey to 254 experienced university voice teachers from 50 of the United States. In an accompanying communication, I introduced myself, explained the purpose of the study, walked the judges through the process of informed consent, and provided instructions for the judges to score each performance using the new rating scale. Judges were asked to listen to each performance only once and take a 10-minute break after the 10th recording, to help mitigate the effects of fatigue on the outcomes. They were to assign an independent overall score for each performance, provide feedback about the research-based rubric, and provide demographic information. I received fully completed research-based rubrics for all 20 performances from 36 of the judges who were solicited.

Phase III: Student perceptions. After the judges completed the research-based rubrics and returned them to me, I shared the results with the 12 of the 14 student performers. Two students moved out of the area prior to the completion of the data collection. I interviewed the students about their reactions and perceptions related to the information contained in the completed research-based rubrics. I asked the students to describe what they liked or did not like about the method of assessment, what they thought the judges heard or did not hear in their performances, their understanding of the strengths and weaknesses of their performances, and what, if anything, they planned to do with the information, or what actions, if any, they planned to take (Appendix G: Student Interview Questions).

I was an instructor in the department and had interactions with the students on many levels: as advisor, course instructor, private lesson teacher, and student teaching supervisor. I conducted the interviews with the experts and with the students myself. Coercion was reduced by ensuring that no data collected during the interviews was attributed to any particular student. Scores from this rubric exercise were not used in an evaluative manner for the predetermined graded course activities or for placement in performance groups.

Instrumentation

I developed the protocol for the expert interviews by including each of the steps outlined by Wesolowski (2012) in his article, *Understanding and Developing Rubrics for Music Performance Assessment*. I attempted to determine the experts' opinions about the levels of achievement in the development of an undergraduate singer, the key characteristics that the experts assess when evaluating singers, and the levels and values

of expected outcomes for each key characteristic at different stages of the singers' development (Appendix A: Expert Interview Questions). The data collected from these expert interviews, in combination with the data collected through the review of the literature, I developed the research-based rubric the judges used to actually assess the recorded performances.

Data Analysis

Inter-judge reliability was tested using the one-way analysis of variance test (ANOVA). This test is usually used to “determine if there is a significant difference among three or more means” (Bluman, 2010, p. 602). Two separate tests were analyzed. One evaluated the difference in mean scores of judges' overall scores, and the other evaluated the judges' scores for each category. The null hypotheses for the difference in means of the overall scores were as follows:

Null Hypothesis #1. When scoring performances using the research-based rubric, there will be no difference in judges' scores.

Null Hypothesis #2. When scoring performances using the research-based rubric one category at a time, there will be no difference in judges' scores.

In addition to application of ANOVA, the judges' scores were tested using the z -test for difference in means to compare the judges' average scores for each category with the judges' average overall scores. The z -test for difference in means is conducted by “selecting pairs of samples and comparing the means of the pairs” (Bluman, 2010, p. 469). The hypotheses for these tests were as follows:

Null Sub Hypothesis #1. There will be no difference in average mean score on the research-based rubric, when comparing individual judge scoring to the overall group mean score.

Null Sub Hypothesis #2. There will be no difference in mean score on the research based-rubric, when comparing individual judge scoring on individual categories to the overall group mean score.

Intra-judge reliability was tested for each judge through analysis of each of the four repeated performances. These scores were also tested with the z -test for difference in means. When z -testing was inconclusive, the chi square test for independence was applied to the comparison of original-to-repeated scores, as well as to the scores for the professional recordings compared to expected scores for the professional recordings. The chi square test for independence

is based on a comparison between expected frequencies and actual, obtained frequencies. If the obtained frequencies are similar to the expected frequencies, then researchers conclude that the groups do not differ. If there are considerable differences between the expected and obtained frequencies, on the other hand, then researchers conclude that there is a significant difference . . . between the groups. (Fraenkel et al., 2012, p. 238)

Null Hypothesis #3 was testing with a z -test for difference in means.

Null Hypothesis #3. There will be no difference in judges' scoring utilizing the research-based rubric, on repeat performance when compared to the same judges' scoring for the original performance.

The null hypotheses for the chi square test for independence was as follows:

Null Sub Hypothesis #3. The event score when applying the research-based rubric is independent upon which judge conducted the rating.

The strength of the potential linear relationship of each category with each of the other categories in the research-based rubric was measured by calculating the Pearson Product Moment Correlation Coefficient (PPMCC). This calculation "expresses the degree of relationship between two categories" (Fraenkel et al., 2012, p. 207). This information was helpful in determining if suggestions to combine categories made by participants in the study were valid.

Null Hypothesis #4. There will be no relationship between each of the ratings of characteristics of breath, tone, accuracy, diction, intonation, vibrato, registration, agility, style, and expression, and judges scores utilizing the research-based rubric.

The judges provided an overall holistic score for each performance that was assigned independent of the rubric score. This holistic score was compared to the rubric score using the chi square test for goodness of fit. The chi square test for goodness of fit is a nonparametric categorical inferential technique (Fraenkel et al., 2012, p. 239) that is used to determine "whether a frequency distribution fits a specific pattern" (Bluman, 2010, p. 573).

Null Hypothesis #5. There will be no difference between holistic scores and calculated rubric-based percentage score.

I collected qualitative data to contribute to validation of the rubric. Feedback from the performers concerning their perceptions of the usefulness of the feedback from the research-based rubric for improving their future performances was coded, analyzed and reported using qualitative methods of analysis. Feedback about the research-based

rubric itself collected from the judges was coded, analyzed and reported using qualitative methods of analysis.

Participants

Experts for the preparation of the research-based rubric were recruited from the research institution's voice faculty who were personally invited to participate. They were selected because they were most familiar with the level of ability of the students at the research location. In keeping with the collaborative nature of participatory action research, they were also selected because they would be key stakeholders in the implementation of any changes.

Expert One was an experienced voice teacher and choir director with a Master of Music in Choral Conducting. In addition to conducting college level choirs and performing as a soloist, he had taught for ten years such college courses as private lessons, class voice, vocal pedagogy and literature, choral arranging. Expert Two was another experienced performer and voice teacher with a Master of Music in Vocal Performance. She had 10 years of collegiate experience teaching private lessons, class voice, and vocal pedagogy and literature. Expert Three was also a voice teacher and soloist with a Doctor of Musical Arts in Vocal Performance. He had taught private lessons for five years as well as singers' diction, vocal pedagogy and literature, and world music. Expert Four was a voice teacher and performer with a Master of Music in Vocal Performance. She had taught private lessons for 10 years, class voice, and singers' diction. Expert Five was a choir director, pianist/accompanist, and voice teacher with a Master of Music Education. She had taught at the college level for five years and had an administrative role in the music department. I interviewed each of the experts

individually at a mutually convenient time and place. Each interview lasted approximately half an hour. I used the Expert Interview Questions (Appendix A) for each interview, recorded each interview on my iPhone, and then transcribed each interview into a Microsoft Word document.

Table 6.

Makeup of Student Participants

Number	Gender	Voice type	Major	Grade
1	M	Tenor	Music performance	Senior
2	F	Mezzo-Soprano	Music performance	Sophomore
3	F	Mezzo-Soprano	Music education	Junior
4	F	Mezzo-Soprano	Music business	Sophomore
5	M	Bass	Music education	Senior
6	M	Tenor	Music education and music perf.	Senior
7	M	Baritone	Music education and music perf.	Junior
8	F	Soprano	Music performance	Junior
9	M	Baritone	Music	Senior
10	M	Tenor	Music performance	Sophomore
11	F	Mezzo-Soprano	Music education	Freshman
12	F	Soprano	Music education	Junior
13	F	Soprano	Music education	Senior
14	F	Mezzo-Soprano	Music education	Sophomore

I recruited student volunteers for the recordings from the research institution's music department students who were enrolled in private voice lessons. They were invited to participate via an announcement at a weekly departmental meeting. They were also key stakeholders in the outcome of the project as any changes to the scoring of juried performances would have a direct effect on their grades and development. As illustrated in Table 6, the group of singers was diverse with regard to gender, voice type, major, and year of study.

There were six male students: three tenors, two baritones, and one bass. There were eight female students: three sopranos and five mezzo-sopranos/altos. One of the sopranos chose to sing the selection in the lower key. Four of the participants were Music Performance Majors. Five of the participants were Music Education Majors. Two of the participants were double majors in both Music Performance and Music Education. The remaining participants included one Music Business Major, one Instrumental Music Education Major, and one student working toward the Bachelor of Arts in Music. Accounting for the age of the participants, there were one freshman, four sophomores, four juniors, and five seniors.

I personally invited professional singers to perform the selection. There was one female singer with a soprano voice classification who performed the selection in the higher key and one male singer with a baritone voice classification who performed the selection in the lower key. One of the professional singers had attained a Doctor of Musical Arts (DMA) in voice and the other had attained a Master of Music (MM) in vocal performance. Both singers were active professionals with many performance credits.

I recruited judges to use the research-based rubric to assess the recorded performances. These judges were not recruited from the research institution's faculty to minimize the possibility that voice teachers would be able to identify their own students among the student performances and to avoid bias in judging based on that recognition. University level teachers of singing were invited via email solicitations via Survey Monkey. Initially, I solicited teachers of singing from two universities chosen from each state in the United States. Responses were minimal, so I began expanding the search, starting with all four-year universities in Missouri and expanding to adjacent states until sufficient participation was realized. Requests were sent to professors at universities in the following states: Montana, Wyoming, Colorado, New Mexico, North Dakota, South Dakota, Nebraska, Kansas, Oklahoma, Texas, Michigan, Iowa, Missouri, Arkansas, Louisiana, Wisconsin, Illinois, Mississippi, Kentucky, Tennessee, Alabama, Georgia, Ohio, West Virginia, Virginia, and North Carolina. Some of the participants notified me that the research-based Survey Monkey requests were automatically sent to the spam folder in their email system. Therefore, I sent out an email from my personal email account to all of the potential participants, and the response was much greater.

There was a pool of 36 judges who completed the scoring. From this pool, 25 of the judges had earned their terminal degree of a Doctor of Musical Arts (DMA) or a Doctor of Philosophy (PhD), nine had earned a Master's degree, and one had earned a Bachelor of Arts (BA) with 30 hours toward a Masters, and one entered "Licentiate" as his qualifications (See Table 7). These judges listened to the recordings and completed the assessments at their convenience.

Table 7.

<i>Judges' Level of Education</i>	
Degree	No. of participants
DMA/PhD	25
MM/MA	9
BA+30	1
Licentiate	1

The judges were also asked about their level of experience. Nine judges had between zero and 10 years of experience, and nine judges had between 11-20 years of experience. Twelve judges had between 21-30 years of experience, and six judges had more than 30 years of experience. Table 8 illustrates the ranges of years of experience for the panel of judges.

Table 8.

<i>Judges' Years of Experience</i>	
Years	n
0-10	9
11-20	9
21-30	12
31-40	4
41-50	1
51+	1

Table 9 is a summary of the judges' job titles. One judge was employed as an instructor, and five were employed as adjunct professors. There were four professors with the rank of assistant professor and five with the rank of associate professor. There were nine 20 full professors. Of these 20, 11 also had administrative roles in their institution including Division Chair, Director of Opera, Dean, Voice Area Director, Choral Director, Coordinator of Vocal Studies, and Department Chair.

Table 9.

<i>Judges' Job Titles</i>	
Job Title	No. of participants
Instructor	1
Adjunct Professor	5
Assistant Professor	4
Associate Professor	5
Professor	9
Professor w/Administrative Role	11
No Response	1

Summary

This participatory action research study was executed in three phases and implemented both qualitative and quantitative methodologies. The first phase was the development of the research-based rubric which was completed by consulting the available literature and interviewing experts. The second phase was the creation of the recordings and the completion of the performance assessments using the newly designed

research-based rubric. Phase III was the collection of feedback from the performers about their perceptions of what they learned from the newly completed research-based rubric.

Chapter Four: Results

As stated in Chapter Three, the methodology utilized to conduct this action research study was mixed-method participatory action research, which was conducted following a series of steps over three phases to complete this research project. The first phase was the preparation of a research-based rubric. The second phase was the implementation of the research-based rubric where the research-based rubric was used by judges to assess the recorded performances. The third phase was the collection of student feedback. Both qualitative and quantitative data were collected during the course of the study. The qualitative data collected during this study were obtained via interviews with experts, interviews with students, and feedback provided by the judges via the survey. The expert interview data were used in conjunction with the literature review to answer the first research question and to develop the research-based rubric. The student interview data were collected and used to answer the second research question. In addition to quantitative measures, feedback from the judges was used to validate the rubric.

Phase I: Rubric Design

To answer the first research question, What are the appropriate performance criteria, learning outcomes, and meaningful descriptors for various levels of proficiency for undergraduate solo vocal performance?, I performed an extensive review of the literature including four textbooks on vocal pedagogy and singing. In addition, I interviewed five experts who had extensive training and experience in the teaching of singing. Each of the interview questions was expressly designed to be aligned with the

first research question. The results of the interviews which were used to construct the rubric were as follows:

Interview question #1: How would you categorize levels of achievement in the development of a singer?

The purpose of this question was to determine the number of levels and the titles of those levels of achievement appropriate for the span of development for an undergraduate student of singing. Expert One felt that using the levels of "Beginning, Medium, Advanced but with room for gray area in between" would be the most effective way to score undergraduate singers. Expert Five also preferred this method because

our freshmen come in with such varying degrees of experience. Some come in with four to five years of voice lessons. Some come in with no voice lessons.

And even with that picture they are still coming in with different levels of reading ability and technical ability so I don't think that classifying it Freshmen,

Sophomores, Juniors, Seniors is as healthy for the singers as having that beginning, intermediate, and advanced with something in between.

Expert Two indicated that freshmen were clearly beginners, but, like Experts One and Five, thought there was some gray area between the upper levels of study. She felt that sophomores and some juniors "know certain things and are still learning others" and that for juniors and seniors it should become "second nature to translate the text."

Expert Three was concerned with the unique development of each individual singer and wanted to ensure that the "stages of development are really recognized." For beginners "the correct notes, rhythms, intonation, dynamics, tempo diction, and style need to learn things correctly and they need the tools, but we don't necessarily expect

them to be able to do the perfect dynamics or the perfect tempo all of the time, but we are trying to help them gain those tools." But then that middle level, "I can't even say that I would apply to all Freshmen, Sophomores, Juniors, Seniors. I wouldn't even apply that to all seniors equally partially because their age and how fast they are going to mature" (Expert Three).

Expert Four wanted to focus on length of study. She concluded that using the number of semesters of study would be too unwieldy because "eight descriptors for each characteristic would be too much." So she preferred to use "year of study, and not separate by semesters." Expert Four went on to say that "I definitely think the expectation should be higher for upperclassmen and we should be more lenient towards the beginners who may have not ever thought about listening and tuning and don't know how to sight read." All of this feedback was critical to answering a portion of the first research question and determining the levels of proficiency that would be used in the rubric.

Interview question #2: When judging performances by vocal students at the undergraduate level, what are the key characteristics for which you are listening?

This question was designed to elicit responses that would help to determine which performance criteria would be included in the rubric. Expert One described the key characteristics as consistent tone, breath control, understanding of the piece itself and interpretation, overall control and consistency, control of all ranges and control of their chest voice that is mixing into their mixed voice that is mixing into a head voice, musicality and musicianship, dynamics, accents, stresses, word stresses, fine details, intonation, diction, and stylistic accuracy. Expert Two valued tone and maturity level of

tone, breath and the way breath is used, the skill that is required to sing the piece, style such as the interaction between voice and piano, knowledge of what they are singing about, vocal expression, and expressing the character of the piece. Expert Three described correct notes and rhythms, intonation, dynamics, tempo, diction, style, consistent vibrato and spin, consistent resonance and focus across all vowels, sing and singers' formant, chiaroscuro, maturing tone quality, increasing ease of production, legato, and style as the key characteristics of good singing. Expert Four listened for breath and how breath is used, the balance of tone and breath, intonation, musicianship including sight reading skills and ear training, phrasing, interpretation and expression, diction, beauty of tone and timbre, placement, registration, style, and vibrato. Expert Five thought tone production and placement, breath support, phrasing, diction, intonation, vowel shape, vibrato, and expression were important factors in assessing singing. The feedback collected from the responses to this interview question was used along with the information collected in the review of the literature to continue to answer the first research question and to determine the performance criteria that would be used in the rubric.

Interview question #3: For each characteristic, how would you describe what you would expect to hear from an expert singer? From an advanced singer? From an intermediate singer? From a beginner?

The purpose of this question was to elicit rich and comprehensive descriptors for each of the performance criteria deemed important enough to include in the rubric. Expert One listened for the following with regard to breath: control, complete opening as you breathe through, consistent airflow, uninterrupted airflow, appoggio or complete

expansion throughout, not running out of breath, not squeezing the sound, not forced.

Expert Two mentioned the use of the breath, and that it should be coordinated, not shallow, not loud, quiet and deep, the singer should use the breath within the context of the body, and use a steady stream of air. Expert Three talked about the singer's use of breath, balance of tone and breath, silent inhalation, open inhalation, and that the singer's phrases are completed.

Descriptors for tone included consistent, not brassy, not harsh, not choppy, connected, legato, semi-covered, not too swallowed, not nasal, and with balanced bright and dark (Expert One). Expert Two used the descriptors mature, balanced, consistent, not breathy, consistent throughout range, not pressed or artificially heavy and not a lot of muscling. Expert Three described good tone as having consistent resonance and focus across all vowels, ping, chiaroscuro, maturity, legato and increasing ease of production. Descriptors for tone used by Expert Four included beauty of tone, timbre, placement, not swallowed, not nasal, and placement. Expert Five used placement, clear, not fuzzy, not airy, not strident, and not throaty to describe tone.

Expert Three used the words "correct notes and rhythms" when describing accuracy. Diction was described by each judge as "word stresses" (Expert One), "clear" (Expert Two), "energy in consonants" (Expert Three), "correct formation of mixed vowels and nasal vowels" (Expert Four), and "correct vowel shape" (Expert Five). Intonation was described as "right on top of the pitch" and "tuning perfectly with every chord" by Expert Four. Expert Three used the words "consistent spin" to describe vibrato, and Expert One used the words "consistency and control" when discussing registration.

When discussing style, Expert One spoke about "dynamics, accents, stresses and accuracy." Expert Two mentioned "skill, interaction between the voice and piano, and historical accuracy." Expert Four spoke about "appropriate" style choices and also spoke about how style is represented in "phrasing" choices. Expression was described with the terms "communicating understanding" (Expert One), "knowledge of what they are singing about" (Expert Two), and "interpretation and presentation" (Expert Four). The feedback provided by the experts in response to this interview question was used along with the information collected in the review of the literature to continue to answer the first research question and to determine the descriptors for each of the performance criteria that would be used in the rubric.

Interview question #4: If values were to be assigned to each level for the purpose of grading, what would be your recommendation?

This intent of this question was to determine a way to attach a numerical value to each descriptor for the purpose of assigning a grade to each student. Expert One expressed agreement that it would be appropriate to assign numbers to each of the increasing levels (beginning, middle, advanced with transitions in between) for the purposes of grading, and that the expected scores for students at various points of study would be different. For example, a student who was completing the fourth semester of study, and therefore halfway through their training, would be expected to earn scores in the middle of the scale. Expert Three was also "comfortable" with the idea that expected scores would be different students at different points in their training. He felt that would allow the scorer to either score by the number or by the descriptor and avoid potential disagreements about the specific descriptors for each level. He said, "I would be OK

with if I don't love your description of [Level Four] for vibrato. I would feel like I could have an argument with you about this, but it is actually not necessary because they shouldn't be at that level anyway. They will be at twos and threes which is all right." Expert Five also expressed agreement with the need to have a sliding scale of expected scores based on a student's length of study. She said, the student "needs to look at the jury sheets and say, 'Well, Freshman year I was a beginner, and two of them scored me as beginner. Then, my Junior year everybody scored me at intermediate, but during my Junior year I am only at intermediate? That should bother me.' That really needs to be the most important thing because if you are a beginner and for a beginner this is your score, that is much better feedback than everyone grading you as a beginner without having that expressed to you." This group of feedback to this interview question was used to continue to answer the first research question and to determine the appropriate scoring scheme that would be used in the rubric.

Comments. Although a question about including comments in the rubric was not included in the interview protocol, two of the experts mentioned the importance of comments in providing thorough and accurate feedback to student singers. Expert Five stated "I think that the comments are more important than the scores." Expert Three described a scoring guide that he had used previously that he was fond of because it used a plus, nothing, minus scale to do the scoring allowing the judges more space and time to write comments. However, Expert One was disenchanted with the comments that his students had received on previous jury scoring sheets. He felt they were very unspecific and addressed very obvious weaknesses such as "work on your middle range," when the student and teacher were fully aware of that weakness and having it pointed out in the

jury scoring was neither new information nor a helpful strategy to improve. Although the experts were not expressly asked about the importance of including room for comments in the rubric, three of the experts expressed their opinions about the inclusion of comments.

Phase II: Rubric Implementation

Once the research-based rubric was developed it was distributed along with the recordings of the student performances to university level teachers of singing who used the research-based rubric to score the performances. There were a total of 36 completed rubrics. The scores from the completed research-based rubrics were calculated and tested to determine inter-judge reliability, intra-judge reliability, the correlation between categories, if the professional singers scored higher than the students, and the relationship between the rubric scoring and holistic scoring.

In addition to the quantitative data collected as part of this study, there was much qualitative data to examine. Each of the judges was invited to provide feedback about the rubric at the end of the scoring session. The quantitative data from the scoring and the qualitative data from the judges' comments are included in this section.

Null Hypothesis #1. The null hypothesis was: When scoring performances using the research-based rubric, there will be no difference in judges' scores. To test this hypothesis, I performed a one-way analysis of variance (ANOVA), with results illustrated in Table 10. The null hypothesis was rejected. This test revealed that some judges demonstrated more agreement with the judge group than did the others ($F = 3.074$; $F\text{-critical} = 1.440$; $df = 35, 677$; $p < 0.05$).

Table 10.

ANOVA Summary for Judges' Scoring

Source of variation	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F_{crit}</i>
Between groups	72.766	35	2.079	3.074	0.00427	1.440
Within groups	457.875	677	0.676			
Total	530.642	712				

Null Sub Hypothesis #1. The results of the subsequent z -test for difference in means (Critical Value = 1.96; $\alpha = 0.05$; $df = 34$) are illustrated in Table 11. This test compared each judge's average scores for each performance with the judges' average overall score. The null hypothesis for each case was: There will be no difference in mean score on the research-based rubric, when comparing individual judge scoring to the overall group mean score. The null hypotheses were not rejected for each case. Therefore, the testing revealed there were no differences in the means of the two groups compared (each test value was less than the critical value). Therefore, there was variance in judges' scorings as noted in the ANOVA above, yet consistency in scoring with the slightly different comparison of individual scores compared to overall average scores.

Table 11.

Results of z-test for Difference in Means of Judges' Scores

<i>Judge</i>	<i>P value</i>
1	0.883
2	0.651
3	0.459
4	0.140
5	0.435
6	0.353
7	0.625
8	0.592
9	0.812
10	0.360
11	0.399
12	0.482
13	0.629
14	0.257
15	0.767
16	0.753
17	0.176
18	0.988
19	0.249
20	0.043
21	0.994
22	0.116
23	0.001
24	0.387
25	0.019
26	0.571
27	0.384
28	0.012
29	0.667
30	0.654
31	0.166
32	0.009
33	0.714
34	0.301
35	0.086
36	0.011

Note. Critical Value = 1.96

Null Hypothesis # 2. Null Hypothesis 2 was also designed to test for inter-judge reliability and was stated as: When scoring performances using the research-based rubric one category at a time, at least one judge will score differently than the others. The null hypothesis was rejected. A single factor analysis of variance (ANOVA) using the average for each category by performer (Table 12) revealed that there was significant difference within the scoring for each category ($F = 2.942$; $F\text{-critical} = 1.929$; $df = 9, 190$; $p < 0.05$). Therefore, some judges scored differently than the others on some categories.

Table 12.

ANOVA Summary for Individual Criteria

<i>Source of variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F_{crit}</i>
Between groups	14.957	9	1.661	2.942	0.002	1.929
Within groups	107.320	190	0.564			
Total	122.280	199				

Null Sub Hypothesis # 2. The results of a subsequent z -test for difference in means (Critical Value = 1.96; $\alpha = 0.05$; $df = 8$) are illustrated in Table 13. This test compared the judges' average scores for each category with the judges' average overall scores. The null hypothesis for each case was: There will be no difference in mean score on the research-based rubric, when comparing individual judge scoring on individual categories to the overall group mean score. The null hypotheses were not rejected for each case. Therefore, the testing revealed there were no differences in the means of the two groups compared (each test value was less than the critical value). Therefore, there was variance in judges' scorings as noted in the ANOVA above, yet consistency in scoring with the slightly different comparison of individual scores compared to overall average scores.

Table 13.

Results of z-test for Difference in Means of Individual Criteria

Group	z-test value
Breath	0.536
Tone	0.045
Accuracy	0.047
Diction	0.835
Intonation	0.081
Vibrato	0.632
Registration	0.610
Agility	0.858
Style	0.315
Expression	0.026

Note. Critical Value = 1.96

Null Hypothesis #3. Null Hypothesis 3 stated, There will be no difference in judges’ scoring, utilizing the research-based rubric, of repeat performance when compared to the same judges’ scoring for the original performance. The intra-judge reliability, or judge consistency, was measured by performing a z-test for difference in means for the four performances presented two times throughout the judges’ listening and scoring. For Events 1 and 2, the null hypothesis was rejected. For Events 3 and 4, the null hypothesis was not rejected. The results, which are shown in Table 14, were inconclusive when considering consistency of scoring when using the research-based rubric.

Null Sub Hypothesis #3. Because the results from the z-test for difference of means were inconclusive, another test, the Chi Square test for Independence, was added to differentiate. The null hypothesis was, The Event score when applying the research-based rubric is independent of which judge conducted the rating. The null hypothesis

was not rejected. This second test examined the differences between scores for each initial event and its corresponding repeated event ($\chi^2 = 0.010449$; critical value = 7.815; $\alpha = 0.05$). It revealed that on the repeated event, whether the score was higher or lower was not dependent upon which judge did the rating. Therefore, the ratings were independent of the judge.

Table 14.

Intra-Judge Reliability for Ratings of Repeated Items

Event/Repeat	Test value	Conclusion
1	-3.102	There is a significant difference
2	-3.394	There is a significant difference
3	-1.564	There is no significant difference
4	-1.307	There is no significant difference

Note. Critical Value = 1.96

Null Hypothesis #4. Null Hypothesis 4 stated, There will be no relationships between ratings of characteristics of breath, tone, accuracy, diction, intonation, vibrato, registration, agility, style, and expression when comparing judges scores utilizing the research-based rubric. The PPMCC matrix shown in Table 15 revealed that all of the 10 categories were significantly correlated with each other ($\alpha = 0.05$; $df = 8$; $\rho = 0.632$). “Registration/Tone” was the strongest correlated pair ($r = 0.993$) followed by “Registration/Agility” ($r = 0.991$). “Breath/Tone” were also strongly correlated ($r = 0.989$). The pairs with the mildest correlation were “Expression/Vibrato” ($r = 0.0682$) and “Expression/Intonation” ($r = 0.0658$). “Agility” ($r = 0.993$) had the strongest correlation to “Overall Score” followed by “Registration” ($r = 0.989$). “Expression” ($r = 0.779$) had the mildest correlation to the “Overall Score.”

Table 15.

Pearson Product-Moment Correlation Matrix for Rubric Categories

	Breath	Tone	Accuracy	Diction	Intonation	Vibrato	Registration	Agility	Style	Expression	Overall score
Breath	1.000										
Tone	0.989	1.000									
Accuracy	0.896	0.871	1.000								
Diction	0.969	0.978	0.908	1.000							
Intonation	0.866	0.871	0.880	0.857	1.000						
Vibrato	0.943	0.947	0.880	0.948	0.872	1.000					
Registration	0.980	0.993	0.882	0.979	0.884	0.957	1.000				
Agility	0.983	0.987	0.909	0.984	0.890	0.945	0.991	1.000			
Style	0.976	0.983	0.875	0.971	0.843	0.931	0.985	0.979	1.000		
Expression	0.712	0.714	0.745	0.751	0.658	0.682	0.737	0.760	0.750	1.000	
Overall Score	0.982	0.984	0.933	0.984	0.910	0.961	0.989	0.993	0.978	0.779	1.000

Note. Critical value = 0.532

Null Hypothesis #5. Null Hypothesis 5 was: There is no difference between holistic scores and rubric-based calculated percentage score. Judges were asked to “assign a holistic score from 0%-100% to each recording. This score should be independent of the rubric and should be the grade you would give if you were not using a rubric at all.” Because judges did not have information regarding the age of the students or the length of their study, there were two who expressed they had difficulty with this task and one judge declined to assign such a score. I observed in the data presented in Table 16 that the holistic scores were consistently higher than the rubric-based calculated percentage (average score/five points possible), so I compared them using the chi square test for goodness of fit.

The null hypothesis was: There will be no difference in the holistic score and the rubric-based score. The null hypothesis was not rejected, ($\chi^2 = 1.4598$; critical value =

22.362; $\alpha = 0.05$) which indicated that there was a good fit between the calculated percentages and the holistic scores.

Table 16.

Comparison of Calculated Percentages vs. Judges' Holistic Scores

Performance	Calculated percentage	Holistic score
1	60.80%	77.97%
2	41.39%	66.11%
3	27.59%	58.71%
4	37.60%	64.60%
5	29.35%	59.09%
6	67.47%	80.88%
7	44.94%	70.86%
8	37.94%	66.69%
9	27.82%	54.76%
10	28.88%	55.89%
11	75.74%	88.57%
12	27.46%	56.15%
13	40.85%	68.91%
14	30.27%	56.06%

Ratings. In Chapter Three, I stated that I expected the scores resulting from the adjudication of the professional singers to be the highest scores attained. Table 17 illustrates that the assumption was correct, for the set of data used in this research study. The professionals did attain the highest average scores. This finding was consistent with the stated expectation and one of the factors that supported the reliability of the instrument.

Table 17.

Average Overall Scores for Student/Professional Performances

Student	Avg. score
1	3.040123
2	2.097222
3	1.545988
4	1.879938
5	1.467901
6	2.255556
7	1.897222
8	1.390794
9	1.443827
10	2.120679
11	1.369136
12	1.513333
13	1.492857
14	1.218095
Pro 1	3.373688
Pro2	3.787037

Feedback from Judges. As part of the electronic survey format by which the research-based rubric was distributed, I was able to ask and collect feedback about the research-based rubric itself. Judges were posed the question: In the space provided below, please provide any feedback regarding the rubric, its levels, categories, characteristics, descriptors, or format. In response to this prompt, 29 of the 36 judges provided feedback about the research-based rubric. I was also able to collect feedback included in the comments provided by the judges for student use.

General comments. There were several positive comments that expressed general approval of the research-based rubric. Judges said things like, “Looks very much like the criteria for our jury exams,” “Love the rubric!” “Like the rubric and levels overall,” “Well-thought out,” and “All made sense to me, good range of choices.” One

judge commented about the benefit of having a break in the middle of the judging experience saying, “I appreciated the ‘please take a 10-minute break’ invitation.” One judge specifically mentioned the opportunity to provide comments about each category saying, “I liked that there was a space to give comments that may clear any confusion about why a particular score was given.”

Constructive comments included sentiments that indicated that the rubric was limited in its usefulness for scoring more advanced singers. They said things like “I suppose this rubric would work for singers who are working to simply learn the very basics. This is not a rubric for singers who have emerging talent,” “Helpful but rudimentary,” and “This is a very good start. I think a lot would be different if we could have seen the performances, too.”

Another challenge mentioned by some of the judges was the fact that the performances were presented with only an audio feed. One stated that “The inability to see facial expressions is a lack,” and at least one potential judge declined to participate because the examples did not include video. Others disagreed with the approach of breaking down the performance into discrete categories for the purposes of assessment saying, “The rubric calls too much attention to the separate/individual components of the technique. While these components, of course, need to all operate optimally, I listen to the voice as a whole, and have difficulty assessing “breath” or “diction” or “style” on their own merits, if, say, the entire thing is sung out of tune . . . one can hardly separate one aspect of singing from another in the total package.”

Finally, another judge felt that one static rubric might not be able to serve as a proper assessment for any given performance. This judge believed that “Not all rubrics will apply to every song.”

In summary, some of the judges provided feedback that indicated that the rubric was consistent with the instruments that they had used in the past and responded positively to the choices made including the space for comments. Judges provided constructive comments about the usefulness of this rubric for more advanced singers, the lack of a visual recording, the difficulty in assessing components of the voice instead of the voice as a whole, and the challenges of adapting one rubric to the multitude of possible song choices.

Song choice. Some of the judges provided feedback about the literature example used in this study. They specifically mentioned the example was not adequate to “illustrate the ability to transition registers” or to assess a student’s overall ability to perform in all languages when only one language was presented, and the example chosen had “no real difficulties to execute.” Other judges felt that the selection was either far too difficult for the level of ability demonstrated by the singers in the study or that the selection was not well suited for some of the singers. In summary, the judges who provided feedback about the choice of song generally expressed that it was not appropriate for either the characteristics being assessed or for the ability level of the group of performers.

Skill level of singers. Judges also mentioned the homogeneity of the sample of singers, and commented on their skill level. “This sample included only one singer that was not either a total beginner or just above that level. It would be hard to really get a

feel for this rubric without a wider variety of skill levels,” and “80% of these singers sounded like beginner of various ages.” Other judges wanted to know exactly what level of vocal training each student had achieved so that they could factor that into the feedback they provided. In general, the judges expressed that they would have liked to have seen more variety in the sample of singers with regard to their level of ability.

Levels of proficiency. The judges provided some feedback with regard to the selection of the five ability levels of beginner, early intermediate, intermediate, early advanced, and advanced. One judge stated that “The rubric has a good differentiation between levels,” but another was concerned that “many of the singers fell in between some of the levels.” Another judge was concerned about the inclusion of the mechanics of singing along with other more sophisticated skills of coordination and understanding all within the same assessment tool. He stated, the mixture of advanced concepts (text interpretation) and very basic elements (breath support) must be accounted for in the final scoring. You cannot get to things like conveying artistically the meaning of the text when breathing and intonation are undeveloped.”

Similarly, another judge suggested that I “take off the intermediate/advanced categories for intonation and accuracy . . . You can be a rank beginner and learn the correct pitches and rhythm. That should be expected of EVERY singer, period.” Another judge concurred that “correct pitches and rhythms should be Early Intermediate and then attention should be focused on phrasing, etc.” The judges expressed that the differentiation between levels was generally appropriate; however, there were times when singers fell between levels. They also provided some feedback about and suggested changes to the expectations for each level of some of the descriptors.

Performance criteria. Judges also had much to say about the performance criteria (breath, tone, accuracy, diction, intonation, vibrato, registration, agility, style, and expression) included in the research-based rubric. There were several comments about the difficulty in drawing a line between pitch accuracy and intonation. There was also concern about combining rhythmic accuracy and pitch accuracy in that “accuracy in pitches and rhythm are separate items and should be such” and one judge stated that scoring was difficult in this category because “many of the singers fell in between some of the levels.” Other criteria that judges mentioned they would like to see connected some way were breath and vibrato as well as style and expression. In fact, one judge thought that criteria should be combined more significantly. He said, “according to the rubric, the scores end up quite low. Maybe combine some [performance criteria] and end up with five.” Another judge was also concerned about overlapping skills and stated,

I thought you differentiated well in defining skill levels. There is always overlap in assessing vocal quality – e.g. breath/support will affect tone-intonation-vibrato, etc., but I think you have done a good job of pulling out the essentials. It might be interesting to break down some of the main categories, such as tone, into smaller sub-categories such as timbre, freedom, etc.

In summary, some of the judges thought that there were too many performance criteria and others thought that there were too few. This feedback was taken into consideration and compared with the feedback from the experts and the student performances when I made revisions to the rubric.

Descriptors. There was also a great deal of feedback about the specific descriptors for each performance criterion. General comments included some about the

benefit of having such descriptors as well as a desire to have statements that were more descriptive and allowed for a little more “wiggle room.” Comments about descriptors for specific categories mentioned the categories of breath, tone, diction, intonation, vibrato, accuracy. Others mentioned specific descriptors that were absent from any category such as “legato,” “chiaroscuro,” “nasality,” “open throat,” and “soft palate position.”

One judge mentioned that the “breath category seemed to be mostly concerned with the quiet inhale” and wondered “what about things like how low the inhale goes, or something related to appoggio, core muscle engagement or connection, etc.” Another judge was complimentary about the use of the “words ring, freedom, and vibrant” and suggested “ability to sustain would also be a good descriptor.” Diction received the most attention and was mentioned specifically by at least five of the judges. The descriptors that included the statement about “consistent and accurate diction in all languages” were confusing since they were asked to evaluate only one example in one language. Judges also wanted more “specifics of the language” to be included in the descriptors and more emphasis to be focused on consonants.

One of the judges was interested in differentiation between the various causes of faulty intonation such as “intonation can be not “hearing” the correct pitch OR not learning the correct pitch OR not being able to support the correct pitch.” Finally, one judge suggested a “possible reference to straight tone in the vibrato category.”

In summary, the judges provided some feedback that was positive and some feedback that was constructive. While some of the judges affirmed the decisions about what was included in the rubric, others suggested that the rubric was too simplistic and rudimentary. Some of the judges would have preferred a video example to assess over

the audio example provided, and there were several comments about the choice of repertoire relative to the ability levels of the singers. Judges also provided feedback about the specific levels, performance criteria, and descriptors.

Phase III: Student Perception Results

Following Phases I and II, during which I developed and implemented the rubric, I was able to meet with the student performers who participated in the study to present the information collected during the first two phases. This section is a summary of the feedback that I collected during these Phase II student interviews. I designed the student interview protocol to answer the research question, How do students perceive their use of the feedback from the solo vocal performance rubric to improve future performances? I scheduled and conducted interviews at the students' convenience. I gave each student a summary of his or her results from the 36 judges and asked each student a series of questions.

5. Intonation			
		Response Percent	Response Count
Beginning: Consistently out of tune		2.8%	1
Early Intermediate: Many out of tune pitches		19.4%	7
Intermediate: Few out of tune pitches		27.8%	10
Early Advanced: Very few out of tune pitches		41.7%	15
Advanced: Consistently accurate on all pitches		8.3%	3
		Comments	1
		answered question	36
		skipped question	0

Figure 1. Sample of aggregated data presented to students

Each student received an aggregate report for each of the performance criteria and a summary of the comments from each of the judges. Figure 1 is an illustration of the feedback provided to one of the students for one of the performance criteria.

Table 18 is an illustration of the summarized comments. All of the feedback was anonymous. The judges were not known by the judges nor were the judges known by the students.

Table 18.

Sample of Comments Presented to Students

Page 8, Q1	Breath	Date
1	Ends of phrases frequently lack consistent breath energy	Sep 12, 2013 11:32 AM
2	Lack of support in upper range	Sep 10, 2013 5:57 PM
3	Except top	Aug 15,2013 7:46 PM
4	Tension in the breathing mechanism. . . coordination of breath with resonance is a bit rigid	Aug 14, 2013 9:43 PM
5	Breath is sometimes noisy, but the tone is always supported	Jul 23, 2013 3:52 PM

To protect the students’ feelings, I reminded them that these performances were atypical of their normal process in that they were not allowed time to fully prepare the piece like they normally would. I also explained to them that each judge would assess the performances based on their own experience and context, and some of these judges were accustomed to working with students in much larger programs than the one in which the students were enrolled. This could result in judges’ ratings that were lower than what the students were accustomed to receiving. I eliminated comments that were deemed mean-spirited or unkind to spare the feelings of the student participants. I also eliminated comments about the research-based rubric itself and not about the student performances. I recorded the interviews on an iPhone audio recording application and subsequently transcribed them into Microsoft Word documents. I used open coding to analyze and report my findings from the transcriptions of the interviews.

Student interview question #1: Describe what you liked or did not like about this method of assessment.

After I analyzed the answers to question one using open coding, it was clear that the answers to this question fell into several groups including the following four: Levels, Performance Criteria, Descriptors, Numerical Scores, and General Comments.

Levels of proficiency. Six of the students commented about the usefulness of having the level labels. Four of the students believed the level labels should be kept because they “describe what the descriptions are stating” and were “more positive than . . . bad or good . . . You can tell that person is not as experienced as some people.” Another student said that this approach made it “more about the age of the voice and where you need to go with it.” Two students felt strongly that the level labels should be omitted. One of those students felt that her peers get “too wrapped up” in where they rank among each other. The other was a student with junior standing and was rated a Beginner in many categories. She indicated that it was hurtful, and if the categories were left off, perhaps the descriptors by themselves would not elicit as much emotion. One student felt that it was possible that singers might fall between levels. She wanted more “gray area” or wiggle room for the judging. Another student noticed the lack of “gray area,” but he thought it was a positive feature of the rubric. He indicated that it would add clarity and consistency to judging. The students provided feedback about the levels of proficiency included in the rubric, they were less concerned that each performance criteria was divided into five levels of proficiency and more concerned about the actual language used to label each of them.

Performance criteria. Students commented that the selected categories of Breath, Tone, Accuracy, Diction, Intonation, Vibrato, Registration, Agility, Style, and Expression were both “consistent” and “appropriate.” There was only one student who took exception with the selected categories. Her opinion was that the inclusion of Accuracy and Intonation as separate categories was confusing and possibly “redundant.” In general, the feedback from the students with regard to the performance criteria selected for the rubric was positive.

Descriptors. Nine of the students who were interviewed had feedback about the rubric's descriptors. They felt that the descriptors added clarity to the feedback they were receiving. They said things like, “I like this because I am able to know what to work towards” or “I could understand where my problem areas were.” Several students thought the step-by-step nature of the descriptors was very helpful. One stated, “This gives me steps to work on, too. I can see this and see the progression from point a to point b, and steps to work on . . . It would be useful as a performer.” One student felt the descriptors were tedious and “redundant” and could be a matter of “personal taste.” The example that she provided was that some people might really appreciate fast vibrato “and think it is amazing and beautiful” while someone else might think it is “awful.”

At least three students also felt that it gave them more specific feedback than the Likert-type scale they were used to using. One student indicated that if he received a seven in a particular category on the Likert-type scale, he wondered, “Okay, so what can I do to fix that other than you just personally think that's a seven?” Two students pointed out that the rubric descriptors could have a leveling effect on the judges scoring. One of them stated, “The descriptors make it very clear what the scoring of each section should

be. It is not your own interpretation of what you see is a beginner or intermediate singer. It is laid out as to what this rubric considers those levels to be." In summary, the students provided feedback that indicated the descriptors were a welcome addition to the assessment process and added more clarity and specificity to the feedback that they were receiving.

Comments. The students also looked to the comments as positive and necessary. And several commented on the importance of the interaction between the descriptors and the comments. Although the descriptors were very detailed, the comments helped them to understand "why you're in this category, or how you made it to this category." One student explained, "I read this, 'consistently shallow and constricted [in the descriptors],' then reading 'breathes in middle of words' [in the comments], that's particularly why you got this. Because some people may have gotten beginning, and not completely understood why you're in this category, or how you made it to this category." One student simply wished for "more comments." The feedback that I collected from the students regarding the comments reinforced just how crucial these comments are to helping the students really understand how they performed and how they can improve their performances in the future.

Lack of numerical score. I did not calculate numerical scores for the students for the purposes of this feedback session. I instead relied on the selected descriptors and accompanying comments. One student was attempting to understand the grading scale immediately upon receiving the feedback document. However, when I asked him at the end if it bothered him that there was not a numerical score he answered, "No, I think this is way better." One other student noticed that there were no numerical scores. He asked,

"This was the numbering system, right? Like, one through five?" I explained that it could be a five point scale for each category. He said, "I think it covers everything the way it should," and we were able to move forward with the rest of the feedback. At the end of the interview, the same student said that he was still "on the fence" about whether or not he wanted the numerical score. He liked the way that the rubric answers the question "What could I have done better?" and would like to have "both" a number and descriptive feedback. In summary, the students were generally concerned about the lack of a numerical score and were uncomfortable about how this type of instrument might ultimately be used to assign a numerical or letter grade to their performance.

Student interview question #2: Describe what you think the judges heard or did not hear in your performance.

Student interview question #3: Describe your understanding of the strengths and weaknesses of your performance

Answers to these two questions tended to overlap; therefore, I have combined the responses into one set in this section. In general, students indicated that the feedback provided to them through the research-based rubric was consistent with their own self-perceptions. They also made statements about their dissatisfaction with their level of preparation and provided insights on their perceptions about recorded performances.

Consistency with self-perception. All of the students mentioned that the feedback was consistent with what they already knew and believed about their abilities and their performances. One student stated, "It just solidified areas that I still need to work on and areas that I know aren't up to par." One student stated, "They heard my nerves . . . because I was really breathy, and my intonation was off; and that is what happens when I

get nervous." Another said, "Weaknesses? I wasn't surprised." Several mentioned the specific categories where the feedback they received was aligned with their own perceptions of their development. "I need to breathe," or "My breathing, that is something I am trying to work on consistently," or "I know I have pitch problems, so they were consistent on that," or "I think that they think I have a good tone. It's sometimes inconsistent, but for the most part, it is a good tone." In summary, all of the students indicated that the feedback that they received from the rubric was consistent with what they already perceived about their abilities. None of the students stated that the comments or selected descriptors were surprising or out of line with what they already believed about himself or herself.

Level of preparation. Six students felt that the limited amount of time that they had to prepare the piece significantly affected their performances and scores. One specifically mentioned "diction" and "style and expression" as areas where he felt he did not have enough time to prepare the song for performance. One student ended the interview stating that the bottom-line take-away from the whole experience was "two weeks is not enough time to get something performance ready." Three students felt that the song was not a good fit for each of their voice types and felt that prevented them from presenting their best possible performances. The feedback from the students indicated that the song selection was too difficult to prepare in the limited amount of time that they were given and that they perceived that negatively influenced the quality of their performances.

Recorded performances. One student stated that she felt her performance would have been better understood if the judges had seen a video recording of the performance

in addition to listening to the audio recording. A second student also expressed his concerns over being judged via recording instead of a live performance. He stated, "In a live performance versus a recording there are so many factors that can change." He also wondered about the recording process and its effects on the final version of a performance. He felt that "if they just come too close to the microphone or too far away, that can affect things too." The feedback from the students indicated that they perceived that the method of presenting the performances to the judges via audio recording was not ideal.

Student interview question #4: Describe what (if anything) you plan to do with the information. What actions (if any) do you plan to take?

In response to this question, all of the students agreed that they could either alone or with the help of their private instructor form an action plan based on the feedback that they received from this rubric. They acknowledged that they would know what to work on but would need the help of their instructor to know how to go about doing that. Some comments included, "I could make a plan with the help of my teacher," and "If I were assessed using this I would know what I would need to do to get better," and "I think it's definitely focused and, um, specific enough that I could look at this and, . . . I would write down, like, all the categories and write 'work on this, work on this.' You can just be able to check that off and kind of keep practicing." One student mentioned its usefulness in short term goal setting as well as long term. She said, "I would obviously strive to be advanced even if it took a little while. I would probably try to each time get to the next level. So if most of the ratings were early intermediate, I would try for intermediate as my smallest goal and kind of go from there. Because

obviously you aren't going to go all the way to advanced from one jury to another." In summary, the students indicated that the feedback that they received would enable them, with the help of their private instructor, to make long term plans to allow them to achieve the stated learning outcomes defined in the rubric and that the descriptors for each level of proficiency would enable them to establish short term goals for their progress toward those learning outcomes.

Student interview question #5: Do you have any other comments about the rubric or have anything else you would like to share?

Most of the responses to this question tended to reiterate earlier points. There were two notable ideas brought out by my asking this question. One student was concerned with the format of the feedback in that the comments were not on the same page as the category to which they referred. One of the most interesting responses took into account the ability of this rubric to capture "someone's personal growth and improvement" by taking periodic snapshots of the person's overall journey of developing as a singer.

Emerging Themes

All of the quantitative data were analyzed using open coding. During the course of this analysis several themes began to emerge across all of the groups of participants that provided data. The seven emerging themes were levels, performance criteria, descriptors, numerical scoring, comments, recording method, and song selection relative to the skill level of the singers will be discussed more fully in Chapter Five.

Summary

The purpose of this mixed-methods participatory action research was to develop and test a comprehensive rubric for assessing undergraduate solo vocal performances. The first phase of the study, rubric development, involved collecting data from five expert vocal music educators. The second phase of the study was the implementation of the research-based rubric in which 36 judges used the rubric to score 20 performances. Feedback from the judges was collected and analyzed, and statistical analysis of the quantitative results indicated that the rubric was both valid and reliable. The third, and final, phase of the study was collecting feedback from students about what meaning they were able to make from the information provided in the completed rubrics. Themes that emerged from the analysis of the qualitative data were levels, performance criteria, descriptors, numerical scoring, comments, recording method, and song selection relative to the skill level of the singers. Chapter Five is a discussion and reflection on the data presented in Chapter Four.

Chapter Five: Discussion

Research about the assessment of musical performances was present in the literature dating as far back as the 1970s and continuing through the time of this writing (Abeles, 1973; Bergee, 1993; Ciorba & Smith, 2009; Cooksey, 1975; DCamp, 1980; Fiske, 1975; Greene, 2012; Horowitz, 1994; Jones, 1986; Latimer et al., 2010; Levinowitz, 1985; Saunders & Holohan, 1997; Wapnick & Eckholm, 1997). I designed this research study in an attempt to address the underrepresented area of assessment of the vocal instrument. I also sought to verify what I found in the literature about the value of the criteria-specific rubric in providing useful feedback to the student (Asmus, 1999; Latimer et al., 2010; Norris & Borst, 2007; Saunders & Holohan, 1997; Wesolowski, 2012). Guiding my research were the following research questions and hypotheses:

Research Question 1: What are the appropriate performance criteria, learning outcomes, and meaningful descriptors for various levels of proficiency for undergraduate solo vocal performance?

Research Question 2: How do students perceive their use of the feedback from the solo vocal performance rubric to improve future performances?

The first research question addressed the development of the tool. The second research question addressed measuring the performers' ability to interpret and use the feedback the tool provided. The following hypotheses were designed to test the reliability and validity of the tool:

Hypothesis #1. When scoring performances using the research-based rubric, at least one judge will score differently than the others.

Sub Hypothesis #1. There will be a difference in average mean score on the research-based rubric, when comparing individual judge scoring to the overall group mean score.

Hypothesis #2. When scoring performances using the research-based rubric one category at a time, at least one judge will score differently than the others.

Sub Hypothesis #2. There will be a difference in mean score on the research based-rubric, when comparing individual judge scoring on individual categories to the overall group mean score.

Hypothesis #3. There will be a difference in judges' scoring utilizing the research-based rubric, on repeat performance when compared to the same judges' scoring for the original performance.

Sub Hypothesis #3. The event score when applying the research-based rubric is dependent upon which judge conducted the rating.

Hypothesis #4. There will be a relationship between each of the ratings of characteristics of breath, tone, accuracy, diction, intonation, vibrato, registration, agility, style, and expression, and judges scores utilizing the research-based rubric.

Hypothesis #5. There will be a difference between holistic scores and calculated rubric-based percentage score.

Review of Methodology

To answer these questions and test these hypotheses, the method that I selected for my study was mixed-methods participatory action research conducted in three phases. As a practitioner in the field and member of the research site's community, I was able to involve other stakeholders from the organization in the planning and execution of this

research project. I employed both quantitative and qualitative methods in collecting and analyzing data.

The first phase was the preparation of a research-based rubric, which was achieved via a review of the literature and expert interviews, and the preparation of recordings, which included both student and professional singers. The second phase was the implementation of the research-based rubric where the research-based rubric was used by judges to assess the recorded performances. The third phase was the collection of student feedback which was facilitated through interviews of the student performers in an attempt to determine what information they learned from the completed research-based rubrics that the judges used to assess their recorded performances.

Phase I: Rubric Development

Research Question 1: What are the appropriate performance criteria, learning outcomes, and meaningful descriptors for various levels of proficiency for undergraduate solo vocal performance?

The first research question was designed to define the appropriate performance criteria and meaningful descriptors for various levels of proficiency for undergraduate solo vocal performance. The interview questions were selected to address each facet of this inquiry. This section will discuss the means by which I compared the results from the expert interviews with the information gathered in the literature review in an attempt to draft the research-based rubric. This rubric was then implemented, and more data were gathered during Phases II and III of the study.

Rubric Organization. Based on the methodology that I followed to construct the rubric outlined in Wesolowski (2012) and the research question that guided this phase of

the research, I needed to organize the rubric to include the following components: levels of proficiency, performance criteria, learning outcomes, and descriptors. I needed to develop a descriptor for each of the levels of proficiency for every performance criteria. These descriptors were based on the progression toward the defined learning outcomes, which were also the descriptor for the highest level of proficiency. Therefore, there needed to be a descriptor at the intersection of level of proficiency and performance criteria and a learning outcome at the intersection of each performance criteria and its corresponding highest level of proficiency. Figure 2 is an illustration of the organization of the rubric.

	Level	Level	Level	Level	Level
Performance	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome
Criteria	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome
Performance	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome
Criteria	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome
Performance	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome
Criteria	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome
Performance	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome
Criteria	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome
Performance	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome
Criteria	Descriptor	Descriptor	Descriptor	Descriptor	Learning Outcome

Figure 2. Rubric organization template

Levels of proficiency. Four of the experts agreed that there should be three main levels of beginning, intermediate and advanced and that there should be stages in between each of those three that represented transition phases between each of the main categories. This finding is consistent with Gordon's (2002) statement, as cited in Latimer

et al. (2010), that "the more descriptors included for each dimension, the more reliable the rubric will become, as long as that number does not exceed five" (p. 170). The fourth expert suggested that the categories be based on grade level such as freshman, sophomore, junior, and senior, but the remainder of the panel believed that the students' progress was not necessarily tied to their age and entering freshmen "come in with such varying degrees of experience" (Expert Five). Therefore the majority of the experts agreed that levels of proficiency would be a more appropriate and useful scale than grade level would be. I concluded that I would label the five levels beginning, early intermediate, intermediate, early advanced, and advanced.

Performance criteria. All five of the experts named breath or breath support and tone or tone quality or tone production as the first two considerations. This was consistent with the review of the literature in which breathing and tone were the first two characteristics discussed by all of the writers (Miller, 1996; Paton, 2006; Vennard, 1949; Ware, 2008). The concept of alignment was included in much of the literature (Miller, 1996; Paton, 2006; Vennard, 1949; Ware, 2008), but was not addressed by the experts. Alignment is extremely important, and it would have been appropriate to include it in any rubric that assessed singing. However, because this assessment was conducted exclusively via audio recordings, and there was no opportunity for the judges to observe the singers' physical posture, it was not included in this rubric. Therefore, the first two criteria included in the rubric were breathing and tone.

The experts were unanimous in their mention of three additional criteria essential to good singing. Those criteria were diction/vowel shape, understanding of the text/interpretation, and understanding of the style. Diction was discussed in great detail

by several of the authors (Paton, 2006; Vennard, 1949; Ware, 2008). Interpretation of text and style were prominent in the writings of Miller (1996) and Ware (2008). These criteria were included in the rubric as diction, expression, and style.

Four experts also considered intonation to be an important criterion. There was also expert consensus about vibrato and registration. After comparing these criteria to Miller (1996), Paton (2006), Vennard (1949), and Ware (2008) I determined that there should be additional criteria for coordination, which would include intonation, vibrato, and registration. These advanced skills can only be performed after the basic breath and tone production skills are mastered and coordinated. The literature (McKinney, 1994; Paton, 2006; Ware, 2008) also supported including agility in this category.

The experts disagreed about the criterion of accuracy. Some felt that the student's ability to sing the correct notes and rhythms was very important. Others felt that it was something that should be expected and should not be assessed. The literature did not address this topic. I felt that it should be included, especially if the subject of the assessment was beginning singers. As one expert stated, "If they are not learning the right notes and rhythms, are they really then able to incorporate these other ideas so that they can express the text the way it is meant to be expressed" (Expert Two)?

One criterion that none of the experts addressed was resonance. All of the authors (Miller, 1996; Paton, 2006; Vennard, 1949; Ware, 2008), however, discussed this as an important criterion. I agreed that resonance was as a hallmark of mature tone, and it was important to include in the rubric. I determined the most appropriate way to include resonance was in the descriptors for tone.

I organized the criteria discussed in this section into three progressive categories which I labeled mechanics, coordination, and understanding. The mechanics category was comprised of the most basic performance criteria that could be mastered and assessed individually and included the performance criteria of breath, tone, accuracy, and diction. The category coordination was made up of performance criteria, which required the mastery of a combination of more than one of the basic performance criteria. This category included the performance criteria of intonation, vibrato, registration, and agility. The category of understanding included the performance criteria of expression and style, which are advanced performance criteria that involve the synthesis of knowledge and skill into an aesthetically pleasing performance. The organization of the early draft of the research-based rubric is illustrated in Figure 3.

		Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
Mechanics	Breath					
	Tone					
	Accuracy					
	Diction					
Coordination	Intonation					
	Vibrato					
	Registration					
	Agility					
Understanding	Style					
	Expression					

Figure 3. Early draft of research-based rubric

Descriptors. Once I determined the criteria that would be included in the rubric, I needed to define the appropriate descriptors for each level of proficiency for each of them. I evaluated the expert responses to an interview question along with the

descriptions available in the literature to determine which descriptors to use in the rubric. It made the most sense to begin with the learning outcome for each performance criteria, to develop its descriptor, and then to develop incremental descriptors for each of the levels that led up to the highest level of proficiency/learning outcome.

Breath. When describing proper breathing the experts used words and phrases when referring to visual cues such as "breathing within their posture" (Expert Four) or "within the context of their body" and "lower expansion" (Expert Two). These visual cues were not included in the rubric because the judges did not experience the performances visually. Other auditory cues included "steady stream of air" (Expert Four), "column of air, not tense or tight, and a quiet, deeper breath versus "shallow and very loud" (Expert Two). Miller (2004) also made many references to the visual aspects of breathing; however, the auditory aspects included descriptors such as "silent" (p. 2), "singing on the inhalation of the breath" (p. 2), and maintaining a feeling of fullness throughout the phrase. I developed the learning outcome for the most advanced singers based on these findings. The descriptor for the learning outcome was, Consistently silent inhalation that is free from tension with a steady stream of air that is consistently present supporting the tone, and the remaining descriptors for each level, as shown in Figure 4, indicated a progression of developing consistency over time toward this ideal.

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
Breath is consistently shallow or constricted and/or is rarely present supporting the tone.	Breath is sometimes shallow or constricted and/or is sometimes present supporting the tone.	Emerging ability to demonstrate silent inhalation that is free from tension with a steady stream of air that is fairly consistently present supporting the tone.	Approaching a consistently silent inhalation that is free from tension with a steady stream of air that is consistently present supporting the tone.	Consistently silent inhalation that is free from tension with a steady stream of air that is consistently present supporting the tone.

Figure 4. Rubric descriptors for breath

Tone. When describing tone, the experts used many adjectives to describe both the desirable sound and the undesirable sound. Descriptors for desirable tone included "clear" (Expert Five), "consistent throughout the range, sounds like the student and not like they are trying to sound like someone else, the right amount of pressure, proper closure of the chords" (Expert Two), "beautifully dark and colorful, warmer or richer, and "mature" (Expert Three). Descriptors for undesirable tone included "fuzzy, airy, strident, throaty" (Expert Five), "brassy" (Expert One), "breathy" (Expert Five), "pressed or artificially heavy or with a lot of muscling, yelling" (Expert Two), and "swallowed or nasal" (Expert One). The literature also provided many descriptors of desirable tone including "freely produced; pleasant to listen to; loud enough to be heard; rich, ringing and resonant; energy flows smoothly from note to note; consistently produced; vibrant, dynamic, alive; flexibly expressive" (McKinney, 1994, p. 77); natural sound, freedom from tension, clear and in tune, elasticity, "ample volume with ringing forward in the mask placement", chiaroscuro, flexibility, and agility (Ware, 2008, p. 59); and audibility, resonance, clarity, intelligibility, pure intonation, dynamic variety, timbre consistency and variety, vibrato, range, and ease of freedom (Paton, 2006, p. 17).

There are many facets of tone and it was difficult to narrow all of these rich descriptors into one statement, but I wanted the descriptor for the most advanced singer to capture the essence of as many of these things as possible. I determined that a clear (Expert Five; Ware, 2008) tone was the best descriptor and essentially ruled out the negative descriptors of fuzzy, airy, or breathy (Expert Five) that the experts provided. A free tone such as the one described by McKinney (1994) and Ware (2008) was one that lacked tension (Ware, 2008), pressing (Expert Two), or throatiness (Expert Five). A rich tone (McKinney, 1994) was one that also included warmth and color (Expert Three). Ringing (McKinney, 1994) addressed the proper placement and the balance of bright and dark tone (Expert Three) or chiaroscuro (Paton, 2006; Ware, 2008).

One characteristic of tone that none of the experts addressed was resonance. However, McKinney (1994) and Paton (2006), discussed this as an important component of tone, and I agreed it was important to include in the descriptors for tone. Finally, experts agreed that maturity of tone was the result of mastering all of these characteristics and executing them consistently in coordination. Therefore, I developed the descriptor for the learning outcome based on these ideals and created the descriptors for the remaining levels of proficiency based on the increasing consistency over time toward this benchmark, the results of which are illustrated in Figure 5.

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
Consistently lacking in clarity, maturity, freedom, richness, ring and/or resonance	Somewhat lacking in clarity, maturity, freedom, richness, ring and/or resonance	Occasionally clear, mature, free, rich, ringing, and resonant	Frequently clear, mature, free, rich, ringing, and resonant	Consistently clear, mature, free, rich, ringing, and resonant

Figure 5. Rubric descriptors for tone

Accuracy. The experts were clear that accuracy, singing the correct pitches and the correct rhythms, was an important basic skill that had to be mastered before any of the more advanced technical coordination or elements of style or expression could be introduced. One expert stated that "accuracy would have to be a precursor" (Expert One) to progressing to more complex skills and abilities. Accuracy was not addressed in the literature; however, the experts were so emphatic about this skill, especially for beginning singers, that I decided it must be included in the rubric. The descriptor for the learning outcome (see Figure 6) was simply "correct pitches and rhythms," and allowances were made for fewer and fewer errors for the less experienced singers.

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
Consistently incorrect pitches and rhythms	Many incorrect pitches and rhythms	Few incorrect pitches and rhythms	Very few incorrect pitches and rhythms	Correct pitches and rhythms

Figure 6. Rubric descriptors for accuracy

Diction. When discussing diction, the experts were clear that it was an important skill on which beginning singers should focus. Diction's impact on other areas of singing was discussed. "Proper vowel formation" (Expert Five) was a consideration for desirable tone and accurate intonation, a notion shared by Ware (2008). The balance between consonants and vowels was a precursor to "consistent legato tone" with energized consonants providing "necessary energy for firm phonation" (McKinney, 1994, p. 156). One expert also mentioned the role of diction in "stylistic understanding and expression" (Expert Two) because "diction needs to be clear, so you are communicating something" (Expert Two). With those points in mind and as seen in Figure 7, I decided to address the

progression from achieving proper vowel formation in beginning singers through executing accurate diction consistently in all performance languages.

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
Inconsistent vowel formation	Consistent vowel formation	Emerging balance of consonants and vowels	Approaching consistent and accurate diction in all languages	Consistent and accurate diction in all languages

Figure 7. Rubric descriptors for diction

Intonation. The experts' opinions on intonation were many and varied. The causes of intonation could be due to lack of breath support, improper tone production, a depressed soft palate, improper vowel formation, tension, inability to distinguish pitch inaccuracy, an issue in only one part of the range (for example, the *passaggio*), or a lack of understanding of how the note functions in the chord or the chord progression. The literature supported the experts' belief that poor intonation was a result of "one or more malfunctioning components of the vocal process" (Ware, 2008, p. 96). The experts also agreed that in the jury setting, it is difficult to determine the cause of inaccurate intonation. They felt that type of diagnosis required a more in depth understanding of the student and was a determination that should be made by the private lesson instructor. They agreed that it was only appropriate to simply describe what was heard in the performance, and that highlighting any noticeable inaccuracies would be a cue to the student and to the private lesson instructor to investigate further. Therefore, when writing the descriptors for intonation in Figure 8, I focused exclusively on if the intonation was accurate or inaccurate.

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
Consistently out of tune	Many out of tune pitches	Few out of tune pitches	Very few out of tune pitches	Consistently accurate on all pitches

Figure 8. Rubric descriptors for intonation

Vibrato. In the literature, a desirable vibrato was usually described as having a regular pattern and being neither too fast or too slow (McKinney, 1994; Miller, 1996; Ware, 2008). The experts discussed vibrato as important, but has less to say about how they would describe vibrato, with one exception. One expert believed that "even and consistent vibrato and spin even in those pickup notes and in runs" (Expert Three) were most likely to "disappear the most in the young singer" (Expert Two). Therefore, I included this (see Figure 9) as part of the descriptor for the learning outcome.

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
Consistently having too fast/too slow speed and/or an irregular pattern	Somewhat having too fast/too slow speed and/or an irregular pattern	Occasionally having too fast/too slow speed and/or an irregular pattern	Frequently having moderate speed and regular pattern	Consistently having a regular pattern even in pick up notes and melismatic passages

Figure 9. Rubric descriptors for vibrato

Registration. As singers advance, and their ranges extend, it is necessary to learn how to sing in multiple registers. Very beginning singers are usually accustomed to singing in only one register either the chest or the head register. The goal is for them to develop consistent tone across all registers of their voices. The experts were split in their opinions about registration. One expert felt that registration was too advanced for undergraduate singers to understand and to master (Expert One). Other experts who believed that registration was an importation skill to address in undergraduate singers

stated that there are major adjustments that must be made by the singer in both air pressure and vowel formation. Miller (1996) would agree that this skill of vowel modification or *aggiustamento* was a very advanced ability, therefore, I only included it in the descriptor for the most advanced level in my rubric (see Figure 10).

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
Ability to sing in only one register	Emerging ability to sing in multiple registers	Ability to sing in multiple registers but with inconsistent tone	Consistent tone quality across all registers	Consistent tone quality across all registers including appropriate vowel modifications

Figure 10. Rubric descriptors for registration

Agility. The experts did not directly discuss agility with the exception of the discussion for the need to have consistent vibrato throughout fast moving passages, which was part of the discussion on vibrato. However, the literature lent enough support for this category for it to be included in the rubric. I chose the descriptors for this category as shown in Figure 11 based primarily on Ware's (2008) description that stated agility is "based on the singer's ability to negotiate musical challenges nimble and quickly, including wide pitch intervals, *coluratura* (fast note) scales and passages, and dynamic variations" (p. 97).

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
Inability to negotiate musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	Emerging ability to negotiate <u>some</u> musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	Ability to negotiate <u>some</u> musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	Emerging ability to negotiate musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	Skillful ability to negotiate musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly

Figure 11. Rubric descriptors for agility

Style. Ware (2008) emphasized the need for singers to have a comprehensive understanding of style periods and their performance practices as well as an ability to apply those elements of style to performances. The experts also agreed that understanding of style periods and stylistic practices was an important, although an advanced, skill. This understanding necessarily comes late in a student's training because they usually do not begin their studies of music history until their sophomore or junior year in school. Elements of style mentioned by the experts included "when and how much vibrato to use," "appropriate tempi" (Expert Four), "the use of ornamentation" (Expert Four), the amount of "interaction between the voice and the piano" (Expert Two), and the "extremes of dynamic contrasts" (Expert Two). Because the elements of style are myriad and subtly applied, I decided to speak of style in general terms trusting that the judges would, by nature of their advanced training in this area, have sophisticated understanding of the stylistic practices that would apply to the piece used in the study and would be able to accurately recognized if the students were able to apply them or not. This approach is illustrated in Figure 12.

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
No evidence of stylistic understanding	Emerging ability to apply a few basic elements of style appropriate to the piece	Ability to apply basic elements of style appropriate to the piece	Emerging ability to employ more sophisticated elements of style appropriate to the piece	Skillfully employs stylistic practices appropriate to the piece

Figure 12. Rubric descriptors for style

Expression. In both the literature and the interviews with the experts, expression was an important topic. The authors and experts agreed that perfect technique is of no use if the singer is unable to communicate meaning. According to one expert, this communication of meaning begins with an "internal understanding of the text" (Expert Two). To begin this journey of understanding, one expert advocated translating the text word for word and then "translate it into how you would say it. Your speak" (Expert Two). Another expert stressed that every action by the singer whether it be dynamics or gestures must also be "meaningful" (Expert Three) to be effective. The list of tools available to singers to accomplish this communication of internal understanding, dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions, were adapted from Miller (1996) taking into consideration the feedback collected from the expert interviews. These choices were illustrated in Figure 13.

Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
1	2	3	4	5
No evidence of internal understanding of the text	Emerging ability to communicate internal understanding of the text	Ability to communicate internal understanding of the text using <u>some</u> of the following: dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions	Emerging ability to communicate internal understanding of the text using <u>all</u> of the following: dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions	Skillful ability to communicate internal understanding of the text using <u>all</u> of the following: dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions

Figure 13. Rubric descriptors for expression

Numerical scoring. The experts agreed that scale of one through five, with one being the value for the beginning singers and five being the value for the advanced singers, was sufficient and appropriate. I adopted this scale for the implementation of the rubric in this study. The selected levels (beginning, early intermediate, intermediate, early advanced, and advanced) were translated into numerical scores for the purposes of conducting the statistical analysis that is discussed as part of Phase II.

Comments. Finally, all experts agreed that there must be space for commenting on each characteristic. They believed that if the purpose of the rubric was to provide the best possible feedback to the students, then the judges must have the opportunity to make comments and expand their feedback beyond what the rubric's descriptors might indicate. One expert even stated, "I think that the comments are more important than the scores" (Expert Five). I agreed with their opinions and included space for the judges to comment within the rubric (See Appendix B: Rubric Draft).

Phase II: Rubric Implementation and Validation

The second phase of the study involved actually scoring student performances using the newly designed research-based rubric. This section will discuss how I used the data gathered during this phase to determine that the rubric was indeed reliable by testing the five stated hypotheses.

The first hypothesis (When scoring performances using the research-based rubric, at least one judge will score differently than the others.) addressed the inter-judge reliability, which was evident in the statistical analyses. The first ANOVA test revealed that some judges demonstrated more agreement with the judge group than did the others, but the subsequent z -test for difference of means revealed that there were no significant differences in the means of the two groups when each judge's individual scores were compared to the overall scores for each performance. In other words, when measuring each judge's agreement with the overall group, even though there were some differences in the levels of agreement, the differences were not significant.

The second hypothesis (When scoring performances using the research-based rubric one category at a time, at least one judge will score differently than the others.) was designed to determine the criteria-specific validity of the rubric. This ANOVA test revealed that there was no significant difference between the judges within the scoring for each category. The subsequent z -test for difference of means also revealed that there were no significant differences in the means of the two groups when the scores for each category were compared to the overall scores for each performance. These findings supported the reliability of the instrument.

The third hypothesis (There will be a difference in judges' scoring utilizing the research-based rubric, on repeat performance when compared to the same judges' scoring for the original performance.) was designed to determine intra-judge reliability (the same judge would score a repeated performance the same way twice). Intra-judge reliability was inconclusive in the first z -test for difference of means. There were significant differences in the scoring of two of the repeated events. However, the chi square test for independence indicated the ratings were independent of the judge. This finding also supported the reliability of the instrument.

The fourth hypothesis (There will be a relationship between each of the ratings of characteristics of breath, tone, accuracy, diction, intonation, vibrato, registration, agility, style, and expression, and judges scores utilizing the research-based rubric.) predicted that there would be relationships between the performance criteria when they were compared. It was not surprising that the categories of "Breath/Tone" were strongly correlated. Vennard (1949) wrote specifically about the importance of the relationship between these two components of singing.

As expected, the professional singers received higher scores than did the student performers, and although the percentages calculated from the rubric scoring were observably different than the holistic scores, when the fifth hypothesis (There will be a difference between holistic scores and calculated percentage score.) was tested, the chi square test indicated there was a good fit between the two types of scores.

Phase III: Perceived Value of Feedback to the Performers

Research Question 2: How do students perceive their use of the feedback from the solo vocal performance rubric to improve future performances?

Based on the results of the interviews with the students, I concluded the presence of the descriptors, which are essential to the construction of a criteria-specific rubric, in conjunction with judge's comments were most helpful to students in validating their self-perceptions, providing specific feedback, and assisting with action planning. The statements made by the students are consistent with Hattie (2012) with regard to the importance of formative evaluation as one of the top 10 influences on student achievement. The feedback in the rubric was specific enough for the students to understand the ultimate goals, where they were in relation to those goals, and what they needed to do to close the gap (Hattie, 2012).

The students also indicated in their responses that the rubric, when utilized regularly over time, would be useful in showing the journey in the development of a singer and would help them to self-assess along the way. This ability to self-assess is also included in Hattie's (2012) top 10 influences on student achievement. If the student is able to self-assess and share that information with the teacher, this feedback loop, which is a third component of Hattie's (2012) list, allows the teacher to see the learning through the eyes of the student and makes learning visible which facilitates planning of the next steps. In summary, the students found the feedback (especially the detailed descriptors) to be affirming, effective for short- and long-term planning, and useful for marking progress over time.

Recommendations

Based on the findings from this study, particularly the significant statistical results that supported the reliability of this instrument and the significant validation from the student performers, I would recommend using this rubric with a few modifications at the

research institution that was the subject of in this study. It would not be appropriate to use this rubric with students from a more selective program. This view is supported by the feedback from the judges that stated that the rubric was “rudimentary” and that it “would work for singers who are working to simply learn the very basics.” The modifications that I would recommend would include contextualizing the numerical scores for the purposes of grading by making some allowances for variance in age of the students, clarifying some of the categories, and revising some of the descriptors. All of the modifications and revisions made to the rubric are included in Appendix H.

Sliding scale for scoring. Students stated that they found the rubric valuable, especially the descriptors and the comments, but some of them were concerned about how the rubric would translate into a numerical score for grading. There was also some discussion among the experts about the possibility of taking into account either the singer’s age or amount of experience in the grading scheme. The experts felt that in practice, assessors actually take into account the level of the singer when assigning a grade. Supporting this assertion, one of the judges indicated that she based her holistic scores on whether she felt the singer was a beginning, intermediate, or advanced singer. Another judge also expressed concerns that the overall scores for the recorded examples ended up being “quite low.” In summary, the experts, students, and judges all desired a method of numerical scoring that would take into account the level of development of the singer.

My recommendation is to use a sliding scale that takes into account the number of semesters/years that a student has been studying voice at the college level. Freshmen (or students who have been studying for one or two semesters) would be expected to have

most of their scores in the Beginner (1) or Early Intermediate (2) level. Therefore, their top scores would be expected to be twos, so 20 would be the basis for the scoring for freshmen. Sophomores (or students who have been studying for three or four semesters) would be expected to have most of their scores in the Early Intermediate (2) or Intermediate (3) level. Therefore, their top scores would be expected to be threes, so 30 would be the basis for the scoring for sophomores. Juniors (or students who have been studying for five or six semesters) would be expected to have most of their scores in the Intermediate (3) or Early Advanced (4) level. Therefore, their top scores would be expected to be fours, so 40 would be the basis for the scoring for juniors. Seniors (or students who have been studying for seven or eight semesters) would be expected to have most of their scores in the Early Advanced (4) or Advanced (5) level. Therefore, their top scores would be expected to be fives, so 50 would be the basis for the scoring for seniors. This sliding scale is illustrated in Table 19.

Table 19.

Sliding Scale to Determine Numerical Grades

Year of study	Semester of study	Scoring basis
Freshman	1-2	20
Sophomore	2-3	30
Junior	3-4	40
Senior	4-5	50

Performance criteria. Several recommendations were made by the judges with regard to the criteria that I selected for the rubric. There were several criteria that the judges suggested that I combine since they were so closely related to each other. The judges recommended that the criteria of breath and tone be somehow combined. They

also recognized a connection between breath and vibrato as well as style and expression. I evaluated their recommendations against the literature reviewed for the study, the data gathered for the study, and the purpose of the rubric itself. In the end, I decided that I would keep the criteria as they were originally selected because keeping some of the more basic criteria, such as Breath, as standalone criteria would serve to provide more specific feedback to the type of students being assessed using this rubric.

Breath and tone. Breath and tone are closely related. In fact, the statistical data showed a strong correlation ($r = 0.989$) between the two. Breath is indeed the motor that causes the instrument of the voice to go; however, there are many other factors that are to be considered in evaluating tone that extend beyond breath such as placement, resonance, and diction. Also considering the context of my study, that most of the singers that I deal with are strict beginners, I thought it would be helpful to keep the basic elements of singing separate for the purposes of delivering the most specific feedback as possible to the students.

Breath and vibrato. For similar reasons, I chose to keep breath and vibrato separate. Again, proper breath support is essential to maintain an appropriate vibrato, but I wanted to keep breath a separate criteria so that the beginning students were again given the most specific feedback as possible with regard to this most essential and basic criterion. Additionally, breath is only one of several considerations when assessing vibrato. Pleasing vibrato is the coordination of many factors, and according to Ware (2008) unpleasant vibrato could be the result of "hyperfunctional or hypofunctional muscular activity, emotional imbalance, physical and vocal fatigue, nervous system

disorders, or vocal-fold injury" (pp. 96-97). For these reasons, I chose to keep these categories separate.

Style and expression. Style and expression are also closely related. In Chapter Two they were discussed in the same section. Both Ware (2008) and Miller (2004) emphasized the importance of style and expression. It is important to understand the definitions of these categories in order to decide if they should be considered separately when assessing young singers. Style is the understanding of the performance practices of a particular musical period. For example, there are different practices with regard to phrasing, articulation, etc. for the Baroque style period than for the Romantic style period. Expression, on the other hand, is the ability of the performer to communicate sentiment (Miller, 1996, p. 202). This requires the performer to have a complete understanding of and connection to the text. The combination of the stylistic elements and the expression of genuine emotion combine to create an aesthetically pleasing and cathartic performance; however, these are different skills that require unique instruction and research to hone. Therefore, since beginning students can begin to progress in one of the areas without mastering the other, I have chosen to keep these two categories separate.

Accuracy and intonation. Both students and judges had a difficult time drawing a distinction between the accuracy (singing the correct pitches and rhythms) and intonation (singing in tune) categories. This topic was also discussed when I interviewed the initial group of five experts. There is really no mention of executing the correct pitches and rhythms in the literature that I reviewed, but it is a concern in a program like ours where students are less prepared than in other more selective programs. In addition,

the pedagogical approaches to resolving each of these issues are separate and distinct. If a student is unable to read music properly (a cognitive concern), a teacher would address that differently than helping them to improve their intonation (a technical concern). Therefore, I believe they should be assessed separately so that they can be properly corrected.

Among the judges, there was support for separating pitch accuracy from rhythmic accuracy as well as only scoring these things for beginning singers and then focusing on other extensions of this skill (i.e. phrasing). I chose to resolve this apparent overlap and create more differentiation between the two categories by changing what I would include in the descriptors for the accuracy category. I agreed with the judges that the focus of singing past the very beginning stages should no longer be executing the proper pitches and rhythms. Therefore, I expanded the scope of the Accuracy category to include phrasing and articulation, which were not directly addressed in any other category. These changes are illustrated in Figure 14.

	Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
	1	2	3	4	5
Original	Consistently incorrect pitches and rhythms	Many incorrect pitches and rhythms	Few incorrect pitches and rhythms	Very few incorrect pitches and rhythms	Correct pitches and rhythms
Revised	Pitches and rhythms are frequently incorrect and there is little evidence of proper phrasing and articulation	Pitches and rhythms are frequently correct but there is little evidence of proper phrasing and articulation	Pitches and rhythms are consistently correct and there is some evidence of proper phrasing and articulation	Pitches and rhythms are consistently correct and there is consistent evidence of proper phrasing and articulation	Pitches and rhythms are consistently correct and there is evidence of flawless phrasing and articulation

Figure 14. Comparison of original and revised descriptors for accuracy

Descriptors. In addition to the feedback collected from the judges and students about the performance criteria that I selected for the research-based rubric, these groups also had feedback about the individual descriptors for some of those performance criteria. Judges' and students' comments as well as references to the literature and a re-examination of the experts' comments influenced me to make slight adjustments to the descriptors of some of the categories including breath, tone, and diction. These changes and the rationale for making them are explained in this section.

Breath. One judge's observation that the "breath category seemed to be mostly concerned with the quiet inhale" and suggested that I include "things like how low the inhale goes, or something related to appoggio, core muscle engagement or connection, etc." was appropriate and aligned with the literature, especially Miller (1996) who advocated appoggio breathing as the standard for proper singing technique. Therefore I revised the descriptors for breath to reflect these concepts. These changes are illustrated in Figure 15.

	Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
	1	2	3	4	5
Original	Breath is consistently shallow or constricted and/or is rarely present supporting the tone.	Breath is sometimes shallow or constricted and/or is sometimes present supporting the tone.	Emerging ability to demonstrate silent inhalation that is free from tension with a steady stream of air that is fairly consistently present supporting the tone.	Approaching a consistently silent inhalation that is free from tension with a steady stream of air that is consistently present supporting the tone.	Consistently silent inhalation that is free from tension with a steady stream of air that is consistently present supporting the tone.
Revised	Breath is consistently noisy, shallow, and/or constricted	Breath is sometimes noisy, shallow, and/or constricted	Breath is sometimes deep and silent and maintains the proper balance ("appoggio") between the inhalation and exhalation mechanisms	Breath is frequently deep and silent and maintains the proper balance ("appoggio") between the inhalation and exhalation mechanisms	Breath is consistently deep and silent and maintains the proper balance ("appoggio") between the inhalation and exhalation mechanisms

Figure 15. Comparison of original and revised descriptors for breath

Tone. One judge's suggestion that "ability to sustain" and another judge's comment that "legato" should be included in the descriptors for tone were also consistent with the literature, particularly Ware (2008) who wrote about the ability to sustain consistent tone throughout an entire phrase. Other judges suggested including the terms "chiaroscuro," "nasality," "open throat," and "soft palate position" within the descriptors for tone. All of these terms were present in the literature as well. Chiaroscuro was specifically discussed by Patton (2006) and Ware (2008) as an important characteristic of desirable tone. The open throat was mentioned by McKinney (1994), Miller (1996), Vennard (1949), and Ware (2008) when they discussed tone. However, the term "free" that is used in the descriptors for this category adequately addresses this concept so I

made no changes to accommodate that specific term. Miller (2004) specifically discussed nasality as an undesirable quality, and recommended the raised soft palate as a solution for correcting this quality. Based on these findings, I revised the descriptors for tone as illustrated in Figure 16.

	Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
	1	2	3	4	5
Original	Consistently lacks clarity, maturity, freedom, richness, ring and/or resonance	Somewhat lacking in clarity, maturity, freedom, richness, ring and/or resonance	Occasionally clear, mature, free, rich, ringing, and resonant	Frequently clear, mature, free, rich, ringing, and resonant	Consistently clear, mature, free, rich, ringing, and resonant
Revised	Tone is consistently lacking in clarity, maturity, freedom, richness, ring, and/or resonance. Possibly nasal at times.	Tone is somewhat lacking in clarity, maturity, freedom, richness, ring and/or resonance. Possibly nasal at times.	Tone is occasionally clear, mature, free, rich, ringing, and legato with an occasional balance of light and dark (chiaroscuro) and/or little nasality	Tone is frequently clear, mature, free, rich, ringing, resonant, and legato with an emerging balance of light and dark (chiaroscuro) and freedom from nasality.	Tone is consistently clear, mature, free, rich, ringing, resonant, and legato with a proper balance of light and dark (chiaroscuro) and freedom from nasality.

Figure 16. Comparison of original and revised descriptors for tone

Diction. Judges also wanted more “specifics of the language” to be included in the descriptors and more emphasis on consonants. This finding is in keeping with both McKinney (1996) and Ware (1996) who found consonants to be important as the impetus for beautiful vowels, and therefore, beautiful tone. Young singers typically sing in four languages: English, French, Italian, and German. It would be impossible to characterize all of the unique and subtle differences of the pronunciation of each language within one rubric; therefore, it was necessary to speak of both vowels and consonants in general

terms. Since the original rubric placed primary focus solely on vowels in the levels one and two descriptors, I revised the descriptors so that judges would consider consonant formation in addition to vowel formation even in the most beginning singers. The revisions to the descriptors for diction that place focus on proper consonant formation within earlier levels of the singers' expected development are illustrated in Figure 17.

	Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
	1	2	3	4	5
Original	Inconsistent vowel formation	Consistent vowel formation	Emerging balance of consonants and vowels	Approaching consistent and accurate diction in all languages	Consistent and accurate diction in all languages
Revised	Vowel and/or consonant formation is inconsistent	Vowel and/or consonant formation is consistent	Balance of consonants and vowels is emerging	Diction in language(s) demonstrated is frequently consistent and accurate	Diction in language(s) demonstrated is consistent and accurate

Figure 17. Comparison of original and revised descriptors for diction

Future Research

Although this study served to answer many questions, it also raised other new questions that should be investigated through future research. Because feedback from the judges and students suggested combining certain performance criteria categories, it would be prudent to recalculate the rubric scores with the specified categories combined. The observation by one of the judges that the scores came out “quite low” and the concerns from other judges and some students about how this scale would translate into a numerical score would suggest that there might be value in determining if there is a way to develop a weighted scale in which some of the categories are weighted more heavily than others. I have discussed one possibility of a sliding scale that should be tested and verified, but other possibilities should be explored and tested.

There was some disagreement between the literature, the experts, and the students about the importance of including/excluding the visual components of a performance. Although I found convincing evidence that scores are more accurate when the visual component is not present (Wapnick & Eckholm, 1997), I would suggest that a future study could investigate if the visual components do indeed change the judge's scores. This could be achieved by implementing the revised rubric in a live setting, perhaps as a complement to the current scoring method to validate the recommended changes to the descriptors, the sliding scale for determining numerical scores, and its application in a live performance setting.

Additionally, two outright errors in the rubric developed should be addressed before an attempt to replicate or build upon this study. First, in the descriptors for diction there is a statement about "consistent and accurate diction in all languages" when the students in the study were clearly only singing in one language. In an actual jury setting where students sing a variety of repertoire in several languages, this would be an appropriate statement; however for this study, that particular statement should be removed from the rubric. Similarly, the descriptors for the expression criterion included a statement about the singer's ability to use "appropriate gestures and facial expressions." Again, while this is an appropriate element to evaluate in a live performance, it was not possible to do as part of a study that provided only audio recordings of the singers' performances. This statement should have been omitted from the rubric as well. Other minor grammatical revisions were made to some of the descriptors to ensure parallel construction and to enable ease of use.

Both judges and students agreed that the song choice was not well suited to some of the singers' voice types. In addition, many students suggested that the amount of time to prepare relative to the difficulty of the selection resulted in lower than usual scores. Therefore, I would recommend selecting a less complex song and allowing more time for preparation in future studies.

I felt that the diversity of the judging panel was a negative factor in this study. They provided feedback that indicated that they expected the beginning students to be at a different level. Some of the judges were from much larger, more selective, and much more mature programs than the one from which the participants were selected. I believe that future studies would benefit from a more homogeneous group of judges who are used to working with students in programs with similar size and scope as the research institution.

Summary

The purpose of this study is to develop and validate a research-based rubric with which to assess undergraduate solo vocal performances and which will enhance the feedback provided to students that they can use to improve future performances. The mixed-methods and participatory action research study included an extensive review of the literature on vocal performance technique and pedagogy, interviews with expert teachers of singing, scoring of recorded student performances by judges who were university level teachers of singing, and collection of feedback from the student performers about the value of the rubric feedback. Results of the study conclusively determined that within this context, the rubric was statistically reliable and the students were able to receive valuable feedback that validated their own self-perceptions and

allowed them to understand what goals they were expected to meet, where they were in relation to those goals, and what they needed to do to fill the gap.

References

- Abeles, H. (1973). Development and validation of clarinet performance adjudication scale. *Journal of Research in Music Education, 126*, 246-255.
- Asmus, E. (1999). Music assessment concepts. *Music Educators Journal, 86*(2), 19-23.
- Austin, S. (2011). Pedagogy from the archives. *Journal of Singing, 67*(3), 343-346.
- Azzara, C.D. (1993). Audiation-based improvisation techniques and elementary instrumental students' music achievement. *Journal of Research in Music Education, 41*(4), 328-342.
- 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. *Dissertation Abstracts International*.
- Bergee, M. J. (1988). The 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 performance. *Dialogue in Instrumental Music Education, 13*(2), 65-81.
- Bergee, M. J. (1993). A comparison of faculty, peer, and self-evaluation of applied brass jury performances. *Journal of Research in Music Education, 41*, 19-27.
- Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education, 51*(2), 137-150.
- Bergee, M.J., & Platt, M.C. (2003). Influence of selected variables on solo and small-ensemble festival ratings. *Journal of Research in Music Education, 51*(4), 342-353.

- Bergee, M. J. (2007). Performer, rater, occasion, and sequence as source of variability in music performance assessment. *Journal of Research In Music Education*, 55(4), 344-358.
- Bluman, A. G. (2010). *Elementary statistics: A step by step approach, a brief version, fifth edition*. New York: McGraw Hill.
- Butt, D. S., & Fiske, D. W. (1968). Comparison of strategies in developing scales for dominance. *Psychological Bulletin*, 70(6), 505-519.
- Ciorba, C., & Smith, N. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *Journal of Research In Music Education*, 57(1), 5-15.
- Cooksey, J. M. (1974). *An application of the facet-factorial approach to scale construction in the development of a rating scale for high school choral music performance*. (Doctoral Dissertation). University of Illinois at Urbana-Champaign.
- Cooksey, J. M. (1977). A facet-factorial approach to rating high school choral music performance. *Journal of Research in Music Education*, 25, 100-114.
- 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). University of Iowa.
- Dunbar, L. (2011). Performance assessment of the masses in 30 seconds or less. *General Music Today*, 25(2), 31-35.
- Fiske, H. E. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education*, 23, 186-196.

- Fiske, H. E. (1983). Judging musical performances: Method or madness? *Update: The Applications of Research in Music Education*, 1(3), 7-10.
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. New York: McGraw Hill.
- Frey, B. B., & Schmitt, V. L. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academics*, 18(3), 402-423.
- Goolsby, T. W. (1999). Assessment in instrumental music. *Music Educators Journal*, 86, 31-35, 50.
- Greene, T. (2012). *An application of the facet-factorial approach to scale construction in development of a rating scale for high school marching band performance* (Doctoral Dissertation). Available from ProQuest Dissertations & Theses A&I. Retrieved from <http://search.proquest.com/docview/1221391425?accountid=12104>.
- Hale, C. L., & Green, S. K. (1999). Six key principles for music assessment. *Music Educators Journal*, 95, 27-31.
- Hattie, J. (2012). *Visible learning for teachers*. New York: Routledge.
- Henschke, J. A. (1998). Historical antecedents shaping andragogy: A comparison of sources and roots. *Proceedings of the Third International Conference on Research in Comparative Andragogy* (pp. 1-12). Radovljica, Slovenia.
- Himonedes, E. (2009). Mapping a beautiful voice: theoretical considerations. *Journal of Music, Technology and Education*, 2(1), 25-54.

- Horowitz, R. A. (1994). *The development of a rating scale for jazz guitar improvisation performance* (Doctoral Dissertation). Available from Dissertation Abstracts International, 55, 3443A.
- Jones, H. (1986). *An application of the facet-factorial approach to scale construction in the development of a rating scale for high school vocal performance* (Doctoral Dissertation). University of Oklahoma.
- Kiesgen, P. (2005). "Well, vocal pedagogy is all subjective anyway, isn't it?". *Journal of Singing*, 62(1), 41-44.
- Latimer, M. E., Bergee, M. J., & Cohen, M. L. (2010). Reliability and perceived pedagogical utility of a weighted music performance assessment rubric. *Journal of Research in Music Education*, 58(2), 168-183.
- Levinowitz, A. (1985). *A criterion-related validity study of three sets of rating scales use for measuring and evaluating the instrumental achievement of first and second year clarinet students* (Doctoral Dissertation). Temple University.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 55.
- Marzano, R. J. (2007). *The art and science of teaching: A comprehensive framework for effective instruction*. Alexandria, VA: Association for Supervision and Curriculum Development.
- McKinney, J. C. (1994). *The diagnosis and correction of vocal faults*. Long Grove, IL: Waveland Press.
- Miller, R. (1996). *The structure of singing: System and art in vocal technique*. Belmont, CA: Wadsworth Group/Thomson Learning.

- Miller, R. (2000). Teaching singing in the new milleneum. *American Music Teacher*, 49(6), 42-43.
- Miller, R. (2004). *Solutions for singers: Tools for performers and teachers*. New York, NY: Oxford University Press.
- Naseth, A. (2011). Constructing the voice: Present and future consideration of vocal pedagogy. *Choral Journal*, 53(2), 39-49.
- National Association for Schools of Music. (2011-2012). *NASM competencies summary*. Retrieved March 1, 2012, from National Association for School of Music Web Site: http://nasm.arts-accredit.org/site/docs/Handbook/NASM_HANDBOOK_2011-12.pdf
- Norris, C. E., & Borst, J. D. (2007). An examination of the reliabilities of two choral festival adjudication forms. *Journal of Research in Music Education*, 55(3), 237-251.
- Paton, J. G. (2006). *Foundations in singing: A guidebook to vocal technique and song interpretation*. New York, NY: McGraw-Hill.
- Saunders, T. C., & Holohan, J. M. (1997). Criteria-specific rating scales in the evaluation of high school instrumental performance. *Journal of Research in Music Education*, 45(2), 259-272.
- Seashore, C. (1938). *Psychology of music*. New York, NY: McGraw-Hill.
- Sell, K. (2005). *The disciplines of vocal pedagogy: Towards an holistic approach*. Burlington, VT: Ashgate Publishing Company.

- 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.
- Stark, J. (1999). *Bel canto: A history of vocal pedagogy*. Toronto, Buffalo: University of Toronto Press.
- Undergraduate Course Catalog. (2013-2014). *Catalogs and schedules*. Retrieved from Lindenwood University Website: <http://www.lindenwood.edu/academics/catalog/catalogs/2013-14UGCatalog.pdf>
- Vennard, W. (1949). *Singing: The mechanism and the technic*. Los Angeles: Edwards Brothers.
- Wapnick, J., & Eckholm, E. (1997). Expert consensus in solo voice performance evaluation. *Journal of Voice, 11*(4), 429-436.
- Ware, C. (2008). *Adventures in singing: A process for exploring, discovering and developing vocal potential, 4/E*. New York, NY: McGraw-Hill.
- Wesolowski, B. C. (2012). Understanding and developing rubrics for music performance assessment. *Music Educators Journal, 98*(3), 36-42.
- Wiggins, G., & McTighe, J. (1997). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education, 50*(3), 245-255.

Appendix A: Expert Interview Questions

1. How would you categorize levels of achievement in the development of a singer? (Example: Expert, advanced, intermediate, beginner)

2. When judging performance by vocal students at the undergraduate level, what are the key characteristics for which you are listening?

3. For each characteristic, how would you describe what you would expect to hear from an expert singer? From an advanced singer? From an intermediate singer? From a beginner?

4. If values were to be assigned to each level for the purpose of grading, what would be your recommendation?

5. Interviewer should address characteristics not mentioned by the expert, but present in the literature:
 - a. Breathing/Breath Support
 - b. Tone Production
 - c. Intonation
 - d. Diction
 - e. Stylistic Attributes/Expression
 - f. Accuracy

Appendix B: Research-Based Rubric

		Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
		1	2	3	4	5
Mechanics	Breath	Breath is consistently shallow or constricted and/or is rarely present supporting the tone.	Breath is sometimes shallow or constricted and/or is sometimes present supporting the tone.	Emerging ability to demonstrate silent inhalation that is free from tension with a steady stream of air that is fairly consistently present supporting the tone.	Approaching a consistently silent inhalation that is free from tension with a steady stream of air that is consistently present supporting the tone.	Consistently silent inhalation that is free from tension with a steady stream of air that is consistently present supporting the tone.
	Tone	Consistently lacks clarity, maturity, freedom, richness, ring and/or resonance	Somewhat lacking in clarity, maturity, freedom, richness, ring and/or resonance	Occasionally clear, mature, free, rich, ringing, and resonant	Frequently clear, mature, free, rich, ringing, and resonant	Consistently clear, mature, free, rich, ringing, and resonant
	Accuracy	Consistently incorrect pitches and rhythms	Many incorrect pitches and rhythms	Few incorrect pitches and rhythms	Very few incorrect pitches and rhythms	Correct pitches and rhythms
	Diction	Inconsistent vowel formation	Consistent vowel formation	Emerging balance of consonants and vowels	Approaching consistent and accurate diction in all languages	Consistent and accurate diction in all languages
Coordination	Intonation	Consistently out of tune	Many out of tune pitches	Few out of tune pitches	Very few out of tune pitches	Consistently accurate on all pitches
	Vibrato	Consistently having too fast/too slow speed and/or an irregular pattern	Somewhat having too fast/too slow speed and/or an irregular pattern	Occasionally having too fast/too slow speed and/or an irregular pattern	Frequently having moderate speed and regular pattern	Consistently having a regular pattern even in pick up notes and melismatic passages
	Registration	Ability to sing in only one register	Emerging ability to sing in multiple registers	Ability to sing in multiple registers but with inconsistent tone	Consistent tone quality across all registers	Consistent tone quality across all registers including appropriate vowel modifications
	Agility	Inability to negotiate musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	Emerging ability to negotiate <u>some</u> musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	Ability to negotiate <u>some</u> musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	Emerging ability to negotiate musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	Skillful ability to negotiate musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly
Understanding	Style	No evidence of stylistic understanding	Emerging ability to apply a few basic elements of style appropriate to the piece	Ability to apply basic elements of style appropriate to the piece	Emerging ability to employ more sophisticated elements of style appropriate to the piece	Skillfully employs stylistic practices appropriate to the piece
	Expression	No evidence of internal understanding of the text	Emerging ability to communicate internal understanding of the text	Ability to communicate internal understanding of the text using <u>some</u> of the following: dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions	Emerging ability to communicate internal understanding of the text using <u>all of</u> the following: dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions	Skillful ability to communicate internal understanding of the text using <u>all of</u> the following: dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions

Holistic Score for this performance (0%-100%) _____

Appendix C: Musical Selection

Appendix D: Phonetic Transcription of Musical Text

Figure 10 Ridente la calma (ridente la kalma)

ridente	la	kalma	nel_lalma	si	desti,	
Ridente	la	calma	nell' alma	si	desti,	
<i>Smiling</i>	<i>the</i>	<i>calm</i>	<i>in the soul</i>	<i>itself</i>	<i>awaken,</i>	
ne	resti	un	seɲo	di	zdeɲo_e	timor.
ne	resti	un	segno	di	sdegno e	timor.
<i>nor</i>	<i>let remain</i>	<i>a</i>	<i>trace</i>	<i>of</i>	<i>anger and</i>	<i>fear.</i>
tu	vjeni	frat:tantɔ	a	strindzer	mi:ɔ	bene
tu	vieni	frattanto	a	stringer,	mio	bene,
<i>You</i>	<i>come</i>	<i>meanwhile</i>	<i>to</i>	<i>tighten,</i>	<i>my</i>	<i>beloved,</i>
le	doltʃe	katene	si	grate_al	mi:ɔ	kɔr
le	dolce	catene	si	grate al	mio	cor.
<i>the</i>	<i>sweet</i>	<i>chains</i>	<i>so</i>	<i>welcome to</i>	<i>my</i>	<i>heart.</i>

(Ware, 2008, p. 227)

Appendix E: ANOVA Data for Judges' Scoring

Judge	Count	Sum	Average	Variance
1	19	38.200	2.010	0.567
2	20	41.444	2.072	0.739
3	20	36.111	1.805	0.617
4	20	46.700	2.335	0.706
5	20	43.755	2.187	0.992
6	20	44.844	2.242	0.977
7	20	42.144	2.107	0.759
8	20	43.400	2.170	0.878
9	20	38.700	1.935	1.145
10	20	36.200	1.810	0.715
11	20	35.411	1.770	0.685
12	20	42.300	2.115	0.564
13	20	42.500	2.125	0.530
14	20	34.300	1.715	0.631
15	18	33.877	1.882	1.037
16	16	30.486	1.905	0.605
17	20	32.400	1.620	0.532
18	20	38.800	1.940	0.516
19	20	45.433	2.271	0.632
20	20	50.277	2.513	0.649
21	20	40.000	2.000	0.891
22	20	50.133	2.506	1.181
23	20	25.972	1.298	0.217
24	20	44.688	2.234	0.781
25	20	49.000	2.450	0.242
26	20	37.977	1.898	0.868
27	20	35.600	1.780	0.672
28	20	28.400	1.420	0.273
29	20	41.500	2.075	0.705
30	20	37.300	1.865	0.527
31	20	33.900	1.695	0.417
32	20	53.400	2.670	0.731
33	20	37.744	1.887	0.703
34	20	45.900	2.295	0.755
35	20	31.300	1.565	0.457
36	20	27.525	1.376	0.450

Appendix F: ANOVA Data for Individual Criteria

Groups	Count	Sum	Average	Variance
Breath	20	35.8719	1.793	0.587
Tone	20	36.5333	1.826	0.725
Accuracy	20	49.3676	2.468	0.754
Diction	20	40.4051	2.020	0.607
Intonation	20	47.6345	2.381	0.600
Vibrato	20	37.1448	1.857	0.633
Registration	20	41.6987	2.084	0.485
Agility	20	40.2077	2.010	0.450
Style	20	34.9134	1.745	0.503
Expression	20	30.5028	1.525	0.298

Appendix G: Student Interview Questions

After reviewing the feedback provided by the rubric as used by expert scorers:

1. Describe what you liked or did not like about this method of assessment.
2. Describe what you think the judges heard or did not hear in your performance.
3. Describe your understanding of the strengths and weaknesses of your performance.
4. Describe what (if anything) you plan to do with the information. What actions (if any) do you plan to take?
5. Do you have any other comments about the rubric or have anything else you would like to share?

Appendix H: Revised Rubric

		Beginning	Early Intermediate	Intermediate	Early Advanced	Advanced
		1	2	3	4	5
Mechanics	Breath	Breath is consistently noisy, shallow, and/or constricted	Breath is sometimes noisy, shallow, and/or constricted	Breath is sometimes deep and silent and maintains the proper balance ("appoggio") between the inhalation and exhalation mechanisms	Breath is frequently deep and silent and maintains the proper balance ("appoggio") between the inhalation and exhalation mechanisms	Breath is consistently deep and silent and maintains the proper balance ("appoggio") between the inhalation and exhalation mechanisms
	Tone	Tone is consistently lacking in clarity, maturity, freedom, richness, ring, and/or resonance. Possibly nasal at times.	Tone is somewhat lacking in clarity, maturity, freedom, richness, ring and/or resonance. Possibly nasal at times.	Tone is occasionally clear, mature, free, rich, ringing, and legato with an occasional balance of light and dark (chiaroscuro) and/or little nasality	Tone is frequently clear, mature, free, rich, ringing, resonant, and legato with an emerging balance of light and dark (chiaroscuro) and freedom from nasality.	Tone is consistently clear, mature, free, rich, ringing, resonant, and legato with a proper balance of light and dark (chiaroscuro) and freedom from nasality.
	Accuracy	Pitches and rhythms are frequently incorrect and there is little evidence of proper phrasing and articulation	Pitches and rhythms are frequently correct but there is little evidence of proper phrasing and articulation	Pitches and rhythms are consistently correct and there is some evidence of proper phrasing and articulation	Pitches and rhythms are consistently correct and there is consistent evidence of proper phrasing and articulation	Pitches and rhythms are consistently correct and there is evidence of flawless phrasing and articulation
	Diction	Vowel and/or consonant formation is inconsistent	Vowel and/or consonant formation is consistent	Balance of consonants and vowels is emerging	Diction in language(s) demonstrated is frequently consistent and accurate	Diction in language(s) demonstrated is consistent and accurate
Coordination	Intonation	Pitches are consistently out of tune	Many pitches are out of tune	Few pitches are out of tune	Very pitches few pitches are out of tune	No pitches are out of tune.
	Vibrato	Vibrato has a consistently too fast/too slow speed and/or an irregular pattern	Vibrato has a somewhat too fast/too slow speed and/or an irregular pattern	Vibrato has an occasionally too fast/too slow speed and/or an irregular pattern	Vibrato has a moderate speed and regular pattern	Vibrato has a consistently regular pattern even in pick up notes and melismatic passages
	Registration	The singer demonstrates the ability to sing in only one register	The singer demonstrates an emerging ability to sing in multiple registers	The singer demonstrates an ability to sing in multiple registers but with inconsistent tone	The singer demonstrates an ability to sing with consistent tone quality across all registers	The singer demonstrates an ability to sing with consistent tone quality across all registers including appropriate vowel modifications
	Agility	The singer demonstrates an inability to negotiate musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	The singer demonstrates an emerging ability to negotiate <u>some</u> musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	The singer demonstrates an ability to negotiate <u>some</u> musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	The singer demonstrates an emerging ability to negotiate musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly	The singer demonstrates a skillful ability to negotiate musical challenges such as wide pitch intervals, coloratura passages, and dynamic variations nimbly and quickly
Understanding	Style	The singer demonstrates no evidence of stylistic understanding	The singer demonstrates an emerging ability to apply a few basic elements of style appropriate to the piece	The singer demonstrates an ability to apply basic elements of style appropriate to the piece	The singer demonstrates an emerging ability to employ more sophisticated elements of style appropriate to the piece	The singer skillfully employs stylistic practices appropriate to the piece
	Expression	The singer demonstrates no evidence of internal understanding of the text	The singer demonstrates an emerging ability to communicate internal understanding of the text	The singer demonstrates an ability to communicate internal understanding of the text using <u>some</u> of the following: dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions	The singer demonstrates an emerging ability to communicate internal understanding of the text using <u>all</u> of the following: dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions	The singer demonstrates a skillful ability to communicate internal understanding of the text using <u>all</u> of the following: dynamics, articulation, phrasing, vocal color, and appropriate gestures and facial expressions

Vitae

Katherine Herrell earned a bachelor of arts in music from Truman State University where she studied voice with R. Paul Crabb and Jacqueline Collett. Her master of business administration studies at Maryville University were completed during her years working for Anheuser-Busch, Inc. where she was involved in project communications, training, and change management. Mrs. Herrell returned to her passion for music and for teaching and earned a master of arts in education from Lindenwood University and worked as a music specialist teaching grades pre-K through Middle School. She has been active in liturgical music for over 30 years. Mrs. Herrell started teaching at Lindenwood University in January 2008 as an adjunct instructor. She served as a part-time faculty member beginning in 2011 before joining the full-time faculty in January 2012. She teaches courses in the areas of music education, music theory, and applied voice. Mrs. Herrell is pursuing a doctor of education degree in instructional leadership from the Lindenwood University School of Education, and she anticipates completion in 2014. She lives in Chesterfield, MO with her husband Ken and two children, Ben and Tess.