Choral Performance Assessment: An Overview of Research to Date

Marvin E. Latimer Jr.
Central Michigan University, Tuscaloosa, Alabama, USA

Abstract

Reliability of adjudicators responsible for choral performance assessment has long concerned music educators. For example, Radocy (1989) argued, "any measure that involves human judgment is inherently subjective because it involves human impressions (p. 30)." He concluded music educators must recognize that all measurement procedures are inherently subjective, either in construction, application, or interpretation. To address such matters, numerous assessment forms have been employed to enhance consistency and reliability of performance adjudication. Recently, performance assessment rubrics, which contain narrative descriptions of various categories, have come into use. The purpose of this study was to longitudinally investigate the validity and reliability of one such rubric employed in the Large Group Choral Festivals in a Midwestern state in the United States. This paper/presentation will report choral adjudicator reliability and validity findings from 45 different adjudicators assessing 350 different choral performances, which occurred over two years. To that end, the following research questions will be addressed: (a) What was the level of agreement among choral adjudicators in assigning the global rating (that is, I, II, III, IV, V) when utilizing this rubric? (b) What was the level of agreement among choral adjudicators in assigning scores for individual performance categories (that is, tone, interpretation, rhythm, and so on) when utilizing this rubric? (c) What was the level of correlation between individual performance categories and global ratings, or specifically, which performance category tended to be the best indicator of the global score? And (d) what is the perceived efficacy of this performance assessment rubric among both adjudicators and directors?

Introduction

Choral music educators at every level of instruction must be prepared to make informed decisions about the content and method of music assessments and their relationship to specific music achievement objectives. According to the Music Educators National Conference (MENC), the reliability of such assessments is tied directly to how well measurements of the same skills or knowledge produce the same results (MENC, 1996). Unlike other content areas (for example, math, reading, social studies, and so forth), which generally focus on assessment of written examples of student work (usually some time after they are written), music assessments frequently involve evaluations of live performances in real time, that is to say, as they are happening and immediately after they are over. Unfortunately, performance based assessment measures have been shown by scholars to possess significant flaws, both in structural design and application.

For purposes of the present discussion, performance assessment can be viewed through two distinct lenses: validity and reliability. Validity is the extent to which an evaluation measures what it purports to measure, while levels of reliability are generally characterized by the consistency of the performance evaluators (for example, adjudicators, professors, teachers, and so forth) and the measurement tool itself (usually an adjudication or assessment form of some kind). In short, an assessment should measure—without bias—a predetermined quality or

characteristic the same way each time it is used, providing the measurement takes place under the same conditions and with the same participants. In the case of music performance assessment, a performance is usually broken into component parts, or musical dimensions (for example, tone, intonation, and so on), which often are scored, totalled, and then presented in aggregate in the form of a global assessment: a numeric score, grade, or rating.

Music assessment validity investigations primarily have focused on the extent to which extra-musical variables influence performance assessments. Reliability investigations, on the other hand, have commonly examined inter-rater reliability (reliability between multiple evaluators), intra-rater reliability (one evaluator's consistency in evaluating different hearings of the same performance), or both. The purpose of this paper is to present an overview of that scholarship, to discuss its pedagogical ramifications, and suggest possible related topics for future study. To that end, it will seek to examine: (a) the extent various performance assessment methods have been shown by previous research to be valid and reliable; (b) whether some assessment tools have demonstrated superior reliability over other assessment tools; (c) certain pedagogical ramifications to be considered from those findings; and (d) suggestions that can be made for future research. It will examine some extant research in all performance media, but will focus, where possible, specifically on research in vocal and choral performance contexts.

Roots of Choral Music Festival Adjudication

In 1906, Peter C. Lutkin founded the first university a cappella choir in the United States. Shortly thereafter, F. Melius Christiansen established the St. Olaf Choir (1912) and John Finley Williamson organized the Westminster Choir (1920). Choruses of all kinds soon became a staple in choral music programs in high schools, colleges, and universities across the United States (Van Camp, 1964). By 1926, choral singing contests existed in at least 12 states. Most of the performance assessment protocols for those contests included a method of scoring performance dimensions and converting the scores to a final global score or rating (Best, 1926).

By the middle of the twentieth century, solo and ensemble music contests for bands, orchestras, and choirs had become an important part of music programs. These "high stakes" evaluative music festivals often were associated—and still are for that matter—with the annual student activities at district, league, state, and, at times, national levels. As adjudicated music contests evolved and became more widespread, participation in them became a driving force in school music programs' perceived levels of success. Several researchers have investigated the extent of those perceptions among parents, administrators, and music program directors (Burnsed & Sochinski, 1983; Rogers, 1983). Their studies included both surveys and panel discussions and represented a cross-section of parents, administrators, and music teachers nationwide. The findings of these scholars suggested that a majority of music program patrons and constituents considered evaluative festivals to be closely associated with the level of success of their particular music programs.

In a professional environment where performance assessment, often by an unknown adjudicator or team of adjudicators, can profoundly influence the success of both choral programs and choral educators alike, fairness naturally becomes a major source of concern. That such concerns are well founded appears to be supported by several evaluative music contest validity investigations, conducted in a variety of adjudicated performance settings (for example, Bergee & Platt, 2003; Bergee & McWhirter, 2005; Elliott, 1995/1996; Ryan & Costa-Giomi, 2004; Wapnick, Mazza, & Darrow, 1998). This line of research produced findings that suggested that extra-musical influences such as gender, ethnicity, performance attire, physical appearance,

conducting skill, director prestige, and time of day could significantly influence music contest ratings. This research also found the adjudicator selection process to have a significant effect on the final rating that was awarded.

Music Assessment

A salient challenge inherent in the measurement of a music performance is the subjective nature of the measurement tool itself. Performance assessment forms, not surprisingly, evolved early in the twentieth century as music contests become more widespread. In 1925, Giddings (1925) suggested a system for vocal performance adjudication that included the following weighted dimensions: (a) intonation, 30 points; (b) beauty of tone, 20 points; (c) balance of parts, 15 points; (d) phrasing, 15 points; (e) enunciation, 10 points; (f) and expression, 10 points. He further recommended that there should be six adjudicators for all vocal ensemble contests. Notably, Giddings advocated a method whereby each adjudicator assessed a different single dimension. He argued that this method allowed adjudicators to serve as a check and balance to each other.

Interestingly, music contest assessment forms, like those described by Giddings (1925), have changed little in nearly a century of use. They tend to list a set of musical dimensions (often down the left side of the form) with very little description of how those dimensions are to be assessed. A common thread among most of them is that they are to be utilized in such a way as to generate one single global score that ostensibly becomes the adjudicators' final assessment of the quality of a particular performance. At this time, the Music Educators National Conference lists 21 recommended forms, each designed for a specific kind of solo or ensemble performance (MENC, 2009).

Many questions exist, however, that are directly related to how these diverse assessment forms are applied in authentic music performance assessment settings. For example, are some dimensions more important than others? Do the dimensions overlap, or can they be considered as discrete, sonic components of the whole? Should the adjudicator take into account such things as size of the group, size of the school, age level of the choristers, or other factors? Arguably, nearly all of these global rating scales provide only a broad categorization of the musical performance. Hence, each adjudicator must utilize his or her own personal criteria to determine the importance and nature of each performance dimension, which may or may not directly relate to the final score.

Traditional Music Festival Adjudication Forms

A growing body of scholarship has purposefully focused on the reliability of traditional forms and adjudicators' use of those forms in a wide range of music performance settings. It includes examinations of such high stakes performances as adjudicated music festivals, university juries, solo and ensemble contests, and festival group auditions, to name only a few. Generally, that research has shown that evaluations of musical presentations—both solo and group—invariably demonstrate limited reliability and validity. Indeed, Fiske (1983), a noted scholar in this area, argued that even highly practiced adjudicators rarely demonstrate adequate reliability. He suggested that a logical way to address such matters was the use of larger adjudicator panels (Fiske, 1978). While such practices might be preferable, and likely would improve reliability, significantly increasing the number of adjudicators for each performance is often not possible due to financial constraints and limited adjudicator pools.

Traditional forms, like those designed by the Music Educators National Conference (MENC) and the National Interscholastic Music Activities Commission (NIMAC), have received attention in numerous reliability investigations (for example, Fiske, 1975, 1977; Burnsed, Hinkle, & King, 1985). Researchers have consistently reported that such forms demonstrate relatively high inter-adjudicator reliability for global scores (final ratings), but generally low inter-adjudicator reliability for individual dimension scores. The finding that dimension scores and final ratings tended to be so closely aligned that they essentially reflected the global performance rating was of particular interest in these studies. Indeed, researchers posited that adjudicators appeared to first evaluate the performance, decide on a final rating, and then fill in the dimension scores to suit their initial impression. Such results are likely attributable, at least in part, to disagreement among adjudicators in how to use the form, especially as it relates to the relative importance of various performance dimensions.

Quality of Adjudicators

As referenced earlier, unlike other content areas, music performance assessments require evaluators to assess live musical events, often with very little indication of what level of proficiency to expect of the given performer(s). A likely consequence of this circumstance is that these assessments tend to possess an unavoidable degree of subjectivity (Radocy, 1989). Efforts to address this matter have yielded mixed results. For example, several researchers have investigated whether adjudicator reliability could be improved through training (for example, Ekholm, 1997; Fiske, 1978; Heath, 1976). Though some reported slightly improved inter-adjudicator reliability levels, their findings, overall, were inconclusive.

The practice of seeking out experienced adjudicators for evaluative music festivals appears to be widespread. Indeed, many state music festival governing agencies such as state MENC organizations, state American Choral Directors Associations, state activities associations, or state vocal associations require an adjudicator application procedure, whereby the most skilled and experienced specialists can be selected for various events. Some evidence suggests that these practices might be warranted. For example, in a study that disaggregated participants by age and musical experience, Towers (1980) reported that both factors significantly increased adjudicator reliability.

Other research, however, has suggested that application and screening methods may do nothing more than unnecessarily limit the available pool of adjudicators. For instance, Fiske (1975) reported findings that suggested that any competent musician could rate performances and produce reliability levels similar to specialists in a specific performance medium. Fiske's conclusions appear to be in line with the bulk of scholarship in this area. Generally those findings, combined with the result that global scores tend to be the most reliable figure in traditional assessment forms, support the notion that it likely is possible to explain final assessments without referencing specific musical techniques, and in a way that can be understood by both specialists and non-specialists alike (Mills, 1987).

All State Choral Group Selection

Like evaluative music performance contests, festival choirs, too, have become an integral part of school music activities in nearly every state. However, the method of adjudicating choristers for such groups remains relatively diverse nationwide (Wine, 1996). Audition components can include assessment of a wide range of musical skills including (a) sight singing,

(b) solo singing, (c) singing in a quartet, (d) director recommendations, and (e) music theory tests. Adjudicators for festival choir auditions can come from the ranks of (a) high school choir directors, (b) university and college choir directors, (c) retired high school choir directors, (d) private voice teachers, (e) church musicians, and (f) doctoral students. Though all audition protocols likely have fairness as a core goal, surveys and questionnaires of participants and directors generally have revealed widespread concerns about audition procedure validity and adjudicator reliability.

An interesting finding related to festival choir selection procedures came from a recent questionnaire of choral directors representing a cross section of choral programs in the United States (Barkey, 2005). That survey found the most often used audition component for state festival choirs to be solo singing. But recent research suggests that such practices, though long considered to be tried and true, may not be the most informative method for ensemble selection. A growing body of evidence, for instance, suggests that data collected in exclusively solo choir auditions may neither be valid nor reliable as it relates to making decisions about choristers' ability to successfully integrate into a given choral ensemble, presumably including ad hoc ensembles such as festival choirs (Daugherty, 2001).

One noteworthy vocal festival choir audition study that relates to the present discussion was an examination of festival choir adjudicator reliability in a Midwest state in the United States (Latimer, 2007). That study examined a criterion specific vocal adjudication form—similar to the traditional adjudication forms discussed above—by comparing it to a simple, researcher designed global score form. Research participants included experienced adjudicators, music educators, and non-music educators who were asked to adjudicate recorded performances of choristers singing the state festival choir audition selection for that year. Like other investigations of traditional assessment forms, both inter-adjudicator and intra-adjudicator results from this investigation suggested that overall reliability of both the adjudication procedure and the assessment form failed to exhibit acceptable levels. Moreover, similar to other previous reliability studies, the single global score tended to be the most reliable. Moreover, no significant differences in reliability were found between participant groups. Such outcomes support earlier contentions that adjudicators' overall impressions of the performance tend to drive their final decision and that any competent listener can serve as a music performance assessor.

Facet-Factorial Rating Scales

That there is no clear cut way to improve adjudicator reliability by focusing on the adjudicators themselves seems to be a common theme in much of the research findings from investigations of traditional forms. Other studies have focused more directly on the development of improved forms, often specifically designed for individual performance media (that is, trumpet, voice, and so forth). In a number of these investigations, researchers have succeeded in creating forms that appear to offer improved reliability and validity when compared with traditional adjudication forms.

Numerous studies, for instance, have used a factor analysis (facet-factorial) method to develop music performance rating scales for a wide range of vocal and instrumental assessment contexts. (for example, Abeles, 1973; Bergee, 1987; Bergee, 1988; Bergee, 1989; Horowitz, 1994; Jones, 1986; Nichols, 1991; Smith, 2007; Zdzinski & Barnes, 2002). The facet-factorial method can be described as follows: (a) a set of dimensions describing characteristics of a music performance in a given medium is analyzed to determine its underlying factor structure; (b)

certain factors then are selected and organized into subscales; (c) they then are paired with Likert-type response scales; and (d) are used to assess various performances (Bergee, 2003).

Though the bulk of this line of research has focused on instrumental media, two notable studies applied the facet-factorial method to the development of rating scales for vocal solos and choirs. Cooksey (1977), for example, constructed a rating scale to evaluate high school choral performances and Jones (1986) developed a rating scale to assess vocal solos. Both of these investigations determined validity by comparing the newly devised facet-factorial scale scores with assessments using MENC and NIMAC adjudication forms. The application of the facet-factorial method proved to be an improved means to assess both vocal and choral performances.

Criteria Specific Rating Scales

Some scholars have suggested that while facet-factorial rating scales produce reliability levels that are consistently better than traditional assessment forms, they also tend to possess some of the negative features of those traditional forms. Specifically, they fail to provide thorough and accurate indications of what causes a particular performance to be either successful or unsuccessful (Saunders and Holahan, 1997). But a prevalent conceptual stance among current music education philosophers holds that the primary role of music assessment is to provide feedback to students about the quality of their musicianship at various levels of development (for example, Elliott, 2005). Presumably, this premise relates to music performance situations as well as music classroom environments.

Researchers have attempted to address such matters by developing criteria-specific rating scales, which are designed to offer more information about a particular music performance (for example, Azzara, 1993; Barnicle, 1993; Levinowitz, 1989; Rutkowski, 1990). These scales represented a significant point of departure from the facet-factorial method (which tended to use Likert-type scales as a feedback tool) by adding written descriptions of specific levels of proficiency for each musical dimension. Though these scales did not consistently demonstrate significantly improved reliability when compared to other forms, researchers suggested that they likely offered superior diagnostic potential; hence, they provided a better teaching tool than previously constructed non-descriptive forms.

Arguably, the facet-factorial and criteria-specific methods of evaluative scale construction provided a bridge between traditional adjudication forms such as the MENC and NIMAC forms and more descriptive rubric forms, which are currently gaining popularity nationally. This transition is perhaps reflective of a general shift in thinking that is directed away from a stance centered primarily on performance assessment as evaluation, and toward a methodology more directly focused on providing learners, or in this specific case performers, with more concise and practical information.

Performance Assessment Rubrics

Recently, the role of assessment in education generally has been the subject of increased scrutiny. Emphases on more performance-based, learner-centered systems have engendered a more task-oriented approach to the design and application of various assessment materials. Such trends likely are based on the principle that learning in all settings should more purposefully reflect learners' specific needs. In music, as in other related disciplines such as physical education, theatre, and art, performance-based tasks require performance-based

assessments. In other words, student performances should be assessed in a way that reflects appropriate levels of relevant task achievement. Assessment rubrics widely have been shown to provide robust means for assessments of this kind (see *Why Rubrics*, 2007).

A rubric is a scoring tool that delineates specific expectations for individual tasks by dividing them into component parts or dimensions. These parts can then be paired with descriptions of any number of acceptable levels of proficiency. Rubrics appeal broadly to teachers and learners alike because they provide vigorous means for both teaching and assessment. Rubrics can improve and monitor student performance by making teachers' expectations clear and by showing students how to better meet those expectations. The result often is marked improvement in the quality of student work and in learning (Stevens & Levi, 2005).

A good deal of discussion and research has focused on the construction of rubrics specifically suited to music performance applications. The key elements of a music performance rubric are the descriptors for what a performance is like within a range of possible performance levels. Unlike traditional non-rubric adjudication forms, rubrics provide those who have been assessed with clearly stated information about how well they performed in each dimension and what they need to accomplish in each dimension to improve the overall performance (Whitcomb, 1999). Important to this discussion, rubrics provide adjudicators, directors, and performers with the characteristics, both necessary and sufficient, for each level of performance in each performance dimension (Asmus, 1999). Notably, researchers have suggested that the number of descriptors included for each dimension can be directly related to the overall reliability of the form (Gordon, 2002).

Recently, such performance assessment rubrics have become popular in various music performance settings such as state ensemble and solo contests, university and college juries, and various applied grading processes. These actions likely are due, at least in part, to numerous policy initiatives by accrediting bodies that require assessment tools to provide evidence of student achievement in a wide range of music skills. Though much has been written (most of it positive) about the pedagogical utility of such assessment techniques, little research has addressed whether they are, in fact, more reliable than the methods discussed above, or whether performers and adjudicators merely perceive them to be an improved music assessment tool. That important research, however, appears to be in its early stages.

One such study, for example, examined grading procedures in a collegiate applied studio setting to establish whether the use of a performance assessment rubric increased or decreased overall satisfaction among students and faculty (Parkes, 2006). It measured satisfaction in three specific subscales: (a) jury process satisfaction, (b) preparedness, and (c) continuous assessment satisfaction. Faculty and students were randomly assigned to a control group that did not use rubrics and an experimental group that did use rubrics. They were given an attitude test at the beginning and end of the semester. Parkes found no statistically significant differences in faculty or student attitudes toward grading after the use of rubric assessment tools.

Another more recent investigation examined the reliability of a multi-dimensional rubric form in undergraduate performance juries (Ciorba & Smith, 2009). Notably, the researcher-designed rubric was used for all performance media (for example, voice, clarinet, piano, and so on). Inter-rater reliability for adjudicators across all dimensions was moderate to high. Perhaps most noteworthy, these researchers found the rubric scores to be significantly related to the students' year in school. Such findings support the notion that rubrics possibly can provide a useful means of tracking student achievement from one term to another.

In another study, more closely related to choral performance contexts, Norris and Borst (2007) compared the reliability of a traditional festival adjudication form, similar to the MENC and NIMAC forms, to a rubric style extension of that same form. They used the same panel of adjudicators to score duplicate choral performances. These researchers reported better reliability in almost every dimension for the more descriptive rubric form. Norris and Borst concluded that the rubric, when compared to the traditional form, offered the adjudicators more guidance on how to score the dimensions. They noted, however, that some dimensions (that is, rhythm and other) tended to be scored less reliably than others.

Finally, a recent study investigated a multi-dimensional weighted performance assessment rubric used in large group festivals in a Midwest state in the United States (Latimer, Bergee, & Cohen, 2009). The researchers analyzed completed rubrics, collected over a period of two years, for internal consistency. They also analyzed adjudicator and director questionnaires collected over that same time period to determine the rubrics' perceived level of pedagogical utility. These researchers reported findings that suggested the reliability of the rubric on the whole was not appreciably better or worse than previously researched music assessment forms. They also found that the most reliable figure was the single global score. The questionnaire results suggested the rubric provided an improved means for justifying ratings and more detailed and accurate descriptions of what constituted acceptable music performances. These researchers concluded while the rubric did not improve reliability, both adjudicators and directors perceived the rubric as having improved pedagogical utility.

Conclusion

In sum, research to date in the domain of music performance assessment suggests that while levels of adjudicator reliability and validity can be improved to some extent, further attention in both areas appears to be warranted. Also, while some efforts have been made to improve various assessment forms as diagnostic tools, more research is needed to assess whether using a more descriptive assessment instrument can improve the overall reliability of the form. Finally, that a music performance cannot successfully be divided into component parts for the purpose of music assessment appears to be a salient thread throughout the bulk of music performance assessment reliability and validity research. Given the probability that performance assessment will likely be a significant part of music education for years to come, it is in the interest of music educators in all performance media and at all levels of instruction to continue to pursue assessment means that show improved reliability, validity, and pedagogical utility.

References

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

Asmus, E. (1999). *Rubrics: Definitions, benefits, history, and type.* Retrieved March 12, 2009 from, http://www.music.miami.edu/assessment/rubricsdef.html

Azzara, C. D. (1993). The effect of audiation-based improvisation techniques on musical achievement of elementary music students. *Journal of Research in Music Education, 41*, 328-342.

Barkey, D. L. (2005). The relationship between choral all-state audition components and selected teaching methodologies (Doctoral Dissertation, Texas Women's University). *Masters Abstracts International, 43,* 383.

Barnicle, S. P. (1993). Evaluating the choral performer. *CMEA News, 44*(2), 24-26.

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

Bergee, M. J. (1989). An investigation of the efficacy of using an objectively constructed rating scale for the evaluation of university-level single reed juries. *Missouri Journal of Research in Music Education, 26,* 74-91.

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. (1987). An investigation of the facet-factorial approach to scale construction in the development of a rating scale for euphonium and tuba music performance (Doctoral Dissertation, University of Kansas). *Dissertation Abstracts International*, *49*, 1086.

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

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,* 342-353.

Best, F. C. (1927). State choral contests. *Music Supervisors' Journal 13*(3), 9-11.

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

Burnsed, V. & Sochinski, J. (1983). Research on competitions. *Music Educators Journal, 70*(1), 25-27.

Ciorba, C. R., & Smith, N. Y. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *Journal of Research in Music Education, 57,* 5-15.

Cooksey, J. M. (1977). A facet-factorial approach to rating high school choral music performance. *Journal of Research in Music Education*, *25,* 100-114.

Daugherty, J. F. (2001). Rethinking how voices work in a choral ensemble. *Choral Journal, 92*(5), 69-75.

Ekholm, E. (1997). Expert consensus in solo voice performance evaluation. *Journal of Voice, 11*(4), 429-436.

Elliott, C. A. (1995/1996). Race and gender as factors in judgements of musical performance. *Bulletin of the Council of Research in Music Education, 127,* 50-55.

Elliott, D. J. (2005). *Praxial education: Reflections and dialogues*. London, UK: Oxford University Press.

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

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

Fiske Jr., H. E. (1977). Relationship of selected factors in trumpet performance adjudication reliability. *Journal of Research in Music Education, 25*, 256-263.

Fiske Jr., H. E. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education, 23,* 186-196.

Giddings, T. P. (1925). Contests: Some pertinent points on the conduct of instrumental and vocal contests. *Music Supervisors Journal, 12*(1), 46-53.

Gordon, E. (2002). *Rating scales and their uses for evaluating achievement in music performance.* Chicago, IL: GIA Publications.

Heath, C. E. (1976). The effect of instruction on the consistency of ratings in the adjudication of trumpet solo excerpts (Doctoral Dissertation, Indiana University, Bloomington). *Dissertation Abstracts International, 37,* 2707.

Horowitz, R. A. (1994). The development of a rating scale for jazz guitar improvisation performance (Doctoral Dissertation, Columbia University Teachers College). *Dissertation Abstracts International*, *11*, 3443.

Jones Jr., H. (1986). An application of a facet-factorial approach to scale construction in the development of a rating scale for high school solo vocal performance (Doctoral Dissertation, University of Oklahoma). *Dissertation Abstracts International, 47,* 1230.

Latimer Jr., M.E. (2007). Adjudicator reliability: A comparison of the use of state festival choir and global score audition forms. *Contributions to Music Education 34*(2), 67-82.

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

Levinowitz, L. M. (1989). An investigation of preschool children's comparative capability to sing songs with and without words. *Bulletin of the Council for Research in Music Education, 100,* 14-19.

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

Music Educators National Conference (2009). *Other solo and ensemble audition sheets.* Retrieved August 3, 2009, from http://www.menc.org/resources/view/other-solo-and-ensemble-adjudication-sheets/

Music Educators National Conference (1996). *Performance standards for music.* Reston, VA: MENC.

Nichols, J. P. (1991). A factor analysis approach to the development of a rating scale for snare drum performance (Doctoral Dissertation, University of Iowa). *Dissertation Abstracts International, 46*, 3282.

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,* 237-251.

Parkes, K. A. (2006). The effect of performance rubrics on college-level applied studio grading (Doctoral Dissertation, University of Miami). *Dissertation Abstracts International, 68*, 8.

Radocy, R. E. (1989). Evaluating student achievement. *Music Educators Journal*, *76*(4), 30-33.

Rogers, G. L. (1983). *Attitudes of high school band directors, band members, parents, and principles toward marching band contests.* Paper presented at an in-service conference sponsored by the Band Committee of the Southern Division of MENC, Louisville, KY.

Rutkowski, J. (1990). The measurement and evaluation of children's singing voice development. *The Quarterly Journal of Music Teaching and Learning, 1*(1 & 2), 81-95.

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

Saunders, T. C. (1990). A preliminary investigation of the suitability of selected rating scales used to measure student music performing skills. *Missouri Journal of Research in Music Education, 27,* 15-29.

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

Smith, B. S. (2007). Development and validation of an orchestra performance rating scale. *Journal of Research in Music Education, 55,* 268-280.

Stevens, D. D. & Levi, A. (2005). *Introduction to rubrics: An assessment tool to save time, convey effective feedback, and promote student learning*. Sterling, VA: Styus.

Towers, R. (1980). *Age group differences in judge reliability of solo voice performances* (Unpublished Masters Thesis). University of Western Ontario, London, ON.

Van Camp, Leonard (1964). The development and present status of a cappella singing in the United States colleges and universities (Doctoral Dissertation, University of Missouri at Kansas City). *Dissertation Abstracts International 28*(2), 716.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (1998). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education, 45*, 47- 479.

Whitcomb, R. (1999). Writing rubrics for the music classroom. *Music Educators Journal, 85*(6), 26-32.

Why Rubrics. Retrieved August 16, 2009 from, http://www.teach-nology.com/Articles/teaching/ rubrics/

Wine, T. R. (1996). All state choruses: A survey of practices, procedures, and perceptions. *Choral Journal, 36*(8), 21-27.

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