Fall 2007

An Examination of the Reliabilities of Two Choral Festival Adjudication Forms

Charles E. Norris

Grand Valley State University, norrisc@gvsu.edu

James D. Borst

East Kentwood High School, james.borst@kentwoodps.org

Follow this and additional works at: https://scholarworks.gvsu.edu/mus_articles

Part of the [Music Commons](#)

ScholarWorks Citation

Norris, Charles E. and Borst, James D., "An Examination of the Reliabilities of Two Choral Festival Adjudication Forms" (2007). *Peer Reviewed Articles*. 4.

https://scholarworks.gvsu.edu/mus_articles/4

This Article is brought to you for free and open access by the Music & Dance Department at ScholarWorks@GVSU. It has been accepted for inclusion in Peer Reviewed Articles by an authorized administrator of ScholarWorks@GVSU. For more information, please contact scholarworks@gvsu.edu.

The purpose of this study was to compare the reliability of a common school choral festival adjudication form with that of a second form that is a more descriptive extension of the first. Specific research questions compare the interrater reliabilities of each form, the differences in mean scores of all dimensions between the forms, and the concurrent validity of the forms. Analysis of correlations between all possible pairs of four judges determined that the interrater reliability of the second form was stronger than that of the traditional form. Moderate correlations between the two forms further support the notion that the two forms measured the dimensions in somewhat different ways, suggesting the second form offered more specific direction in the evaluation of the choral performances. The authors suggest continued development of language and descriptors within a rubric that might result in increased levels of interrater reliability and validity in each dimension.

Charles E. Norris

Grand Valley State University, Allendale, Michigan

James D. Borst

East Kentwood High School, Kentwood, Michigan

An Examination of the Reliabilities of Two Choral Festival Adjudication Forms

In the United States, curriculum development programs sanctioned by state and federal departments of education include specific standards or benchmarks that define learning outcomes for teachers and students. Likewise, music education curriculum includes specific standards to be achieved by students (MENC, 1994). These goals and objectives for teaching and learning remain the foundation on which educators base the delivery of curricula. Moreover, standards define the intent of teaching and learning and provide the impetus

Charles E. Norris is an associate professor of music and coordinator of music education at Grand Valley State University, 1320 Performing Arts Center, 1 Campus Drive, Allendale, MI 49401; e-mail: norrisc@gvsu.edu. James D. Borst is director of choral activities at East Kentwood High School, 6230 Kalamazoo NE, Kentwood, MI 49508; e-mail: james.borst@kentwoodps.org. Copyright © 2007 by MENC: The National Association for Music Education.

for measuring and evaluating student achievement. Specifically, the exit goals of Standard 1 of the National Standards in Music Education ("singing, alone and with others, a varied repertoire of music") at the high school proficient level are stated as (p. 18):

1. Sing with expression and technical accuracy a large and varied repertoire of vocal literature with a level of difficulty of 4, on a scale of 1 to 6, including some songs performed from memory.
2. Sing music written in four parts, with and without accompaniment; demonstrate well-developed ensemble skills.

Because the implementation of the standards is recent in the field of music education, assessments may seem nebulous and speculative among music educators. Consideration of the appropriate manner in which music education standards are measured is important because without guidelines for measuring these completed standards, music education learning outcomes lose their intent and purpose. Reliable and valid evaluation of music students' achievements provides teachers, students, parents and administrators with diagnostic feedback that helps to assess the extent to which someone is musically educated.

Assessing the performance ensemble creates challenges unlike those in general education disciplines and other music classes because of the corporate nature of performing with others. Furthermore, because of the elusive and esoteric nature of aesthetics, measures of performance outcomes can be questionable. Despite these challenges, music education literature suggests that a detailed rating scale, or rubric, may be the best means of assessing performance (Asmus, 1999; Gordon, 2002; Radocy & Boyle, 1989; Whitcomb, 1999).

A rubric provides guidance to music educators about how to accomplish and assess learning standards in performance (Whitcomb, 1999). This is done with the use of criteria that describe several achievement levels for each aspect of a given task or performance. With criteria that describe the component parts of given tasks and performance, music educators are bound to specificity, not conjecture. Both music teachers and their students tend to prefer this type of specificity over more global evaluation (Rader, 1993).

The key elements of a rubric are its dimensions and the descriptors. The dimension is a musical performance outcome to be assessed, while the descriptors serve to define the range of achievement levels within the dimension (Asmus, 1999). Within this range, specific criteria are used to rank performance from the lowest level to the highest level of achievement.

Statistics warrant the use of more than one dimension in any rubric. The reliability of one dimension improves when combined with others, while combining two or more dimensions guarantees more reliability on the composite score. Moreover, the more descriptors included in a dimension of a rubric (up to five), the more reliable it will become (Gordon, 2002). It seems that a balance of dimensions with an optimal number of criteria for each dimen-

sion is most desirable when developing rating scales.

Historically, there has been steady interest in the reliability of musical performance evaluation. Fiske's (1975) study of the ratings of high school trumpet performances revealed similar reliabilities among judging panels of brass specialists, nonbrass specialists, wind specialists, and nonwind specialists. Having examined relationships between the traits of technique, intonation, interpretation, rhythm, and overall impression, Fiske noted that technique had the lowest correlations to the other traits. He concluded that it would be more practical and time-efficient if performances were rated from an overall perspective, since most traits were highly correlated with the overall impression.

A number of studies attempted to create and establish reliable and valid solo performance rating scales, the majority of which focused on criteria-specific scales. With these devices, judges indicate a level of agreement with a set of statements regarding a variety of musical performance dimensions. Jones (1986) developed the Vocal Performance Rating Scale (VPRS) to evaluate five major factors of solo vocal performance: interpretation/musical effect, tone/musicianship, technique, suitability/ensemble, and diction. Interrater reliability estimates of judges' levels of agreement to 32 specific statements yielded a strong correlation for total score (.89) but various strengths of agreement, from weak to strong, for the other aforementioned dimensions.

The development and study of Bergee's (1988) Euphonium-Tuba Rating Scale (ETRS) for collegiate players necessitated reliability checks of 27 specific statements regarding low brass performance. Using Kendall's Coefficient of Concordance to analyze judges' degrees of agreement with these statements revealed strong reliability in the four major dimensions (W values: interpretation/musical = .91, tone quality/intonation = .81, technique = .75, rhythm/tempo = .86) and the total score (W = .92). Bergee's (1989) follow-up study of interrater reliability for clarinet performance resulted in less impressive but significant W values in five of six factors and total score: interpretation = .80, tone = .79, rhythm = .67, intonation = .88, articulation = .70, and total score = .86. Analysis of the factor, tempo, yielded a W value of .38, much lower than the other five dimensions.

Bergee (1993) continued study of criteria-specific rating scales with the Brass Performance Rating Scale, adapted from his earlier ETRS. Analysis of ratings of college applied juries showed significant average Pearson correlations within and between groups of both faculty and collegiate student judges for overall ratings (.83-.96). Among the four dimensions, strong correlations between and within faculty and peer groups were observed for interpretation/musical effect (.80-.94), tone quality/intonation (.83-.95), and technique (.74-.97). Rhythm correlations were lower and less consistent (.13 to .81).

Bergee (2003) extended his research of interrater reliability in college juries to brass, voice, percussion, woodwind, piano, and strings.

In this study, raters again used criteria-specific scales unique to each of the instrument families, containing broad dimensional and subdimensional statements to which the jurors responded using a Likert scale, using the number "1" for strong disagreement and the number "5" for strong agreement. Significant correlations, albeit varying from moderate to moderately strong, were noted in nearly all subscales for all instruments (.38-.90), total scores for all instruments (.71-.93), and jury grades for all instruments (.65-.90).

In an earlier study, Bergee (1997) deviated from criteria-specific scales, instead using MENC (1958) solo adjudication forms, almost identical to those commonly used at solo/ensemble festivals nationwide. Instead of assigning numeric ratings of 1-5 in each of five broad dimensions (tone, intonation, interpretation, articulation, and diction), judges were asked to use a scale from 1-100. Correlations between judges' scores of voice, percussion, woodwind, brass, and string college juries varied greatly in the individual categories as well as in the total scores, ranging from .23-.93.

Cooksey (1977) applied the principles of criteria-specific rating scales in developing an assessment device for choral performance with which he obtained strong interrater reliabilities (.90-.98) with both overall choral performance ratings and within traditional categories such as tone, intonation, and rhythm. Similar to aforementioned studies, judges used a 5-point Likert scale to indicate their agreement with 37 statements about high school choral performances. Although these reliability estimates are superficially impressive, Cooksey's reliability estimates are statistically unsurprising, as 20 judges' scores were used in his analyses.

Most closely related to the current investigation, two studies examined the reliability of traditional large ensemble adjudication forms that use 5-point rating scales (using descriptors from excellent to unacceptable) to evaluate various musical dimensions (i.e., tone, intonation, technique, etc.) and to provide a total score and a final rating. Having browsed state music education association Web sites and affiliate activities associations, the authors of the current study found that 17 of the 19 states that publish their adjudication forms are currently using this "traditional" judging format. Burnsed, Hinkle, and King (1985) studied this traditional form by examining agreement among band judges' ratings at four different festivals. Using a simple repeated-measures analysis of variance (ANOVA), the authors found no significant differences among judges' final ratings at any of the festivals; however, significant differences were noted in various dimension scores at each festival. The dimension of tone was rated differently at three festivals, and intonation was rated differently at two festivals, while balance and musical effect were rated differently at one festival each.

Garman, Barry, and DeCarbo (1991) continued study of the traditional form in examining judges' scores at orchestral festivals in five different years. The authors found correlations in dimensions ratings to vary from as low as .27 to as high as .83, while overall rating corre-

lations ranged from .54 to .89. The authors of this study strongly advocated an examination of the descriptors that appear under each category heading so that meaning might be similar for all adjudicators.

Other studies, whose primary foci were not necessarily the reliability of rating scales, have demonstrated strong interrater reliabilities (correlation coefficients of .90 and above) when evaluating specific musical capabilities. In each of these studies the authors used rubrics that contained descriptors to designate specific levels of achievement for one or more musical dimensions. The scores of Azzara's (1993) four raters resulted in strong interrater reliability with strong correlations (ranging from .90 to .96) in the areas of tonal, rhythmic, and expressive improvisational skills. Levinowitz (1989) noted strong interrater reliability between two judges of children's abilities to accurately perform rhythmic and tonal elements in songs with and without words (.76–.96). Norris's (2000) three adjudicators achieved high agreement ($r = .98$) when using a descriptive 5-point scale to measure accuracy in sung tonal memory.

The above review of literature demonstrates that music education research has adequately explored interrater reliability of criteria-specific rating scales. In these studies, researchers analyzed adjudicators' levels of agreement with numerous statements about solo and group musical performances. Although these investigations examined descriptive statements, they were not able to evaluate descriptions of specific standards of achievement. Additional research studied interrater reliabilities with the use of traditional large-group festival adjudication forms, which, like the criteria-specific scales, lack the same description of specific achievement standards. The third component of the related literature focused on studies that used scales with specific descriptors for levels of achievement for one or more musical dimensions. The latter group of studies provided inspiration for the present study.

Each spring, students from more than 40 states of the United States are adjudicated in choral festivals (Norris, 2004), with judges most commonly using the traditional adjudication form. Although the music education profession purports the use of rating scales as beneficial for measuring performances, there is no research that explores the interrater reliability of an instrument that has a balanced combination of dimensions with descriptors—those beyond the vague “excellent, good, satisfactory, poor, and unsatisfactory.”

With the intent of improving assessment in music education, the purpose of this study is to compare the reliability of a traditional festival adjudication form that is commonly used to assess performance of school choirs with that of a second tool that is an extension of the first. The second tool, a rubric, contains both the dimensions found in the common form and descriptors that define the various achievement levels within each of the dimensions. Specific research questions are:

1. Do the mean scores of specific musical dimensions, total scores, and overall ratings differ between the two forms?

2. How do the reliabilities of these forms compare in regard to specific musical dimensions, total score, and overall rating?
3. Do the forms appear to be measuring the same constructs and therefore provide concurrent validity for one another?

METHOD

A format was designed in which a panel of four highly regarded choral music educators, as evidenced in their ensemble achievements and election as conductors of state honors choirs, evaluated two performances of the same choirs with the traditional and more detailed adjudication forms: the traditional (Form A, Figure 1) for a morning session and the detailed rubric (Form B, Figure 2) for an afternoon session. The order in which the forms were used—Form A followed by Form B—was purposely determined based on the perception that using Form B first, replete with achievement-level descriptors for each dimension, would bias the ratings assigned with Form A that was devoid of descriptors. Furthermore, because of its common use, evaluations made with Form A afforded the most logical benchmark from which comparisons with any other type of rating form could be made.

Both forms contained the dimensions of tone, diction, blend, intonation, rhythm, balance, and interpretation. The category “other factors” (items such as facial expression, appearance, choice of literature, and poise) was not used because music teachers and adjudicators have identified “other factors” as least important in performance evaluation (Stutheit, 1994). Moreover, research suggests that “other factors” weakly correlates with other musical dimensions as well as total scores and overall ratings (Burnsed, Hinkle, & King, 1985; Garman, Barry, & DeCarbo, 1991). As is customary, both forms had places for total score and rating as well as a scale to determine the overall rating.

The language of Form B was based on that from a rubric used at one time in Washington State (Brummond, 1986). This rubric was piloted as a possible replacement for the traditional Form A in Michigan, but due to adjudicators’ objections that there were too many dimensions and achievement levels (10 for each), its use was not continued. The language and dimensions from the Washington rubric were adapted to the dimensions of Form A, resulting in the creation of Form B.

Because numerous investigations (Wapnick, Mazza, & Darrow, 1998; Wapnick, Mazza, & Darrow, 2000; Ryan & Costa-Giomi, 2004) indicate that appearance and behavior may bias evaluation of musical performance, the adjudicators were presented with audiorecordings from an actual Michigan School Vocal Music Association high school district choral festival. The recordings were engineered with two Audio Tech 4040 individual microphones, one Royers SF24 stereo microphone, and a Mackie Onyx 1620 Series mixing board. From the engineer’s master CD, the

CHOIR ADJUDICATION FORM

Group Name _____

Selection _____

Adjudicator _____

Rate each category on a five-point scale.

1- Excellent 2-Good 3-Satisfactory 4- Poor 5-Unsatisfactory

Category	1	2	3	4	5	Comments
Tone Quality						
Diction						
Blend						
Intonation						
Rhythm						
Balance						
Interpretation						

TOTAL _____ RATING _____

I = 7 - 10 II = 11 - 17 III = 18 - 24 IV = 25 - 31 V = 32 - 35

Figure 1. Form A: A Traditional Choir Adjudication Form.

CHORAL FESTIVAL RUBRIC

Adjudicator

Selection

Group

Category	1	2	3	4	5
Tone Quality	A pleasing, free vibrant tone in character with the age of the singers and the sections. Proper support, placement and focus resulting in the ultimate in resonance without being forced or pushed.	A well-supported, energized tone with good placement and focus; lapses in resonance infrequent; usually occurring in extreme ranges or difficult rhythmic passages.	A satisfactory tone quality with adequate support, vibrancy and resonance present most of the time. Lack of breath energy and/or tension occasionally creates a forced and/or breathy, and/or spread tone	Tone lacks consistent support and focus; tension and a forced tone quality and/or breathy, and/or spread tone is present most of the time.	Tone is devoid of proper support, space and focus resulting in a non-energized and/or very forced unacceptable sound.
Diction	Properly formed vowels and consonants in all registers result in a clear linear tone line in all sections with very clear diction.	Good vowels and tone line with slight inconsistencies in extended registers; some consonants inaudible.	Basic vowels and consonants usually formed correctly. Some problems with diphthongs and some inconsistencies in vowels; most words are easily discernible.	Basic vowels are usually incorrect; diphthongs lack definition and precision. Words are difficult to discern.	Basic vowels incorrect, diphthongs and tone line non-existent, words cannot be identified.
Blend	Individuals reach the ultimate in tone production, while remaining in character with the other singers; sections display a cohesive tone, while matching the tone of the other sections.	There is a strong consistency of tone by individuals and sections; adjustments are made rapidly to correct problems.	An adequate blend, most of the time. Individuals are occasionally heard due to a different tone, unique vibrato and/or volume.	Individual voices are heard much of the time. Sections have some similarities of tone.	Singers are "on their own" resulting in an unacceptable tone quality. The individual sounds range from inaudible to over singing.
Intonation	Superb clarity of pitch in melodic and harmonic intervals in all registers; necessary adjustments are made instantly.	Pitch is correct; good intonation, few problems in extreme ranges and/or difficult sections; adjustments are made rapidly.	Pitch is correct most of the time; occasional problems with intonation, wide interval relationships and/or range extremes; adjustments are made.	Frequent pitch problems result in conflicts with the accompaniment and/or inappropriate modulations.	Tonal center is lacking; problems with melodic and harmonic intonation occur throughout.
Rhythm	Exemplary consistency of the pulse; rhythmic precision is nearly flawless.	Good control of the pulse, minor inconsistencies are corrected immediately. Minor flaws in rhythmic precision.	Rushing and/or dragging affects the rhythmic precision and/or inaccurate rhythms occur.	Rhythmic precision is lacking and/or basic rhythms are incorrect and/or the pulse is inconsistent.	Inconsistent pulse and rhythmic inaccuracies present throughout; corrections seldom made.
Balance	Appropriate balance between sections, during varying dynamics, textures, ranges and accompaniments render a perpetual sense of ensemble.	Proper balance predominates; minor lapses occur during extended registers and/or extreme dynamic levels; adjustments are made.	Adequate balance, most of the time; problems occasionally occur in extended ranges, at changing dynamic levels and textures.	Balance problems frequently occur; some adjustments are made; the sense of ensemble is usually lacking.	Balance problems throughout; adjustments seldom made; very little feeling of ensemble.
Interpretation	A thorough display of an artistically shaped musical line containing an accurate use of dynamic levels based on music markings; rise and fall of the line and word stress perpetuating a feeling of forward motion.	A good demonstration of forward motion and dynamic contrast. Dynamic ranges usually achieved and feeling of forward motion present most of the time.	Adequate volume variation with a fairly consistent musical line; occasional inaccurate dynamic levels are present.	The range of dynamics is very limited and/or too much contrast is used. The phrasing is very mechanical; many words are performed separately.	An inaccurate use of dynamic ranges; most words are performed separately; no feeling of form and motion.

TOTAL RATING I = 7 - 10 II = 11 - 17 III = 18 - 24 IV = 25 - 31 V = 32 - 35

researchers chose one selection from each of 15 randomly selected SATB choirs and copied them onto two Maxell CDRs with Roxio 6.0 software in two different random orders. All performances were presented through a stereo sound system consisting of a Mackie 1201-VL preamplifier, Tascam CD-150 compact disc player, and two Event 20-20 speakers.

Before the start of the morning session, the adjudicators were presented with packets containing scores, pencils, and copies of Form A. In the absence of specific descriptors, the adjudicators were told to rate the dimensions for each performance according to the standards they would use in a live judging situation. Following the morning session, the adjudicators took a 1.5-hour lunch break. To minimize any memory about the morning performances, conversation during the break was deliberately steered away (by the authors) from both discussion of the performances that had just been heard and what was to happen in the afternoon session. Unbeknownst to the panel, they would be adjudicating the same choirs in the afternoon session with a different form.

At the beginning of the afternoon session, the adjudicators received packets with new scores and copies of Form B. Following a 3-minute "study period," the panel was instructed to judge as closely as possible to the language found in the rubric. The adjudicators were also informed that the performances that they were about to evaluate were the same as those they heard in the morning, but would be presented in a randomly different order. In both sessions, there was a 2-minute period of silence after each performance; researchers collected forms after every performance.

Means and standard deviations for each of the dimensions on both forms were calculated. To determine any differences in means between forms, paired-sample *t*-tests were computed for each dimension. Prior research studies have examined interrater reliability using statistics with certain limitations. For instance, while Pearson correlation averages do provide information about consistency, they do not indicate actual agreement among judges. Kendall's Coefficient of Concordance certainly provides information about interrater agreement, but is used only with ordinal or rank-order data. Because the judges' scores of the current study were perceived as interval data (the distances between 1 and 2, 2 and 3, etc., were considered equal or intervallic), interrater reliability estimates in the current study were derived from an intraclass correlation coefficient (ICC), which not only indicates consistency, but also accounts for actual agreements of ratings among judges (McGraw & Wong, 1996). Because three-judge panels are the norm at high school choral festivals, the ICCs were computed using all four judges' scores as well as each of the four possible combinations of three judges. Intraclass correlation was also used to examine the concurrent validity of the two rubrics' dimension scores, total scores, and overall ratings.

Table 1
Means and Standard Deviations of All Dimensions for Adjudication Forms A and B

Dimension	Form A		Form B	
	Mean	Std. Dev.	Mean	Std. Dev.
Tone	2.98	1.07	3.20	1.00
Diction	2.68	1.12	2.95	1.05
Blend	2.75	1.10	3.13	1.07
Intonation	3.00	1.11	3.28	1.12
Rhythm	2.05	1.04	2.40	.98
Balance	2.70	1.12	3.15	1.18
Interpretation	2.88	1.18	3.10	1.20
Total Score	18.95	6.89	21.22	6.72
Rating	2.70	.94	3.00	1.06

RESULTS

Mean scores and standard deviations for Forms A and B are reported in Table 1. Each dimension on Form B was rated lower than its counterpart on Form A (since the number "1" is considered the "best" score, lower scores or ratings are indicated by higher numbers). Paired-samples *t*-tests revealed significant differences between forms at the .05 level or lower in the following dimensions: tone ($t = -2.27, p = .027$), diction ($t = -2.40, p = .02$), blend ($t = -3.36, p = .001$), intonation ($t = -2.34, p = .023$), rhythm ($t = -2.80, p = .007$), balance ($t = -4.09, p < .001$), total score ($t = -3.94, p < .001$), and rating ($t = .323, p = .002$). The computed *t* value for the dimension of interpretation was statistically nonsignificant ($t = -1.79, p = .079$).

The questions pertaining to whether one of the two forms exhibits stronger reliability than the other can be answered by examining the ICCs for the various combinations of judges for all dimensions on both Form A and Form B (Table 2). The weakest ICCs overall occurred in the dimension of rhythm. Examining each dimension by form, the ICCs on Form B are stronger in every comparison except in the dimension of rhythm. In the 45 comparisons found in Table 2, interrater reliability on Form B was .10 or higher than interrater reliability of Form A in 34 instances. Agreement on Form B was .15 or higher than that on Form A in 24 instances.

An examination of concurrent validity of the two forms with regards to each dimension, total score, and rating reveal significant ICCs (all at $p < .001$), albeit varying in strength: tone (.73), diction (.67), blend (.63), intonation (.63), rhythm (.51), balance (.68), interpretation (.68), total score (.75), and rating (.71).

Table 2

Intraclass Correlation Coefficients (ICCs) for All Dimensions by Groupings of Judges and Judging Form

Dimension	Form	Judge Groupings				
		123	234	124	134	1234
Tone Quality	A	.69	.53	.48	.54	.57
	B	.76	.71	.65	.64	.69
Diction	A	.48	.52	.39	.39	.49
	B	.63	.62	.73	.70	.71
Blend	A	.54	.48	.45	.40	.48
	B	.71	.71	.72	.72	.72
Intonation	A	.48	.49	.42	.53	.50
	B	.70	.52	.72	.67	.60
Rhythm	A	.60	.45	.42	.34	.44
	B	.52	.39	.36	.46	.43
Balance	A	.50	.49	.31	.38	.41
	B	.70	.72	.62	.71	.69
Interpretation	A	.60	.58	.51	.51	.55
	B	.71	.68	.56	.59	.62
Total Score	A	.72	.62	.52	.54	.62
	B	.80	.74	.79	.69	.77
Rating	A	.67	.58	.45	.476	.54
	B	.75	.69	.75	.66	.70

Note. Coefficients were computed using a two-way random effects model (absolute agreement). All coefficients $p < .001$.

DISCUSSION

A cursory review of Table 2 reveals that the ICCs for both Forms A and B fall within or slightly below the overall ranges of Pearson averages and Kendall's coefficients reported in other investigations (Garman, Barry, & DeCarbo, 1991; Bergee, 1988, 1989, 1993, 1997, 2003). In line with these studies, analysis of agreement among judges on both Forms A and B resulted in stronger ICCs and therefore better reliability for total score and overall ratings, but weaker ICCs and lower reliabilities for individual dimensions. Since four judges can hardly be representative of an entire population of judges, it would be erroneous to draw any other conclusion except that these judges agree more or less like judges in prior studies.

The differences in ICCs between Forms A and B can be viewed with considerable interest. Because the ICCs between judges with Form B were in almost every instance stronger than those of their

Form A counterparts, the authors suggest that Form B provided a clearer delineation of what constituted a particular achievement level in a given dimension. Analysis of the dimension of rhythm revealed the only contradiction to the apparent superiority of Form B, resulting in not only unimpressive ICCs on both forms, but also little difference in ICC strengths between forms. Clearly, neither form provided more favorable clarity about the various levels of achievement with regard to rhythm. While these observations do not encourage us to make vast generalizations about judging reliability, they do urge music educators to consider devising performance evaluation forms similar to Form B.

The additional analysis of the means of dimensions, total score, and overall ratings corroborates the above comments (see Table 1 and above *t*-test results). Form B yielded significantly different ratings in every dimension except interpretation, suggesting that the adjudicators in this setting rated the choirs more severely when using Form B. This finding contradicts previous research (Bergee & Platt, 2003; Bergee & Westfall, 2005) in which ratings grew higher or better as the judging day progressed. The authors suggest that the more specific descriptors of Form B more clearly defined what constituted a given rating in a particular dimension, resulting in not only the noted differences, but also in the stricter ratings. This reflects the fact that the judges were bound to specificity, thereby more closely aligning their individual standards. It is also worthy of note that Form A, the form with which most music educators are familiar, yielded atypically low scores in all dimensions, total score, and overall rating. The mean overall rating of Form A was 2.7, a number that would likely never occur in a "real world" festival situation where a preponderance of Division I and Division II ratings are awarded (Bergee & Platt, 2003; Bergee & Westfall, 2005).

Cognizant of this distribution of ratings, the subjects commented that they felt as though they were rating the dimensions in too strict a manner not only on Form B, but also on Form A. While the judges concurred that they did their best to comply with the words "excellent, good, satisfactory, poor, and unsatisfactory" as they evaluated each of the dimensions on Form A, they shared the perception that their ratings would have been inflated (better) in a live festival situation in which directors would be able to review the judging forms. The candor of the subjects, coupled with the typical distribution of ratings at choral festivals, implies that festival scores often do not reflect the descriptors assigned to the actual ratings.

A final consideration was to determine if the two forms could establish concurrent validity for each other. The ICCs reported in the results section indicate that the forms are not always measuring the dimensions in the same way. The moderate (albeit significant) ICCs may have resulted due to the differences in specificity provided in the two forms. Consistent with the discussion thus far, we suggest that Form B offers more guidance in how to score the various dimensions of choral performance. While both Form A and Form B contain the

necessary dimensions to adequately assess choral performance, Form B supplies choral music educators with additional comprehensive information, enabling a more thorough conclusion than what transpires during actual performances.

The design of the current study facilitated comparison of the reliability of ratings of the same performances with two forms in a predetermined order. The researchers intentionally used Form A first because it was not only that with which the adjudicators were familiar, but also that which the researchers believed would exert minimal influence in how ratings were assigned with Form B. Despite both the rationale for the present study's design and the fact that ratings were lower with the afternoon session's Form B (contradicting previous research findings), the possible effects of the morning session on the afternoon session warrant consideration. It is unknown whether or how much of what judges may have recalled from the morning session influenced afternoon results. Future study using a counterbalanced condition order (two panels of judges, one judging with Form A first, the other judging with Form B first) would allow investigation of not only this type of order effect, but also the effects of time of day.

Additionally, the dimensions' overlapping relationships must be considered. For example, a singer's diction and tone quality are arguably interdependent to some degree. To what degree is a rating in diction somehow connected to a rating in tone quality? Investigations are needed that provide analysis of testing tools that contain a variety of sentence content and syntax as descriptors. A study that simultaneously explores two or more such adjudication forms would prove valuable for this much needed strand of research. While the current study addresses issues of reliability, future work, such as has been just briefly described, must also address the validity of each dimension.

Taking into consideration the above discussion of possible confounding variables, the authors suggest that rubrics containing dimension-specific descriptors could be better suited for the purposes of evaluating performances than instruments containing scant language (words such as excellent, fair, unsatisfactory, etc.) as the descriptors, whereby the adjudicators assign evaluative numbers based on their individual standards. The results of this study support the favorable opinions regarding the use of dimensions and descriptors in rubrics (Asmus, 1999; Gordon, 2002; Whitcomb, 1999). The goal of all assessment research should focus on the development of reliable and valid tools that are specific enough to provide diagnostic feedback for conductors and performers, yet global enough to allow for artistic expression.

REFERENCES

- Asmus, E. (1999). *Rubrics: Definitions, benefits, history and type*. Retrieved July 25, 2005, from the University of Miami Frost School of Music Web Site: <http://www.music.miami.edu/assessment/rubricsdef.html>

- Azzara, C. D. (1993). The effect of audiation-based improvisation techniques on the music achievement of elementary music students. *Journal of Research in Music Education*, 41, 328–342.
- Bergee, M. J. (1988). Use of an objectively constructed rating scale for the evaluation of brass juries: A criterion related study. *Missouri Journal of Research in Music Education*, 5 (5), 6–15.
- Bergee, M. J. (1989). An investigation into 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*, 6, 74–91.
- 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. (1997). Relationships among faculty, peer, and self-evaluations of applied performances. *Journal of Research in Music Education*, 45, 601–612.
- Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education*, 51, 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, 342–353.
- Bergee, M. J., & Westfall, C. R. (2005). Stability of a model explaining selected extramusical influences on solo and small-ensemble festival ratings. *Journal of Research in Music Education*, 53, 358–374.
- Brummond, B. (1986). *WMEA choral adjudication form*. Edmunds, WA: Washington Music Educators Association.
- Burnsed, V., Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert festivals. *Journal of Band Research*, 21 (1), 22–29.
- Cooksey, J. M. (1977). A facet-factorial model to rating high school choral music performance. *Journal of Research in Music Education*, 25, 100–114.
- Fiske, H. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education*, 23, 186–196.
- Garman, B. R., Boyle, J. D., & DeCarbo, N. J. (1991). Orchestra festival evaluations: Interjudge agreement and relationships between performance categories and final ratings. *Research Perspectives in Music Education*, 2, 19–24.
- Gordon, E. (2002). *Rating scales and their uses for measuring and evaluating achievement in music performance*. Chicago: GIA Publications.
- Jones, H. (1986). An application of the facet-factorial approach to scale construction in the development of a rating scale for high school vocal solo performance (Doctoral dissertation, University of Oklahoma, 1986). *Dissertation Abstracts International*, 47 (4A), 1230.
- 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.
- MENC. (1958). *Choral large group adjudication score sheet*. Retrieved April 23, 2005, from http://www.menc.org/information/prek12/score_sheets/chorallg.gif
- MENC. (1994). *The school music program: A new vision*. Reston, VA: Author.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intra-class correlation coefficients. *Psychological Methods*, 1 (1), 30–46.

- Norris, C. (2000). Factors related to the validity of reproduction tonal memory tests. *Journal of Research in Music Education*, 48, 52–64.
- Norris, C. E. (2004). A nationwide overview of sight singing requirements at large group choral festivals. *Journal of Research in Music Education*, 52, 16–28.
- Rader, J. L. (1993). Assessment of student performances for solo vocal performance rating instruments (Doctoral dissertation, University of Houston, 1993). *Dissertation Abstracts International*, 54 (8A), 2940.
- Ryan, C., & Costa-Giomi, E. (2004). Attractiveness bias in the evaluation of young pianists' performances. *Journal of Research in Music Education*, 52, 141–154.
- Stutheit, S. A. (1994). Adjudicators', choral directors', and choral students' hierarchies of musical elements used in the preparation and evaluation of high school choral contest performance (Doctoral dissertation, University of Missouri–Kansas City, 1994). *Dissertation Abstracts International*, 55 (11A), 3349.
- Wapnick, J., Mazza, J. K., & Darrow, A. A. (1998). Effects of performer attractiveness, stage behavior and dress on violin performance evaluation. *Journal of Research in Music Education*, 46, 510–521.
- Wapnick, J., Mazza, J. K., & Darrow, A. A. (2000). Effects of performer attractiveness, stage behavior, and dress on evaluation of children's piano performances. *Journal of Research in Music Education*, 48, 323–336.
- Whitcomb, R. (1999). Writing rubrics for the music classroom. *Music Educators Journal*, 85 (6), 26–32.

Submitted May 17, 2006; accepted April 4, 2007.