Using a prescreening rubric for all-state violin selection: Influence on performance and teaching experience.

By: Michael L. Allen, Rebecca B. MacLeod, Laurie Scott and John M. Geringer

Geringer, J. M., Allen, M. L., MacLeod, R. B., & Scott, L. (2009). Using a prescreening rubric for all-state violin selection: Influence on performance and teaching experience. *Update: Applications of Research in Music Education, 28(1), 41-46.*

Abstract:

Performance assessment is an integral part of young musicians' development. Students enrolled in music programs frequently participate in adjudication festivals and many audition for select ensembles. Outcomes of such events are deemed consequential among all concerned: students, parents, teachers, and administrators. Furthermore, the number of all-state applicants for some individual instruments in many states exceeds 100 and in some states approaches or exceeds 200. This is an inordinate number of recordings or live auditions for individuals or judging panels to evaluate. It seems paramount to utilize an efficient yet fair and accurate audition process to assess large numbers of performances in a short period of time. To explore ways of addressing this issue, the authors designed a prescreening rubric with the goal of reducing the number of full-length recordings heard by judges to a more manageable number. The authors then compared ratings of listeners who used the rubric and heard only the etude portion of the audition to scores of trained and experienced judges who heard the entire audition.

Keywords: music education | performance evaluation | string performance | string music teaching | rubrics | music adjudication | auditions | violin | violin performance | violin education

Article:

In an extensive survey published in Update, Barnes and McCashin (2002) were able to obtain responses regarding practices and procedures of all-state orchestras in all 50 states. They found that 19 states used recorded auditions and 18 states used live auditions, most of which were behind a screen. Scales were the most common component used in auditions (86% of the states), followed by sight reading (56%). Approximately equal numbers required études (20%), solo literature (24%), and/or orchestral literature (23%). Barnes and McCashin noted that open-ended responses from a majority of states (60%) indicated a desire to change some aspect of their all-state system, the most common of which was discontent with the adjudication process itself.

Because of their important role in music education, the accuracy and consistency of judging systems used in competitions and all-state auditions have been questioned (Bergee, 2003; Burnsed, Hinkle, & King, 1985; Fiske, 1975, 1983; Saunders & Holahan, 1997; Thompson & Williamon, 2003). Researchers have found, for example, that increasing the number of judges to five or seven has increased the overall reliability of the judging process (Bergee, 2003; Burnsed et al, 1985), as has the use of an Olympic-style judging system (Dugger, 1997; Perkins & Allen, 1991). Investigators have used criteria-specific rating scales in the attempt to increase interjudge reliability (Saunders & Holahan, 1997). In addition, a number of studies have constructed and tested factor analytic techniques in the development of rating scales for specific instruments, such as solo clarinet (Abeles, 1973), euphonium and tuba (Bergee, 1988), strings (Zdzinski & Barnes, 2002), voice (Jones, 1986), and snare drum (Nichols, 1991). Such an approach has extended to ensemble performance as well (Cooksey, 1977; Smith & Barnes, 2007).

Most evaluation tools that have been developed employ some type of numerical system to help the listener evaluate the performance. Fiske found in 1975, however, that overall rankings of judges were related to a significant extent to each of the subcategories. Fiske further noted that rating individual components and summing or averaging scores was unnecessary in selection-rejection situations (e.g., where specific feedback to participants was not necessary). An investigation by Wapnick, Flowers, Alegant, and Jasinskas (1993) showed that using rating scales for professional piano performances did not increase consistency in performance evaluation, nor did use of the musical score. Furthermore, an examination of ratings assigned by adjudicators in an international string competition indicated that judges were able to reliably rank recordings of the performers in a screening procedure for entry to the final round of live auditions using a global 5-anchor, 9-point scale (Smith, 2004). Thus, a global or single overall score derived from a panel appears to be an acceptable measure in situations wherein an accept or reject decision only is needed, such as some competitions and selection auditions.

Researchers have explored possible effects of judges' musical background on evaluation of performance. Results from a number of studies have shown that the amount of experience generally seems to have little or no effect on performance ratings (Bergee, 2003; Doerksen, 1999; Hewitt & Smith, 2004; Mills, 1987; Wapnick et al., 1993; Winter, 1993). For example, musicians' and nonmusicians' ratings were not significantly different when evaluating piano performances, and both groups were influenced similarly by descriptions of the performances (Duerksen, 1972). Musical ratings of experts were similar to those of evaluators who did not have subject matter expertise (Byo & Crone, 1989), and expert and nonexpert ratings were similar for error detection skills (Byo, 1993). In addition, no relationship was found between an individual's ability to perform and his or her ability to reliably adjudicate (Fiske, 1977). Others have noted that raters need not be experienced performers on a particular instrument (Bergee, 2003; Fiske, 1975; Hewitt, 2007).

Previous research regarding effects of the amount of music heard or duration of examples on performance ratings generally has found evidence of relatively small influences. Vasil (1973) studied effects of excerpt duration on ratings of high school clarinet performances. Performances varied from just over 1 min to 5 min in length. Neither ratings nor reliability of judges were different across the durations. Subsequent research has found that duration of music examples can exert some influence on listeners' ratings. In a study of high-level piano performances, listeners rated longer and faster excerpts more consistently than other combinations of tempo and duration (Wapnick et al., 2005). Recently, Geringer and Johnson (2007) found that performances of high school groups were rated lower in the longer duration excerpts as compared to shorter versions. In contrast, listeners rated longer durations of professional performances slightly higher than short performances. The additional information present in the longer examples had significant but relatively minor effects on evaluation of performances. Geringer and Johnson speculated that as long as the performance quality remains at least somewhat consistent, reasonably accurate assessments can be made quickly.

Investigations of music performance evaluation in recent years have incorporated the use of rubrics. Rubrics are intended to provide clear descriptions of graduated levels of achievement, for example, keyboard performance competencies in a music fundamentals course (Price, 2007). Rubrics have been used as measures of achievement in solo and small-ensemble festivals (Bergee & Platt, 2003) and in choral festival adjudications (Norris & Borst, 2007). In the latter study, correlations among judges were higher when they used rubrics with specific performance descriptors compared to a traditional adjudication form.

There is continued need to investigate the development of valid and reliable tools for evaluating music performance, particularly in festival and all-state auditions where large numbers of performers must be judged. The purpose of the present study is to investigate the use of a prescreening rubric that was intended to increase efficiency in the process of selecting high school violin students for an all-state orchestra. Possible influences of the performance and teaching experience of the judges were also investigated. The following questions were posed: (a) What is the relationship between rankings using the prescreening rubric and rankings based on traditional ratings by expert judges? and (b) Is there a relationship between the performance and teaching experience of the judges and their performance ratings?

Method

Traditional Procedure

Three all-state orchestras are selected each year in Florida: Grades 7 and 8, Grades 9 and 10, and Grades 11 and 12. Recordings of the individual auditions are made at multiple sites across the state, given a code to preserve anonymity, and sent to a central location for adjudication. Ratings from five judges are used for each of the individual string instruments for each of the grade levels. Independent judges (public school and university string teachers) are trained in the system and give global scores for each of the following audition components (using violin Grades 11 and 12 as an example): étude (60 points), orchestral music excerpt (60 points), three-octave major scale (30 points), three-octave minor scale (30 points), and sight reading (30 points). To determine audition placement for the all-state ensembles, the component scores are first added together per judge to determine the total raw score for each student. Individual judges' raw scores are then converted to rank-order scores. When the rank-order scores for each student have been determined, the highest and lowest rankings across the five judges are eliminated, and the scores (rankings) of the three remaining judges are summed (this is often called "Olympic scoring"). The individual with the lowest composite ranking is assigned first chair for that level and that instrument, the next lowest ranking is assigned second chair, and so on until the 36 violin section chairs have been assigned. This follows the general procedure described by Perkins and Allen (1991).

Prescreening Procedure

A prescreening procedure was initiated in 2006 to reduce the number of complete violin auditions that judges were asked to hear (more than 150 violin auditions were submitted per level). This procedure was developed to reduce judge fatigue and was based on a rubric system to eliminate the weakest performers, performances that had no realistic chance of being selected for the all-state ensemble. Judges listened to the étude-only portion of the audition recording (a section of an étude by Kreutzer), with a duration of approximately 1 min.

The prescreening rubric was defined by five levels of performance: a score of 1 indicated a superior performance likely to compete among the higher chairs, a score of 2 indicated a very good performance likely to be selected for membership in the all-state orchestra but unlikely to compete near the top, a score of 3 was defined as an average performance that although good is not likely to be selected to the all-state ensemble, a score of 4 was defined as a flawed performance unacceptable for all-state, and a score of 5 meant that the tape should not have been submitted.

Participants and Procedures in Present Study

Recorded performances of violinists in Grades 11 and 12 on the étude only submitted in the previous year were used to orient listeners to the levels of the rubric: performances ranked 1, 36, 72, 100, and 140 (out of 175 submissions) were presented as exemplars of each of the five performance levels. For material to be judged in the present study, we selected 15 performances, 3 at each level of the rubric (1 from the top, middle, and bottom ranks at each level), based on the composite scores (using all audition components) as judged by the five-judge all-state panel using the traditional procedure from the previous year.

We then collected performance adjudication data in three different ways. First, we asked 120 participants to listen to the 15 étude-only performances and make judgments using the five-level rubric. These volunteer auditors were university students or public school teachers who had string, wind ensemble, or choir/piano performance experience (40 in each group). Participants had either contractual teaching experience (n = 47) or no formal teaching experience (n = 73).

Second, we collected traditional ratings (using the global point system) on the étude alone from three experienced string adjudicators (each of whom had a minimum of 5 years of public school orchestra teaching experience, 3 years of university-level string methods teaching, and 2 years of state-level adjudication experience). As a third step, we obtained traditional composite rankings using the regular five-member all-state judging panel. These rankings were based on all components of the audition (étude, orchestral music excerpt, scales, and sight reading). This set of judges heard all violin performances submitted that year. Using the Olympic scoring method described above, the highest and lowest rankings per performance were eliminated. Thus, we used the scores of the middle three judges to determine rankings based on the entire audition.

Results

The Olympic scores for all components of the audition were compared with the three experienced judges' global rating of the étude only. This revealed a Kendall coefficient of concordance (W) of .93. The expert judges agreed among themselves highly on the étude-only ratings (W = .97). Comparisons were then made between the three expert judges' ratings on the étude only and the three groups of listeners also hearing only the étude but who used the rubric system. Spearman rank correlations with the expert judges were high for the string group (r^sub s^ = .96), the band group (r^sub s^ = .92), and the choir/piano group (r^sub s^ = .94; see Table 1). Correlations between the traditional process judges' rankings of the entire recorded performances and the 120 listeners using the rubric with the étude only were also high (r^sub s^ = .94, .93, and .94 for strings, band, and choir/piano groups, respectively), as were rank associations for the participants with teaching experience (r^sub s^ = .95) and no experience (r^sub s^ = .94). Table 2 shows comparative rankings of the 15 violin performances forjudges hearing the entire audition, the judges hearing the étude only, and the string, wind, and choir/piano performance experience groups.

An additional comparison was made to determine whether there were differences between the three performance experience groups and two teaching experience groups in the rubric scores. There were no significant differences between the groups of listeners with string, wind, or choir/piano performance experience. There were also no significant differences in rankings between those with teaching experience versus no teaching experience, nor did teaching and performance experience interact significantly.

Discussion

One purpose of select groups (e.g., all-state, all-district, etc.) in music education is, presumably, to give advanced students the opportunity to perform literature beyond the scope of the normal classroom and to perform and interact with other high-achieving students. As such, each association sponsoring these types of ensembles has the responsibility to design and administer a fair and reliable system of membership selection. We found in this study that a prescreening procedure utilizing a five-level rubric can effectively distinguish between the auditioning students who are legitimate candidates for membership and those with no realistic chance of being selected to an all-state group.

Rubric scores of the 120 volunteer listeners based on the étude only were essentially the same as expert judges' global scores on the étude, and furthermore they correlated highly with overall ratings obtained from the Olympic judging procedure that included all segments of the audition. Thus, the adoption of a five-level rubric appears to hold promise for reducing the number of full-length auditions that must be heard by judges as a prescreening mechanism to accurately and reliably select the most proficient performers for all-state placement. Given the reality of adjudicating a large number of student auditions, this procedure would seem to reduce adjudicator fatigue and the overall cost to the association administering the selection process.

Another issue related to adjudicator fatigue is the length of material to be heard. When determining audition requirements, associations are frequently concerned with choosing materials that are difficult enough to distinguish among performance abilities and long enough for the adjudicators to make reliable assessments. In the present study, the evaluations of the Kreutzer étude (approximately 1 min) correlated highly with the evaluation of the entire audition (approximately 4 min). In regard to length, it would appear that material as short as 1 min in length is sufficient to produce accurate and reliable results (also see Vasil, 1973). However, it should be noted that this procedure occurred in the context of a recorded audition, and individual students were unaware of how much of the audition adjudicators actually heard. In a live audition, to stop a student who has been working for months after 1 min would not seem the best way to encourage practice and preparation for auditions in the future.

We found that performance background and teaching experience had no effect on the ability of adjudicators to reliably use the prescreening rubric. Earlier studies have also noted that raters need not be experienced performers on the particular instrument being adjudicated (Bergee, 2003; Fiske, 1975). Therefore, it seems that the criteria for selecting adjudicators may be related to fundamental musicianship rather than the number of years taught or major performance instrument. The financial implications here are substantial, allowing state music associations to select adjudicators from a smaller geographic region, thus reducing the expenses of travel and housing. Although the present study found no difference in the teaching experience variable, future studies should continue to investigate the influence of judging experience in both global and rubric-based assessment contexts. Although we found that the étude-only evaluation correlated highly with the overall evaluation, future studies should investigate other dimensions of the audition process, including the components themselves (scale, orchestral excerpt, sight reading, etc.).

It would also seem useful for future studies to investigate the constructs of a "superior" performance, at least in the context of all-state selection. Perhaps novice and experienced judges agree on superior performance as opposed to substandard performance. What roles do intonation, tone, rhythm, and phrasing play in making these judgments? Does one aspect of musicianship dominate others, or do adjudicators respond to various combinations when evaluating performance? Are these constructs applicable across instruments in judging musical performance, or do they vary according to instrument and voice?

Most associations in selecting audition material attempt to provide adjudicators with sufficiently varied material (lyrical vs. rhythmical, loud vs. soft, fast vs. slow) with which to make judgments. The particular Kreutzer étude used in the present study was essentially rhythmical in nature, moderately fast, with little or no dynamic contrasts. Yet rankings of teachers, students, and experienced judges correlated highly with the five expert judge Olympic-scoring system used to listen to the entire audition. Would the results be similar if other excerpts, for example, one that was primarily lyrical in nature, had been selected?

The goal in any performing situation is to perform at one's highest level of ability or perhaps more importantly to exceed one's previous standard. In timed events such as track and field or timed swimming trials, a clock provides an objective measure. As "fair" as we attempt to make music competitions, quantifying music performance is at best a messy endeavor. What students, teachers, and parents may label as "unfair" is simply the result of subjective judgment. Rubrics, Olympic scoring, and rating scales are attempts at providing a fair and reliable assessment of a musical performance. Results in the present study indicate that agreement is possible among expert judges, teachers, and college music students in establishing a rank order of performance ability. Much additional research with performance rubrics seems appropriate to aid in the development of both fair and efficient methods of choosing all-state participants, particularly when adjudicating large numbers of applicants.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Financial Disclosure/Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

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

Barnes, G. V., & McCashin, R. (2002). All-state orchestras: A survey of practices and procedures. UPDATE: Applications of Research in Music Education, 20(2), 16-20.

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. (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.

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

Byo, J. L. (1993). The influence of textural and timbral factors on the ability of music majors to detect performance errors. Journal of Research in Music Education, 41, 156-167.

Byo, J. L., & Crone, L. J. (1989). Adjudication by nonmusicians: A comparison of professional and amateur performances. Missouri Journal of Research in Music Education, 26, 60-73.

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

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

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

Dugger, R. (1997). Inter-judge reliability for the 1994 Oklahoma all-state band auditions based on an Olympic style judging system. Journal of Band Research, 32(2), 66-75.

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

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

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

Geringer, J. M., & Johnson, C. M. (2007). Effects of excerpt duration, tempo, and performance level on musicians' ratings of wind band performances. Journal of Research in Music Education, 55, 289-301.

Hewitt, M. P. (2007). Influence of primary performance instrument and education level on music performance evaluation. Journal of Research in Music Education, 55, 18-30.

Hewitt, M. P., & Smith, B. P. (2004). The influence of teaching career level and primary performance instrument on the assessment of music performance. Journal of Research in Music Education, 52, 314-327.

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

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

Nichols, J. P. (1991). A factor analysis approach to the development of a rating scale for snare drum performance. Dialogue in Instrumental Music Education, 15, 11-31.

Norris, C. E., & Borst, J. E. (2007). An examination of the reliabilities of two choral festival adjudication forms. Journal of Research in Music Education, 55, 237-251.

Perkins, D. W., & Allen, M. L. (1991). An investigation of interjudge reliability of the Texas music educators association all-state orchestra string auditions. Texas Music Education Research, 21-23.

Price, H. E. (2007). Effect of keyboard ownership on keyboard performance in a music fundamentals course. International Journal of Music Education, 25(1), 49-54.

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. P. (2004). Five judges' evaluation of audiotaped string performance in international competition. Bulletin of the Council for Research in Music Education, 160, 61-69.

Smith, B. P., & Barnes, G. V. (2007). Development and validation of an orchestra performance rating scale. Journal of Research in Music Education, 55, 268-280.

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Music performance assessment as a research tool. Music Perception, 21(1), 21-41.

Vasil, T. (1973). The effects of systematically varying selected factors on music performing adjudication. Unpublished doctoral dissertation, University of Connecticut, Storrs.

Wapnick, J., Flowers, P., Alegant, M., & Jasinskas, L. (1993). Consistency in piano performance evaluation. Journal of Research in Music Education, 41, 282-292.

Wapnick, J., Ryan, C., Campbell, L., Deek, P., Lemire, R., & Darrow, A. A. (2005). Effects of excerpt tempo and duration on musicians' ratings of high-level piano performances. Journal of Research in Music Education, 53, 162-176.

Winter, N. (1993). Music performance assessment: A study of the effects of training and experience on the criteria used by music examiners. International Journal of Music Education, 22, 34-39.

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