Assessment


The goal of this study was to create and pilot a multi-dimension, statistically reliable assessment tool for judging anonymous honor band auditions that have minimal control over the adjudicator pool. Every year, thousands of middle and high school instrumentalists throughout the nation audition for a possible placement in regional and statewide honor bands or orchestras. Acceptance to an honor group is not only a validation of a student’s hard work and ability, but also provides a high-level musical experience that cannot be duplicated in the student’s personal educational setting. Furthermore, all-state ensemble recognition is often a crowning achievement in one’s secondary music career and can be an important supplement on scholarship applications. But how valid are these audition results? The majority of auditions are held in double blind, anonymous settings as a protection of privacy as well as a measure of preventing corruption. This level of security creates a situation in which pretesting and post testing results is a virtual impossibility; thus, knowing if the correct students were actually selected and ranked accurately is impossible to validate.

Because of this particularity, numerous studies have been administered with the goal of creating an ideal, fail-proof audition assessment model arming adjudicators with tools to produce the most accurate results possible. Various studies have focused on the topics of rubrics, rating scales, ordinal rankings, audition materials, plus external factors such as school types, student types, socio-economics, etc. However, realities common to most honor group audition situations and often ignored are adjudicator competency, experience, and training. In fact, most studies typically control for the adjudicator as a means to produce valid, quantifiable measurements of the sample group’s audition data. In the case of artistic measurement of human behavior, the reality of creating a panel of volunteer adjudicators of equal merit is simply impossible for most secondary school arenas. Judging panels typically consist of a collection of three to five directors from the district, local musicians, or regional college faculty; and all with distinctively different experience, training, and backgrounds.

The assessment tool created for this project focused on the expectation of a varied adjudicator panel. It is a single ballot with five rubric-type point zones that is implemented for each auditionee. Because most auditions are expected to consist of at least 20 auditioners, the point zones are divided based off of percentages of normal distribution. A single column worth 100 points is used to score each portion of the audition and each portion is later weighted by a district’s point specifications (i.e., etude are worth 50 percent of total score, sight-reading is worth 15 percent of total score). The descriptors for each point zone are purposely subjective and vague to allow a judge’s to utilize her existing knowledgebase and skillset, regardless experience or training.

The adjudication ballot was piloted by Midwestern, large-school all-region band during their 2012 auditions with 23 different adjudicator panels (n=23). The results were checked for reliability using the Cronbach’s Alpha. In every scenario, the reliabilities measured above a .90

Festival ratings are usually assumed to be ordinal-level measurement. That is, a Superior (I) rating is considered “higher” than an Excellent (II) rating, an Excellent “higher” than a Good (III), and so forth. In research contexts, however, festival ratings often are analyzed as either interval- or categorical-level data. This study constitutes an attempt to analyze festival ratings using techniques specifically designed for ordinal-level data. In particular, I used ordinal regression techniques to determine the extent to which certain variables influenced the ratings awarded to two years’ worth of state-level festival entrants. I found that type of event (solo/ensemble), time of day (morning/afternoon), geographical district (metropolitan/non-metropolitan), school size (highest classification/four lowest classifications combined) and the time of day by school size interaction predicted the rating. School district expenditure per average daily attendance (ADA) and performing medium (vocal/instrumental) did not predict the rating. Ratings displayed clear ordinal characteristics in the geographical district, time of day, type of event, and school size variables. They did not display clear ordinal characteristics, however, in the expenditure per ADA and performing medium variables. The overall model failed the test of parallelism, which implies that festival ratings should be thought of as categorical rather than ordinal.

Dekaney, Elisa. Syracuse University, NY. The Effect of Thin-Slice Methodology on the Perception of General Music Classroom Attractiveness, Teacher’s Values, And Music Pedagogies.

The effect of thin-slice methodology on the perception of general music classroom attractiveness, teacher’s values, and music pedagogies Thin-slice methodology has been used in psychology research is defined as the ability a person may have to observe a small selection of an interaction, in a time range from two seconds to five minutes draw accurate conclusions in the emotions and attitudes of people. According to Ambady, LaPlante, and Johnson (2001), who investigated the issue of interpersonal sensitivity in thin-slice judgments, research in this area has indicated that individuals show higher degrees of consensual accuracy when they are judging others. According to them, intuitive judgments about others based on thin slices can be used to evaluate social acuity and to form the basis of predictions regarding social and emotional outcomes.

Intuitive judgment is based on the observer’s perception. Leeuwenberg (2003) proposed that, “Perception is affected by context, knowledge, attention, motives, and planning actions, and perception should not be studied in isolation, but in relation to all these high level of cognition” (p. 389). In music education, physical attractiveness has been perceived as a sign of better performance. For instance, Wapnick, Darrow, Kovacs, and Dalrymple (1997) concluded that those who were rated higher on the physical attractiveness scale were also rated higher in their performances. Likewise, VanWeelden (2004) investigated the perceptions of performance of racially stereotyped music and conductor race and found conductor race to be a significant factor in evaluations of ensemble performance. These studies have investigated people’s attractiveness, but not environments. Would attractiveness of a classroom also determine how observers perceive the quality music teaching in that particular classroom, teacher’s values, and student’s success? The purpose of this study was to use thin-slice methodology to investigate observer’s perception of three general music classrooms.
Participants were undergraduate and graduate students (N=66) enrolled in the music education program at a private university in the northeastern United States. Three diverse general music settings were chosen for this study: urban, suburban, and rural. Several high definition pictures, showcasing different angles of these music classrooms, were taking using a Nikon 5000 DSRL camera to provide an accurate overview of the environment. The pictures were taken after the school year ended and prior to classroom summer rearrangement. Using Qualtrics survey building, three pictures from each classroom were randomly chosen and grouped together as one time segment. Participants were randomly assigned to one of two treatment groups: Group 1 (n=33) viewed each of the three pictures for 10 seconds and group 2 (n=33) viewed each of the three pictures for 3 minutes. After viewing each segment, participants were asked to write what information they could gather from the pictures regarding school setting (urban, rural, or suburban), what is being taught in that classroom (guitar, popular music, classical music, women composers, world music), music pedagogies (Orff, Kodaly), and student body population (high achieving students, low achieving). Although there were some guided suggestions of what to look for in the picture, participants’ answers were in an open format. Participants were allowed three minutes to write their answers for each segment. Responses were analyzed for each treatment group and post-hoc analysis looked at class status (graduate, senior, junior, sophomore, and freshman) to determine if graduate and upper classmen students provided more elaborated answers than sophomore and freshman. Results for this study revealed that the use of thin-slice methodology in music research could facilitate successful explorations of sensitivity and accuracy in the field of music education. Furthermore, this methodology could help the understanding of interpersonal sensitivity exposing biases and preconceived ideas of teacher’s values and learning environments.

References:


The Instrumental Music Error Detection Test, or IMEDT, was developed to examine error detection ability regarding pitch, rhythm, and articulation errors in recordings of a wind ensemble. This test was designed to simulate an authentic rehearsal situation. The eight musical excerpts were selected from grade three band literature and performed with full instrumentation. Excerpts were selected that depicted a variety of tonalities (major and non-major), tempi (fast and slow) and meters (simple and non-simple). A total of 30 errors was inserted into the recordings; 12 pitch errors, 12 rhythm errors, and 8 articulation errors. A university wind ensemble recorded the excerpts, first as written, or what was considered to be a “model performance,” and a second time with the errors inserted. The completed IMEDT contained two recordings of each of the eight musical excerpts, the first as written and the second with inserted errors.

The IMEDT was administered in six different test administration variations to determine the method that was most valid and reliable and had the highest internal consistency. These six administrations contained studying the score prior to hearing the error version of the recording with controlled and non-controlled time, only hearing the model performance recording with no score study time with controlled and non-controlled time, and finally a combination of both, score study and listening to a model performance recording with controlled and non-controlled time. The six methods were rotated between musical excerpts to create six test variations.

Each test was administered in an individual setting with the participant and me and took approximately 45 minutes to an hour to complete. Sixty-two participants completed this first phase of test administration. Using Cronbach’s alpha to estimate the reliability and internal consistency, it was empirically decided that the test administration variation of score and recording with non-controlled time (S&R/N) had the highest alpha level. The order of musical excerpts was also determined empirically through this statistical test. Twenty additional participants completed the second phase of test administration of the IMEDT in the S&R/N method, again in a individual setting, taking approximately 40-45 minutes to complete. After data collection was complete, it was determined that the IMEDT was both reliable and internally consistent (α = .72).

Latimer, Marvin and Hancock, Carl. University of Alabama, Tuscaloosa. The Effect of Non-Performance Variables on Alabama Vocal Association Choral Performance Evaluation Ratings.

The purpose of this study was to measure the effect of non-performance variables on Alabama Vocal Association (AVA) Choral Performance Evaluation (CPE) ratings. CPE data were all adjudicator ratings and scores (N = 1215) from five consecutive years of AVA CPEs. Demographic data were collected from all participating schools (N = 68) using National Center for Educational Statistics data from the 2008-09 Common Core of Data. Non-performance variables included (a) year of the performance evaluation, (b) voicing, (c) grade level, (d) public school or private school, (e) median housing income, (f) per student expense, (g) percentage of free and reduced lunch, (h) percentage of white students in the school, and (i) performing in the morning or afternoon. Research questions were: (a) To what extent do non-performance variables influence the outcomes of the various ratings systems? (b) To what extent do non-performance variables influence sight-reading assessment? And (c) which system reduces the bias
of non-performance variables by demonstrating the least influence over the final rating and score? Two analysis techniques were used: Multiple Regression and MANOVA. Three primary dependent variables were examined: Total Score, Score Rating, and Rating. Secondary dependent variables were created by removing the influence of the final score and final rating. All models were statistically significant \( (p < .001) \) indicating that the significant predictors when combined together predicted the dependent variable. Generally, however, results indicated that while the examined variables were significantly related to assessment outcomes, the overall influence was weak. Conclusions suggest that AVA can minimize the influence of non-performance variables by using the Score Rating Dependent Variable (i.e., converting raw score into a rating).

Madsen, Clifford; Fredrickson, William and Kawakami, Ed. Florida State University, Tallahassee. Student Versus Faculty Assessments in Relationship to Effective Teaching.

Abstract: This investigation was an ex post facto descriptive study of education modules constituting the “capstone course” taken just before students’ final internship. Previous research indicated that all of these activities were deemed important and directly related to effective teaching. Forty five \( (N = 45) \) students’ scores were analyzed across these twelve modules that comprised the course and results were compared to a final assessment of effectiveness made after graduation. All undergraduates were in the same course and previously completed their regular four year curriculum leading to the baccalaureate degree in music education. Research questions were: (1) what, if any Learning Activities correlate to a summative rating of effective teaching made after graduation? And, (2) is there a relationship between students’ final ranked evaluation of these same Learning Activities and instructors summative evaluations?

The twelve modules were evaluated using a four level grading rubric: Superior, Excellent, Adequate and Below Expectations. Learning Activities were: LA # 1 translating ideas into observable and measurable behaviors. The first module required the prospective teacher to specify observable and measurable behaviors for ideas; LA # 2 specifying schedules of student/teacher interactions. This module goes through all extant schedules of reinforcement and includes a transfer task using effects of partial reinforcement; LA #3 listing the sequential steps for teaching a self-determined music task. An additional and important temporal aspect was the ability to analyze a music learning task included the temporal direction of each instructional unit; LA # 4 was an extended essay covering ethical and moral issues involving teaching.

For example, is corporal punishment legal in your state? Can humans really avoid controlling others (through an ongoing exchange of rewards and punishments) even if one wishes to do so? LA # 5 constituted a summative task consisting of questions covering the entire course. For example, “What would be the long-term effect(s) of a student receiving ONLY academic and social teacher approval? What would be the long-term effect(s) of receiving ONLY academic and social teacher disapproval? LA # 6 every student completed a self-shaping activity. This module covered 27 days of self-modification—a five-day baseline and at least 21 days of attempts to change one’s own behavior. This activity required keeping daily records with a reliability observer. LA # 7 was a rank ordering of the initial willingness of
each person to immediately volunteer leading a self-determined 1-2 minute music activity. LA # 8 encouraged students to initiate an original music product or process done outside of class but directly relating to the class. LA # 9 was an evaluation of the final teaching episode at the end of the course. LA # 10 required every student to keep a Time Log for 24/7 and record within every ½ hour period what he/she was doing at that time. LA# 11 all students completed four detailed observations of music activities and also computed reliability on one of these activities with another person. For LA# 12 each student completed 12 transfer tasks one for each of the specific research examples chosen from the textbook. After students had graduated from the program and completed State Teacher Certification requirements they were assessed for future success. The assessment came from the course instructor and the internship supervisor. The agreement between these two independent ratings was extremely high (r=.97).

Correlations for all Learning Activities appear below:

<table>
<thead>
<tr>
<th></th>
<th>LA 1</th>
<th>LA 2</th>
<th>LA 3</th>
<th>LA 4</th>
<th>LA 5</th>
<th>LA 6</th>
<th>LA 7</th>
<th>LA 8</th>
<th>LA 9</th>
<th>LA 10</th>
<th>LA 11</th>
<th>LA 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA 1</td>
<td>0.508</td>
<td>0.515</td>
<td>0.673</td>
<td>0.563</td>
<td>0.469</td>
<td>0.569</td>
<td>0.514</td>
<td>0.707</td>
<td>0.495</td>
<td>0.647</td>
<td>0.291</td>
<td>0.332</td>
</tr>
<tr>
<td>LA 2</td>
<td>0.291</td>
<td>0.132</td>
<td>0.332</td>
<td>0.096</td>
<td>0.204</td>
<td>0.151</td>
<td>0.334</td>
<td>0.182</td>
<td>0.305</td>
<td>0.154</td>
<td>0.233</td>
<td>0.275</td>
</tr>
<tr>
<td>LA 3</td>
<td>0.344</td>
<td>0.194</td>
<td>0.284</td>
<td>0.435</td>
<td>0.144</td>
<td>0.128</td>
<td>0.258</td>
<td>0.010</td>
<td>0.265</td>
<td>0.142</td>
<td>0.171</td>
<td>0.451</td>
</tr>
<tr>
<td>LA 4</td>
<td>0.416</td>
<td>0.266</td>
<td>0.432</td>
<td>0.324</td>
<td>0.291</td>
<td>0.235</td>
<td>0.088</td>
<td>0.088</td>
<td>0.063</td>
<td>0.142</td>
<td>0.483</td>
<td>0.489</td>
</tr>
<tr>
<td>LA 5</td>
<td>0.289</td>
<td>0.487</td>
<td>0.247</td>
<td>0.030</td>
<td>0.237</td>
<td>0.313</td>
<td>0.044</td>
<td>0.080</td>
<td>0.063</td>
<td>0.088</td>
<td>0.234</td>
<td>0.460</td>
</tr>
<tr>
<td>LA 6</td>
<td>0.357</td>
<td>0.151</td>
<td>0.101</td>
<td>0.443</td>
<td>0.191</td>
<td>0.234</td>
<td>0.460</td>
<td>0.332</td>
<td>0.151</td>
<td>0.063</td>
<td>0.234</td>
<td>0.460</td>
</tr>
</tbody>
</table>

Learning activities where students had to specify actual classroom behaviors instead of ideas along with answering summative questions covering ethical issues and the entire course combined with being able to specify appropriate temporal reinforcement evidence the highest correlations. Surprisingly, while students’ highest ranking was for their Self-shaping Activity, followed by their Time Log assignment and for their individual daily transfers, all of these evidenced very low correlations when compared to the instructors’ final assessments or to other course activities.


Literature Review: Participation in large group contests and festivals for bands is often a very important event for many schools, band directors, and students. Often, these successes of band programs are correlated with their achievements at contests and festivals. Frequently, schools, band directors, and students feel undue pressure to achieve high scores from judges. This has led many to question the value and purpose of contests and festivals in general. Rohrer (2002) wrote that this controversy has been prevalent since the inception of music contests in the United States.

Methods: The study reported herein examined three potential influences on the 2005 Oklahoma Junior High/Middle School and High School large group band contest judge ratings in concert playing and sight-reading along with school class size (Oklahoma UIL classification). A total of 149 Junior High/Middle School and High School bands participated in music contests in one of seven different Oklahoma large
group band contest sites in 2005. The seven contest sites employed 24 judges: 12 concert judges and 12 sight-reading judges. Twelve of the judges adjudicated at two different sites, whereas 12 adjudicated at only one site. This accounts for 18 different times judges adjudicated at the 2005 Oklahoma band contests. Of the 12 judges who judged at two different sites, each adjudicated in the same category (e.g. either concert or sight reading at both sites). The qualifications of the judges were not provided, thus, unknown.

Results: Using the statistical measurement, MANOVA, a General Linear Model Multivariate procedure, a multivariate analysis of variance on the class size and judges results indicated that There is significant interaction between Concert Judge 1 * Concert Judge 2 F (12,372) = 1.056, p<.05; Concert Judge 1 * Concert Judge 3 F (9,372) = 1.514, p<.05; Concert Judge 2 * Concert Judge 3 F (12,372) = 1.289, p<.05; and C Concert Judge 1 * Concert Judge 2 * Concert Judge 3 F (6, 246) = 1.324, p<.05. The Pillai’s Trace interaction between the following main effects accounts for the variances between class size and concert judges to indicate that Class Size is 3.6%, Concert Judge 1 is 1.1%, Concert Judge 2 is 1.4%, and Concert Judge 3 is 1.9%. The Wilks’ Lambda interaction between the following main effects accounts for the variances between class size and concert judges are Class Size is 3.6%, Concert Judge 1 is 1.1%, Concert Judge 2 is 1.4%, and Concert Judge 3 is 1.9%. The Hotelling’s Trace interaction between the following main effects accounts for the variances between class size and concert judges are Class Size is 3.6%, Concert Judge 1 is 1.2%, Concert Judge 2 .1.4%, and Concert Judge 3 is 1.9%. The Roy’s Largest Root interaction between the following main effects accounts for the variances between class size and concert judges are Class Size is 3.6%, Concert Judge 1 is 2.2%, Concert Judge 2 .2.3%, and Concert Judge 3 is 2.9%. The F distribution main effects for Class Size and Concert Judge 1, Concert Judge 2, and Concert Judge 3, along with interactions Concert Judge 1 * Concert Judge 2, Judge 1 * Concert Judge 3, Judge 2 * Concert Judge 3, and Judge 1 * Concert Judge 2 * Concert Judge 3 are very small (.011 to .077), it can be concluded that they do not contribute much to the model. The Box’s Test of Equality of Covariance was used to determine if the dependent variables follow a multivariate normal distribution, and if the variance-covariance matrices are equal across the cells formed by the between-subjects effects. The significance value of the test is greater than .05. Thus, suggesting that the assumptions are met. The Levene’s Tests the equality of the error variance across the cells defined by the combination of facet levels. Further results will be discussed.

Discussion: Thus, it can be said that holding all else constant, the judges’ assessments for the 2005 Oklahoma band contests results are similar and there does not appear to be any bias toward class size or between concert and sight-reading events than would be expected by chance. After looking at different facets that could effect the statistical data, and after removal of the effects of covariate(s), it still does not appear that there is any significant any bias in judging of the 2005 Oklahoma large-group band contest. Because performing groups in every state want fairness and consistency in their contest results, the implications of this study assert that music contests can be impartial and reliable

Poché-Rodriguez, Kelley. Texas Tech University, Lubbock. An Analysis of SAB vs. SATB Sight-Reading Contest Literature for Middle School Choirs: A Replication and Extension.
Between 2006 and 2013, there were 10,567 middle school choral entries in the Texas University Interscholastic League (UIL) Concert and Sight-Reading Contest. Fully 91.15% of those entries (n=9,632) were comprised of single-gender choirs, while only 8.85% (n=935) constituted mixed choirs. In contrast to the rise in overall annual entries (from 1,130 in 2006 to 1,477 in 2013), the number of mixed choir entries has steadily declined (from 138 in 2006 to 92 in 2013). Despite this trend, mixed choirs are still a part of many middle school choral programs, posing unique pedagogical problems for directors in their preparation for choral competitions.

One such challenge lies in the determination of whether to perform Soprano – Alto – Baritone (SAB, 1 male voice part) or Soprano – Alto – Tenor - Bass (SATB, 2 male voice parts) literature. While many directors recommend taking into consideration the number, strength, and vocal independence of adolescent boys in the choir, the range of the part or parts and the students’ progress in the voice change process may be even more critical factors.

The purpose of this study was to conduct an analysis of Texas UIL SATB and SAB middle school sight-reading literature from 2006 – 2013 and the corresponding sight-reading contest ratings. This study is a replication of an earlier research project, including the addition of data from the 2013 contest results. Each choral piece was analyzed to determine the ranges of individual voice parts. The Baritone range in SAB literature spans from a Perfect 4th to a Major 7th, while the Tenor and Bass parts in SATB literature only range from a Minor 3rd to a Major 6th. Additionally, for all but one year of this study, the lowest “D” pitches of the SAB Non-Varsity Baritone parts contained optional alternate notes an octave higher, which could be sung by boys for whom the lower notes would be out of range. Paradoxically, however, none of the Varsity Baritone parts contained optional alternate notes.

The official UIL Concert and Sight-reading Contest ratings for all mixed choir entries in the 28 Texas Music Educators Association (TMEA) regions, available to the public on the official UIL website, were examined from 2006 to 2013. A determination of a choir’s sight-reading voicing as SAB or SATB was made, based upon the majority voicing of the concert repertoire selections, according to the official UIL contest rules. Sight-reading ratings from the three judges were added together to produce a composite rating, with 3 being the highest (corresponding to a Sweepstakes rating of three “1’s”), and 15 the lowest. While only 35.19% of all mixed choirs sight-read SATB literature, the mean score (3.90, SD=2.05) was 3.28 points better than that of all choirs reading SAB literature (7.18, SD=3.85). Results of a t-test for independent samples showed this difference in scores to be significant (t [14, 16] = -14.05, p<.0001).

Results indicate that, though it might seem counterintuitive, middle school mixed choirs seem to perform better at UIL contest when sight-reading SATB, rather than SAB, literature. Future research might further examine the differences in middle school sight-reading contest ratings between single-gender and mixed-gender choirs.
The purpose of this study was to examine the intersection of curriculum and participation in adjudicated events in three instrumental music classrooms in Kansas. The key research question is: How do participants, colleagues, students, and administrators perceive the preparation and execution of adjudicated events within the curricular framework of the academic year? The phenomena of curricular decisions made by three experienced instrumental music teachers in order to prepare for and participate in large group adjudicated experiences were of particular interest.

**Conceptual Framework:** Empirical research on adjudicated events has examined multiple perspectives, stakeholders, and questions centered on validating the worthiness of this experience for student musical growth. These events have the potential to advance performance standards, shape morale within a musical program, increase interest, and provide both student and directors recommendations for improving ensemble quality (Rohrer, 2002). Researchers interested in adjudicated events have focused many investigations on assessing student and director attitude (Austin, 1987, 1988; Brunsed, Sochinski, & Hinkle, 1983; Rogers, 1985; LaRue, 1986; Hamann, Mills, Bell, Daughtery, & Koozer, 1990; Howard, 1994; Meyers, 2011), improving the reliability of the contest (Brakel, 2006; Hash, 2012), examining the extramusical influences on student success at adjudicated events (Bergee & Platt, 2003; Bergee & McWhiter, 2005; Bergee & Westfall, 2005, Bergee, 2006; Cheek, 2007), and supporting increased musical achievement as a result in participating in an adjudicated event (Austin, 1987, 1988; Jarrell, 1971; Temple, 1973).

**Method and Sources of Data:** A multiple case design (Stake, 2006) using a primary data set of phenomenological interviews (Seidman, 2012) with experienced instrumental music teachers served as the primary data set. Additional participants (music colleague and administrator) were interviewed and student volunteers were asked to complete an open-ended survey to understand what impressions students have about the topic adjudicated events and how their teachers (study participants) prepare and participate within the curricular framework of the academic year. In this study, the single case of the music classroom is of interest because it captures the descriptions and perceptions of participants and stakeholders about the purpose of the study. All participants are members of the Kansas Music Educators Association (KMEA) and have served on curriculum design committees as well as attend adjudicated events through the year. Three information-rich classroom cases were selected. In order to gain insight from experienced instrumental music teachers (See Table 1) who have past involvement with the curriculum and adjudicated event processes in Kansas, I chose a criterion sampling strategy for this multiple case study. The teachers at the three schools held similar beliefs, practices, and goals regarding adjudicated event preparation and participation. Transcripts were made of all interview data sets. To insure accuracy, teacher participants received electronic copies of their interview transcripts and had an opportunity to verify their thoughts, opinions, and observations. The process of member checks allowed participants the opportunity to rephrase and rearticulate their thoughts, should they feel differently about what was initially stated. Through the interview process, the investigator used clarifying questionings to confirm the details of the interview(s). The multiple data sources provided several lenses, through which the researcher was able to analyze the phenomenon presented in this study, in addition to presenting a rich description of the lived experiences of the participants. After coding the transcripts, the researcher asked another peer researcher to code the transcripts to
corroborate findings from the transcripts. Merriam (2009) confirms peer examination involves “… asking a colleague to scan some of the raw data and assess whether the findings are plausible based on the data” (p. 220). The researcher is familiar with the process of adjudicated events in Kansas. Previous to conducting the study, the researcher had experience being an inservice music educator in Kansas. The researcher does meet all of the criteria used in screening potential participants for this study; however, by the definition provided above, is not considered an experienced music teacher. Moreover, the researcher has experience working in a professional capacity with the teacher participants.

Findings:

Presented in three classroom profiles, findings support past discussions of curriculum in music education (cite). Participants described that planning short- and long-term curricular goals is an ongoing process for optimal student learning. Unique intersections with philosophy of music education and programming repertoire emerged throughout the interviews. Students, administrators, and music colleagues confirmed the importance of attending adjudicated events as an important curricular event for learning. Utilitarian and aesthetic examples for attending adjudicated events were listed during data collection. The final paper includes a full discussion of the findings and implications for the field


Ideally, classroom assessment should be used to inform instruction and improve student learning (Colwell & Hewitt, 2010). In reality, teachers base student grades on many other factors such as participation, attendance, attitude, and other non-musical criteria (McCoy, 1991; Simanton, 2000; Russell & Austin, 2010). McCoy (1991) found that, attitude included as a non-musical criteria, both band and choral directors weighted non-musical criteria more heavily than cognitive and performance criteria combined. Almost a decade later, Simanton (2000) found that band directors reported that participation/attitude, attendance, and other non-musical criteria accounted for 63.5% of a student’s grade. In a more recent study, Russell and Austin (2010) discovered that 91% of secondary music educators reported using attendance as a grading factor with an average grade weight of 25%, and 93% of those teachers used attitude as a factor describing an average weight of 27% for that criterion. In addition, 61% described using home practice for grading purposes with an average grade weight of 8% making the total average grade weight for non-achievement criteria 60% (Russell & Austin, 2010). This translates to a decades-long continued practice of music teachers assigning high grade weight to non-musical and non-achievement-based criteria. It is problematic, however, to grade using non-musical factors because the reliability and validity of those assessments are difficult to measure (Kotora, 2005). In order for assessments of musical achievement to be considered reliable and valid, they must actually measure musical achievement (Salvador, 2010). Standards-Based Grading (SBG), or the evaluation of students based on how close their classroom performance is to meeting curriculum standards or specific learning targets, is a way to provide students and parents with growth-producing feedback about classroom achievement in a valid and reliable way. However, Russell and Austin (2010) found that only 2% of school districts are using SBG practices. The purpose of this exploratory study is to examine the
use of SBG practices among practicing music educators to describe the prevalence of the practice as well as report the rationale teachers provide regarding their practice. The purpose of this study will be to estimate the level of familiarity that practicing music teachers have with SBG in order to answer the following research questions: 1) Are practicing music teachers aware of SBG and can they accurately describe what SBG is, 2) If so, to what extent are SBG practices being used by these teachers?


For many public school instrumental and vocal ensembles, the adjudication process plays a key role in the curricula teachers develop for their courses (Hash, 2012). In the current culture of high-stakes testing in education, administrators are turning towards standardized test scores to assess teacher effectiveness, even incentivizing pay for higher test scores (Harms, 2012). As school districts and federal legislation seek to tie teacher evaluations to student achievement (Downey, 2013), questions begin to arise regarding how educators of non-tested subjects will be evaluated. It is becoming increasingly evident that administrators will likely reach towards contest and competition ratings to evaluate music educators because these contest scores provide decision makers with third-party data that can be used to compare schools against each other (National Association for Music Education, 2011). Research on the effect of performance order on the adjudication of music festivals is sparse. However, several researchers have concluded that the way a contest is presented to the adjudication panel may have implications for how an ensemble is evaluated.

The purpose of this study is to determine the effect of repertoire performance order on adjudicated evaluations. Five recordings of different ensembles playing an identical 30-second excerpt from the final movement of Lincolnshire Posy by Percy Grainger were presented to 20 evaluators. These recordings were taken from performances representing a wide range of ensemble skill as determined by a panel of experts. Among the ensembles recorded were a professional wind band, three high achieving high school bands, and one middle school band. The evaluators (N = 20) consisted of public school music educators (n = 14), higher education music educators (n = 4), and music education graduate students (n = 2). Evaluators listened to the groups in one of five different orders and were asked to rate the quality of the performance on a scale of 1-10 and to rank the performances from highest quality to lowest quality. Research questions included: (1) Are the rankings of each ensemble consistent between the five performance orders? (2) Are the mean ratings of each ensemble consistent between different performance orders? (3) Are the mean rankings of each ensemble consistent within the different performance orders? and (4) Are the ratings of each ensemble consistent within the different performance orders?

Descriptive data were compiled for all the ratings and rankings for each ensemble in each of the five performance orders. Due to the ordinal nature of these data, non-parametric tests were used for analysis. Results indicated: (1) A significant difference was found in the rankings of the professional wind band in one of the orders. In the order in which significance was found, the professional band was heard after the best high school band (as determined by a panel of three experts before the
In this single case, the professional band was ranked below two of the high school bands, indicating an order effect. There were no other significant differences found in any of the other orders. In all other performance orders, the professional band was ranked first. (2) When analyzed as a whole, no significant difference was found in the ratings for all the ensembles. (3) When analyzed as a whole, no significant difference was found in the rankings of all the ensembles. (4) No significant difference was found in the ratings of each ensemble across the different performance orders.

When looking at the differences in the mean rankings, there was a negative correlation between ensemble skill and difference in the mean rankings. This appears to refute the findings of Brakel (2006), who found that more advanced groups enjoyed a higher rate of adjudicator reliability than did less advanced groups.

Also of interest, two of the high school bands received their lowest mean rankings when they followed the middle school band. A similar effect occurred with the professional band. Two of the high school bands received their highest mean rankings when following the professional band.

This study attempted to identify a possible confounding variable in the evaluation and assessment of school music groups. Attempts to identify and control for various confounding variables in the adjudication process will help music educators improve the quality of the contests and competitions they hold and should make for a better student experience. Further research should be conducted on the fairness of the adjudication process of public school music contests.

Wapnick, Joel. McGill University, Montreal, Canada. Darrow, Alice-Ann, Florida State University, Tallahassee. Evaluations of Half, Intact, and Composite Performances of Two Chopin Etudes as Played by Two Expert Pianists.

Presenters: Joel Wapnick, McGill University

The main purpose of this study was to investigate possible order effects in the ratings of performances by two expert pianists of two Chopin etudes (op. 25, #8 and op. 25 #9). Secondary purposes were to determine the interjudge reliability of undergraduate and graduate music majors when judging these performances, and to see whether preferences for the performances of one performer over the other would be revealed, given that both performances were note-perfect and musically appropriate. Participants were pianist and nonpianist music majors from two universities (n=131). They listened to trials consisting of (1) intact performances from two different performers, (2) the first halves of each performance, (3) the second halves of each performance, (4) two composite intact performances consisting of the first half played by one pianist and the second half played by the other pianist, and (5) a second presentation of one of the original intact performances, presented to assess intrasubject reliability. The performers differed from each other in that one was an internationally acclaimed concert pianist ("eminent pianist," abbreviated EP) and the other was a less well-known, but highly accomplished, winner of international competitions ("accomplished pianist," and abbreviated AP). Listening sessions were held in both small and large groups. Participants were told that the purpose of the study was to determine the degree to which a performance of part of a short piece of music predicts
ratings for the performance of the piece when played in its entirety, and were informed that each musical excerpt would consist of an entire performance the etude, the first half of the etude only, or the second half of the etude only. Participants were asked to rate performances on a seven-point scale, and to hold off from making a rating until the entire trial example was played.

Although ratings for EP were higher than they were for AP, no significant differences between the two composite performances were found. Ratings thus were not more dependent on what was heard at a performance's beginning (a primacy effect) versus what was heard at a performance's conclusion (a recency effect). Differences between the two composite pairs were small. Moreover, with only one exception, the mean ratings for the component performances tended to fall close to halfway between the ratings for EP's and AP's intact performances. Also, correlations between first half and intact ratings were neither higher nor lower than correlations between second half and intact ratings. Second halves of etudes were rated higher than first halves, and intrasubject reliability was low.

Results from this study were consistent with results from previous research regarding whole-versus-part ratings. In the earlier work, excellent performances were rated significantly higher than ratings for sections of those performances—ratings of the whole was higher than ratings of any of its parts. Similarly, in the present research EP’s intact performance of the op. 25, no. 8 Etude was rated significantly higher than his first-half performance (though not higher than the second-half performance), and in the op. 25, no. 9 Etude, EP’s intact performance was rated significantly higher than either of EP’s half-performances. Also, AP’s second intact performance (but not the first performance) was rated significantly higher than AP’s sectional performances for the op. 25, no. 9 Etude. Although data from the current study are not uniformly in line with the supposition that, for high-quality performances, “the whole is greater than the sum of its parts,” the fact that this conclusion appears to have held more strongly for the more highly rated of the two performers is consistent with previous research showing that the superiority for whole over part ratings was greatest for the best performance than for other performances.

There are an almost limitless number of ways in which sectional versus intact ratings might be investigated in future research. Of particular interest would be studies in which performances under consideration are longer in duration and more varied in musical content than was the case in the present study. Another possibility would be to compare fixed sectional ratings with ratings obtained continuously by using a continuous digital response interface or a computer-based display. A deeper understanding of how responsive evaluators are to changes in performance as a function of musical structure, and to changes in performance quality, might lead to an increased knowledge concerning the process of musical evaluation.

Whitaker, Jennifer. University of North Carolina at Charlotte. Orman, Evelyn and Yarbrough, Cornelia (Emerita Professor). Louisiana State University, Baton Rouge. **Content Analyses of Responses to Music Teaching and Performance Videos on YouTube: A Pilot Study.**

Researchers’ interest in YouTube for music teaching and learning has increased. Recent YouTube related studies have focused on how musicians use YouTube for promoting their music (Cayari, 2011),
the culture of online folk music communities engaged in teaching and learning (Waldron, 2012, 2013; Waldron & Veblen, 2008; Veblen & Waldron, 2012), instructional characteristics of selected folk music videos (Kruse & Veblen, 2012), and descriptive characteristics of music therapy (Gooding & Gregory, 2011) and music education videos (Whitaker, Orman, & Yarbrough, in press). While these studies provide a general idea of extant music related videos and musicians’ use of YouTube, little is known about user’s reactions to music education related videos on YouTube.

This pilot study was designed to determine whether a coding system for analyzing comments used in a previous analysis of YouTube viewer responses in another discipline (Lewis, Heath, Sornberger, & Arbuthnott, 2012) could be effectively applied to viewer responses to music teaching and performance videos. Specific questions included what are the general characteristics and musical topics of a) initial responses and b) responses to initial comments.

A total of 513 comments (n = 208 teaching, n = 305) were gathered from randomly selected music education teaching and performance videos identified in previous research. Comments were transferred to a database created specifically for this investigation. Analysis categories included Self-Disclosure, Feedback, Factual Information, Help Related, and Irrelevant/Indecipherable (Lewis et al., 2012). We developed a list of musical topics discussed within responses using an inductive approach. Both general and musical coding rubrics were applied to each comment as a whole and were not mutually exclusive. Frequency counts were calculated and converted to percentages. Inter-coder reliability on all 513 comments averaged across all categories equaled 93% with a range of 89% to 99% for subcategories.

Results indicated the largest percentage of initial responses to teaching videos gave hostile, angry, insulting, or negative feedback (38%). In fact, 12% included profane language. Twenty-six percent shared personal experiences, 18% expressed personal validation, acceptance, or admiration, 9% thanked the uploader, 5% used humor, 4% referenced a specific occurrence in the video, 3% invalidated the pedagogy used in the video, 7% asked factual questions or requested additional information, and 4% commented on factual information. Two percent were Indecipherable or Irrelevant. The most frequent musical topic was range (3%); followed by embouchure, style, and vibrato (2% each).

Initial responses to performance videos revealed 31% of commenters shared personal experiences, 34% expressed personal validation, acceptance, or admiration toward the uploader; 10% percent used humor; 7% referenced a specific occurrence in the video; 4% asked factual questions or requested additional information; 3% gave hostile, angry, insulting, or negative feedback; and 2% each thanked the uploader, provided general words of encouragement, and compared the performer to another musician. Three percent were indecipherable or irrelevant. Responses to performance videos contained no musical topics.

Similar to initial responses, replies to comments of teaching videos frequently shared personal experiences (9%); gave hostile, angry, insulting, or negative feedback to the commenter (11%); expressed personal validation, acceptance, or admiration (7%); used humor (4%); and contained profanity, thanked the commenter, acknowledged a compliment, and commented on factual information (2% each). Replies to comments containing musical topics totaled 1% or less. Replies to comments made about performance videos shared personal experiences (4%); gave hostile, angry,
insulting, or negative feedback (7%); expressed personal validation, acceptance, or admiration (4%); disapproved of video quality (2%); commented on factual information (4%); asked factual questions or requested additional information (3%); and were indecipherable or irrelevant (2%). No replies contained musical topics.

Given these results, the modified coding rubric initially developed by Lewis et al. (2012) functioned well for analysis of responses to instructional and performance videos. Since this was a pilot study, the number of comments analyzed was small and may not provide an accurate picture. However, if music educators choose to use YouTube videos either in the classroom or as “homework,” they should be aware of highly negative comments and the potential for student exposure to profanity. Many positive responses contained one word such as “great” or “cool” that could be functioning as online “applause.” Given the lack of music specific comments, it seems National Standards 6 and 7 may not be transferring to online situations. Since responses to online videos may be a measure of the general public’s musical knowledge, music educators may want to incorporate activities involving online videos as part of their curriculum.