ANALYSIS OF THE RATINGS AND INTERRATER RELIABILITY AT HIGH SCHOOL CHORAL FESTIVALS IN INDIANA

BY

SHEH FENG NG

Submitted to the faculty of the
Jacobs School of Music in partial fulfillment
of the requirements for the degree,
Master of Music Education
Indiana University
May 2015

Accepted by the faculty of the
Indiana University Jacobs School of Music,
in partial fulfillment of the requirements for the degree
Master of Music Education.

Master's Committee

_____

Patrice Madura Ward-Steinman, Chair

_____

Lissa Fleming May

_____

Peter Miksza

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to many people, without whom this research would not have been possible.

Firstly, my utmost gratitude goes to Dr. Patrice Madura Ward-Steinman, my thesis committee chair – Thank you for tissues, kind encouragement, and putting me back on the right track numerous times. This thesis would not have happened without her guidance and time spent poring over draft after draft of my writing. She has been so wonderful even during a personally stressful period that I can't ever thank her enough.

To my thesis committee members Dr. Peter Miksza and Dr. Lissa Fleming May, for their willingness to share with me their extensive knowledge and experience as educators and researchers. I am indebted to Dr. Miksza for his invaluable help with the statistics, and for helping to clarify my thinking and approach to the data analysis. Thanks also extend to Dr. Christopher Dye, who gave me a foundation in statistical analysis and time spent explaining complex stats, without which I would not have the confidence to embark on this topic, and Dr. Katherine Strand, who stepped in at the last minute for my defense.

To the wonderful and generous staff at the Indiana State School Music Association, in particular Mr. Charles "Rusty" Briel, Executive Director, and Mr. Mick Bridgewater, Assistant Executive Director, for spending time patiently explaining the ins and outs of the ISSMA festival and to Tama Poncar, Office Manager at ISSMA, for answering my queries and providing me with the data for this study.

Lastly, thanks to my family and friends, who have supported me in this journey. Thank you all for your love and friendship, which have sustained me through challenging moments. Without you, I would not be here today.

ABSTRACT

This study investigated the interrater reliabilities of large-group choral festivals and provides data for comparison with similar events. Data for this study included ratings and points awarded by a total of 58 panels (of three adjudicators each), to 925 choir performances by 689 discrete high school choirs at the Indiana State Schools Music Association-sponsored choral festivals in 2012, 2013, and 2014. The research investigated (a) frequency distributions for ratings, types of choirs, and group level self-selection; (b) pairwise interrater correlations of ratings and points awarded; (c) interrater reliability and panel internal consistency, and; (d) differences in points awarded between adjudicators in each panel.

Results indicated that a significantly higher proportion of choirs were awarded Gold (77%) and Silver (22%) ratings than other types of awards. There were significantly more mixed (60%) and treble (34%) than there were mens (6%) choirs. There were also more choirs entering at Group I (39%) and Group III (30%) levels than Group II (17%) and Group IV (15%) levels. Percentage agreements of ratings were mainly high, with 41 out of the 58 panels (71%) having a mean percentage agreement of >70%. 155 out of 174 pairs of adjudicators (89%) had pairwise percentage agreement of >60%. While mean Interrater Reliability Coefficients (IRCs) were almost all positive, there was a large range for correlation coefficients from weak ($r_s = .155$) to very strong ($r_s = .939$). Internal consistency ranged from moderate ($\alpha = .48$) to high ($\alpha = .96$) over the three years. A majority of the Intraclass Correlation Coefficients (ICCs) were in the strong (0.7 - 0.8) to almost perfect (> 0.8) agreement ranges, indicating very good agreement by panel. While there were instances of significant differences ($p < .01$) found over the three years, in general the panel members seemed to agree on points awarded.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

STATEMENT OF THE PROBLEM

A great number of high school choirs in the United States participate in local, district, state, regional, national, or even international music festivals and competitions each year. The students in these choirs are usually instructed and prepared for such festivals and competitions by their school music teachers. Some of these festivals and competitions are considered by choirs to be "high-stakes" (high-risk or with important consequences), due to the financial cost of participation and need for accountability to the schools that support the choirs. Decisions on the employment of choral teachers can also be made based on the achievement of choirs at previous festivals. Results attained at these festivals and competitions thus become not only a matter of pride for the choirs and provide a concrete measure of choral achievement, but also act as a means of making related decisions affecting the choirs (including allocation of school budget). Thus, there is much interest by stakeholders in finding out what affects the performance of choirs at these festivals and competitions, as well as what affects the adjudication itself.

The history of choral adjudication goes back almost a century. Probably one of the earliest studies of state choral contests was done by Florence Best and published in the *Music Supervisor's Journal* in 1927. In a table that summarizes the adjudication practices of 12 states (no state contests existed or no data were received from the other states), Best (1927) looked at the state contest eligibility criteria, adjudication panel, grading practice, scoring dimensions and weighting, and other relevant factors of adjudication.

Since then, there have been many more studies on what affects performance and adjudication in various contexts. Advantages of participating in festivals and contests are

manifold, including their role in helping music gain strong acceptance and support as part of the total school curriculum. Music education majors who had previously participated in festivals and contests also cite these as an important factor in their decision to pursue a music career (Bergee, Coffman, Demorest, Humphreys, & Thornton, 2001). Teachers of choirs that participated in choral competitions and festivals surveyed opined that participation in competitions or festivals motivated their choirs, provided opportunities for musical growth and learning, and served as a vehicle to reinforce their teaching and improve their choirs' standards (Rittenhouse, 1989; Battersby, 1995). Research has also shown that there are positive relationships between music achievement gains and participation in rated festivals (Austin, 1988), as well as a strong positive correlation between participation in music festivals and increased musicianship (Howard, 1994).

On the other end, there are also many criticisms of festivals and contests. Many teachers feel pressure to enter these evaluative events in order to justify their programs (Rogers, 1985). Rittenhouse (1989) found that school administrators (principals) not only favored competition and winning awards (and the resultant honor that the awards brought to the school) at higher levels than did choral directors, but that these administrators believed that the ratings or rankings received at the festivals or contests were evaluative of the choral group for a given year. Miller (1994) detailed a number of negative effects of competitions in music, including the inevitable comparison between different competing groups (those who receive better ratings are deemed superior to those receiving poorer ratings), job security of directors based on their groups' ratings, and the neglect of much of the individual student's musical development in favor of preparation for the contests. The nature of evaluative performances in festivals and contests,

however, demands that groups are judged based on a single performance instead of over a period of time. This artificial and stressful situation may not be the most effective method of judging a group. For example, the group may be fatigued from traveling to the performance venue, or members may be suffering from performance anxiety. The limited time during which each group is being evaluated (usually less than 20 minutes) is also rarely the best indicator of the merit of a group or the work they have put in over a much longer period of time.

Nonetheless, with their advantages and criticisms, festivals and contests are an integral part of the musical landscape across the United States, and most schools have participated in some form of competitive or evaluative music festival or contest. Many studies have been conducted on factors related to performances and adjudication, in various contexts such as large-ensemble performances, small-ensemble or solo performances, or even individual jury assessments.

Existing Research on Music Performance Assessment

Due to the multifaceted nature of music performance assessment, the existing research can be found in a wide range of performance situations, on a still wider range of topics. Research on performance assessments can be found in the following areas: (a) interrater reliability (e.g., Bergee, 2003; Brakel, 2006; Fiske, 1977; Garman, Boyle, & DeCarbo, 1991; Hash, 2012; Latimer, Bergee, & Cohen, 2010; Smith, 2004) and intrarater reliability (Kinney, 2009), number of adjudicators on a panel (e.g., Bergee, 2003; Brakel, 2006; Fiske, 1975, 1977, 1983); (b) effect of various factors on reliability of performance evaluation (e.g. Fiske, 1977; Geringer & Johnson, 2007; Hewitt, 2005; Rickels, 2012); (c) development and validation of assessment tools (such as rating scales

or rubrics used in music performances (Ciorba & Smith, 2009; Greene, 2012; Latimer, Bergee, & Cohen, 2010; Norris & Borst, 2007; Saunders & Holahan, 1997; Smith & Barnes, 2007; Zdzinski & Barnes, 2002); and (d) nonmusical factors affecting adjudication, such as adjudicator experience or expertise (e.g., Brakel, 2006; Fiske, 1975, 1977; Kinney, 2009; Rogers, 2004), adjudicator training (Fiske, 1978, 1983; Winter, 1993), adjudicator bias (e.g., Cassidy & Sims, 1991), time of day in which the performance took place (Bergee & McWhirter, 2005; Bergee & Platt, 2003; Bergee & Westfall, 2005; Flores & Ginsborgh, 1996), number of hours adjudicators worked in a day (Barnes & McCashin, 2005), excerpt duration and score use (Napoles, 2009a, 2009b), and other non-performance variables such as school size, or funding received (e.g., Howard, 2012; Rickels, 2012).

Existing Research on Interrater Reliability

Interrater reliability studies have mainly been done in the solo and small-ensemble context (e.g., Bergee, 2007; Bergee & McWhirter, 2005) or in the context of assessments of solo performances such as juries (e.g., Bergee, 2003; Ciorba & Smith, 2009; Kinney, 2009). There has been some research on reliability in the context of large-group festivals (Brakel, 2006; Burnsed, Hinkle, & King, 1985; Garman et al., 1991; Hash, 2012; King & Burnsed, 2009; Latimer, Bergee, & Cohen, 2010), in particular for concert bands and orchestras. Brakel (2006) looked at three-member panels at the festivals run by the Indiana State School Music Association (ISSMA) in 2002 and 2003, and found higher interrater reliability for Group I versus Group III high school bands and orchestras in Indiana. Burnsed et al. (1985) found a lower interrater reliability for certain judging criteria (tone, intonation, balance, and musical effect), but high interrater

reliability on global scores given to concert bands in Virginia contests. Garman et al. (1991) looked at interjudge agreement and relationships between performance categories and final ratings in the context of orchestra festival evaluations in Dade County, Florida over five festivals in a seven-year period, and found a wide range of interrater reliability coefficients and for individual judging criteria. Latimer, Bergee, & Cohen (2010) investigated the reliability among adjudicators when looking at individual judging dimensions as well as global scores using a multidimensional weighted performance assessment rubric in Kansas state high school large-group (bands, choirs, and orchestras) festivals. Hash (2012) examined interrater reliability for senior division concert band contests sponsored by the South Carolina Band Directors Association from 2008 to 2010, and found that two-member panels judging sight-reading were more reliable than were three-member panels assessing the concert performances. Interestingly, these results are in contrast with earlier studies by Fiske (1975, 1977, 1983) and Bergee (2003), who recommended a minimum of seven and five adjudicators per panel, respectively, for acceptable reliability.

As can be seen from the above, there have been few studies on interrater reliabilities on choirs. The exception to this is a study by Napoles (2009) who studied the effect of excerpt duration and music education emphasis on ratings of recordings of children's choral performances. Napoles (2009) found that ratings of independent dimensions all correlated highly with global scores, and that there were very slight differences in the ratings given by instrumental majors and choral majors, and for 20-second excerpts and 60-second excerpts. The scarcity of research in the interrater

reliability in the context of live-performed choral adjudication means more investigation needs to be done in these areas.

Statement of Problem

There are many gaps in the research on interrater reliability, especially in live-performed, large-ensemble adjudication, and in particular for choirs. The current research suggests that interrater reliability tends to be high on global ratings of performances, but less robust on ratings of separate judging criteria such as rhythm or blend (Bergee, 2003; Ciorba & Smith, 2009).

Purpose of the Study

Therefore, the purpose of this study was to examine the interrater reliability of choral festival adjudication in order to add to the existing body of research on choral adjudication.

Research Questions

The specific questions addressed in this study are:

1. What were the descriptive statistics for types of choirs, group levels, and ratings awarded?
2. What were the interrater reliability coefficients for choral festival adjudication over a period of three years?

Delimitations

The study was delimited to the festivals and contests sponsored by the Indiana State School Music Association (ISSMA), and only to performances by high school choirs in the state of Indiana that participated in ISSMA Organization, State Qualification, and State Finals events in 2012, 2013 and 2014.

Definitions of Terms

Throughout this study, the following terms and their definitions will be used:

- Interrater/ Interjudge reliability: the concordance, or the degree of agreement among raters/ judges. In this particular setting, interrater reliability refers to how well the individual judges in a panel agree with each other.
- Ratings: the type of award given to the groups, i.e., Gold, Silver, Bronze, or Participation
- Points: the number score given to the groups based on the strength of their performance, e.g. 85 points out of a total possible of 90 points.

CHAPTER II

CRITICAL REVIEW OF RELATED RESEARCH LITERATURE

Assessment Tools and Interrater Reliability

Interrater reliability has been studied in relation to the type of assessment tool used in adjudication. Ciorba and Smith (2009) investigated the effectiveness of a multidimensional assessment rubric administered in instrumental and vocal undergraduate performance juries. The instrument (rubric) was crafted by a faculty panel from a small Midwestern university over a six-month period prior to the study. The panel comprised four experienced university faculty members with performance expertise in brass, woodwinds, keyboard, and voice. The panel identified three common dimensions of music performance that were applicable across all instrumental and vocal areas: (a) musical elements, (b) command of instrument, and (c) presentation. They then crafted five graduated descriptors outlining various levels of achievement for each dimension, into a five-point Likert-like scale. The rubric was piloted over two semesters under jury conditions, and changes to the rubric were made for the main study.

Students at the same Midwestern university ($N = 359$) were assessed using this rubric. The 359 student performances were assessed by 28 panels of judges ($N = 37$) who listened to each participant play for about 10 minutes, then independently scored the students using the rubric on pieces, etudes, scales, and sight-reading material, depending on the requirements of their performance area and their current level of performance expertise. In addition to assigning a score, judges also provided written comments, and a summative grade (based on a holistic impression of the students' performance). Rubric scores and grades awarded by each judge were averaged together to provide an overall

score for each scale dimension, a composite scores, and a grade for each student (Ciorba & Smith, 2009).

Interjudge reliability (calculated using Cronbach's alpha) was found to be moderate to high across all dimensions for all groups (except for one woodwind panel and one voice panel), and internal reliabilities were consistent within each performing area. There was a significant level of agreement among members of the panels (with only two exceptions of the woodwind and voice panels mentioned previously), with reliability coefficients for each scale dimension at above .70 (elements .70 to 1.0, command .71 to .97, presentation .70 to .98). Reliability coefficients for the composite scores ranged from .66 to .99, while reliability coefficients for grades ranged from .56 to 1.0. Based on a 4-point scale, the overall mean score for grades was relatively high (3.31), reflecting a negative skew. Scale dimensions, which were based on five-point scales, were distributed normally. In addition, standard deviations were narrower for grades than they were for scale dimensions in most groups (Ciorba & Smith, 2009).

A one-way multivariate analysis of variance (MANOVA) was carried out for scale scores and grades by participants' year in school. Results show a significant difference in scores by year, Wilks's ^ = .75, $F(6, 704) = 18.33$, $p < .01$. Analyses of variance (ANOVAs) on each dependent variable were conducted as follow-up tests to the MANOVA. Using the Bonferroni method, each ANOVA was tested at the .025 level. The ANOVA on the scale scores was significant, $F(3, 353) = 25.27$, $p < .01$, whereas the ANOVA on grades by year was nonsignificant, $F(3, 353) = .95$, $p < .42$. Pearson correlations among scale dimensions, composites, and grades were also calculated, with correlation among scale dimensions and composities high at .81-.89, and moderate

correlations among scale dimensions and grades at .64 to .72. Students at higher grade levels performed better on average than students at lower grade levels, and scores derived from the rubric were significantly correlated to students' year in school. This allows for the multidimensional assessment rubric to be applied to different grade levels to determine performance achievement over time (Ciorba and Smith, 2009).

Investigations of interrater reliability and rating scales included the use of rating scales in adjudication. Saunders and Holahan (1997) investigated the suitability of criteria-specific rating scales in selecting high school students for participation in an honors ensemble. Students (N = 926; 546 female and 380 male) enrolled in Grades 9-12 at public and private high schools in Connecticut served as subjects for the study. Only students who performed with woodwind and brass instruments were examined as their performances were assessed using the same evaluation form. Thirty-six judges, who were instrumental music specialists recruited from among Connecticut elementary, secondary, and college-level instrumental music teachers, attended a standardization session before evaluating the students. The judges all viewed a 15-minute videotape, followed by a question-and-answer session to clarify procedures for using the evaluation form.

The evaluation form included solo performance dimensions (tone, intonation, technique/ articulation, melodic accuracy, rhythmic accuracy, tempo, interpretation), scales (technique, note accuracy, musicianship) and sight-reading (tone, note accuracy, rhythmic accuracy, technique/ articulation, interpretation). Each criterion described a specific level of music skill, content, and performance-technique achievement. The sum of the scores for each of the performance dimensions made up the overall score for each student assessed with the form (Saunders and Holahan, 1997).

Pearson correlations for the seven prepared-piece dimensions were low to moderate (.46 to .65), for the scale performances were moderate (.58 to .75), and for the sight-reading were low to moderate (.36 to .61). Cronbach's alpha intrarater correlations were moderately high to high (median reliability = .915). Correlations of prepared-piece tone and sight-reading tone was .76, and correlations of prepared-piece interpretation and sight-reading interpretation was .71. Correlations between each performance dimension and the total score ranged from .54 to .75. Stepwise multiple regression indicated that student total scores could be predicted from scores of five individual dimensions (tone, technique/articulation, rhythmic accuracy, interpretation, and sightreading – interpretation). The results show that criteria-specific rating scales can be used to evaluate student woodwind and brass performances with substantial reliability (Saunders & Holahan, 1997). While my study is focused on interrater reliability using a standard rating scale provided by ISSMA, looking at interactions between each of the dimensions on the scale and reliability coefficients may provide clues as to how effective the current rating scales used by ISSMA are.

In a more relevant study to my research, Latimer, Bergee, and Cohen (2010) investigated the reliability and perceived pedagogical utility of a multidimensional weighted music performance assessment rubric used in Kansas state high school large-group festivals. The rubric was designed by an ad hoc committee appointed by the Kansas Music Educators Association (KMEA), and consisted of a committee chair, three choir directors, three band directors, three orchestra directors, the KMEA president, executive director, an at-large board member, and the person in charge of state music activities. The rubric was piloted at several district large-group festivals before a three-

year trial in Kansas. It consisted of nine point-weighted dimensions: Tone (15), Intonation (15), Expression (15), Technique or Diction (10), Rhythm (5), Note Accuracy (5), Balance (5), Blend (5), and Other (5). Each dimension was described on a five-point scale, and a total score for the performance was converted into ratings from I (outstanding) to V (ineffective). Adjudicators were also surveyed for years of experience as judges, whether they found the rubric effective, and on the weighting for each dimension on the rubric. Directors were surveyed on their teaching experience and their opinions of the rubric and weighting scale. The rubric was found to be internally consistent (Cronbach's alpha = .88), and correlations between each dimension and the total mark was moderate (r = .62) to moderately high (r = .87), with the exception of Other, which was moderately low (r = .46), as might be expected due to its vagueness.

In a similar study, Norris and Borst (2007) examined the reliabilities of two choral festival adjudication forms. Four choral music educators were asked to evaluate two performances of the same set of choirs, using two different adjudication forms. Form A, the "traditional" choir adjudication form, had a five-point scale (1- Excellent, 2- Good, 3- Satisfactory, 4- Poor, 5- Unsatisfactory), with no descriptors for each of the criteria of tone quality, diction, blend, intonation, rhythm, balance, and interpretation. Form B, an author-designed form adapted from a rubric used in Washington State, had descriptors on a five-point scale for the same criteria.

The four adjudicators used Form A in a morning session, then Form B in an afternoon session on the same day, to rate audio recordings of randomly-selected SATB choirs taken from a Michigan School Vocal Music Association high school district choral festival. The recordings of the same set of choirs were copied in two different random

orders; the first order was used for the morning session, and the second order was used for the afternoon session. Adjudicators were provided with scores, pencils, and copies of Form A (morning session) or Form B (afternoon session) (Norris & Borst, 2007).

The authors calculated means and standard deviations for each of the seven criteria on both forms, and computed *t*-tests for each dimension. They derived interrater reliability from an intraclass correlation coefficient (ICC), and computed ICCs using all four judges' scores as well as each of the four possible combinations of three judges. Results showed that each dimension on Form B was rated lower than its equivalent in Form A. Significant differences were found in favor of the rubric form for all measures except interpretation. Paired-sample *t*-tests showed significant differences between forms 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 ($t = .323$, $p = .002$) (Norris & Borst, 2007).

The ICCs on Form B were also stronger than their corresponding dimensions on Form A for every dimension except rhythm. Interrater reliability on Form B was .10 or higher in 34 out of 45 instances, and agreement on Form B was .15 or higher in 24 instances. The authors concluded that:

> rubrics containing dimension-specific descriptors could be better suited for the purposes of evaluating performance 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 (Norris & Borst, 2007, p. 249).

Factors Affecting Reliability

The number of adjudicators on a panel and the experience level of judges have also been investigated in relation to interrater reliability. Bergee (2003) investigated the faculty interjudge reliability of music performance evaluation on end-of-semester applied music solo performances. Prior to the study, a number of performance rating scales were found and adapted for the purpose of the study: a brass rating scale by Bergee (1988), a percussion rating scale by Nichols (1991), a woodwinds rating scale by Abeles (1973), a voice rating scale by Jones (1986), a researcher-developed rating scale for piano, and a strings rating scale by Zdzinski and Barnes (2002). Due to time constraints, subscales were limited to three items, which were then paired with Likert-type scales with response categories for Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree. Evaluators also had to assign a grade to each performance using a letter scale from A+ (an *excellent performance in all respects*) to F (an *exceedingly poor performance in all respects*).

Brass (n = 4), percussion (n = 2), woodwind (n = 5), voice (n = 5), piano (n = 3), and string (n = 5) instructors at a large university evaluated graduate and undergraduate music majors and minors in one semester. Full-panel interjudge reliability was found to be consistently good regardless of panel size (ranging from n = 2 to n = 5). All subscale interjudge reliabilities for all groups (except percussion) were statistically significant, with the exception of the suitability subscale in voice. All rating scale total score interjudge reliability coefficients were statistically significant, as were all for the global letter grade assessment. There was no loss of average reliability as group size incrementally decreased. No reliability coefficients were reported (Bergee, 2003).

Interjudge reliability in this study was stable and consistent, and good for rating scale total scores, subscales, and the global letter grade, especially among the larger panels. The amount of prior experience of the adjudicator (whether they were more or less experienced or whether they were teaching assistants or faculty members) had no apparent effect on reliability. As the reliability for larger panels was consistently found to be higher in this study, Bergee (2003) recommended the use of a minimum of five adjudicators for performance evaluation in this context. While the number of adjudicators used at the ISSMA festivals are standard at three adjudicators per panel, I am curious whether this has an effect on interrater reliabilities, given the findings by Bergee (2003).

In a study by Fiske (1977), the relationship between reliability of music performance adjudication, judge performance ability, and judge nonperformance music achievement was analyzed. Thirty-three subjects rated an audition tape recording of performances by 20 trumpeters, with the same recording used in the retest for intrarater reliability. Subjects rated the performances using five criteria: intonation, rhythm, technique, phrasing, and overall. Each performance was rated on a five-point scale for each of the five criteria. Data related to music knowledge and performance ability of the subjects were also obtained.

To measure intrarater reliability, correlations were run between the test scores and retest scores for each judge for each of the five criteria. A *t*-test was computed to compare ratings by brass versus non-brass judges. A trait intercorrelation matrix was created to examine relationships between average reliability coefficients, applied music grades, music history grades, and music theory grades, for individuals as well as whole panels (Fiske, 1977).

Individual judge stability was found to range from .32 to .82, with an average of .60. When brass judge reliability coefficients were compared with those of nonbrass judges, a *t* value of 2.113 was found, which was significant beyond the .05 level, suggesting judge reliability improves through teaching experience, particularly with instruments outside of the individual's specialty. Results of the statistical analyses showed no significant relationship between judge performing ability and judge reliability; no significant relationship between judge performing ability and judge nonperformance music achievement; and a statistically significant inverse relationship ($r = -.33$, $p < .05$) between judge reliability and judge nonperformance music achievement (as measured by music history and theory grades) (Fiske, 1977).

The adjudicators' experiences and abilities also present another aspect of variability in adjudication, in particular when many panels consist of adjudicators with mixed backgrounds and years of experience with adjudication or with the art form itself. Rogers (2004) investigated whether a select group of professional choral directors agreed on good choral tone, and whether there were differences in (a) the ratings given by novice directors and experienced directors, (b) the ratings given by high school choral directors and college choral directors, (c) the identification of choral problems when listening to the same taped examples, and (d) the remedies for perceived problems of choral tone.

Rogers (2004) prepared anonymous taped examples of choral selections of the music and types usually required for festival or contest participation. Recordings of 12 choirs were compiled into a master compact disc recording. A panel of 12 adjudicators of varying backgrounds and levels of experience (four college choral directors (COLL), four experienced high school choral directors with at least eight years of experience (HS8),

and four relatively new high school choral directors with less than four years of experience (HS4) was selected to evaluate the recordings. These adjudicators were selected based on their professional reputation as determined by their choirs' ratings at state and regional choral festivals and contests, or invitations received for their choirs to perform at professional divisional and national conferences. Adjudicators received a Choral Tone Evaluation Form and listened to the compact disc on a high-quality audio system. Prior to the adjudication session, information from a review of the literature, interviews with selected choral conductors at the high school and college levels, and a focus group of three choral directors (two college and one high school) helped to determine a list of components that characterized good choral tone: balance, breath support, flexibility, intonation, placement/resonance, relaxation of tone, uniform vowel sound, and vibrato. These same components were then used by the panel of adjudicators to evaluate the recordings using a five-point Likert-type scale for each component. Adjudicators were additionally asked to make recommendations for correcting identified problems in each performance.

ANOVAs on the Likert-type scores given by all 12 adjudicators produced non-significant results, indicating no significant differences in the scores given by each group (HS4, HS8, COLL). The resulting chi-square tests confirmed that the COLL group produced significantly more responses than the other two groups (except for the component "relaxation of tone"), and that those responses represented more discriminating hearing at a more refined level, indicative of adjudicators with a higher level of education and more extensive experience, both as a choral director and as adjudicator. The number of problems and solutions for each component by group also

confirms that experience appears to be a very strong indicator for adjudicators providing solutions to perceived problems. Chi-square comparisons of the totals for each component by category was significant (except for the component "relaxation of tone") (Rogers, 2004).

Other results from this study showed that choral directors agreed on examples of good choral tone, and evaluated choral tone in a consistent manner. Their experience, education, or teaching level (high school or college) did not produce any significant results when only the scores given by the adjudicators were compared. They also agreed on the top one or two content statements describing appropriate solutions to perceived problems for each component. Neither the amount of experience of the adjudicators (HS4 vs HS8) nor their teaching level (HS4, HS8, vs COLL) affected the actual scores provided for each component and the overall score. There was a very high level of agreement among all three groups of adjudicators on all components (except for "relaxation of tone"), and they also tended to use similar terminology in conveying their solutions. The college directors did, however, produce a significantly higher number of responses for the identified content statements, and more similar problems were identified by more of the college directors than in the other two groups (Rogers, 2004).

In a more directly relevant study to my research, Kinney (2009) investigated the effects of music experience and excerpt familiarity on the internal consistency of performance evaluations. Participants included undergraduate nonmusic majors (n = 63), undergraduate music majors (n = 42), graduate music majors (n = 17), and music faculty (n = 9). These were further categorized into nonparticipants (n = 28) (nonmusic majors who had no previous formal training in music beyond typical elementary/middle general

music curricula), and ensemble participants (n = 35) (nonmusic majors who had at least two years of formal study in a high school performing ensemble), based on their past music experiences and music training. Participants were played keyboard performances of three pieces, one of which was considered an unfamiliar excerpt. Each participant heard 45 excerpts, 15 of which were exact repetitions of a previous excerpt so that internal consistency could be calculated for each participant.

Participants responded to the excerpts by rating them on forms that included two 7-point Likert-type scale items for each stimulus: accuracy and musical expression. Internal consistency for each individual rater was calculated through Pearson product-moment procedures (r), correlating each individual participant's evaluations on the 15 repeated stimuli. Significant main effects were found for the variables of excerpt familiarity, $F(1, 92) = 55.54$, $p < .001$, and expertise, $F(2, 92) = 399.28$, $p < .0001$. Internal consistency means were significantly higher for familiar excerpts on the whole, although the difference between these means was not large (M = .38 to .33 respectively). Use of post hoc Scheffe procedures for multiple comparisons of expertise found that music majors' internal consistency was strongest (M = .62), followed by ensemble participants' (M = .35), and then nonparticipants' (M = .10). There was no significant main effect for order of stimuli presentation. There was also a significant two-way interaction between expertise and familiarity, $F(2, 92) = 8.32$, $p < .001$. Additionally, although all groups' internal consistency decreased when evaluating unfamiliar excerpts, differences between familiar and unfamiliar internal consistency means were smaller for music majors (mean difference = .02) than for ensemble participants and nonparticipants

(mean difference = .07 and .08 respectively). The nonparticipants' internal consistency mean for presentation order was also low at $M = .05$ to .13 (Kinney, 2009).

Results of this study (Kinney, 2009) suggest that internal consistency of performance evaluation is related to music experience and training, with more experienced groups demonstrating greater internal consistency across both accuracy and expression evaluations. Greater expertise was also associated with higher internal consistency, and with an ability to evaluate separate components of a music performance as opposed to a global rating.

Interrater Reliability in the Adjudication of Large Ensembles

Several studies looked at interrater reliability with particular focus on large-group adjudication. Brakel (2006) studied the reliability for the Indiana State School Music Association (ISSMA) Instrumental Festival (Bands and Orchestras) using the 2002 and 2003 population of adjudicators (n=43) and events (n=840). Prior to the adjudication, ISSMA conducted adjudicator training sessions "on a periodic basis" (Brakel, 2006) and sent a CD recording of the top three ensembles at the state festival from the previous year to the members of the panels in advance of the festival date. Adjudicators for the festival were selected based on a criteria of a minimum of three years' teaching experience. If the panel was adjudicating an orchestra, at least one member of the panel would be a string instrument specialist. Each panel consisted of three adjudicators; no panels consisted of two or more of the same members.

Results of this study suggest that a training session for adjudicators improved the overall reliability of the adjudication in 2003 as compared with the 2002 reliability. In the 2002 festival, reliability of all the panels ranged from α=.44 to α=.94, with a mean of

α=.82. In 2003, the reliability of all the panels ranged from α=.76 to .94 (mean α=.87). Judges within each adjudication panel in 2002 were found to have a positive correlation with at least one other judge, while judges within each adjudication panel in 2003 were all found to have significant correlations with the exception of one panel. In general, the strength of the correlations between pairs of judges was found to be lower than with the three judges combined. Inter-judge reliability was generally acceptable, especially in group I events, but some low and negative correlations were also found. Inter-judge reliability (Pearson *r* coefficient) according to the group level adjudicated ranged from *r* = -.12 to 1.00, while inter-judge reliability according to the type of organization adjudicated indicated fairly consistent reliability ($\overline{x}$ = .82) for band organizations, and less consistent reliability ($\overline{x}$ = -.23 in 2002 and $\overline{x}$ = .58 in 2003 for string orchestras; $\overline{x}$ = .79 in 2002 and $\overline{x}$ = .83 for full orchestras). Group I ensembles were found to have the highest degree of reliability, while Group III ensembles showed the lowest reliability. Contest point totals appeared to show greater inconsistency between judges when the performance was poor (Brakel, 2006).

In the most relevant study for large-group adjudication, Hash (2012) examined procedures for analyzing ratings and interrater reliability of high school band contests. Data from festivals sponsored by the South Carolina Band Directors Association (SCBDA) from 2008 to 2010 were collected, and analyzed for distribution of ratings among the bands, and for reliability of individual judging panels. Performance and Sight Reading ratings for senior division bands participating in SCBDA concert festivals from 2008 to 2010 (*N* = 353) were analyzed. The ratings were I (superior) to V (poor), which were then converted to points and added together to get a total score for each band. The

data included individual and final ratings by 45 adjudicators (27 concert performance, 18 sight-reading) from 18 judging panels (nine concert performance, nine sight-reading) at nine contest locations over the three-year period. Analysis involved nonparametric statistics as contest ratings were considered ordinal data.

Interrater reliability (IRC) was calculated through Spearman's rank order coefficient (to measure the extent to which individual judges' ratings moved in the same direction), Cronbach's alpha (to measure internal consistency for both concert-performance and sight-reading panels), and interrater agreement (IRA) between individual judges. All calculations of mean IRC involved Fischer's z transformation in order to control for underestimation (Hash, 2012).

Mean final ratings by site varied from 1.87 (SD = 0.72) to 1.51 (SD = 0.54) for an average of 1.73 (SD = 0.70) for all bands (N = 353) over the three-year period. Most of the bands (86.7%, n = 306) earned a final rating of I/ Superior (40.8%, n = 144) or II/ Excellent (45.9%, n = 162). Only 13.3% (n = 47) of the groups earned a III/Good (12.7%, n = 45) or IV/ Fair (0.6%, n = 2), and no bands earned a V/ Poor. Individual judges' scores also reflected a low variability in both concert performance and sight-reading ratings. Of the total number of individual judges' ratings issues in each event (concert performance, N = 1,059; sight-reading, N = 706), most (concert performance: 81.9%, n = 867; sight-reading: 88.4%, n = 624) were either a I/ Superior (concert performance: 37.7%, n = 399; sight-reading: 52.3%, n = 369) or II/ Excellent (concert performance: 44.2%, n = 468; sight-reading: 36.1%, n = 255). Only a small number of individual judges' ratings resulted in a III/ Good (concert performance: 15.7%, n = 166; sight-reading: 11.0%, n = 78), IV/ Fair (concert performance: 2.4%, n = 25; sight-reading:

0.6%, $n = 4$), or V/ Poor (concert performance: 0.1%, $n = 1$; sight-reading: 0.0%, $n = 0$). The average final ratings and the percentage of bands earning a I/ Superior were higher for each advancing classification with the exception of bands in Class 3 (Hash, 2012).

Interrater reliability for concert performance ranged from (Spearman) .44 to .86, with an average of .75. Internal consistency (Cronbach's alpha) ranged from .70 to .94, with an average of .89. Interrater reliability for sight-reading ranged from (Spearman) .65 to .95 with an average of .85. Internal consistency (Cronbach's alpha) ranged from .82 to .97, with an average of .91 among the nine contest locations (Hash, 2012).

Using Friedman ANOVAs, significant differences were found among individual judges' ratings within 8 of the 18 adjudication panels, indicating that some adjudicators graded at a higher degree of severity than others did. No significant differences were found between the mean final ratings for contests held in 2008, 2009, or 2010. Significant differences were also found in the mean final ratings for different classifications ($N = 353$, $df = 4$, $x^2 = 69.67$, $p < .001$). Thus, a post hoc analysis using a series of Mann-Whitney U tests was carried out to identify significant differences among these groups, and Bonferroni correction applied to control for the greater chance of Type I error that would result from multiple comparisons. The analysis revealed that Class 3 bands scored significantly lower than ensembles in Classes 4, 5, or 6 ($p < .001$), and that Class 6 bands received significantly higher ratings than all other classifications ($p < .001$). There was also a moderately low but significantly negative correlation between classification and final rating ($r = -.42$, $p < .001$), with the final rating higher for each advancing classification. No comparisons were done for Class 1 vs Class 2 bands (Hash, 2012).

My study will be looking at a similar type of festival in Indiana, for events of high-school mixed, treble, and men's choirs. I intend to loosely replicate the studies by Brakel (2006) and Hash (2012), by looking at three-year data at the ISSMA high school Organization, State Qualification, and State Finals events, and analyzing for interrater reliability.

Summary

This review of literature has included the topics of assessment tools and interrater reliability and factors affecting reliability. Some of the important findings are summarized as follows: (a) the use of criteria-specific assessment rubrics were shown to be reliable tools for adjudication (Ciorba & Smith, 2009; Latimer, Bergee, & Cohen, 2010; Norris & Borst, 2007; Saunders & Holahan, 1997); (b) there are currently conflicting results on adjudication panel sizes and their impact on interrater reliability, with some studies citing little difference between increases or decreases in panel size (Brakel, 2006), and others citing a need for a minimum number of adjudicators on the panel (Bergee, 2003; Fiske, 1977); (c) there are conflicting results on adjudicator expertise or familiarity with the music, with some studies citing that adjudicators' prior experience had no apparent effect on reliability (Bergee, 2003; Rogers, 2004), while others suggest that music experience and training had positive associations with internal consistency and increased ability to evaluate separate components of music performances (Kinney, 2009). However, research is still lacking on interrater reliability in large choral ensembles. Some of these discrepancies will be addressed in this study.

CHAPTER III

METHODOLOGY

Purpose

The purpose of this study was to examine the descriptive data and interrater reliability of choral festival adjudication in order to add to the existing body of research on choral adjudication.

Participants

In this study, I analyzed ratings and points awarded to 925 performances (689 discrete high school choirs) at the choral festivals sponsored by the Indiana State Schools Music Association (ISSMA) in 2012, 2013, and 2014. Ratings and points were awarded by a total of 58 panels over the three years of the festivals.

The choirs either registered for organization events (at the district level), or state qualification events. Choral directors or schools registered their choirs under one of the following Group Levels: I, II, III, IV, or V, based on the difficulty of the choir's repertoire (refer to Table 1).

Table 1
Classifications for ISSMA Organization and State Qualification Events Based Upon Grade Level, Difficulty of Music Performed, and Experience:

Group	Grade Levels	Difficulty of Music	Further Classification
V	5-8	Easy	First time performers. Minimum of 2 vocal parts for half of composition
IV	5-9	Easy	Minimum of 2 vocal parts
III	5-12	Medium Easy to Medium	Minimum of 3 vocal parts
II	5-12	Medium to Medium Difficult	Minimum of 3 vocal parts
I	5-12	Difficult	Music from current Required List

Vocal organizations joining Organization Events at the district level performed three numbers (pieces). Group I organizations performed one piece selected from the required list and two pieces of their own choice; one of the pieces was required to be a cappella. Groups II and III organizations performed one piece selected from the required list and two pieces of their own choice, while Group IV organizations performed three pieces of their own choice in the concert segment. One of the own-choice pieces for each group was required to be of the same grade level as the pieces in the required list.

Organizations that entered Group I, II, or III at the district High School level were required to sight-read. Group IV organizations had the option to sight-read for comment only. Organizations were to sight-read, a cappella, the designated rhythmic, melodic unison, and harmonic exercises.

Adjudicators for the Organization events and State Qualifying events were selected from current or retired choral directors with at least three years' teaching experience. Many of the adjudicators also had several years of adjudication experience. Each panel would consist of adjudicators with varied experience levels, so that participating groups would receive feedback from different perspectives. Adjudicators for the State Qualifying events would be trained using actual adjudication forms with sample audio recordings from the previous year's festival. Adjudicators at the State Finals were selected from experienced choral directors or university choral faculty members from outside of the state of Indiana.

Measure

This study was conducted using mainly quantitative collection tools. A face-to-face interview was conducted with the two head festival organizers – the Executive

Director and Assistant Executive Director of ISSMA – on adjudication-related processes and practices. Three-year data on the adjudication (individual judges' scores, total scores, adjudication procedures, rubrics or scales used) were compiled from each festival and analyzed for interrater reliability on concert ratings and points awarded scores for each choir. Interrater reliability was calculated from the adjudication data by individual sites.

ISSMA revised and copyrighted their organization rating form in 1999 (Brakel, 2006), with more minor revisions made to the form in intervening years between 1999 and 2014 (Briel, C., personal communication, May 27, 2014). In the revisions, the number of categories on the form was reduced from eleven to nine, with more equal emphasis/weight on each category. The ISSMA organizers reported that the interjudge reliability has increased tremendously since the rating form was revised. The forms remained the same during the data collection period for this study.

Organization events. A panel of three adjudicators were used for the concert segment of Organization Events. One judge provided recorded (audio) comments only and the other two provided written comments only. Concert segment adjudicators assigned between one (outstanding in nearly every detail) to four (unacceptable in nearly every detail) marks each to nine categories of musical criteria: Intonation, Tone Quality and Blend, Breathing Technique, Note Accuracy, Rhythmic Accuracy, Diction and Enunciation, Dynamics and Balance, Interpretation and Musicianship, and Other Factors. The total marks awarded were then converted to ratings: 9 – 13.5 marks for Gold Division, 14 – 18 marks for Silver Division, 18.5 – 22.5 marks for Bronze Division, and 23 or more marks for Participation. The three adjudicator's ratings were then converted into one resultant concert rating for the group via a conversion table (see Table 2).

Only one adjudicator was used for the sight-reading segment. Sight reading adjudicators assigned between one (outstanding in nearly every detail) to four (continuous major flaws) marks each to 11 criteria in four categories: Rhythmic Exercise, Melodic Exercise, Harmonic Exercise, and General Effect (overall). This was done on a seven-point scale, with marks ranging from 1 to 4 in half-point intervals. The total marks awarded were then converted to ratings: 11 – 16.5 marks for Gold, 17 – 22 marks for Silver, 22.5 – 27.5 marks for Bronze, and 28 or more points for Participation. The final rating was determined by a combination of the final concert rating and the sight-reading rating, which was computed according to the conversion table (see Table 2).

Table 2
Table for Converting Ratings at ISSMA Choral Contest

For Concert Rating Three Judges, Four Ratings – Every Possible Combination				For Final Rating Column C refers to concert rating, column SR refers to sight-reading rating							
Gold	**Silver**	**Bronze**	**Participation**	**Gold**		**Silver**		**Bronze**		**Participation**	
GGG	GSS	GBB	GPP	**C**	**SR**	**C**	**SR**	**C**	**SR**	**C**	**SR**
GGS	GSB	GBP	SPP	G	G	G	B	S	P	P	B
GGB	GSP	SBB	BPP	G	S	G	P	B	S	P	P
GGP	SSS	SBP	PPP			S	G	B	B		
	SSB	BBB				S	S	B	P		
	SSP	BBP				S	B	P	G		
						B	G	P	S		

State qualification events. State Qualification Events were open to Group I organizations only. For State Qualification Events, each choir performed two required numbers (from the current ISSMA Group I required list for the type of organization) and one piece of their own choice, one of which had to be a cappella. Adjudication of the concert segment was by three judges. All three adjudicators provided audio recorded comments only. A separate sight-reading judge also provided recorded comments. Each choir sight-read, a cappella, the designated rhythmic, melodic unison, and harmonic

exercises. Concert segment adjudicators awarded up to 10 marks for each of nine categories of musical criteria: Intonation, Tone Quality and Blend, Breathing Technique, Note Accuracy, Rhythmic Accuracy, Diction and Enunciation, Dynamics and Balance, Interpretation and Musicianship, and Other Factors. The total marks awarded were then converted to ratings: 54 or more points (out of a total possible of 90 points) for Gold, 41 – 53 points for Silver, 32 – 40 points for Bronze, and 31 or less points for Participation. The three adjudicator's ratings were then converted into one final concert rating for the group via the same conversion table as in Table 2.

Sight-reading adjudicators assigned between one (Participation level performance) to four (Gold level performance) marks in half-point intervals (a 7-point rating scale) to four categories: Rhythmic Exercise, Melodic Exercise, Harmonic Exercise, and General Effect. The total points awarded were then converted to ratings: 44 – 38.5 points for Gold, 38 – 27.5 points for Silver, 27 – 16.5 points for Bronze, and 16 or less points for Participation. It is worth noting that the sight-reading marking rubrics for the State Qualification events are reversed, with 1 being the most desirable and 4 being the least desirable mark awarded. This could potentially lead to errors in marking sight-reading events, especially since many of the adjudicators for sight-reading at the State Qualification events had also adjudicated at the State Organization events.

State Finals. Choirs for the State Finals are selected from the State Qualification events each year, where all groups registered and performed as Group I ensembles. The top 16 choirs in the Mixed choirs category, and the top eight choirs in the Treble and/or Men's choirs category would compete at the State Finals. The State Finals utilizes the

following criteria and process until all the top 16 Mixed and eight Treble/ Men's choirs have been selected:

a) Best composite score from 4 judges (3 concert and 1 sight-reading judge)
b) Best composite score from 3 concert judges
c) Best two concert scores
d) Score from the head judge only
e) The flip of a coin

A draw would determine the order of performance of these selected State Finals choirs.

Each of these 24 choirs performed two required numbers (from the current ISSMA Group I required list for the type of organization) and one piece of their own choice. These pieces may or may not be the same as the pieces used at the State Qualification performance. Adjudication of the concert segment was by three judges. All three adjudicators provided audio recorded comments as well as a written summary of the performance. No sight-reading is required at the State Finals. Judges conferred after hearing four (4) organizations to establish a standard along national lines. Thereafter, each judge adjudicated independently without further conferring with the other judges.

Adjudicators awarded up to 30 marks for each of three categories of musical criteria: Technique (Intonation, Tone Quality, Blend, and Breathing), Accuracy (Note Accuracy, Rhythmic Accuracy, Diction and Enunciation), General Musicianship (Dynamics, Balance, and Interpretation), and up to 10 marks for the category of "Other Factors" (Stage Presence, Poise, Posture, and Concert Decorum), for a total of 100 points. Each adjudicator's raw scores were then converted to ranks, with the best score ranked

"1"; the lowest score would be given the lowest rank of 16 (in the Mixed choirs category) or 8 (in the Treble/ Men's choirs category). Adjudicators were provided with a tote sheet as well as index cards to ensure that they did not give the same point total to two different choirs. The three rankings for each choir (one from each adjudicator) were then totaled to determine the final rank score for each choir, with the lowest total being the best rank score. For example, a choir that received rankings of 3, 3, and 2 would have a total final rank score of 8, which is a better final ranking than another choir with rankings of 1, 5, and 6 (resulting in a total rank score of 12). In the event of a tie, the judges' rank preference will be used to establish a "best two out of three" comparison between the two affected organizations. In the event of a three-way tie where the ranking preference will not resolve the tie, the best raw score will be used to determine placing. Any remaining ties will result in the duplication of the award.

Procedure

I arranged for one face-to-face meeting with the festival organizers, who then sent adjudication data via a secure web file delivery to my receiving account. The adjudication data collected were then stored in a portable hard disk drive and categorized by festival type, year, and group level of adjudication. SPSS 22 was used to generate results for interrater reliability and correlation coefficients.

Descriptive data were compiled for ratings and points awarded to each performance at individual sites in each year, all sites across each year, as well as all sites across three years of the festival. Each adjudicator's ratings were converted to points (Participation = 0, Bronze = 1, Silver = 2, Gold = 3) for data analysis. Statistics were

calculated for interrater reliability for each individual site each year, as well as interrater reliability for all the festivals sites each year.

In the ISSMA high school choral contests, choirs register either for Organization events (at the district level) or State Qualification events. Choral directors or schools register their choirs under one of the following Group Levels: I, II, III, IV or V, based on the difficulty of the music performed. Some schools had more than one choir that were registered under different Group Levels (for example, one of their choirs would be registered as a Group I choir, while another choir would be registered as a Group III choir). Some schools also had more than one type of choir that were registered under different categories (for example, one choir would be an SATB choir while another would be a SSA choir). Adjudicators were in panels of three (performance events) or one (sight-reading). There was a mix of experienced and "new" adjudicators, although the criteria for being an adjudicator was at least three years of choral teaching experience. Some adjudicators judged at more than one site, while some adjudicated both performance events and sight-reading at different sites in the same year of the contests.

Data Analysis

In analyzing the interrater reliability of the adjudicators, I made the assumption that the number of adjudicators would not affect the interrater reliability. Because ISSMA uses three-adjudicator panels as a standard in both their regional and state festivals, the number of adjudicators for all their festivals remains constant and would not impact the analysis. As suggested in the existing literature, interjudge reliability for three-person panels was acceptable, and there was no significant variation in reliability between three or five judges; increasing the number of adjudicators on a panel would only result in

marginal increases in interrater reliability (King & Burnsed, 2009). No reliability analysis was done for the sight-reading adjudicators as ISSMA uses only one sight-reading adjudicator per site. Data for this study included individual ratings and points awarded by a total of 58 panels of adjudicators over the three years of the festivals. Nonparametric statistics were used in the analysis of the data because contest ratings are considered to be ordinal data (Bergee & Westfall, 2005; Phillips, 2008). Interrater reliability were calculated for the Organization, State Qualification, and State Finals events using several methods:

1) Because adjudicators often exactly agreed in their ratings of choirs, pairwise interrater correlations (IRC) based on concert points awarded were calculated. Spearman's rank order coefficients (r_s) were calculated for Judge 1-Judge 2, Judge 2-Judge 3, and Judge 1-Judge 3 pairings, then the average r_s for the three pairings was calculated via a z-transformation.
2) Reliability for each three-member panel as determined using Cronbach's alpha (α), using points awarded by each adjudicator.
3) Intraclass correlation coefficient (ICC) (2-way random) for each three-member panel of judges for points awarded. ICC provides a composite of interrater and intrarater variability, and provides an estimate of the panel's agreement.
4) Friedman's Chi-Square analysis examined differences in points awarded among judges in each site. This method was selected due to the number of adjudicators on each panel (three), and the abnormal distribution of points (Brakel, 2006; Hash, 2012) awarded in a contest setting.

An additional analysis was done on the ratings from Organization and State Qualification events only; there were no ratings at the State Finals;

1) Percentage agreement between the ratings awarded by pairs of judges. This was calculated for Judge 1-Judge 2, Judge 2-Judge 3, and Judge 1-Judge 3 pairings, as well as Judge 1-Final Rating, Judge 2-Final Rating, and Judge 3-Final Rating pairings.

CHAPTER IV

RESULTS AND DISCUSSION

Results

Frequency distributions. Data from the three years of ISSMA high-school choral contests were analyzed for their frequency distributions. Table 3 shows that the number of panels, performances, and groups were approximately equal over the three years.

Table 3
Descriptive Statistics for Numbers of Panels, Performances, and Groups by Year

Year	No. of panels (Org, Quals, Finals)	No. of performances (Org, Quals, Finals)	No. of discrete groups (Org, Quals, Finals)	No. of groups (Finals only)
2012	19	306	226	24
2013	17	295	218	24
2014	22	324	245	24
Total	58	925	689	72

As detailed in Table 2, final ratings awarded to each choir are arrived at using a conversion table. There was a higher proportion of choirs awarded final Gold ratings (77%, n = 712) and Silver ratings (22%, n = 202) than other ratings (see Table 4). Only 11 choirs (1%) were awarded Bronze ratings, and no choir received Participation ratings during the three years.

Table 4
Descriptive Statistics for Composite Final Ratings (Type of Award) by Year

Year	Gold	% of total	Silver	% of total	Bronze	% of total	Participation	% of total
2012	239	78.10	66	21.57	1	0.33	0	0
2013	227	74.18	61	19.93	7	2.29	0	0
2014	246	80.39	75	24.51	3	0.98	0	0
Total	712	76.97	202	21.84	11	1.19	0	0

Similarly, in sight-reading assessments, there was a high proportion of choirs in Groups I, II, and III over the three years that received Gold ratings (55%, *n* = 508), and Silver ratings (27%, *n* = 249). Only 3% of choirs received Bronze ratings (*n* = 30), and less than 1% received Participation ratings (*n* = 4). A proportion of the choirs in Group IV (14%, *n* = 134) elected not to do sight-reading, or did sight-reading for comments only (see Table 5).

Table 5
Descriptive Statistics for Organization Events Sight-Reading Ratings (Type of Award) by Year

Year	Gold	%	Silver	%	Bronze	%	Partici pation	%	Comment Only	%	No SR	%
2012	168	54.9	87	28.4	7	2.3	0	0.0	19	6.2	25	8.2
2013	158	53.6	78	26.4	17	5.8	3	1.0	11	3.7	28	9.5
2014	182	56.2	84	25.9	6	1.9	1	0.3	25	7.7	26	8.0
Total	508	54.9	249	26.9	30	3.2	4	0.4	55	5.9	79	8.5

Frequency distributions for type of choirs and group self-selection levels were also calculated. There were more mixed and treble choirs than there were men's choirs. Of the 925 performances, almost 60% were by mixed choirs, 34% were by treble choirs, and very few (6%) were by men's choirs (see Table 6).

Table 6
Descriptive statistics for frequencies by type of choir

Choir Type	Frequency	Percent
Mixed	551	59.6
Treble	317	34.3
Men's	57	6.2
Total	925	100.0

About a third of the choirs entered at Group I level (39%, *n* = 358), and another third of the choirs entered at Group III level (30%, *n* = 276). Notably fewer entered at the Group

II (17%, *n* = 157) and Group IV (15%, *n* = 134) levels (see Table 7). No choir registered at the Group V level.

Table 7
Descriptive statistics for frequencies by group level

Group Level	Frequency	Percent
I	358	38.7
II	157	17.0
III	276	29.8
IV	134	14.5
Total	925	100.0

*Interrater reliability: Pairwise percentage agreement in **ratings** from Organization and State Qualification events.* Mean percentage agreement was calculated by taking the average of the three pairwise percentage agreements in the panel. Forty-one out of the 58 panels (71%) had a mean percentage agreement of > 70%, indicating greater than moderate agreement on ratings within the panel. The less than moderate agreement on ratings in the other panels was usually a result of one, but in some cases two, adjudicators in the panel whose ratings were disagreeing with the others. Pairwise percentage agreements were calculated for the following pairs: Judge 1-Judge 2, Judge 2-Judge 3, Judge 1-Judge 3. Percentage agreements were also calculated for Judge 1-Final rating, Judge 2-Final rating, and Judge 3-Final rating. A hundred and fifty-five out of the 174 pairs (89%) had percentage agreements of > 60%. Percentage agreement for ratings between pairs of adjudicators, and between individual adjudicators' ratings and the final ratings were mostly moderate (ranging from agreements of 70% - 79%) to high (agreements of 80% - 89%). Some pairs of adjudicators achieved excellent to perfect agreement (agreements of 90% - 100%). The 19 pairs of adjudicators that showed low

agreement (< 60% agreement) were mainly in the Organization Events and not the Qualification Events (see Table 8).

Table 8
Percentage agreement between pairs of judges' ratings, and between each judge's rating and final rating

2012 Organization Events								
Site	*n*	% J1, J2	% J2, J3	% J1, J3	Mean %	% J1, Final	% J2, Final	% J3, Final
1	22	95.45	90.91	86.36	90.91	90.91	95.45	86.36
2	15	100.00	73.33	73.33	82.22	100.00	100.00	73.33
3	18	61.11	61.11	61.11	61.11	77.78	77.78	83.33
4	11	*54.55	72.73	63.63	63.64	72.73	81.82	90.91
5	15	86.67	*53.33	60.00	66.67	93.33	86.67	66.67
6	12	100.00	91.67	91.67	94.45	100.00	100.00	91.67
7	9	77.78	77.78	77.78	77.78	88.89	88.89	88.89
8	17	70.59	82.35	82.35	78.43	88.24	82.35	94.12
9	22	95.45	90.91	95.45	93.94	95.45	90.91	90.91
10	15	80.00	73.33	86.67	80.00	93.33	80.00	93.33
11	22	*54.55	72.73	63.63	63.64	54.55	81.82	72.73
12	17	70.59	70.59	76.47	72.55	88.24	82.35	88.24
13	13	*53.85	*46.15	69.23	56.41	92.31	61.54	76.92
14	18	72.22	83.33	77.78	77.78	83.33	88.89	94.44

2013 Organization Events								
Site	*n*	% J1, J2	% J2, J3	% J1, J3	Mean %	% J1, Final	% J2, Final	% J3, Final
1	17	*47.06	*58.82	76.47	60.78	82.35	64.71	94.12
2	16	75.00	81.25	68.75	75.00	81.25	87.50	81.25
3	15	66.67	73.33	66.67	68.89	80.00	86.67	86.67
4	12	83.33	83.33	66.67	77.78	83.33	100.00	83.33
5	5	100.00	*40.00	*40.00	60.00	80.00	80.00	40.00
6	14	78.57	78.57	71.43	76.19	85.71	92.86	85.71
7	17	70.59	70.59	64.71	68.63	70.59	64.71	58.82
8	18	100.00	83.33	83.33	88.89	100.00	100.00	83.33
9	20	75.00	70.00	95.00	80.00	100.00	75.00	95.00
10	24	91.67	95.83	87.50	91.67	91.67	100.00	95.83
11	10	80.00	90.00	70.00	80.00	80.00	100.00	90.00
12	26	76.92	69.23	92.31	79.49	73.01	57.69	73.08
13	24	100.00	66.67	66.67	77.78	100.00	100.00	66.67

Note: *less than 60% agreement

(Table continued on next page)

(Table 8 continued)

2014 Organization Events

Site	*n*	% J1, J2	% J2, J3	% J1, J3	Mean %	% J1, Final	% J2, Final	% J3, Final
1	25	80.00	80.40	80.00	80.13	92.00	88.00	88.00
2	11	*36.36	63.64	*45.45	48.48	63.64	72.73	72.73
3	13	*53.85	84.62	61.54	66.67	61.54	92.31	92.31
4	9	88.89	88.89	77.78	85.19	88.89	100.00	88.89
5	10	80.00	100.00	80.00	86.67	80.00	100.00	100.00
6	16	62.50	62.50	100.00	75.00	100.00	62.50	100.00
7	7	85.71	*42.86	*57.14	61.90	100.00	85.71	57.14
8	16	75.00	87.50	62.50	75.00	75.00	100.00	87.50
9	7	85.71	85.71	71.43	80.95	71.43	85.71	71.43
10	13	61.54	*53.85	*38.46	51.28	69.23	84.62	69.23
11	15	66.67	*53.33	86.67	68.89	100.00	66.67	86.67
12	17	64.71	64.71	88.24	72.55	94.12	70.59	94.12
13	21	85.71	80.95	85.71	84.12	95.24	90.48	90.48
14	17	94.12	82.35	88.24	88.24	100.00	94.12	88.24
15	12	91.67	66.67	75.00	77.78	91.67	83.33	83.33
16	11	63.64	81.82	63.64	69.70	63.64	81.82	81.82
17	25	68.00	64.00	72.00	68.00	88.00	80.00	84.00

2012 Qualification Events

Site	*n*	% J1, J2	% J2, J3	% J1, J3	Mean %	% J1, Final	% J2, Final	% J3, Final
1	28	96.43	96.43	92.86	95.24	96.43	100.00	69.43
2	17	100.00	94.12	94.12	96.08	100.00	100.00	94.12
3	8	100.00	100.00	100.00	100.00	100.00	100.00	100.00
4	27	81.48	88.89	85.19	85.19	88.89	100.00	96.30

2013 Qualification Events

Site	*n*	% J1, J2	% J2, J3	% J1, J3	Mean %	% J1, Final	% J2, Final	% J3, Final
1	28	96.43	100.00	96.43	97.62	96.43	100.00	100.00
2	10	60.00	60.00	90.00	70.00	90.00	60.00	100.00
3	15	100.00	100.00	100.00	100.00	100.00	100.00	100.00
4	24	91.67	91.67	91.67	91.67	95.83	95.83	95.83

Note: *less than 60% agreement

(Table continued on next page)

(Table 8 continued)
2014 Qualification Events

Site	*n*	% J1, J2	% J2, J3	% J1, J3	Mean %	% J1, Final	% J2, Final	% J3, Final
1	29	89.66	93.10	89.66	90.81	93.10	96.55	96.55
2	8	87.50	87.50	87.50	87.50	75.00	87.50	75.00
3	13	100.00	92.31	92.31	94.87	100.00	100.00	92.31
4	29	*58.62	*51.72	86.21	65.52	96.55	62.07	89.66

Note: *less than 60% agreement

It is worth noting that the number of groups adjudicated at each site were mostly fewer than 20 groups, with some sites seeing less than 10 groups. Thus the percentage agreement might seem disproportionately low or high for fewer differences in ratings between the pairs of adjudicators, or between each adjudicator's rating and the final rating. Furthermore, the probability of adjudicators agreeing on ratings is rather high due to the range of points available within each category of ratings. However, percentage agreements at less than 70% are still considered to be moderately low, and percentage agreements less than 50% indicate low or unacceptable agreement in ratings between adjudicators.

*Interrater reliability: Pairwise interrater reliability correlations (IRC) on **points**.* Interrater reliability correlations (IRC) for concert points awarded were calculated for pairs of judges. These pairwise correlations were then put through z-transformations in order to find the mean IRC for the three pairs of judges. While mean IRCs were all positive, there was a large range for correlation coefficients from weak (r_s = .155) to strong (r_s = .939). The mean pairwise interrater IRC for concert points awarded ranged from a Spearman (r_s) coefficient of .170 to .865 in 2012 Organization events; .047 to 1.000 in 2013 Organization events, and; -.102 to .982 in 2014 Organization events. IRC

for concert points awarded ranged from a Spearman (r_s) coefficient of .000 to .964 for all three years of Qualification events, and from -.120 to .826 for all three years of State Finals Events. IRC for Qualification events were more consistently in the moderate ($r_s > .40$) to strong ($r_s > .80$) ranges, while the Organization events had more instances of IRCs in the weak and very weak ranges ($r_s < .40$). IRC for State Finals events were generally in the moderate range ($r_s = .40$ to .60), indicating only moderate interrater reliability at the State Finals (see Table 9).

Table 9
Pairwise Interrater Reliability Correlations (IRC) (using Spearman's r_s) by Site or Category

2012 Organization Events					
Site	*n*	J1, J2	J2, J3	J1, J3	Mean[1]
1	22	.170	.384	.321	.294
2	15	.860	.711	.714	.772
3	18	.255	.347	.337	.314
4	11	.586	.557	.584	.576
5	15	.063	.418	.283	.260
6	12	.338	.534	.410	.431
7	9	.658	.872	.524	.718
8	17	.718	.827	.674	.747
9	22	.721	.788	.835	.786
10	15	.653	.865	.555	.720
11	22	.287	.711	.559	.541
12	17	.553	.399	.767	.595
13	13	.731	.402	.280	.499
14	18	.351	.462	.273	.364

Note: Mean[1] : calculated using the Fisher *z*-transformations of r_s

(Table continued on next page)

(Table 9 continued)

2013 Organization Events

Site	*n*	J1, J2	J2, J3	J1, J3	Mean[1]
1	17	.560	.680	.563	.604
2	16	.838	.822	.839	.833
3	15	.733	.584	.578	.638
4	12	.643	.795	.737	.731
5	5	.667	1.000	.667	.491
6	14	.723	.582	.718	.679
7	17	.589	.520	.397	.506
8	18	.874	.527	.567	.696
9	20	.612	.559	.721	.636
10	24	.784	.603	.505	.647
11	10	.803	.629	.856	.779
12	26	.721	.770	.824	.775
13	24	.391	.047	.301	.251

2014 Organization Events

Site	*n*	J1, J2	J2, J3	J1, J3	Mean[1]
1	25	.614	.551	.635	.601
2	11	-.102	.850	.098	.394
3	13	.434	.625	.627	.568
4	9	.712	.281	.646	.571
5	10	.452	.637	.738	.622
6	16	.497	.568	.745	.615
7	7	.734	.982	.667	.877
8	16	.840	.846	.852	.846
9	7	.908	.495	.718	.757
10	13	.302	.138	.299	.248
11	15	.713	.682	.572	.660
12	17	.825	.768	.705	.770
13	21	.744	.760	.721	.742
14	17	.561	.732	.690	.667
15	12	.661	.683	.296	.568
16	11	.875	.896	.953	.915
17	25	.686	.590	.506	.599

Note: Mean[1] : calculated using the Fisher *z*-transformations of r_s

(Table continued on next page)

(Table 9 continued)

2012 Qualification Events

Site	*n*	J1, J2	J2, J3	J1, J3	Mean[1]
1	28	.825	.548	.656	.695
2	17	.810	.766	.792	.790
3	8	.491	.503	.738	.591
4	27	.656	.712	.768	.715

2013 Qualification Events

Site	*n*	J1, J2	J2, J3	J1, J3	Mean[1]
1	28	.711	.554	.570	.617
2	10	.596	.632	.725	.654
3	15	.544	.331	.000	.308
4	24	.747	.667	.791	.739

2014 Qualification Events

Site	*n*	J1, J2	J2, J3	J1, J3	Mean[1]
1	29	.702	.710	.790	.737
2	8	.934	.903	.964	.939
3	13	.625	.841	.642	.720
4	29	.409	.729	.669	.619

2012 State Finals Event

Category	*n*	J1, J2	J2, J3	J1, J3	Mean[1]
Mixed	16	.579	.826	.659	.704
Treble/ Men's	8	.667	.452	.571	.570

2013 State Finals Event

Category	*n*	J1, J2	J2, J3	J1, J3	Mean[1]
Mixed	16	.474	.697	.426	.544
Treble/ Men's	8	.455	.405	-.120	.261

2014 State Finals Event

Category	*n*	J1, J2	J2, J3	J1, J3	Mean[1]
Mixed	16	.556	.635	.453	.552
Treble/ Men's	8	.310	-.048	.452	.155

Note: Mean[1] : calculated using the Fisher *z*-transformations of r_s

Since the likelihood of adjudicators agreeing on ratings (Gold, Silver, Bronze, Participation) is high, calculations of pairwise IRC on points awarded by adjudicators give a clearer picture of true agreement between adjudicators.

Interrater reliability: Panel internal consistency, intraclass correlation coefficient (ICC) and interrater differences (Friedman's Chi-Square analysis). Due to the high likelihood of adjudicators awarding the same ratings to performances at festivals, I also examined the internal consistency for each **panel** as determined using Cronbach's alpha (α) for points awarded by each adjudicator. In the Organization and State Qualification events, internal consistency was moderate (α= .55) to high (α= .94) in 2012, high (α= .73 to .94) in 2013, and moderate (α= .50) to high (α= .96) in 2014. In the State Finals events, internal consistency was moderate (α= .48) to moderately high (α= .86) over the three years (see Table 10).

In addition to looking at internal consistency for each panel, I also calculated the Intraclass correlation coefficient (ICC) (2-way random) for each three-member panel of judges for points awarded, as an estimate of the agreement in each panel. By using a fully-crossed (Rater x Choir), 2-way ANOVA design, I considered the three adjudicators in each panel to be a random sample from a population of all the adjudicators in the ISSMA contests, thus estimating the reliability of the larger population of adjudicators. ICC (2,3), using average measures for each panel at a confidence level of .95, ranged from a low of .381 (fair agreement) to a high of .923 (almost perfect agreement) in the three years of events. A majority of the ICCs were in the strong (0.7 - 0.8) to almost perfect (> 0.8) agreement ranges, indicating very good agreement by panel. It is worth

noting that the Treble/ Men's category in the State Finals events showed only fair agreement within the panel in 2013 (.495) and 2014 (.409) (see Table 10).

A further Friedman's Chi-Square (χ^2) analysis examined the differences in the mean ranks of points awarded among individual judges in each site (degrees of freedom = 2). Due to relatively small sample sizes, I set the criterion of $p < .01$ as being statistically significant in order to account for random error. In both Organization and Qualification events, significant differences ($p < .01$) were found among individual judges' points awarded within 3 out of 18 sites in 2012; within 4 out of 17 sites in 2013, and; within 5 out of 21 sites in 2014. In the State Finals events, significant differences ($p < .01$) were found among individual judges' points awarded in the Mixed choirs category in 2012 and 2014 (see Table 10). These significant differences suggest that some of the adjudicators were harsher or more lenient than others in their assessment of the choirs. Due to the relatively small number of sites with significant differences, no further tests on this data were required.

Table 10
Interrater reliability: Panel internal consistency using Cronbach's alpha (α), intraclass correlation coefficient (ICC) and interrater differences (Friedman's Chi-Square analysis)

Organization Events

Year	Site	*n*	α	ICC	χ^2	Sig
2012	1	22	.615	.586	3.949	.139
	2	15	.927	.923	4.075	.130
	3	18	.552	.460	10.971	.004
	4	11	.807	.683	12.293	.002
	5	15	.720	.658	8.373	.015
	6	12	.752	.749	2.311	.315
	7	9	.870	.848	4.667	.097
	8	17	.833	.812	5.828	.054

Note: *Significant at $p < .01$

(Table continued on next page)

(Table 10 continued)

Year	Site	*n*	α	ICC	χ^2	Sig
	9	22	.903	.891	14.000	.001
	10	15	.858	.855	4.440	.109
	11	22	.800	.708	18.667*	.000
	12	17	.805	.805	3.781	.151
	13	13	.758	.589	15.469*	.000
	14	18	.562	.396	17.768*	.000
2013	1	17	.864	.866	.646	.724
	2	16	.898	.892	1.458	.482
	3	15	.788	.780	3.949	.139
	4	12	.868	.862	1.378	.502
	5	5	.824	.624	9.000	.011
	6	14	.938	.862	16.510*	.000
	7	17	.734	.727	.406	.816
	8	18	.865	.786	14.613	.001
	9	20	.840	.804	8.553	.014
	10	24	.874	.863	6.432	.040
	11	10	.808	.813	.514	.773
	12	26	.919	.872	23.526*	.000
	13	24	.507	.381	24.261*	.000
2014	1	25	.806	.807	3.889	.143
	2	11	.556	.559	3.619	.164
	3	13	.810	.765	4.531	.104
	4	9	.820	.709	11.636	.003
	5	10	.853	.657	15.846*	.000
	6	16	.833	.774	10.262	.006
	7	7	.934	.844	11.185	.004
	8	16	.935	.910	10.475	.005
	9	7	.869	.728	9.478	.009
	10	13	.496	.466	2.980	.225
	11	15	.837	.485	27.138*	.000
	12	17	.863	.804	12.594	.002
	13	21	.899	.901	.800	.670
	14	17	.873	.879	.769	.681
	15	12	.923	.922	.585	.746
	16	11	.961	.873	17.714*	.000
	17	25	.730	.736	1.326	.515

Note: *Significant at $p < .01$

(Table continued on next page)

(Table 10 continued)

State Qualification Events

Year	Site	*n*	α	ICC	χ^2	Sig
2012	1	28	.877	.870	5.679	.058
	2	17	.939	.939	1.164	.559
	3	8	.940	.895	7.548	.023
	4	27	.861	.859	1.431	.489
2013	1	28	.801	.799	.716	.699
	2	10	.887	.705	16.667*	.000
	3	15	.827	.805	2.528	.282
	4	24	.904	.895	8.600	.014
2014	1	29	.896	.876	14.000	.001
	2	8	.959	.756	14.000	.001
	3	13	.770	.782	1.755	.416
	4	29	.794	.737	18.294*	.000

State Finals Events

Year	Category	*n*	*α*	ICC	χ^2	Sig
2012	Mixed	16	0.864	0.799	11.079*	0.004
	Treble/ Men's	8	0.815	0.702	6.250	0.044
2013	Mixed	16	0.785	0.782	4.871	0.088
	Treble/ Men's	8	0.476	0.495	0.839	0.657
2014	Mixed	16	0.811	0.700	13.875*	0.001
	Treble/ Men's	8	0.519	0.409	6.250	0.044

Note: *Significant at $p < .01$

Discussion

Distribution of choir types. The proportion of men's choirs (6.2%) at the ISSMA choral contests possibly indicates a difficulty with recruiting enough male singers to make up more men's choirs, or that male singers preferred to be part of mixed choirs (59.6%). However, these statistics are probably not dissimilar at other such contests. Several authors have discussed the negative perceptions and struggles of adolescent boys

in choirs (Demorest, 2000; Eshelman, 1992; Freer, 2011, 2012), including how choral singing is perceived to be a less "masculine" activity (Dibben, 2002; Hall, 2005; Lucas, 2011), and how boys' changing voices make them self-conscious or possibly experience difficulties in pitch-matching during this time. Even students who liked to sing might not want to sing in a choir, and interest in choir participation declined with students' grade level/age (Mizener, 1993).

Distribution of choirs' self-selection to group levels. The ISSMA's mission is "to provide educationally evaluated music performance activities for the students and teachers of the State of Indiana, to assist in the development of performance oriented assessment of state and national musical academic standards, and to offer educational support to fulfill this mission" (ISSMA Music Festivals Manual, 2013 – 2014, p. 1). The group levels available at the ISSMA high-school choral contests are a means for schools to participate at their choir's current level for gaining an assessment and feedback for improvement. The group levels available (I, II, III, IV and V) are based on the difficulty of repertoire requirements, and in the case of Group I choirs, whether or not they are participating competitively in the State Qualification contests. While it is impossible to speculate the myriad reasons for schools' selection of group levels, it is interesting to note that in the 2012 – 2014 ISSMA high-school choral contests, there was a larger percentage of choirs that elected to participate in Group I (38.7%) and Group III (29.8%) as compared with Group II (17.0%) and Group IV (14.5%). No school opted to participate in the Group V category. Perhaps some choirs opted to take part in a higher-level category in order to challenge themselves with more difficult repertoire requirements, and/or in the case of some Group I entrants, to take part in the State

Qualification events. There is no prior research studying reasons and the impact of choirs' self-selection to group levels. Recommendations for more research in this area can be found in Chapter V.

Distribution of ratings and points awarded. As described earlier, in the Organization and State Qualification events, the ISSMA's judging system uses a conversion table to decide on final rankings. As an example, the final ranking awarded to a group is decided by taking the two out of three rankings that agree, or, in the case where all three rankings disagree, the middle ranking. Thus, a group that was awarded Gold, Bronze, and Participation ratings by the three adjudicators would end up with a final ranking of Bronze (the middle ranking). Within the three years of ISSMA high-school choral adjudication studied, adjudicators generally agreed on ratings awarded to choirs that received a final rating of Gold, mostly within a narrow point range. However, inconsistencies were found in their ratings and points awarded to choirs that obtained Silver or Bronze awards, with points between adjudicators in the same panel differing as much as 10.5 points in the Organization events. Out of the 925 performances heard over the three years, in ten cases all three adjudicators ended up with different ratings of Gold, Silver, and Bronze for the same choir. In another four cases, one out of the three adjudicators had given a rating that was two ratings lower than the other two adjudicators (e.g., Gold, Gold, Bronze). In the State Qualification events, points given to choirs tended to vary widely between adjudicators, often differing by ten points or more between adjudicators in a panel; in the most extreme case there was a difference of 28 points between two adjudicators in the panel. The largest difference in points between

adjudicators in a panel was found in a State Finals event, with a difference of 35 points in the Mixed choirs category.

The ISSMA organizers were aware of instances in the contests where the same panel of adjudicators awarded high, middle, and low points to the same group, a phenomenon that they termed the "rainbow effect," due to the presence of three different colors of awards. They speculated that what caused this "rainbow effect" was groups that did not use the adjudication system appropriately; for example, a choir that performed music of a high difficulty level while being entered in a low group level (i.e., singing a Group I difficulty piece while competing at the Group III or Group IV level, or vice-versa). The adjudicators could then be disagreeing on points awarded if they were considering the difficulty of the pieces in their adjudication. However, this would not adequately explain vast differences in points awarded in State Qualification or State Finals events, where all choirs are registered as Group I choirs and performing music of equivalent difficulty.

The change at the ISSMA contests from the traditional Division ratings (I and II ratings) to Gold, Silver and Bronze designations (Brakel, 2006) served to ensure a better distribution of scores between ratings. However, from the distribution of awards seen in this study, it would appear that the ratings have migrated upwards. The majority of awards given to choirs at ISSMA being Gold or Silver may point to several possible reasons: (1) that the standard of the choirs is genuinely high, and meet the judging criteria and standards expected for Gold or Silver awards at each Group Level; (2) that the adjudicators employed at ISSMA contests, being educators themselves, may be more understanding or sympathetic of the challenges of high school performers in a stressful

setting; and (3) that the grading rubrics and/or adjudicator training may be affecting how adjudicators judge at the contest.

In reference to point (3) above, the ISSMA adjudicator training emphasizes a three-step adjudication process: (1) Impression, in which adjudicators give a global assessment of the choir formulated on their personal experiences, training, and taste, (2) Analysis, in which adjudicators justify reasons behind the impression and how they translate into points or ratings, and (3) Comparison, in which adjudicators compare the performance with general performance standards and rate each category against the rubric. The use of a rubric could account for more groups being awarded Gold and Silver ratings, because adjudicators would be required to justify their grading, and the equal weight given to each category of music criteria would control for any adjudicator bias.

However, there are some issues to consider with contest results that are skewed to mainly Gold and Silver awards. Even though there are other ratings (Bronze and Participation) available, the heavy skew of the ratings awarded may mean that adjudicators are considering choirs to be only in one of two categories of results – Gold or Silver. This makes data analysis on ratings challenging, because not only would analysis be based predominantly on only two categories of ratings, there is also a high probability of chance agreements between adjudicators' ratings. While the author has made every effort to determine interrater reliability by several different and complementary methods, discretion is advised when considering interrater reliability or percentage agreement between adjudicators' ratings.

There is a second layer of complication to ratings that are awarded based on Group Levels. A Gold awarded to a Group I choir is vastly different to a Gold awarded to

a Group IV choir. While many of the sites try to arrange the order of appearance for choirs, either from Group IV (easiest repertoire; often beginner groups) to Group I (most challenging repertoire; often experienced groups) or vice-versa, there are some sites where choirs appear in random order, or where Group Levels are not in order (for example, Group II, followed by Group I, followed by Group IV, then Group III). In such cases, adjudicators have to adjust their judging standards from Group Level to Group Level, or even from choir to choir. At the ISSMA contests, adjudicators for the High School Organization festivals are issued with the following reminders:

> Keep in mind that you are judging High School Students. Please be realistic. The organization should however, be judged on how well they have performed the music which the director has selected for them. If a group enters in Group IV, but selects Group I level music, you have every right to expect that they will be able to perform the chosen musical program (ISSMA, Instructions for high school organization festival, 2013-2014).

Notwithstanding these reminders, the adjudication forms used are the same regardless of Group Level. This means that adjudicators would need to have in their own minds four different sets of standards for Group I, II, III and IV choirs, and be able to call up each standard at will during the adjudication process. The State Qualification events have a slightly different set of instructions for their adjudicators:

> It is important that while using the rubrics established for this event, that you use the standard of a State Finals performance, not the standard of the site that you are assigned. In order for the scores to have a successful degree of relativity from site to site, the State Level Performance must be the guide in determining the

appropriate numbers to assign to the various categories (ISSMA, Instructions for state qualification festival, 2014).

Thus, at the State Qualification events, adjudicators (who may or may not have adjudicated at the High School Organization festivals), would have to adjudicate choirs based on an entirely different, absolute standard.

The State Finals, which are based on a slightly different adjudication premise, present different issues. Adjudicators at the State Finals are instructed:

> A judge may not give the same point total to two organizations. We provide you with a tote sheet so you can keep a running tabulation of your scores to enable you to avoid scoring two groups the same. We have also provided index cards to assist you with the ranking process. Be sure to give a point score in each category. Range of points – We suggest a range of half the total for each box. After you have judged half of the total groups, you may use decimals beginning with .5 decimal. Judges may confer after the first four groups to establish a standard. After that point we ask that you release your forms following each performance. The most important factor is that your point totals remain consistent for you. (ISSMA, Instructions for State Finals, 2014).

The aim at the State Finals is ranking, rather than the selection at State Qualification events, or the awarding of ratings at the Organization events. Thus, adjudicators are compelled to vary their scores even if they feel that the choirs are on par with each other. While there is a rubric used in the adjudication, the categories are broader (each with a 30-mark range instead of the 10 marks in the State Qualification rubrics or the four marks at the Organization events), thus affording more variability in the grading at the State

Finals. In analyzing the raw scores given by adjudicators at the State Finals, two broad types of judges appeared to emerge:

- Type 1: those that utilized scores within a very narrow, high range (e.g. 80 points – 95 points), with generally small scores gaps between each group (1-2 points difference);
- Type 2: those that utilized scores within a large range (e.g. 42 points – 96 points), and with larger score gaps between each group (5-10 points difference).

The problem with this is that a Type 1 judge might score a mid-rank choir at 85 points, while a Type 2 judge might score the same choir at 55 points. The issue stems from different expectations at the State Finals level. A Type 1 adjudicator might feel that all the choirs performed at a high standard and award scores in the higher end of the spectrum of available marks, while a Type 2 adjudicator might be prioritizing ranking and giving a wider range of marks in order to clearly differentiate the groups. While ranking is the aim and adjudicators' scores are only taken into consideration in the event of a tie, the fact that there is such a large difference in scores for the same choir could be reflective of the lack of standardization at the State Finals.

Percentage agreement between adjudicators. The percentage agreement of ratings between pairs of adjudicators, and between the ratings of each adjudicator and the final ratings were mostly very good, with mainly high, or in some cases, even perfect percentage agreement on ratings. Thirty-nine out of the 56 panels (70%) had mean percentage agreement of $> 70\%$ (good agreement), and 149 out of 168 pairs (89%) had pairwise percentage agreement of $> 60\%$ (acceptable agreement). The 19 pairs of

adjudicators that showed low agreement (< 60% agreement) were mainly in the Organization events and not the Qualification events.

However, this approach is not without its issues, as the calculation of percentage agreement may appear to be highly reliable even if adjudicators were to be scoring completely at random. This issue is compounded because in the ISSMA contests, the number of groups that were awarded Bronze or Participation ratings was so negligible that we are essentially looking at dichotomous ratings (Gold and Silver ratings) between pairs of adjudicators. Even if the percentage agreement appears very high, there is a very high probability that these ratings could have been arrived at purely by chance, since in dichotomous ratings, there is a possibility of attaining much higher "chance" agreements between two raters (Wood, 2007, p.5). However, in this study, there are also a significant number of judge-judge and judge-final rating pairings with lower percentage agreements (< 60%) that reflect true disagreements in their ratings. These reveal one or more adjudicators within the comparison that were "off" in their assessment of the choirs, at least in the context of their fellow adjudicators or with the final amalgamated rating (based on all three concert adjudicators' ratings). Of course, one needs to also consider that the final ratings might not be truly indicative of a choir's performance, since they are arrived at through the conversion as detailed in Table 2. As an example, a choir that received a Gold, Silver and Bronze from the panel would receive a final rating of Silver, even though its performance might well be of a Gold or Bronze standard.

In general, high percentage agreements on ratings are not unusual in festival settings, because of the wide mark range for each rating. In this study, the adjudicators generally tended to agree on Gold ratings, but not so much on Silver or Bronze ratings. It

is also worth noting that the number of groups adjudicated at each site were mostly fewer than 20 groups, with some sites seeing less than 10 groups. Thus, the percentage agreement might seem disproportionately low or high for fewer differences in ratings between the pairs of adjudicators, or between each adjudicator's rating and the final rating. There are pros and cons to this method of assessing groups, which will be discussed more in Chapter V.

Interrater reliability: Pairwise interrater reliability correlations (IRC). The results of the pairwise interrater correlations (IRC) for concert points by site showed a large range for correlation coefficients from weak (r_s= .155) to strong (r_s= .939). IRC for Qualification events were more consistently in the moderate (r_s > .40) to strong (r_s > .80) ranges, while the Organization events had more instances of IRCs in the weak and very weak ranges (r_s < .40). IRC for State Finals events were generally in the moderate range (r_s = .40 to .60). A speculation for this result could be that choirs participating in the Qualification events were of a more uniform standard (being all Group I choirs), and thus it was easier for the panels of adjudicators to agree on the ratings. Organization events had choirs registered in various Group Levels, and adjudicators might have had difficulty adjusting their marking to the different Group Levels, especially if they appeared in mixed order (e.g., a Group I choir followed by a Group IV choir, then a Group II choir), as was the case at several sites. Another probable explanation is that adjudicators at the Qualification events were more experienced, with some already having had adjudication experience that same year with the Organization events, and thus were more familiar with the standards of the competing choirs or with the adjudication process in general. However, these two reasons do not adequately explain why the IRC for State Finals

events were only in the moderate range. As discussed previously, since the purpose at State Finals is to differentiate between as well as rank choirs, perhaps some adjudicators try to make this differentiation clearer by utilizing a large range of points (from points in the 40s to the 90s), while others are working within very narrow point scores in the higher point ranges (points given mainly in the 80s and 90s) in order to more accurately reflect the absolute standard of the choirs.

Interrater reliability: Panel internal consistency. Internal consistency was mainly good in all three years of the Organization and State Qualification events, with moderate (α = .55) to high (α = .94) reliabilities in 2012, high reliabilities (α = .73 to .94) in 2013, and moderate (α = .50) to high (α = .96) reliabilities in 2014. In the State Finals events, internal consistency was moderate (α= .48) to moderately high (α= .86) over the three years. Since the range of available points for each rating (Gold, Silver, Bronze, or Participation) is rather large, with the Gold rating consisting of the largest range of points, it is very likely that adjudicators would end up giving the same rating to any particular choir. In analyzing contest data, high levels of internal consistency may be found even if the ratings did not agree. For example, adjudicators may have awarded points that were close in actual number, but that landed in different rating categories. However, Fiske (1978) suggests that ratings with low internal consistency may mean that evaluators applied inconsistent standards from one group to the next. Examining the internal consistency for points awarded by adjudicators gave a better indication of internal consistency, and showed that adjudicators did mainly give points in similar ranges. The few instances of unacceptable or negative correlations between adjudicators' ratings may have been due to certain sites having adjudicators that did not use the rating

system appropriately, or in the case of the State Finals, adjudicators that may have used different standards of adjudication. Thus, contrary to the choral teachers' perception of poor interrater reliability (Madura Ward-Steinman, 2014), the high internal consistencies of the panels indicate generally rather good interrater reliability at the ISSMA contests for Organization and State Qualification events, and moderate to good interrater reliability at the State Finals. A possible speculation for this negative perception could be due to conflicting comments written or recorded by each adjudicator, or significant differences in the points or ratings awarded by some panels.

Interrater reliability: Intraclass correlation coefficient (ICC). The ICC (2,3) ranged from a low of .381 (fair agreement) to a high of .923 (almost perfect agreement). A majority of the ICCs were in the strong (0.7 - 0.8) to almost perfect (> 0.8) agreement ranges, indicating very good agreement by panel. Again, this runs counter to the perception by choral teachers of low interrater reliability at the ISSMA choral contests (Madura Ward-Steinman, 2014). As discussed previously, high percentage agreements on ratings is not unusual in festival settings, because of the wide mark range available for each rating. In this study, it was found that judges at the Organization and State Qualification events generally tended to agree on Gold ratings, but not so much on Silver or Bronze ratings. This mirrors the findings by Brakel (2006), who looked at the ISSMA band and orchestra contest data and found that reliability was higher for Group I groups than for Group III groups. However, it is worth noting that the Treble/ Men's category in the State Finals events showed only fair agreement within the panel in 2013 (.495) and 2014 (.409). A possible explanation for this could be that there are two categories (Treble choirs and Men's choirs) being adjudicated in the same session, and judges might not be

able to grade and rank different types of choirs as effectively as if they were to grade and rank a homogeneous category of choirs.

Interrater reliability: Interrater differences (Friedman's Chi-Square analysis). Friedman's Chi-Square analysis was used to examine the differences in the mean ratings and points awarded among individual adjudicators at each site. Significant differences ($p < .01$) among individual adjudicators' ratings and points awarded imply low internal consistency. Significant differences ($p < .01$) were found in 4 out of 18 sites in 2012; 4 out of 17 sites in 2013, and; 5 out of 21 sites in 2014. These low figures indicate mainly good interrater reliability for the majority of the panels. Significant differences were probably due to one or two adjudicators (in the panel of three) whose ratings or points awarded were considerably different from the others, which then affected the analysis. The organizers might be interested in looking more closely at their data to determine which particular panels/combination of adjudicators/individual adjudicators might be causing this significant difference, and recommend them for further training or reconsider their use in future years of adjudication.

CHAPTER V

SUMMARY, CONCLUSIONS, IMPLICATIONS, AND RECOMMENDATIONS

Summary

The purpose of this study was to examine the descriptive data and interrater reliability for the Indiana State School Music Association (ISSMA) high school choral contests over a period of three years in order to add to the existing body of research on choral adjudication.

Data for this study included ratings and points awarded by a total of 58 panels (of three adjudicators each) to 925 choir performances by 689 discrete high school choirs at the ISSMA-sponsored choral festivals in 2012, 2013, and 2014. Choirs either registered for Organization events (at the district level) or State Qualification events. Choral directors or schools registered their choirs under one of the following Group Levels: I, II, III, IV or V, based on the difficulty of the choir's repertoire. Three-year data on the adjudication (individual judges' scores and ratings, and final ratings) were compiled and analyzed for descriptive frequencies. Interrater reliability was calculated from the adjudication data by individual sites.

Descriptive frequencies for ratings (Gold, Silver, Bronze, Participation), type of choir (SATB, Mens, Treble), and group level self-selection (Group I, II, III, IV or V) were calculated. Interrater reliability on ratings were calculated via adjudicator pairwise percentage agreement. Interrater reliability on points awarded were calculated via pairwise interrater correlations (IRC) (r_s), reliability for each three-member panel (Cronbach's alpha (α)), intraclass correlation coefficient (ICC) (2-way random) for each

three-member panel, and Friedman's Chi-Square analysis to examine difference in points awarded among adjudicators at each site.

The results of this study are summarized as follows:

Descriptive frequencies

1. A higher proportion of choirs were awarded final Gold ratings (77%) and Silver ratings (22%). Only 11 choirs (1%) were awarded Bronze ratings, and no choir received Participation ratings from 2012 – 2014.
2. There were more mixed (60%) and treble (34%) than there were mens (6%) choirs.
3. There were more choirs entering at Group I (39%) and Group III (30%) levels. Notably fewer choirs entered at Group II (17%) and Group IV (15%) levels. No choirs entered at the Group V level from 2012 – 2014.

Interrater reliability

1. Percentage agreements of ratings at Organization and State Qualification events were mainly high. Some panels and pairs of adjudicators had very high to even perfect percentage agreement on ratings. Forty-one out of the 58 panels (71%) had a mean percentage agreement of > 70%. One hundred and fifty-five out of 174 pairs of adjudicators (89%) had pairwise percentage agreement of > 60%. The 19 pairs of adjudicators that showed low agreement (< 60%) were mainly in the Organization events and not the State Qualification events.
2. While mean IRCs were almost all positive (except for one instance of negative correlation), there was a large range for correlation coefficients from weak

(r_s = .155) to strong (r_s = .939). Qualification Events had IRCs more consistently in the moderate (r_s > .40) to very strong (r_s > .80) ranges, while Organization Events had more instances of IRCs in the weak and very weak ranges (r_s < .40). IRC for State Finals Events were generally in the moderate range (r_s = .40 to .60).

3. Internal consistency in 2012 was moderate (α = .55) to high (α = .94); in 2013 was high (α = .73 to .94), and; in 2014 was moderate (α = .50) to high (α = .96). In the State Finals events, internal consistency was moderate (α= .48) to moderately high (α= .86) over the three years.
4. ICC (2,3) ranged from a low of .381 (fair agreement) to a high of .923 (almost perfect agreement). A majority of the ICCs were in the strong (0.7 - 0.8) to almost perfect (> 0.8) agreement ranges, indicating very good agreement by panel. However, the Treble/ Men's category in the State Finals events showed only fair agreement within the panel in 2013 (.495) and 2014 (.409)
5. Friedman's Chi-Square analysis showed the differences in the mean ranks of points awarded among individual judges in each site. Significant differences (p < .01) were found in 3 out of 18 sites in 2012; in 4 out of 17 sites in 2013, and; in 4 out of 21 sites in 2014. In the State Finals events, significant differences (p < .01) were found among individual judges' points awarded in the Mixed choirs category in 2012 and 2014.

Conclusions

While there are scant studies on interrater reliabilities of adjudicators judging large instrumental ensembles (bands and orchestras), the existing literature lacks such studies on choral ensembles. This study of the ISSMA choral contests is important in

adding to the dearth of knowledge on large choirs, and to the limited research on interrater reliability by authors such as Brakel (2006), Burnsed, Hinkle, & King (1985), Garman et al. (1991), Hash (2012), King & Burnsed (2009), Latimer, Bergee, & Cohen (2010).

As found by other researchers, there is a trend towards higher ratings (Boeckman, 2002; Brakel, 2006) that may suggest that adjudicators are not considering the full range of ratings available to them. This creates issues with interrater reliability, since adjudicators may only be fully utilizing and considering groups to be in one of two categories (e.g., Gold or Silver). Studies that have looked at number of adjudicators on a panel and their effect on interrater reliability (Bergee, 2003; Brakel, 2006; Fiske, 1977) seem conflicting, with some advocating for a minimum number of five or seven adjudicators on a panel, while others suggesting that a panel of two or three adjudicators would also result in high interrater reliability. However, in this study, I found that pairwise correlations were generally lower than the alpha estimates for panels of three adjudicators, suggesting that the use of three judges is more reliable than two judges in a panel in such contexts.

Interrater reliability in this study was generally high, and suggests that choral teachers' perceptions of low interrater reliability at the ISSMA contests (Madura Ward-Steinman, 2014) may be tainted by conflicting ratings, differences in points awarded, or contrasting comments by the panel of adjudicators. More investigation needs to be done in order to determine if this is true.

Implications

The findings reported in this study suggest several implications for music educators and organizers of choral festivals.

The proportion of men's choirs in the ISSMA contests possibly indicates a difficulty with recruiting enough male singers to make up more men's choirs, or that male singers preferred to be part of mixed choirs. Music educators and directors could include more strategies for recruiting and retaining male singers from the elementary school up to high school levels. On a general level, this could involve more aggressive recruitment strategies, selecting suitable repertoire for male singers, and improving boys' perceptions towards choral singing. On a personal level, educators may need to overcome their personal biases towards boys in choirs, in particular when the boys undergo challenging vocal changes during puberty. Choral directors who are anxious to win awards or have their choirs perform their best at contests may need to address their priorities in music education, and help their male singers succeed even in high-stakes situations.

Additionally, the larger percentage of choirs taking part in Group I and Group III as compared with Group II and Group IV assessments could indicate that some choirs may have wanted to challenge themselves in a higher category, or, in the case of some Group I entrants, wanted to take part in the State Qualification events. Choral teachers have to weigh many factors in deciding which group level to register their choirs in for assessment purposes. Schools and directors need to consider the difficulty of the repertoire at the chosen group level, and whether their choir is able to perform the repertoire successfully. Since registration happens months before the actual contests, they

need to know their choir's strengths and weaknesses well enough and estimate the level they can reach in those few months of preparation. Choosing a too-high group level may mean that the choir is pushed beyond their capabilities and may emerge from the contests with a disappointing low rating and lose interest in choral singing or music altogether. An astute choice of group level, and deep understanding of their choir is required for a successful, educationally-supportive outcome from these contests.

A study by Madura Ward-Steinman (2014) on high-achieving secondary school choral music teachers in Indiana found that one of the points of discomfort by the most successful choral directors participating in the ISSMA choral festivals was the perceived lack of adjudicator reliability. In this current study, while some of the panels exhibited low (or, in two cases, negative) interrater reliability, the analysis showed good to excellent interrater reliability for ratings, and good to high interrater reliability for points awarded by the adjudicators. What, then, is causing the perceived lack of adjudicator reliability at the ISSMA contests? Perhaps the organizers of music festivals could address these issues more openly with their participants and adjudicators. Publishing information about adjudicators, such as their teaching or adjudication experience, may help improve participant confidence in the adjudication process or outcome. Tracking their adjudicators' reliabilities closely over the years may also help the organizers identify and retrain or eliminate any adjudicators with suspect judging abilities. Perhaps some of the negative perceptions on interrater reliabilities stem from the large point difference between adjudicators on the panel, or the "rainbow effect," where groups are awarded different ratings (e.g., Gold, Silver, and Bronze from the three adjudicators in the same panel). More investigation needs to be carried out to find out why this happens in the

adjudication setting, and what can be done to mitigate it. Meanwhile, contest organizers can examine their adjudication procedures more closely to see if their training of adjudicators, procedures, or assessment rubrics and forms are clear. Choral directors can also educate themselves more on the issues in adjudication, and perhaps volunteer to be judges at festivals themselves in order to better understand the difficulties in adjudication.

A predominance of Gold and Silver ratings awarded in the ISSMA choral contests also has implications for contest participation and adjudication. While a predominance of high ratings at the ISSMA contest may increase festival participation and encourage students and directors, this practice may not adequately differentiate levels of achievement between groups, and therefore may actually weaken the validity of these ratings (Hash, 2012). The results in this study mirror those found in Brakel's (2006) study, in which Group I ensembles were found to have the highest degree of interrater reliability. While adjudicators had little issue agreeing on Gold-rating performances, contest point totals also appeared to show greater inconsistency among judges when the performance was poor.

Grade inflation at music festivals (Boeckman, 2002) might pose a problem to groups genuinely wanting to be evaluated accurately so that they can find out where they stand in relation to other groups, or receive feedback for improvement. While this study did not have enough data to look at trends, it is certainly a cause for concern that 99% of the choirs at the ISSMA contest received Gold and Silver ratings in the three years of the contest from 2012-2014. A plausible explanation for this could be the larger mark range available to Gold and Silver ratings, or the different adjudication standards for different Group Levels. Festival organizers could consider revising their grading scheme to reflect

a more even spread of scores across the various ratings, or developing a multidimensional assessment rubric, which has shown to be applicable to different grade levels to determine performance achievement over time (Ciorba & Smith, 2009).

Although interrater reliability at the ISSMA high school choral contests over the three years was generally high, there were certainly panels with low interrater reliability. Some panels had significantly large differences in their points and ratings awarded (including some panels with three different ratings for the same group, or adjudicators who had given two ratings lower than their peers in the same panel). A close inspection of these unusual cases could be useful to organizers and help them to decide which adjudicators to re-train or exclude from further adjudication duties.

The Treble/ Men's category in the State Finals events showed only fair agreement within the panel in 2013 (.495) and 2014 (.409), as compared with the Mixed category, which showed moderately high agreement within the panel (.700 to .799 over the three years). A possible explanation for this could be that there are two categories (Treble choirs and Men's choirs) being adjudicated in the same session, and judges might not be able to grade and rank different types of choirs as effectively as if they were to grade and rank a homogeneous category of choirs. Festival organizers might consider splitting up different categories and having them adjudicated in separate sessions.

Two other issues relate to ratings given to choirs that participate in the ISSMA contests. Firstly, final ratings are arrived at based on a conversion table (see Table 2). The pros of this system are that outliers – very strict or very lenient adjudicators – would be eliminated from the final ratings. However, the cons are that a choir (for example, one that received Gold, Silver, and Bronze ratings from the panel) might end up with a rating

that is not indicative of its true performance standard. The contest organizers might want to review the effectiveness of this conversion system, or seek feedback from adjudicators and choral directors on its usefulness. Secondly, ratings are given based on the Group Level that each choir has registered for. A Gold rating given to a Group I choir is vastly different to a Gold rating given to a Group IV choir, and does not really indicate to choirs where they stand in relation to an absolute standard. Perhaps it is the intention of the ISSMA to provide feedback to schools for educational purposes in a comparative setting, in which case, it would be useful to provide schools and adjudicators with a benchmark or descriptors for standards for each rating at each Group Level. Alternatively, the organizers of similar contests could allow time for discussion between adjudicators in each panel, with allowances for adjustments of points and ratings awarded, in order to eliminate the effect of outliers.

Recommendations for Future Research

Based on the results of this study, the following recommendations for future research are made:

1. More investigation needs to be done to ascertain the reasons behind the smaller number of men's choirs, and whether or not more could be done to encourage a larger number of men's choirs to form or participate in the contests.
2. Little research exists on group self-selection in contest settings. It would be interesting to study choirs' perceptions of their level, and what their considerations are (apart from the difficulty of the repertoire) when deciding which group level to register for.

3. A replication of this study in other states or in comparative large-scale choral contest settings would help to uncover similarities or differences in descriptive contest data and interrater reliabilities. In particular, it would be interesting to see if there were trends in types of choirs, awarded ratings, and contest grade inflation through longer-range studies of contests with available historical contest data.
4. Comparative investigations can be done to see if similar findings appear in other high-school choral contests in other states or other countries that also use three-adjudicator panels in a tiered group-level system. In particular, it would be interesting to see what standards were being employed at each stage of the contest (e.g., regional v.s. state qualification v.s. state finals), and what impact each type of standard had on the judging.
5. Comparative studies on interrater reliability in different types of contests might be interesting for organizers wanting to find the most effective methods of contest organization, adjudicator training, or use of rubrics and scoring or feedback forms.
6. More investigation needs to be done on adjudicator reliability that also takes into account the adjudicator experience and expertise (e.g., Brakel, 2006; Fiske, 1975, 1977; Kinney, 2009), effects of adjudicator training (Fiske, 1978, 1983; Winter, 1993), and the reliability of rubrics used in judging (Latimer et al., 2010; Norris & Borst, 2007). Other sources of data, such as school demographics, school size, experience and expertise of choral directors, could also be taken into account for future research.

7. Future research can look at related issues in adjudication of choirs in a contest setting. It would be interesting to investigate adjudicator training in terms of length and content of training and their impact on adjudication. More research can also be done on adjudicator processes, such as improving adjudicator reliability by increasing the number of adjudicators per panel, removing the highest and lowest scores to account for bias, and taking the average of the remaining scores to arrive at final ratings.

To conclude, this study looked at descriptive data and interrater reliability over three years of the ISSMA high school choral contests and found generally high interrater reliabilities based on a three-adjudicator panel, but also a number of other interesting phenomena such as heavily skewed ratings tending towards Gold and Silver ratings, and significant differences in points or ratings awarded between adjudicators in the same panel. These findings confirm that three-member adjudication panels are generally reliable and suggest that less-than-acceptable interrater reliabilities are probably caused by adjudicator disagreements in how to grade choirs that had registered in an unsuitable Group Level at the contest, or by a less clear grading system with no absolute standard of grading. These can easily be mitigated by organizers or choirs taking the necessary steps to ensure that groups are enrolled in suitable group levels or by a systematic grading system and adjudicator training. Contest organizers and participating choirs should take into account the many issues affecting adjudication, as well as remind themselves of the purpose of participation in choral contests, so as to reap the maximum benefits of the experiences and learning that can be gained from contest participation.

REFERENCES

Abeles, H. F. (1973). A facet-factorial approach to the construction of rating scales to measure complex behaviors. *Journal of Educational Measurement*, *10*, 145-151. doi:10.1111/j.1745-3984.1973.tb00792.x

Austin, J. R. (1988). The effect of music contest format on self-concept, motivation, achievement, and attitude of elementary band students. *Journal of Research in Music Education, 36*(2), 95-107. doi: 10.2307/3345243

Battersby, S. (1995). Benefits of competitions/contests for choral directors and students in the tri-state area. *Dissertation Abstracts International, 55*(11), 3344A. (University Microfilms No. AAC95-11030)

Bergee, M. J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education, 51*(2), 137-150. doi:10.2307/3345847

Bergee, M. A. (2006). Validation of a model of extramusical influences on solo and small-ensemble festival ratings. *Journal of Research in Music Education*, *54*, 244-256. doi:10.1177/002242940605400307

Bergee, M. A. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education*, *55*(4), 344-358. doi:10.1177/0022429408317515

Bergee, M. J., Coffman, D., Demorest, S. M., Humphreys, J. T., & Thornton, L. P. (2001). *Influences on college students' decision to become a music teacher*. Available at http://www.menc.org

Bergee, M. A., & 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*(3), 358-374. doi:10.1177/002242940505300407
Retrieved from: http://www.jstor.org/stable/3648433

Best, F. C. (1927). State Choral Contests. *Music Supervisors' Journal, 13*, 9 + 11
doi: 10.2307/3383201
Retrieved from http://www.jstor.org/stable/3383201

Boeckman, J. (2002). Grade inflation in band contest ratings: A trend study. *Journal of Band Research, 38*(1), 25-36.

Brakel, T. D. (2006). Inter-judge reliability of the Indiana State School Music Association high school instrumental festival. *Journal of Band Research, 42*(1), 59-69.
Retrieved from: https://www.questia.com/library/journal/1P3-1270331351/inter-judge-reliability-of-the-indiana-state-school

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

Ciorba, C. R., & Smith, N. Y. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *Journal of Research in Music Education, 57*, 5-15.
doi:10.1177/0022429409333405

Cooksey, J. M. (1977). A facet-factorial approach to rating high school choral music performance. *Journal of Research in Music Education*, *25*, 100-114.
doi:10.2307/3345190

Dibben, N. (2002). Gender identity and music. In R. MacDonald, D. Hargreaves & D.

Miell (Eds.), *Musical Identities*, 117–33. New York: Oxford University Press.

Demorest, S. M. (2000). Encouraging male participation in chorus. *Music Educators Journal, 86*(4), 38-41.

Retrieved from: http://www.jstor.org/stable/3399604

Eshelman, D. (1992). Leading a renaissance in training adolescent boy singers. *The Choral Journal, 33*(3), 23-27.

Retrieved from: http://www.jstor.org/stable/23548821

Fiske, H. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education, 23*, 186-196. doi:10.2307/3344643

Fiske, H. E. Jr. (1977). Relationship of selected factors in trumpet performance adjudication reliability. *Journal of Research in Music Education, 25*, 256-263. doi:10.2307/3345266

Retrieved from: http://www.jstor.org/stable/3345266

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

Freer, P. K. (2011). Weight lifting, singing, and adolescent boys. *The Choral Journal, 52*(4), 32-41.

Retrieved from: http://www.jstor.org/stable/23560601

Freer. P. K. (2012). The successful transition and retention of boys from middle school to high school choral music. *The Choral Journal, 52*(10), 8-17.

Retrieved from: http://www.jstor.org/stable/23560678

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.

Gates, J. T. (1989). A historical comparison of public singing by American men and women. *Journal of Research in Music Education, 37*(1), 32-47. doi: 10.2307/3344951

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. doi: 10.1177/0022429408317366

Greene, T. (2012). An application of the facet-factorial approach to scale construction in development of a rating scale for high school marching band performance. *Dissertation Abstracts International, 50*(06), 1447A. (University Microfilms No. AAC89-19644)

Hall, C. (2005). Gender and boys' singing in early childhood. *British Journal of Music Education, 22*(1), 5-20. DOI: http://dx.doi.org/10.1017/S0265051704005960

Hash, P. M. (2012). An analysis of the ratings and interrater reliability of high school band contests. *Journal of Research in Music Education, 60*, 81-100. doi:10.1177/0022429411434932

Hewitt, M. P. (2000). Marching band show customization and director involvement: Their relationship to performance scores. *Bulletin of the Council for Research in Music Education, 146,* 18-30. Retrieved from: http://www.jstor.org/stable/40319031

Howard, K. K. (1994). A survey of Iowa high school band students' self-perceptions and

attitudes toward types of music contests (Doctoral dissertation, Northwestern State University of Louisiana, 1979). *Dissertation Abstracts International, 55*, 2201.

Howard, S. A. (2012). The effect of selected nonmusical factors on adjudicators ratings of high school solo vocal performances. *Journal of Research in Music Education, 60*, 166-185. doi: 10.1177/0022429412444610

Johnson, C. M., & Geringer, J. M. (2007). Predicting music majors' overall ratings of wind band performances: elements of music. *Bulletin of the Council for Research in Music Education, 173*, 25-38.

Retrieved from: http://www.jstor.org/stable/40319468

Killian J. N. (1998). Characteristics of successful choirs in a contest setting. Texas Music Education Research, 39-43.

Retrieved from: http://www.tmea.org/assets/pdf/research/Kill998.pdf

Kinney, D. W. (2009). Internal consistency of performance evaluations as a function of music expertise and excerpt familiarity. *Journal of Research in Music Education, 56*, 322-337. doi:10.1177/0022429408328934

Retrieved from http://www.jstor.org.stable/40204937

Latimer, M. E., Bergee, M. J., & Cohen, M. L. (2010). Reliability and perceived pedagogical utility of a weighted music performance assessment rubric. *Journal of Research in Music Education, 58*, 168-183. doi:10.1177/0022429410369836

Lucas, M. (2011). Adolescent Male Attitudes about Singing in Choir. *Update: Applications of Research in Music Education 30*(I), 46-53.

doi: 10.1177/8755123311418623

Madura Ward-Steinman, P. (2014). A Descriptive Study of High-Achieving Secondary School Choral Music Teachers in Indiana. (Invited research paper presented at the Indiana Choral Directors Association conference, Indianapolis, July 2, 2014)

Miller, R. E. (1994). A dysfunctional culture – Competition in music. *Music Educators Journal, 81*, 29-33. doi:10.2307/3398761
Retrieved from: http://www.jstor.org/stable/3398761

Mizener, C. P. (1993). Attitudes of children toward singing and choir participation and assessed singing skill. *Journal of Research in Music Education, 41*(3), 233–245. doi: 10.2307/3345327

Napoles, J. (2009). The effect of excerpt duration and music education emphasis on ratings of high quality children's choral performances. *Bulletin of the Council for Research in Music Education, 179*, 21-32.
Retrieved from: http://www.jstor.org/stable/40319327

Norris, C. E., & Borst, J. D. (2007). An examination of the reliabilities of two choral festival adjudication forms. *Journal of Research in Music Education, 55*, 237-251. doi:10.1177/002242940705500305

Rickels, D. A. (2012). Nonperformance variables as predictors of marching band contest results. *Bulletin of the Council for Research in Music Education, 194*, 53-72. doi:10.5406/bulcouresmusedu.194.0053

Rittenhouse, J. H. (1989). Competitive and noncompetitive choral festivals at the secondary level. *Dissertation Abstracts International, 50*(06), 1447A. (University Microfilms No. AAC89-19644)

Rogers, D. M. (2004). The level of agreement among adjudicators concerning problems

and solutions when analyzing taped examples of choral tone. (Doctoral dissertation, University of South Carolina, 2004). *Dissertation Abstracts International, UMI number 3142849.*

Rogers, G. (1985). Attitudes of high school band directors and principals toward marching band contests. *Journal of Research in Music Education, 33*, 259-267. doi:10.2307/3345252

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. doi:10.2307/3345585

Smith, B. P., & Barnes, G. V. (2007). Development and validation of an orchestra performance rating scale. *Journal of Research in Music Education, 55*, 268-80. doi:10.1177/002242940705500307

Wood, J. M. (2007). Understanding and computing Cohen's Kappa: A tutorial. *WebPsychEmpiricist.* Retrieved from http://wpe.info/papers_table.html

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. doi:10.1177/025576149302200106

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. doi:10.2307/3345801

VITA

Sheh Feng Ng pursued her Master of Music Education at the Jacobs School of Music from 2013 – 2015 on a full scholarship from the Ministry of Education, Singapore. She completed her BM (with Honors) at the University of Birmingham in the United Kingdom, followed by a Post-Graduate Diploma in Education at the National Institute of Education in Singapore. Prior to joining the Jacobs School, Sheh Feng was an Arts Education Officer (Music) at the Student Development Curriculum Division, Ministry of Education in Singapore, where she led teams in large-scale national projects such as the Singapore Youth Festival and the Music Talent Development Centre. Apart from being involved in educational policy and curriculum reforms, she also led in-service training of music teachers, organized workshops and master classes for music directors and instructors, and processed grants to school music programs. In her free time, she enjoys road trips, adventuring, and the great outdoors. Her favorite things are good food, great music, dogs, and people.

Sheh Feng can be contacted at shehfeng@gmail.com