DOCUMENT RESUME

ED 360 204 SO 022 784

AUTHOR Garman, Barry R.; And Others
TITLE Orchestra Festival Evaluations: Interjudge Agreement and Relationships between Performance Categories and Final Ratings.
PUB DATE 91
NOTE 8p.
PUB TYPE Journal Articles (080)
JOURNAL CIT Research Perspectives in Music Education; n2 p19-24 Fall 1991

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Bands (Music); Music Activities; *Music Education; *Orchestras; Secondary Education; Singing; Statistical Analysis; *Student Evaluation
IDENTIFIERS Music Festivals; *Music Performance

ABSTRACT

Band, orchestra, and choir festival evaluations are a regular part of many secondary school music programs, and most such festivals engage adjudicators who rate each group's performance. Because music ensemble performance is complex and multi-dimensional, it does not lend itself readily to precise measurement; generally, musical performances are evaluated subjectively, that is, reflecting either consciously or subconsciously the criteria that an individual evaluator considers most important. Allowing individual adjudicators to employ their own criteria in evaluating performance festivals, however, presents some potential problems. To help alleviate these problems, most performance festivals do two things: (1) employ more than one adjudicator, and (2) ask adjudicators to consider a common set of performance categories in arriving at a final rating. The purpose of this study was to examine the "interjudge" reliability for five groups of judges on seven rating categories on a band/orchestra adjudication form and determine the extent to which category ratings are interrelated. Interjudge reliability coefficients for three sets of judges were found to be marginally acceptable (in the .80s); those for the other two sets of judges (.67 and .54) were not. Interjudge reliability coefficients for the various category ratings were generally much lower than those for the final ratings. Two performance categories (technique and intonation) were the best predictors of final ratings. The categories "selection" and "general effect" contributed nothing toward predicting the final ratings. (Author/DB)

Orchestra Festival Evaluations: Interjudge Agreement and Relationships Between Performance Categories And Final Ratings

By

Barry R. Garman, J. David Boyle, Nicholas J. DeCarbo,
School of Music, Univeristy of Miami

SO-AR

Orchestra Festival Evaluations: Interjudge Agreement And Relationships Between Performance Categories And Final Ratings

By Barry R. Garman, J. David Boyle, Nicholas J. DeCarbo,
School of Music, University of Miami

Abstract

The purpose of the study was to (a) examine the interjudge reliability for five groups of judges on seven rating categories on a band/orchestra adjudication form and (b) determine the extent to which category ratings are interrelated. Specific questions addressed in the study were:

1 .What are the interjudge reliability coefficients for each category and for the final ratings?

2. What are the correlation coefficients between each of the performance category ratings and the final ratings for individual judges and for the combined judges?

3. To what extent do ratings for the performance categories influence the final ratings?

4. Are there differences in judges' mean ratings for each of the performance categories and for the final ratings?

Interjudge reliability coefficients for three sets of judges were marginally acceptable (in the .80s); those for the other two sets of judges (.67 and .54) were not. Interjudge reliability coefficients for the various category ratings were generally much lower than those for the final ratings.

Two performance categories (technique and intonation) were the best predictors of final ratings. The categories "selection" and "general effect" contributed nothing toward predicting the final ratings.

Band, orchestra, and choir festival evaluations are a regular part of many secondary school music programs, and most such festivals engage adjudicators who rate each group's performance. Whether such festivals are actually competitions may be a matter of perspective. If a festival is structured so that adjudicators rate a performance in relation to the performances of other groups, say for a first place award, perhaps the festival is a competition. On the other hand, if each group is evaluated in relation to some fixed standard irrespective of the evaluations of other groups participating in the festival, perhaps it should not be considered a competition. Of course there is always the question whether adjudicators can totally isolate a group's performance from that of others. Despite efforts to keep the focus on providing feedback for improving the performance of the participating ensembles, perhaps there will always be an element of competition in secondary school music festivals.

Because music ensemble performance is complex and multidimensional, it does not lend itself readily to precise measurement; generally, musical performances are evaluated subjectively, that is, reflecting either consciously or subconsciously the criteria that an individual evaluator considers most important. Allowing individual adjudicators to employ their own criteria in evaluating performance festivals, however, presents some potential problems. To help alleviate these problems, most performance festivals do two things: (a) employ more than one adjudicator and (b) ask adjudicators to consider a common set of performance categories in arriving at a final rating.

Most adjudication forms include about six performance categories or "standards" against which adjudicators are asked to provide ratings. Typical performance categories are tone, intonation, technique, balance, interpretation, musical effect, and "other factors." Adjudicators usually are provided with descriptions of the various standards and then asked to rate each performance against the standard. Most festivals require adjudicators to employ a five-point, or five-category, rating scale. The five rating categories usually are designated by Roman numerals, I through V, and they may be given parallel verbal labels such as superior (I), excellent (II), good (III), fair (IV), and poor or unacceptable (V). Others, as in the case of the present study, may use letter ratings.

An assumption of this procedure is that each adjudicator's ratings of a performance relative to the performance categories or standards will more-or-less provide the basis for the respective adjudicator's rating of the performance, which in turn is averaged with those of two other adjudicators to determine a final rating for each performance. Whether ratings for the respective performance categories indeed provide the real basis for overall ratings is questionable. Furthermore, research on the effects of the variables in the adjudication process is inconclusive.

Related Literature

Several studies have examined the relationship between certain judge characteristics and evaluation ability as measured by the reliability of individual judges. Other measures examined in some of these studies are the internal consistency of evaluation forms and agreement among judges. Other studies have tested various procedures for their effectiveness in increasing judge reliability.

Vasil (1973) and Massel (1978) examined the differences in judges' rank orderings between tape recorded versions of performances and live (Vasil) or videotaped (Massel) versions. The difference was found to be minimal for the top fifteen of thirty-three performances (Vasil) and all of twenty-two performances (Massell).

Fiske (1977) investigated the relationships among reliability of music performance, adjudication, judge performance ability, and judge nonperformance music achievement. Thirty-three recent music education graduates rated a series of tape recorded trumpet performances twice. Results showed no relationship between performing ability and judge reliability or between performing ability and judge nonperformance music achievement. There was a significant inverse relationship between judge reliability and nonperformance music achievement.

Fiske (1978) investigated the use of training sessions to increase judge reliability. Although no significant effects were found, Fiske

suggests that training sessions may prove effective with instructional revisions and the use of stronger motivating factors.

Mullin (1979) examined the relationship between musical quantitative decision-making ability as measured by subtests of aptitude tests by Seashore (Seashore, Lewis, and Saetveit, 1939, 1960) and Gordon (1965) and qualitative decision-making ability as measured by investigator-designed reliability measures. Results showed no relationship between correct response to aurally presented test items and evaluation reliability. It was concluded that two different independently functioning problem-solving strategies are present in the evaluation process.

Towers (1980) investigated the effect of judge age and musical experience on reliability. For a set of judges ranging in age from seven years to "adult," there was a significant trend towards increased improvement in reliability with increasing age. Musical experience also had a significant effect on judge reliability.

The purpose of a study by Burnsed, Hinkle, and King (1985) was to determine (a) the internal consistency of a performance evaluation form, (b) the interjudge reliability of groups of judges, and (c) the significance of performance rating categories as predictors of final ratings at selected adjudication festivals. The reliabilty of the adjudication form used was found to be very high. Out of four groups of judges, agreement was low for three groups on ratings of tone and for two groups on ratings of intonation and balance. A significant colinear relationship was found between all seven ratings (tone, intonation, technique, balance, interpretation, musical effect, and final). Of the six performance categories, musical effect correlated highest with the final rating. It was concluded that judges tend to evaluate performances from a global perspective and that performance categories may not represent separate entities.

Burnsed and King (1987) continued the 1985 investigation, adding data from additional ensembles and groups of judges to their previous data. Interjudge agreement in all nine groups was again high for final ratings, but low in six groups on ratings of tone and in five groups on ratings of intonation. A correlation of all performance ratings revealed that performance category ratings and final ratings were so closely related as to represent a global rating. Again, musical effect correlated highest with the final rating. As with the 1985 study, the investigators concluded that certain category ratings, including tone, intonation, and balance, may be viewed with some skepticism, since judges appear to base their ratings on a single, global evaluation.

Purpose

The purpose of the present study was to (a) examine the interjudge reliability for five groups of judges regarding seven rating categories found on a band/orchestra adjudication form published by the Music Educators National Conference, and (b) determine the extent to which category ratings are interrelated. Specific questions addressed in the study were:

1. What are the interjudge reliabilty coefficients for each category and for the final ratings?

2. What are the correlation coefficients between each of the performance category ratings and the final ratings for individual judges and for the combined judges?

3. To what extent do ratings for the performance categories influence the final rating for combined judges?

4. Are there differences in judges' mean ratings for each of the performance categories and for the final ratings?

Methodology

Data for the study are based on the adjudication results for the Dade County Orchestra Festivals held during the 1980s. Forms for 1980 through 1990 were originally examined, but substantial differences in the forms for 1980 and 1981 precluded the use of those data in the study; also, complete data for all three adjudicators were not available for 1984, 1985, and 1988. Consequently, the present study is based on data for the 1983, 1986, 1987, 1989, and 1990 Dade County Orchestra Festivals. Ratings were available for 13 orchestras in 1983, 13 in 1986, 15 in 1987, 18 in 1988, and 13 in 1990. Although the ratings were originally given on a five-letter scale, they were converted to a five-number scale for statistical analysis, i.e. A (Superior), B (Excellent), C (Good), D (Fair) and E (Poor) ratings were treated numerically as 4, 3, 2, 1, and 0 respectively.

Four basic statistical analyses were conducted on each year's results. Pearson product-moment correlations between the independent judges' final ratings were used to provide a measure of interjudge reliability. Interjudge reliabilities were determined for all category ratings and for the final ratings for each year's judges by calculating Pearson product-moment correlations between each pair of judges and averaging the three coefficients.

Pearson product-moment correlations also were calculated between each judge's ratings of the respective performance categories and the final rating. Thus, each judge's ratings and the mean of the three judges' ratings for the performance categories of tone, intonation, technique, selection, interpretation, and general effect were correlated with the final overall rating.

A step-wise multiple regression analysis was calculated to ascertain the extent to which the combined judges' ratings for the various performance categories "accounted for" or predicted the final rating. The six performance categories were the independent variables and the final overall rating was the dependent variable.

A repeated-measures ANOVA was used to compare the judges' mean ratings for each of the performance categories and for the final ratings.

Essentially, the data from these analyses provided information regarding the extent to which the three judges differed in their overall ratings of the orchestra performances for the six performance categories and for the final rating.

Results

The results are presented as they pertain to the four basic questions of the study.

Interjudge Reliability

As shown in Table 1, the interjudge reliability coefficients for the final ratings ranged from a low of .54 in 1989 to a high of .89 in 1987. Generally, reliability coefficients for the final ratings were higher than for the various categories.

Although not tested statistically, considerable year-to-year differences were apparent in the interjudge reliability coefficients for all categories and the final ratings. The year-to-year disparities in the interjudge reliability coefficients were greatest for the categories tone, general effect, and intonation. For tone, intonation, technique, interpretation, and final ratings, the interjudge

reliability coefficients for 1989 were much lower than for other years. In 1986 and 1987, however, the interjudge reliabilities for general effect were very low (.07 and .27, respectively).

Table 1

Interjudge Correlations on Category and Final Ratings, 1983, 1986, 1987, 1989, and 1990

	Correlation			
Category/Year	*J1/J2*	*J1/J3*	*J2/J3*	*Mean*
Tone				
1983	.80	.86	.83	.83
1986	.77	.79	.78	.78
1987	.72	.89	.75	.79
1989	.25	.57	.14	.32
1990	.75	.76	.75	.75
Intonation				
1983	.67	.51	.70	.63
1986	.80	.72	.62	.71
1987	.60	.79	.75	.71
1989	.13	.51	.53	.39
1990	.92	.85	.75	.84
Technique				
1983	.77	.80	.59	.72
1986	.51	.70	.63	.61
1987	.73	.67	.65	.68
1989	.32	.51	.49	.44
1990	.67	.65	.79	.70
Selection				
1983	.52	.77	.29	.53
1986	.65	.76	.55	.65
1987	.67	.54	.55	.59
1989	.62	.80	.50	.64
1990	.70	.61	.83	.71
Interpretation				
1983	.79	.53	.61	.64
1986	.68	.68	.76	.71
1987	.67	.72	.80	.73
1989	.16	.66	.59	.47
1990	.79	.73	.65	.72
General Effect				
1983	.55	.53	.61	.56
1986	.20	.00	.00	.67
1987	.81	.00	.00	.27
1989	.20	.53	.40	.38
1990	.87	.67	.71	.75
Final				
1983	.80	.61	.61	.67
1986	.79	.87	.90	.85
1987	.89	.95	.83	.89
1989	.42	.71	.49	.54
1990	.77	.78	.88	.81

Correlations Between Categories and Final Ratings

With few exceptions, correlations between the individual judges' ratings and the final rating and between the combined judges' ratings and the final rating were statistically significant. With two exceptions, all of the correlation coefficients between judges' ratings for tone and the final rating were above .70, with most in the .80s. All but four correlations between intonation and the final rating were above .70, for technique all but three, and for interpretation all but three. For selection, however, the correlations with final ratings were generally lower, with ten of the 20 correlation coefficients below .70. The correlations for general effect were both lower and more disparate. (see table 2)

Table 2

Correlations Between Independent Judges' Ratings of Categories & Final Ratings, 1983, 1986, 1987, 1989, and 1990

	Correlation with Final Rating			
Category/Year	*Jdg 1*	*Jdg 2*	*Jdg 3*	*Combined*
Tone				
1983	.78	.71	.80	.88
1986	.83	.82	.86	.82
1987	.95	.70	.82	.81
1989	.79	.30 ns	.84	.69
1990	.92	.80	.79	.83
Intonation				
1983	.72	.77	.69	.88
1986	.74	.84	.72	.79
1987	.85	.68	.83	.82
1989	.76	.41 ns	.69	.68
1990	.88	.90	.88	.89
Technique				
1983	.81	.93	.66	.93
1986	.81	.74	.94	.83
1987	.82	.84	.83	.86
1989	.64	.64	.84	.86
1990	.75	.95	.87	.90
Selection				
1983	.78	.74	.59	.74
1986	.72	.61	.77	.64
1987	.64	.79	.85	.74
1989	.75	.57	.66	.61
1990	.48 ns	.66	.75	.55 ns
Interpretation				
1983	.84	.87	.63	.90
1986	.79	.82	.90	.82
1987	.76	.90	.87	.85
1989	.66	.55	.88	.82
1990	.71	.74	.93	.84
General Effect				
1983	.67	.83	.61	.86
1986	.50 ns	.61	.00 ns	.47 ns
1987	.93	.76	.00 ns	.60
1989	.72	.53	.87	.72
1990	.70	.82	.70	.81

Prediction of Final Ratings from Category Ratings

The multiple regression analyses sought to determine the extent to which combined judges' category ratings predicted the final ratings. With separate analyses conducted for each year's ratings, technique proved to be the best predictor for all years. (See Table 3.) Generally, the addition of other category ratings did not increase the multiple correlation coefficient greatly. The greatest increases in the adjusted R^2 values after all variables were added in the regression analysis were for 1986 (from .69 to .86) and for 1987 (from .74 to .89); the least change was for 1983 (from .87 to .93).

Table 3

Stepwise Multiple Regression Analysis: Prediction of Final Rating from Category Ratings

Year/Categories	*R*	R^2	R^2 *adj.*	*F*	*DF*	*P*
1983						
Tech	.93	.87	.87	248.78	1,37	.001
Tech., Inton.	.95	.91	.90	175.38	2,36	.001
Tech., Inton., Inter.	.96	.93	.92	147.44	3,35	.001
Tech., Inton., Inter., Select.	.97	.94	.93	124.12	4,34	.001
Tech., Inton., Inter., Select., General	.97 (All)	.94	.93	101.80	6,33	.001
1986						
Tech.	.83	.70	.69	85.14	1,37	.001
Tech., Inter.	.90	.81	.80	75.50	2,36	.001
Tech., Inter., Gen.	.92	.85	.83	63.97	3,35	.001
Tech., Inter., Gen., Tone	.93	.87	.86	57.68	4,34	.001
Tech., Inter., Gen., Tone, Inton.	.94	.88	.86	49.26	5,33	.001
Tech., Inter., Gen., Tone, Inton., Select.	.94	.88	.86	40.58	6,32	.001
1987						
Tech.	.86	.75	.74	125.73	1,43	.001
Tech., Inter.	.92	.85	.85	123.06	2,42	.001
Tech., Inter., Tone	.94	.88	.87	101.42	3,41	.001
Tech., Inter., Tone, General	.95	.90	.88	85.37	4,40	.001
Tech., Inter., Tone, General, Inton.	.95 (All)	.90	.89	73.88	6,39	.001
1989						
Tech.	.86	.74	.74	149.82	1,52	.001
Tech., Inter.	.91	.83	.82	128.62	2,51	.001
Tech., Inter., Select.	.92	.85	.84	93.35	3,50	.001
Tech., Inter., Select., General	.93	.86	.85	74.23	4,49	.001
Tech., Inter., Select. General, Inton.	.93	.86	.85	61.34	5,48	.001
Tech., Inter., Select., General, Inton.,	.93	.87	.85	50.47	6,47	.001
1990						
Tech.	.90	.80	.80	151.00	1,37	.001
Tech., Tone	.93	.87	.87	124.77	2,36	.001
Tech., Tone, Inter.	.95	.90	.89	102.20	3,35	.001
Tech., Tone, Inter., General	.95	.91	.90	85.18	4,34	.001
Tech., Tone, Inter., General, Inton.	.96 (All)	.91	.90	69.51	6,33	.001

Table 4

Comparison of Judges' Mean Ratings for Categories and Final Ratings

Category/ Year	*No. of Orch.*	Mean Ratings *Judge 1*	*Judge 2*	*Judge 3*	*F*	*P*
Tone						
1983	13	2.31	2.83	3.03	2.37	.11
1986	13	2.52	3.08	2.67	1.82	.18
1987	15	2.72	2.80	2.56	0.53	.59
1989	18	3.01	3.46	3.28	1.73	.19
1990	13	2.75	2.85	3.03	0.32	.77
Intonation						
1983	13	1.66	2.48	2.63	4.58	.02*
1986	13	2.47	2.82	2.17	1.92	.15
1987	15	2.38	2.51	2.31	0.29	.75
1989	18	2.20	2.78	3.04	4.56	.02*
1990	13	2.59	2.69	2.34	0.33	.72
Technique						
1983	13	2.36	3.03	2.83	1.93	.16
1986	13	2.57	3.31	2.72	3.18	.05(4)
1987	15	2.63	2.53	2.74	0.37	.69
1989	18	3.23	3.16	3.15	0.05	.95
1990	13	2.52	2.77	2.62	0.20	.82
Selection						
1983	13	3.39	3.37	3.39	0.00	.99
1986	13	2.77	3.46	3.08	2.73	.08
1987	15	3.55	2.95	3.20	2.64	.08
1989	18	3.44	3.68	3.61	0.43	.65
1990	13	3.85	3.54	3.34	2.68	.08
Interpretation						
1983	13	2.71	3.03	3.03	0.68	.51
1986	13	3.05	3.08	2.77	0.85	.44
1987	15	3.23	2.69	2.92	2.38	.11
1989	18	3.28	3.17	3.33	0.27	.77
1990	13	3.00	2.92	2.69	0.33	.72
General Effect						
1983	13	2.69	3.18	3.32	2.56	.09
1986	13	3.31	3.46	4.00	5.15	.01*
1987	15	2.97	3.15	4.00	10.60	.01*
1989	18	3.43	3.72	3.37	1.30	.28
1990	15	3.15	3.08	3.13	0.02	.98
Final						
1983	13	2.43	3.00	2.92	1.58	.22
1986	13	2.69	2.85	2.85	0.17	.85
1987	15	2.87	2.69	2.93	0.42	.66
1989	18	3.32	3.08	3.44	1.19	.31
1990	13	2.77	2.92	2.85	0.08	.92

The greatest increases were after the addition of the second variable, which for 1986, 1987, and 1989 was interpretation. Although there were again year-to-year differences, other variables seemed to add little to the predictive strength of technique and interpretation. Selection ratings added virtually no predictive value for any of the five years.

Comparison of Judges' Mean Ratings

Whereas the interjudge reliability data reflected relationships between judges' ratings, the repeated-measures ANOVA was used to compare the judges' mean ratings for each category and their respective final ratings. As shown in Table 4, there were no statistically significant differences among the judges' mean ratings during any year for tone, technique, selection, interpretation, or final rating. The only statistically significant differences in the judges' mean ratings were for intonation in 1983 and 1989 and for general effect in 1986 and 1987.

Discussion

As apparent from the data, the interjudge reliabilities for both the final ratings and the category ratings were unsatisfactory. Certainly, interjudge reliability coefficients for final ratings should be much higher than those observed in the present study. The reliability coefficients for 1986, 1987, and 1990, which were in the .80s, were marginally acceptable, but the reliability coefficients for 1983 (.67) and 1989 (.54) were clearly unacceptable. While it is expected that judges will vary some in the standards they apply in adjudication festivals, and hence the level of ratings, there should be a consistency of rating from orchestra to orchestra.

Interjudge reliability coefficients for the various category ratings were even lower than those for the final ratings; they also reflected gretaer year-to-year variation than the final ratings. General effect reliability ratings were the lowest, suggesting that individual judges have differing interpretations of general effect. A further examination of the data revealed that for both 1986 and 1987 at least one judge failed to make any distinctions among the orchestras for the general effect category. Why no distinctions were made is unclear, but an examination of the verbal descriptors on the adjudication form may have been a factor. The two verbal descriptors for the general effect category were "stage presence" and "artistry," which are seemingly incongruous. General effect apparently has various meanings to different judges and may be redundant with the categories and their descriptors. The descriptors "stage presence" and "artistry" provide little guidance to the adjudicators.

Other categories reflecting very low (less than .40) interjudge reliability coefficients for at least one year were tone and intonation. Further research revealed that the judges for the years with the very low reliabilities included relatively inexperienced adjudicators who also had different musical training and experience. One was an experienced public school strings/orchestra teacher, one was a performer and applied music teacher, and the other was a composer/theorist who had some experience as a youth orchestra conductor. Perhaps their varied professional backgrounds were contributing factors to the way they judged tone and intonation of the orchestras. For the other years, however, interjudge reliability coefficients for tone and intonation were generally in the .70s, which is considerably better, but still less than desired.

While not reflected in the data, another factor may have had some bearing on the ratings. This factor, which was learned from talking with some of the adjudicators, concerned information extraneous to the musical performance which had been conveyed to judges. Apparently judges were informed that some orchestra programs were in early stages of development, and perhaps the judges varied in the extent to which they sought to be "encouraging" or "accommodating" in their ratings of these programs.

Judges' individual and combined ratings for the tone, intonation, technique, and interpretation categories tended to correlate well with the final rating, while correlations for the selection and general effect categories were much lower. The lower correlations undoubtedly are a partial reflection of the inconsistencies among the judges' ratings for the general effect and selection categories, but they may also be reflections of the different nature of these two categories. The problems with the descriptors for the general effect category were noted above, but the fact that both the selection and the general effect categories take into consideration variables other than performance also may contribute to the low correlations. Perhaps the profession should re-examine the need for these categories on adjudication forms; at the very least, the descriptors for the categories need re-thinking.

The data from the regression analysis are an extension of the correlational data between the categories and the final ratings. For three of the five years examined, the four performance categories (technique, intonation, interpretation, and tone) accounted for most of the predictive value. Selection and general effect contributed virtually nothing toward predicting the final ratings.

Data from the comparison of the judges' mean ratings of the orchestras for the respective categories, however, are somewhat more encouraging. The judges' mean ratings were not significantly different for any year's final ratings for tone, technique, interpretation, selection, general effect, and the final ratings. Apparently, judges are in general agreement with respect to the levels at which they rate the groups overall.

Implications

1. The profession needs to re-examine some of the categories, especially general effect and selection, on such adjudication forms. There is a particular need to re-think the descriptors for the various categories and to include descriptors that will have a common meaning for all adjudicators.
2. Guidelines for adjudicators need to provide more and better information regarding the use of the categories in arriving at a final rating. Descriptors for the various categories should be well defined.
3. Some type of adjudicator orientation should be developed to ensure that adjudicators have a common understanding of the terms, the categories, and their use in arriving at the final ratings.
4. Festival managers should be careful to avoid any comments either prior to or during the evaluation festival that might inadvertently bias judges toward leniency in their ratings of orchestras from programs that are in early stages of development. Evaluation festivals should provide as accurate and objective ratings as possible.

REFERENCES

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

Burnsed, V., & King, S. (1987). How reliable is your festival rating? Update, 5(3), 12-13.

Fiske, H. E. (1978). The effect of a training procedure in musical performance evaluation on judge reliability. Ontario Educational Research Council report.

Fiske, H. E. (1983). Judging musical performances: Method or madness? Update, 1(3), 7-10.

Gordon, E. (1965). Musical Aptitude Profile. Boston: Houghton Mifflin.

Massel, P. (1978). The influence of voice quality and the visual element on vocal adjudication. Unpublished master's thesis, University of Western Ontario.

Mullin, A. (1979). Melodic/rhythmic decision-making by senior secondary music students. Unpublished master's thesis, University of Western Ontario.

Seashore, C. E., Lewis, D., & Saetveit, J. (1939, 1960). Seashore Measures of Musical Talents. New York: The Psychological Corporation.

Towers, R. (1980). Age-group differences in judge reliability of solo voice performances. Unpublished master's thesis, University of Western Ontario.

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