

Running head: RATER

Judging The Judges:  
Improving Rater Reliability at Music Contests

David Conrad  
Manteno Community Unit School District #5  
Manteno, Illinois

August 16, 2003

## Abstract

Rater reliability affects music contest results. Research has shown that judge training, double-judging, having valid categories, blind-scoring, defining the scoring criteria, and designing clear rating forms can reduce judging errors. Educators need additional research to examine the effects of judge conferencing. Music educators should use research findings to evaluate and refine the rating forms used in music contests.

## **Judging The Judges: Improving Rater Reliability at Music Contests**

Music contests are an important part of most school music programs. For many music educators, they are integral to the public's perceptions of our programs and of our professional competence (LaRue, 2001). With so much weight and importance placed on the outcome, can we really trust the judges' evaluations at music contests?

At some time in our careers, we have all complained about music judging – someone was too harsh, the results were unexpected, or the judges gave wildly different scores. Setting aside the philosophical debate of whether music contests are good or bad for us, researchers have studied music judging and suggest several improvements to address these problems.

### Measuring Reliability

We can't judge music performances using a simple multiple-choice or matching test. Educators have developed various rating forms to aid music judging. However, designing reliable performance assessments has been challenging, as even the best rubrics require some subjective evaluation from the grader.

The term *test reliability* describes whether a test and its results are consistent over time. If the same judging panel judges the same performance on two different days, the results should be the same. Likewise, if two different judging panels judge the same performance on the same day, the

results should also be the same. If the same group played an identical performance twice – with the same errors and successes – their scores should be identical both times they grade them.

One specific measure of test reliability is the consistency of the judges, also known as *rater reliability*. Two types of rater reliability are usually associated with music judging:

1. *Interrater reliability* is the extent to which each judge agrees with one another in their evaluation of the same performance. According to the Illinois State Board of Education (1995), “high interrater reliability indicates that the raters used the same criteria to evaluate a performance and that they understood and applied the criteria similarly” (p. 57). How important is interrater reliability? Approximately 50% of the error variances in rating scales can be attributed to disagreement between raters, resulting in low interrater reliability (Anastasi, as reported in Tesse et al., 1994).

Machine graded tests should have perfect interrater reliability. The machine reads each cards same objective criteria. Yet as Americans learned during the 2000 election controversy in Florida, machine punch cards to are subject to other errors.

2. *Intrarater reliability* is the extent to which outside factors may influence a judge’s scoring. Did the group perform first or last in the day? Was the judge daydreaming? If the same judge evaluated the same performance at two different

times, would the results be the same? When a judge's scoring process changes over time, the results may suffer from low intrarater reliability.

Do judges make honest mistakes? They are humans, and are therefore subject to certain flaws that can skew their judging. They bring their own unique experiences, knowledge, and skills to their work. Common rater errors include:

1. Personal bias: when a judge rates a series of performances using a limited part of the rating scale, such as grading all performances too generously or too severely, or only grading near the middle of the rating scale (a central tendency error).
2. Logical error: when separate elements of a rating scale are confused, such as when a judge instinctively evaluates *tone quality* and *intonation* the same because of an assumption that out-of-tune singers always have poor tone quality.
3. Halo effect: when overall perception affects the final rating, such as a judge who knows the past reputation of a band and allows that knowledge to affect the scoring of the present performance.

### Improving Rater Reliability

Research suggests several methods to increase rater reliability, and by that increasing the effectiveness of music contest judging:

- **Include the category of *musicality* or *musical effect*.** In two studies, *musicality* had the highest criterion-validity in relationship to the rank order of students (Owen, 1969) and to the final global rating (Abeles, 1973a). Another research study found that the category of *musical effect* had the highest correlation to the final rating (Burnsed, Hinkle, and King, 1985).
- **Eliminate or clarify ambiguous categories such as *other factors*, *general effect*, and *selection*.** Burnsed, Hinkle, and King (1985) found that judges did not rate the *other factors* category or always gave it the highest score. In fact, that category was so unreliable that they excluded it from the data in their final analysis. Garman (1991) studied interrater reliability for the Dade County Orchestra Festival, focusing on the correlations between the individual category scores and the final global rating of each judge. He found that the categories of *technique* and *intonation* had the highest reliability. He also concluded that the category *general effect* had low reliability and should be removed, while *selection* had low interrater reliability. Vague or nonexistent criteria typically accompany these categories.

Having clear descriptions of all categories will increase the intrarater reliability.

According to Moskal & Leydens, “well-designed scoring rubrics respond to the concern of intrarater reliability by establishing a description of the scoring criteria in advance” (2000). To reduce bias, judges should constantly refer to these criteria.

- **Decide which categories are most critical.** What are the essential behaviors that should be measured? Abeles studied one design method, employing a facet-factorial analysis to select six categories for a clarinet performance rating scale (1973a & 1973b). He asked college music students to write an essay that described a given clarinet performance. From these essays, the author used factor analysis to establish six categories: interpretation, tone, intonation, rhythm, musicality, and articulation. Boyle supports the facet-factorial design process, as it is the most frequently used tool by researchers in music assessment (as reported in Colwell, 1992). It provides a framework that music educators can use when redesigning rating forms.
- **Eliminate overlapping categories.** Burnsed, Hinkle, and King (1985) concluded that separate categories are often judged as interrelated factors. While concluding that rating scales *can* have high internal consistency, they argued that most music judges evaluate performances from a global perspective. Separate categories may not represent distinct entities in the minds of judges. Fiske (as reported in Colwell, 1992) concluded that using one global rating – *instead of several separate categories, averaged together* – would yield reliable results.

For example, a concert band that plays out-of-tune may score low in the intonation category, and the offending intonation may also pull-down scores in the blend, tone quality, and musicality categories. Specific criteria should clarify the differences between each category.

- **Use multiple judges.** Averaging the scores of two or more raters can reduce the likelihood of error. One Maryland school district trained math and language arts teachers to grade performance-based exams. The district found that many teachers were simply lousy test graders! The district felt it could not forbid these teachers from grading their own students' exams. They solved this by using double-scoring, where two teachers would grade each test and average their scores (Myerberg, 1978).
- **Use blind scoring.** Bias errors can be reduced with when the identity of the performers are hidden from the raters. Colwell (1992) challenges music teachers to employ these controls, whenever practical to do so.
- **Provide judge training.** Rating errors can be reduced when raters are trained in the use of a particular rating instrument (Gordon, 1970). Training should include practice scoring sessions, allowing judges to compare and discuss their evaluations with one another. According to Garman (1991),

“Some type of adjudicator training should be developed to ensure that adjudicators have a common understanding of the terms, the categories, and their use in arriving at the finals ratings” (p. 23).
- **Design clear and effective rating scales.** Ables (1973a) argues that printed rating scales force the assessor to use an established set of criteria. He emphasized that rating sheets should list the specific criteria for evaluation, with uniform judging instructions.

Owen (1969) sought to reduce the effects of judge subjectivity by designing new rating forms. He studied three designs and found that graphic-continuum scale was the most effective, showing high interrater reliability (.90). Other researchers have extended his design model to clarinet performances (Ables, 1973a), high school bands (DCamp, 1974), choral ensembles (Cooksey, 1980), and tuba/euphonium performers (Bergee, 1987). The Ables study achieved interrater reliability above the .90 level (1973a, 254), while the Bergee study achieved inter-reliability between .94 and .97 (as cited in Colwell, 1992).

- **Consider judge conferencing.** Some music contest rules forbid judges from talking to one another about their evaluations. However, other disciplines have found that diverse scoring interpretations can be resolved through discussion among the judging panel, usually lead by a supervisor or head judge (Illinois State Board of Education, 1995 & Myerberg, 1978). Educators need research to examine the effects of judge conferencing in music contests.

### Conclusion

Music contests can provide greater benefit when we reduce the chances of rater error. Since well-designed rating forms can have high rater reliability, music educators should use research findings to refine the judging process.

## References

Ables, H.F. (1973a). Development and validation of a clarinet performance scale. Journal of Research in Music Education, 21(3), 246-255.

Ables, H.F. (1973b). A facet-factorial approach to the construction of rating scales to measure complex behaviors. Journal of Educational Measurement, 10(2), 145-151.

Bergee, M.J. (1987). An application of the facet-factorial approach to scale construction in a rating scale for euphonium and tuba music performance. Unpublished doctoral dissertation, University of Kansas, Lawrence.

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

Colwell, R. (Ed.). (1992). Handbook of research on music teaching and learning. New York: Schirmer Books.

Cooksey, J.M. (1974). An application of the facet-factorial approach to scale construction in the development of a rating scale for high school choirs. Unpublished doctoral dissertation, University of Illinois, Urbana.

DCamp, C.B. (1980). An application of the factor-factorial approach to scale construction in the development of a rating scale for high school band performance. Unpublished doctoral dissertation, University of Iowa, Ames.

Garman, B.R., et. al. (1991). Orchestra festival evaluations: Interjudge agreement and relationships between performance categories and final ratings. Research Perspectives in Music Education, (2), 19-24.

Illinois State Board of Education (1995). Assessment handbook: A guide for developing programs in Illinois schools. Springfield, IL: State of Illinois.

LaRue, P. (2001). Ma, we got a II at contest! NFIMA Journal, 18 (1), 15-17.

Moskal, B. & Leydens, J. (2000). Scoring rubric development: validity and reliability. Practical Assessment, Research & Evaluation, 7(10). [Available on-line at <http://edresearch.org/pare/getvn.asp?v=7&n=10>].

Myerberg, N.J. (1978). Inter-rater reliability on various types of assessments scored by school district staff. Unpublished paper presented at the Annual Meeting of the American Educational Research Association: New York. [ERIC Document Reproduction Service No. ED 400 291].

Owen, C.D, Jr. (1969). A study of criteria for the evaluation of secondary school instrumentalists when auditioning for festival band. Office of Education, Bureau of Research: Washington, D.C.

Tasse, M.J., et.al. (1994). Test-retest and interrater reliability of the Quebec Adaptive Behavior Scale. Unpublished paper presented at the Annual Conference of the American Association of Mental Retardation. [ERIC Document Reproduction Service no. ED 379 873].

### Biography

David Conrad teaches band, chorus, and show choir in the Manteno, Illinois schools. He holds music education degrees from the *University of Illinois* and *Illinois State University*, and is completing course work in school administration at *Governors State University*. He has worked as an assessor and a pilot tester for the National Board for Professional Teaching Standards and has created music rubrics used by teachers throughout North America. His popular lecture program on school media relations has produced a training video and companion web site used in university graduate courses. His writing has been published in *The Irish Music Teacher* and *School Community Relations*.