



## MENC: The National Association for Music Education

---

Development and Validation of a Clarinet Performance Adjudication Scale

Author(s): Harold F. Abeles

Reviewed work(s):

Source: *Journal of Research in Music Education*, Vol. 21, No. 3 (Autumn, 1973), pp. 246-255

Published by: [Sage Publications, Inc.](#) on behalf of [MENC: The National Association for Music Education](#)

Stable URL: <http://www.jstor.org/stable/3345094>

Accessed: 17/09/2012 23:44

---

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at

<http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).



*Sage Publications, Inc.* and *MENC: The National Association for Music Education* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of Research in Music Education*.

<http://www.jstor.org>

## DEVELOPMENT AND VALIDATION OF A CLARINET PERFORMANCE ADJUDICATION SCALE

Harold F. Abeles

One of the main difficulties in the evaluation of a complex behavior such as music performance is that the measures employed are typically subjective judgments based on irregular and uncontrolled observations.<sup>1</sup> A consensus among judges concerning the adequacy of a performance is difficult to obtain. One method of improving the evaluation of music performance is through replacement of judges' general impressions by ratings arrived at by more systematic procedures. Rating scales improve evaluation because adjudicators must use a common set of evaluative dimensions rather than develop their own subjective criticisms. If the evaluative dimensions (for example, items in the scale) adequately sample the content area under investigation, the scale should have satisfactory content validity. An additional benefit of a common set of evaluative dimensions should be an increase in interjudge reliability of performance ratings.

Individual music teachers have unique objectives, particularly at the more advanced levels. However, a basic assumption of this study is that there are generally agreed upon performance standards as evidenced by the use of adjudicators for evaluations at contests and festivals. Assuming there are common standards, it seems desirable to develop an instrument to enable raters to measure effectively those aspects of performance that have common standards of proficiency.

<sup>1</sup> William K. Whybrew, *Measurement and Evaluation in Music* (Dubuque, Iowa: William C. Brown Company Publishers, 1962), p. 63.

## Procedures

The first phase of the study was to obtain a sample of the aspects of clarinet performance that concern instrumental music teachers in their evaluation of performance. Seventeen instrumental music teachers enrolled in graduate music education courses at the University of Maryland in the fall of 1970 were asked to write a one- or two-page essay describing the auditory aspects of a music performance by a junior high school clarinetist. The researcher made a content analysis of the essays. Fifty-four different statements descriptive of clarinet performance were identified. The statements were generally left in the words of the respondents, with minor changes made in cases of ambiguity. To understand better the aspects of music performance included in the fifty-four statements, an a priori theoretical structure of music performance was developed based on a structure used by Hosmer to base music performance ratings.<sup>2</sup> The structure includes seven factors of music performance: tone, intonation, interpretation, technique, rhythm, tempo, and general effect (for example, spirit).

The fifty-four statements developed by content analysis were listed under categories of the adopted a priori structure by the researcher and a faculty member of the University of Maryland music education department. To broaden the sample of clarinet performance descriptions, a list of adjectives used to describe music performance was developed from an examination of several studies.<sup>3</sup> Forty additional statements were developed by putting the adjectives into sentence form similar to the statements generated by the descriptive essays.

The ninety-four statements were next transformed to items so that they could be used by instrumental music teachers to rate actual clarinet performances. The items were phrased both positively and negatively in order that adjudicators would not develop a response set when evaluating. The items were then randomly ordered, and a five-point Likert-type response scale was developed. Responses ranged from "highly agree that the statement is descriptive (of the performance)" to "highly disagree that the statement is descriptive." A panel of three instrumental music teachers reviewed the items independently and judged them to be positive or negative statements about performance. The three teachers agreed on all items. The following are examples of items from the performance description item pool: (1) The attacks and releases were clean. (2) Musical communication was effective. (3) The clarinetist played with a natural tone. (4) He played flat in the low register. (5) He played too slowly.

<sup>2</sup> Helm Hosmer, "As the Adjudicator Hears It," *The Etude*, Vol. 67 (April 1949), pp. 224, 262.

<sup>3</sup> Robert Fleury, *Objective Measurement of Group Instrumental Music* (doctoral dissertation, University of California at Los Angeles, 1964); Victor Grabel, "Educational Value of Contests," *The Etude*, Vol. 51 (June 1933), pp. 379, 413; Max T. Krone, "Toward a More Helpful Adjudication," *Music Educators Journal*, Vol. 23 (March 1937), pp. 28-29; Lee M. Lockardt, "Constructive Criticisms for Contests and Festivals," *Music Educators Journal*, Vol. 21 (October 1934), pp. 17-20.

Samples of one hundred clarinet performances were collected on tape. The performers were fifty clarinetists enrolled in four junior high schools (grades seven, eight, and nine) in Prince George's County, Maryland, during the fall of 1970. They had been studying clarinet in the public schools from two to five years. Each clarinetist recorded two different performances. Fifty instrumental music teachers from Prince George's County were each asked in January 1971 to evaluate two different randomly assigned performances by rating them on each of the ninety-four items in the pool. One hundred different performances, each of a different piece of music, were evaluated. The judges were asked to "as accurately as possible describe the performance which you have just heard," employing the response scale and statements of the item pool. Each judge was told that the performer was an eighth grade student who had been enrolled in public school instrumental music for three years. Each performance was played back several times to ensure that the judge had the aural stimulus of the performance present during the evaluation of each item.

To examine empirically the factor structure of clarinet music performance as defined by the ninety-four items, a factor analysis was performed on the results of the administration of the item pool. The usefulness of the structure produced by factor analysis may ultimately depend on the ease of interpreting the relation of the structure to clarinet music performance. In this study, from four to ten factors were rotated employing varimax rotation. This enabled the researcher to examine the number of factors that best agreed with the a priori theoretical structure previously described. The factor analysis solution that best agreed with the a priori structure was a six-factor rotation. The six factors were interpretation, intonation, rhythm continuity, tempo, articulation, and tone.

Items were selected for the Clarinet Performance Rating Scale (CPRS) based on the factor matrix produced by the analysis of the one hundred performance evaluations employing the ninety-four-item pool. Thirty items were selected to ensure that the rating scale would not be too long for practical use. Five items were chosen to represent each of the six factors. The items chosen had relatively high factor loadings on the factor they were selected to define and had relatively low correlations with the five other factors. On factors where many items met these two criteria, the diversity of the items in sampling apparently different content of a factor was considered in the selection. The thirty statements were grouped by factors and paired with a five-point Likert-type scale, as in the item pool previously described. Both positive and negative statements were selected from the item pool.

To estimate the interjudge reliability of the CPRS and obtain initial data on criterion-related validity, instrumental music teachers were asked during spring 1971 to use the instrument in rating a set of recorded performances. The items ratings obtained were also factor-analyzed to examine the stability of the structure developed from the original pool of ninety-four items. Thirty-two instrumental music teachers who were enrolled in University of Maryland graduate music education courses were divided into three adjudication groups. Group one consisted of nine

judges, group two of eleven, and group three of twelve. Three sets of ten performances each were randomly selected from the original hundred. Members of each adjudication panel were asked to evaluate independently one set of performances, employing the response scale and statements of the CPRS. Each performance lasted approximately ninety seconds and was heard only once by each panel as a whole in order to simulate an actual adjudication situation.

Results of administration of the CPRS were used to provide an interjudge reliability estimate for the rating scales. Estimates of the interjudge reliability for the factor scores as well as the total scores were obtained for each of the three adjudicator groups. To examine the effect of employing different numbers of judges, reliability estimates were obtained for adjudication panels varying in size from one through ten.

To examine the criterion-related validity of the CPRS, each set of ten performances was rated using the method of paired comparisons. The paired-comparison ratings were obtained by placing forty-two instrumental music teachers into three equal-size groups. Each judge was asked: "In each pair select the performance which demonstrates the best overall musical ability." Each performance in a set was paired with the nine other performances, yielding a total of forty-five pairs. Presentation of the pairs was randomly ordered. Interjudge reliability estimates were obtained for the paired-comparison ratings. These estimates were based on the rank order (from the paired comparisons) of the performances by the judges. Within each of the three performance sets, zero-order correlation coefficients were obtained between the results of the administration of the CPRS and both ranks and normalized scale values of the paired-comparison data. Correlations were obtained with CPRS factor scores as well as the total scores.

## Results

The final form of the rating scales was based on the six-factor rotation of the ninety-four-item pool evaluations. The factor names applied to the six-factor matrix described the items that were highly loaded on each factor. Factor 1 clearly seemed to represent the interpretation element of a performance. Factor 2 seemed to be easily interpreted as a tone factor, as all the items with loadings above .60 dealt with some aspect of tone quality. The third factor was somewhat more difficult to interpret, as it seemed to combine items from the a priori categories of rhythm, technique, and interpretation. Since all items on this factor seemed descriptive of the flow of the music, the term rhythm continuity was considered appropriate. Items weighted highly on factor 4 involved the fast-slow aspects of the performance, and therefore this factor was labeled tempo. Factor 5 encompassed the items involving aspects of intonation and flatness and was termed intonation. The sixth factor, which was difficult to interpret, was a kind of articulation variable, for the highly loaded items on it dealt with aspects of tonguing.

The thirty items selected for the CPRS appear in Table 1, grouped under the factors that they were selected to measure. Items were selected based upon their loadings on the six-factor matrix of the ninety-four items selected. When one of these items was also highly loaded on another factor, it was replaced by an item that had lower correlations with other factors. If more than five items met the criteria of high primary factor loadings and factor simplicity, items were chosen that appeared by their content to sample more varied aspects of the factor.

Table 1  
Factor Loadings for Items Selected for the CPRS

Items	Factors					
	1	2	3	4	5	6
<b>1. Interpretation</b>						
Effective musical communication	752		310			
The interpretation was musical	778		334			
The piece was played in character	722		269			
Played with musical understanding	698		339			
Played with traditional interpretation	673				271	
<b>2. Tone</b>						
Thin tone quality		-654	205			
Played with a natural tone	212	-670				321
There was a lack of tonal color		-613				
The quality of the tone was rich	307	-721				
Sounded shallow		-705	262			
<b>3. Rhythm/Continuity</b>						
Uneven rhythm	344		694	238		
Smoothness in execution	398		570			399
Melodic continuation	463		633			
Insecure technique	212	-204	642			
The rhythm was distorted	218		538	221		
<b>4. Intonation</b>						
Played out of tune	338				549	
Flat in low register		-216	272		450	
The intonation was good		-251			659	
Played overall flat		-404			516	
Tended to be flat		-244			534	
<b>5. Tempo</b>						
Played too fast				553		
Seemed to drag			258	-603		
Hurried repeated notes				604		
Played too slowly		-281	327	-634		
Rushed			266	643		231
<b>6. Articulation</b>						
Squeaked	233		212		204	591
Free from tonguing noise	226					540
Attacks and releases were clean		-257	263		227	632
Tonguing produced thunkie sound						531
Accents were played as indicated	227	-352				669

The factor matrix produced by factor analysis of the evaluations, using the thirty items of the CPRS and forcing a six-factor structure, is shown in Table 2. These results produced a factor structure similar to the one based on the factor analysis of the ninety-four-item pool results. The results also indicated a pattern of item loadings that supported the item selection procedure for the six-factor subscales.

Table 2  
Varimax Six-Factor Matrix of the CPRS

Variables	1 Intona- tion	2 Rhythm/ Con- tinuity	3 Tempo	4 Articu- lation	5 Tone	6 Inter- pre- tation
Effective musical communication	-.09	.37	-.09	.12	.13	-.78
The interpretation was musical	-.09	.29	-.13	.11	.16	-.82
The piece was played in character	-.08	.24	.00	.14	.06	-.80
Played with musical understanding	-.08	.23	-.09	.19	.06	-.81
Played with traditional interpretation	-.08	.31	.00	.21	.06	-.74
Thin tone quality	-.20	.15	.03	.21	.82	.06
Played with natural tone	-.35	.03	.01	.21	.67	-.21
There was a lack of tonal color	-.20	.14	-.01	.10	.67	-.16
The quality of the tone was rich	-.17	.14	.02	.24	.69	-.14
Sounded shallow	-.23	.12	.09	.23	.81	-.02
Uneven rhythm	-.05	.74	-.06	.06	.10	-.26
Smoothness in execution	-.15	.73	-.06	.14	.07	-.30
Melodic continuation	-.11	.77	-.09	.12	.09	-.29
Insecure technique	-.12	.79	.05	.15	.17	-.25
The rhythm was distorted	-.12	.82	-.06	.05	.17	-.23
Played out of tune	-.67	.22	.07	.16	.33	-.23
Flat in low register	-.80	.10	-.15	.15	.21	-.06
The intonation was good	-.65	.12	.00	.15	.34	-.29
Played overall flat	-.86	.05	-.06	.19	.16	-.01
Tended to be flat	-.82	.12	-.02	.15	.20	.00
Played too fast	-.08	.13	-.79	.05	.00	-.08
Seemed to drag	-.03	.11	.75	-.05	.18	-.12
Hurried repeated notes	-.01	.12	-.69	.04	.18	-.19
Played too slowly	.02	.07	.73	-.05	.14	-.02
Rushed	-.07	.17	-.86	.00	.05	-.15
Squeaked	-.24	.15	-.14	.56	.09	-.04
Free from tonguing noise	-.12	.03	-.07	.83	.21	-.09
Attacks and releases were clean	-.25	.12	.05	.65	.24	-.22
Tonguing produced thunkie sound	-.09	.12	-.07	.76	.15	-.17
Accents were played as indicated	-.10	.09	.00	.68	.27	-.24

Interjudge reliability coefficients were obtained for the three groups of instrumental music teachers who served as judges for the administration of the CPRS. The Hoyt Analysis of Variance procedure was employed. The interjudge reliability estimates for CPRS total scores were as follows: group one (n=9), .939; group two (n=12), .949; and group three (n=11), .978. Table 3 lists the interjudge reliability estimates for each of the five item scales and with each of the adjudication groups. An examination of these tables shows all but two of the scale reliability coefficients were above .70. All three of the coefficients for the total score were above .90.

Table 3  
Interjudge Reliability Estimates for CPRS Scale Scores

Factors		Groups		
		1 (n=9)	2 (n=12)	3 (n=11)
Interpretation	1	.899	.952	.945
Tone	2	.659	.766	.716
Rhythm/Continuity	3	.956	.876	.939
Intonation	4	.578	.835	.980
Tempo	5	.786	.922	.847
Articulation	6	.779	.883	.892

To estimate interjudge reliability for adjudication panels of different sizes, a generalized Spearman-Brown prophecy formula was applied to the interjudge reliability estimates.<sup>4</sup> The reliability estimate of the total scores was above .80 for as few as three judges. The scale score's reliability estimates for six or more judges were generally above .70 except for the tone subscale that had coefficients of about .60. Table 3 also indicates reasonable consistency in the reliability estimates across the three sets of evaluations. An examination of the scale reliability estimates shows the scale of intonation and tone yielded lower average estimates than did the other four scales.

In order to examine the criterion-related validity of the CPRS, the ten performances in each of the three sets used for the reliability analysis were given a global rating by instrumental music teachers, utilizing a paired-comparison procedure. The interjudge reliability of the paired-comparison criterion was .993 for group one, .985 for group two, and .978 for group three (n=14 for all groups). All three of the coefficients demonstrate a high degree of agreement among the judges, as they are above .97. The zero-order correlation between the CPRS scores and both the rank order and the normalized scale scores for the global performance rating appear in Tables 4, 5, and 6. Coefficients for both the CPRS total and scale scores are reported. Inspection of these tables indicates strong (above .80) relationships between total scores of the CPRS and the global

<sup>4</sup> Harold Gulliksen, *Theory of Mental Tests* (New York: John Wiley and Sons, 1950), p. 78.



performance ratings. The interpretation scale seems to have the strongest relationship (above .80) with the criterion.

## Conclusions

In the development of the CPRS the most substantial result concerned the structure of clarinet music performance. The six-factor structure produced from the factor analysis of the results of performance descriptions seemed essentially the same as the a priori theoretical structure based on a literature search. This six-factor structure for clarinet performance would also seem appropriate for classifying music performance in general, as none of the factors seem to reflect idiosyncratic clarinet characteristics. Additional investigation would be needed to support this generalization.

The interjudge reliability estimates for the CPRS were consistently high (.90). These results were similar to those obtained in other evaluations of music performance.<sup>5</sup> The resulting reliability estimates might have been partly a result of two sources of inflation: (a) the heterogeneity of the sample and (b) the added cues provided by all performances involving different pieces of music. Although the performers employed did represent the intended junior high school ability level, the subjects from four different schools may have increased the sample heterogeneity and therefore inflated the results. The use of different pieces of music by each performer created a problem in determining sources of variance. The total variance was due both to the differences among pieces as well as to the differences among performers. These two sources could not be examined separately. Additional studies that require judges to evaluate several different performers playing the same pieces seem warranted. The criterion-related estimates of the CPRS with the paired-comparison criterion seemed quite substantial (>.80). These results again paralleled validity coefficients reported in other studies.<sup>6</sup> But the results also might have been influenced by the heterogeneity of the sample or the use of different pieces.

The study results also demonstrated that the technique of scale construction employed to evaluate music performance seemed to produce an evaluation instrument that was both reliable and valid. The scale demonstrated good interjudge reliability, although these results must be interpreted in light of the number of judges employed. The interjudge reliabilities of the subscales (factors) were also high, but were even more dependent on the number of judges employed because there were so few items (five) contributing to each scale score. The construct validity of the rating scale was supported by the procedures of gathering performance

---

<sup>5</sup> Fleury; Kenneth V. Gutsch, "Objective Measurement in Instrumental Music Performance," Co-op Research Project No. 1734, U. S. Department of Health, Education and Welfare, Office of Education (Hattiesburg: University of Southern Mississippi, 1964); John G. Watkins, *Objective Measurements of Instrumental Performance* (New York: Bureau of Publication, Teachers College, Columbia University, 1942).

<sup>6</sup> Watkins.

descriptions supplemented by a literature search, and the item selection process based on the factor structure of clarinet performance. The predictive utility of the scale was demonstrated by the relationship between the rating scale and a global measure of performance.

There are four areas not touched on by this study that should be investigated in the future.

1. An investigation of the CPRS reliability and validity as a function of piece variance and performer by piece interaction should be attempted.

2. The construction of scales for the evaluation of clarinet performance at different age and ability levels, and the evaluation of music performance on instruments other than the clarinet should be attempted.

3. Further investigation seems warranted to resolve the problem of sample heterogeneity by applying the CPRS to performers within a class in a single school.

4. An examination of the generalizability of the structure of clarinet music performance generated in this study to music performance in general should be attempted.

■ University of North Carolina at Greensboro  
Greensboro, North Carolina