sempre:

Society for Education, Music
and Psychology Research

Article

Psychology of Music
2016, Vol. 44(3) 324–339
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0305735614567700
pom.sagepub.com
SAGE

Assessing jazz big band performance: The development, validation, and application of a facet-factorial rating scale

Brian C. Wesolowski

Abstract

The purpose of this study was to develop a valid and reliable rating scale to assess jazz big band performance and to evaluate the psychometric properties of jazz big bands at three performance achievement levels (e.g., low, moderate, high). The pool of initial scale items ($N = 22$) was gleaned from jazz big band research and instructional literature. Using a four-point Likert scale, rating responses ($N = 102$) were gathered for jazz ensemble performances ($N = 102$) from volunteer raters ($N = 102$). A factor analysis produced a reduced 18-item scale with a four-factor structure: blend/balance, time-feel, idiomatic nuance, and expression. The four factors accounted for 63.32% of the variance and had a total alpha reliability of .84. Discriminant function analyses revealed that four specific items contributed most to identifying ensembles with low and moderate performance ratings. The factor structure was able to predict group membership with 88.5% accuracy.

Keywords

assessment, ensemble, jazz, pedagogy, performance

Students are assessed for three broad reasons: (a) to promote learning; (b) to certify achievements; and (c) to provide data that can be used for program quality assurance (Yorke, 2008). Asmus (1999) noted that educational reform and related matters concerning teacher accountability provided grounds for the widespread demand of new assessment strategies. Most recently, the use of achievement data as a means to evaluate teacher effectiveness and student growth has required music educators to rethink how they collect and interpret evidence of achievement in the classroom (Crowe, 2010; Wesolowski, 2014). Conventional methods of performance evaluation tend to include a subjective mode of data collection and analysis, where raters and/or teachers set predetermined criteria based upon aesthetic value judgments. This often results in unreliable and invalid assessment outcomes (Lehman, 2007; Whybrew, 1973).

The University of Georgia, USA

Corresponding author:
Brian C. Wesolowski, The University of Georgia, Hugh Hodgson School of Music, 250 River Rd, Athens, GA 30602, USA.
Email: bwes@uga.edu

The growth of measurement development and research since the 1970s has verified that certain statistical methods can be applied to music performance assessment data that yield both reliable and valid results. Specifically, facet-factorial methods of scale construction attempt to systematically establish common items to a particular performance area. Facet-factorial approaches to rating scale construction have been created to accommodate instrument-specific performances (Abeles, 1971; Bergee, 1987; Jones, 1986; Nichols, 2005; Pazitka-Munroe, 2002; Russell, 2010a; Zdzinski & Barnes, 2002), instrumental jazz improvisation (Horowitz, 1994; D. T. Smith, 2009), jazz improvisation achievement (Pfenninger, 1990), and ensemble performance (Cooksey, 1977; DCamp, 1980; B. P. Smith & Barnes, 2007). An evaluation of existing facet-factorial scale development studies indicated a total of eight aurally perceived factors: (a) tone; (b) intonation; (c) rhythmic accuracy; (d) articulation; (e) dynamics; (f) timbre; (g) interpretation; and (h) tempo (Russell, 2010b). Russell's proposed model of music performance assessment specified two component factors that affect the perception of music performance: technique (e.g., tone, intonation, rhythmic accuracy, and articulation) and musical expression (e.g., dynamics, timbre, interpretation, and tempo).

A scale developed specifically for the measurement of jazz big band performance achievement is necessary as the central purpose of a measurement instrument is predicated on the notion that it measures a unique latent construct (Wilson, 2005). Although facet-factorial scale construction has demonstrated overlap of some items and domains (i.e., factors) between instrument-specific musical performances (Russell, 2010b), an intention should be established by the developer of any measurement tool that the results of all items working together 'can be interpreted to help make a decision as the measurer intended them to be' (Wilson, 2005, p. 5). A unique scale developed to measure the construct of jazz big band performance achievement does not currently exist, and its development is therefore necessary for reliable and valid measurement of such a unique construct.

Although facet-factorial rating scales exist for jazz improvisation (Horowitz, 1994; D. T. Smith, 2009), the shortcoming is that they focus solely on one component of the jazz big band performance paradigm. Pre-service music educators are often underprepared for successful jazz teaching in secondary schools (Ellis, 2007; Prouty, 2012; Treinen, 2011; West, 2011). More specifically, many public school educators are deficient in recognizing and diagnosing characteristic aural indicators of jazz performance (e.g., articulation, nuance, style, and timbre) (Baker, 1989). This study may aid educators in directing and teaching jazz big bands by adding to the overall scholarly content of jazz pedagogy and related fields, clarify aspects of evaluating the swing style, and help to identify and organize important evaluative criteria in jazz big band performance. The development of such a measurement scale may also provide insight into the unique and complex performance parameters of jazz big band performance, the most common medium of jazz education in secondary educational institutions (Dunscomb & Hill, 2002).

The purpose of this study was to develop a valid and reliable rating scale to assess jazz big band performance and to evaluate and compare the psychometric properties of jazz big bands at three performance achievement levels. The following research questions guided this study:

1. What items describe jazz big band performances?
2. What factors contribute to the assessment of a jazz big band performance?
3. What differences among jazz big bands exist at three performance achievement levels?
4. What criteria best predict group membership into three performance achievement levels (i.e., low achieving, moderate achieving, and high achieving)?

Method

Research Question 1

Item and scale development. In order to glean items that describe jazz big band performances, research and instructional literature that reference jazz big bands was thoroughly evaluated (Berry, 1990; Coker, 1978, Dunscomb & Hill, 2002; Goins, 2003; Jarvis, Beach, & Wiest, 2002; Kernfeld, 1995; Kuzmich & Bash, 1984; LaPorta, 1965; Lawn, 1981; Miles & Carter, 2008; Pizer, 1990; Wheaton, 1975; Wiskirchen, 1966). A total of 66 descriptive statements were collected. The statements were edited to produce short, concise, and useful item stems appropriate for a Likert rating scale. Any criteria indicating visual aspects of musical performance and items reflecting redundant criteria were combined or removed from the item-pool. Remaining items ($N = 37$) were categorized into performance factor groupings ($N = 8$) based upon similarity in content. The performance factors included the following categories: balance and blend, intonation, sonority, phrasing, articulation, stylistic interpretation, time-feel, and rhythmic accuracy.

To provide validity evidence of the item pool and content domains, the final pool of items ($N = 37$) organized into specified content domains ($N = 8$) was given to a panel of expert collegiate jazz educators ($N = 5$) to evaluate. The panel was asked to independently provide feedback on wording and relevance of items, verify the aural nature of the item description, mark redundant items, and label each item as having a neutral, positive, or negative connotation (Spector, 1992).

After receiving feedback from the initial judging panel, finalized statements chosen to be included in the final item pool ($N = 24$) were paired with a four-point Likert scale. Response options included ‘Strongly Agree,’ ‘Agree,’ ‘Disagree,’ and ‘Strongly Disagree.’ Previous research applying facet-factorial approaches to scale construction utilized five-point Likert scales (Abeles, 1971; Bergee, 1987; Horowitz, 1994; Russell, 2010a; Zdzinski & Barnes, 2002), and seven-point Likert scales (D. T. Smith, 2009). Although all studies have provided reliable results with the respective samples and chosen response sets, a four-point response set was chosen in order to eliminate a neutral response and provide a better measure of the intensity of participants’ attitudes and opinions (Asmus, 1981; Dumas, 1999). The items were removed from their related content domains and randomized in order to prevent possible rater effects related to specific content domain areas.

Pilot study. To test the overall functionality of the rating scale and provide further validity evidence of the item-pool ($N = 24$), the scale was piloted using jazz big band performances ($N = 8$) of varying ability levels. The rating panel ($N = 4$) consisted of full-time university jazz studies faculty members ($n = 2$) and professional performers with expertise in jazz big band performance ($n = 2$). The raters were each instructed to listen to the eight performances as many times as needed and to evaluate each ensemble using the specified measurement tool. Initial data was analysed and feedback regarding the measurement tool was solicited from the panel. The item-pool was revised and edited based upon panelist suggestions. The results were the elimination of two items from the pool. A finalized pool of 22 items was utilized in this study (see Table 1). Raters and recordings used in the pilot study were not used in the full study.

Selecting raters and performances. Volunteer raters ($N = 102$) were selected based upon their performance and teaching experience in the jazz idiom. The raters were drawn from a pool of professional musicians adept in the jazz idiom ($n = 29$), university professors holding positions

Table 1. 22-item Jazz Big Band Performance Rating Scale.

1. A steady tempo was kept throughout the performance	SD	D	A	SA
2. Articulations are consistent with a good concept of jazz phrasing	SD	D	A	SA
3. Background figures are well-balanced to the soloist during solo sections	SD	D	A	SA
4. Dynamic extremes are controlled	SD	D	A	SA
5. Eighth note values are given appropriate duration	SD	D	A	SA
6. Ensemble accents figures in an appropriate manner	SD	D	A	SA
7. Ensemble demonstrates a good concept of jazz phrasing	SD	D	A	SA
8. Ensemble demonstrates a uniform feeling of pulse	SD	D	A	SA
9. Ensemble is balanced to the lead trumpet player during ensemble passages	SD	D	A	SA
10. Ensemble maintains a steady time feel	SD	D	A	SA
11. Ensemble performs composition at an appropriate, idiomatic tempo	SD	D	A	SA
12. Ensemble performs with a time feel appropriate to the composition	SD	D	A	SA
13. Ensemble performs with understanding of the swing eighth note concept	SD	D	A	SA
14. Ensemble plays with a balanced sound in full passages	SD	D	A	SA
15. Ensemble plays with a large, full sound	SD	D	A	SA
16. Good overall balance between winds and rhythm section	SD	D	A	SA
17. Good overall blend between brass and saxophones	SD	D	A	SA
18. Lead players perform with appropriate and idiomatic nuances	SD	D	A	SA
19. Melodic lines end with an appropriate amount of emphasis	SD	D	A	SA
20. Phrasing of eighth note lines is executed smoothly	SD	D	A	SA
21. Rhythm section makes appropriate balance adjustments between ensemble and solo sections	SD	D	A	SA
22. The rhythm section and winds share a common feel for the pulse	SD	D	A	SA

in jazz studies departments (n = 27), and undergraduate (n = 19) and graduate (n = 27) students enrolled in jazz studies programs (Fiske, 1979). All raters were supplied an instructional packet that detailed the purpose of the study, specific use of the supplied measurement tool, task expectations, and related ethical policies and guidelines.

The anonymous recordings were full performances drawn from a pool of middle school, high school, collegiate, and professional jazz big band performances. Following the methodology of D. T. Smith (2009), raters were supplied two anonymous reference recordings that served as anchors to demonstrate: (a) a weak performance deserving the lowest ratings on the measurement tool; and (b) a strong performance deserving the highest ratings on the measurement tool. Anchor recordings were selected from the results of the pilot study. Each rater was provided one distinct jazz big band recording performed in a medium-tempo, swing style and instructed to evaluate the full ensemble performance using the provided measurement tool within the range of skill illustrated by the two anchor recordings.

Research Question 2

Statistical procedures. A factor analysis of the rater responses was conducted in order to provide content validity evidence of items and content domains within the scale. Factor analysis is a particularly helpful method for establishing construct validity of the scale by simultaneously reducing items that are redundant or duplicated in a set of correlated variables (Harman, 1976). Content validity evidence is of importance as the items and domains define the latent construct being measured by the rating scale. In this study, the latent construct can be defined as jazz big band performance achievement. Factor analysis is a popular methodology for establishing convergent and divergent validity of newly constructed measures, whereby highly correlated items loading together form single factors with results that provide a clear specification of items and content domains within the latent construct of interest (Keith, 2006).

Research Question 3

Statistical procedures. A multivariate analysis of variance (MANOVA) was conducted in order to differentiate how groups of varying performance achievement levels (e.g., high, moderate, and low) performed on the rating scale. Task behavior is affected by performance achievement due to the psychological processes of restructuring cognitive skills and knowledge at varying achievement levels (Nicholls, 1984; Weiner & Kukla, 1970). Therefore, educational achievement measures must take into account context-dependent tasks and influences that may affect the measurement of performances (i.e., construct-irrelevant validity) (Messick, 1984). Inter-class grouping of high-, moderate-, and low-ability levels has been shown to have an effect on student achievement (C. L. Kulik & Kulik, 1982; J. A. Kulik & Kulik, 1989; Slavin, 1990). Therefore, ensembles were separated into high-, moderate-, and low-ability levels based upon *z*-probability in order to evaluate construct-irrelevant validity of the measure. Additionally, the obstacle of raters disagreeing on item difficulty as a continuous variable further supports the collapsing of data into categorical variables (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Research Question 4

Statistical procedures. Groups of varying ability levels often perform differently on items (Lucas & Beresford, 2010). Therefore, documenting what criteria best divide group membership into three performance achievement levels can better describe the validity and appropriate use of the scale. As a follow up to the MANOVA, a descriptive discriminant function analysis (DFA) was performed on the response set in order to determine which scale items maximally discriminate the three performance achievement levels (i.e., What items contribute to maximally separating the three performance achievement groups?). DFA can offer a better understanding of the overall structure of the items by providing empirical evidence of the major differences between group levels (Stevens, 2002, p. 285). Predictive discriminant function analysis was also performed on the response set in order to provide predictive validity of the measure by describing how much power and accuracy the items (as indicated by the DFA) had in discriminating between levels.

Results

Research Questions 1 and 2

A factor analysis was conducted in order to describe what items best describe jazz big band performance and to determine what factors contribute to the assessment of a jazz big band performance. Prior to conducting the factor analysis, six assumptions were considered in order to evaluate the appropriateness of utilizing factor analysis as a statistical method in this study: (1) most of the rating scale items correlated at least .30 with the other items (Field, 2012); (2) the Kaiser-Meyer Olkin measure of sampling adequacy was .88, above the recommended value of .50 (Rummel, 1970); (3) the Bartlett's test of sphericity was significant (χ^2 = 1351.83, $p < .001$) (Rummel, 1970); (4) the communalities were all above .30 (Harman, 1960); (5) the sample size was in excess of 100 (n =102) and the subject-to-variable ratio was above the minimum of 3:1 (in this study, the ratio was 4.6:1) (Asmus, 1989; Kerlinger & Pedhazur, 1973); and (6) independence was assumed as 102 raters evaluated 102 distinct jazz big band performances. Factor analysis was therefore found to be a suitable method for examining the data.

A common factor analysis (principal axis factoring) was utilized in order to focus only on the variance in common with the variables (Asmus, 1989; Asmus & Radocy, 1992). Initial rotation based upon Eigenvalues greater than 1 generated a five-factor solution, explaining 68.30% of the variance. Factors 5 through 22 explained a total of 31.70% of the variance, all with Eigenvalues under 1.0. Oblimin oblique rotations factor solutions ranging from two to six were examined. Oblimin oblique rotations were utilized because of the interrelationship of variables chosen (i.e., all items are related to the construct of jazz big band performance) (Asmus & Radocy, 1992). A four-factor solution was utilized after examining the leveling off of data on the scree plot. Additionally, the other examined factor solutions became problematic in their interpretations. Four items were removed from the factor structure. Item 18, 'Lead players perform with appropriate and idiomatic nuances,' was removed because it cross-loaded above .32 between two factors: factor 1 (−.31) and factor 3 (.47) (Tabachnick & Fidell, 2001). The following three items did not load with any of the factors and were therefore removed from the analysis: Item 7, 'Ensemble demonstrates a good concept of jazz phrasing;' Item 19, 'Melodic lines end with an appropriate amount of emphasis;' and Item 20, 'Phrasing of eighth note lines is executed smoothly.'

Factor solutions ranging between two and five were then examined using oblimin oblique rotations with coefficients under .40 suppressed with the remaining 18 items. Following the same parameters and assumptions as discussed with the previously conducted exploratory factor analysis, a final four-factor solution utilizing all 18 items was kept. The four-factor rotation yielded a simple structure explaining 63.32% of the variance (see Table 2). The labels assigned to each of the factors were inferred by the author based upon the content of the correlating items. The factors were labeled as: (a) Factor 1, Blend/Balance (n = 8); (b) Factor 2, Time-feel (n = 5); (c) Factor 3, Idiomatic Nuance (n = 2); and (d) Factor 4, Expression (n = 3). Factor loadings for this scale were clear, with no cross-loading items and high to moderate factor loadings ranging from .88 to .45 (Factor 1, Blend/Balance), .89 to .51 (Factor 2, Time-feel), .78 to .73 (Factor 3, Idiomatic Nuance), and .88 to .71 (Factor 4, Expression).

Multicollinearity was assessed using variance inflation factor (VIF) values for each item. No VIF values exceeded 10, indicating insignificant variable overlap (Cohen, Cohen, West, & Aiken, 2003). Internal consistency was examined for the utilized sample using Cronbach's alpha. The alpha reliability for the 18 items was estimated at .84. Alpha was estimated for each of the four factors: (1) Blend/Balance, .87 (n = 8); (2) Time-feel, .65 (n = 5); (3) Idiomatic Nuance, .80 (n = 2); and (4) Expression, .87 (n = 3). Alpha was also estimated for each of the rater groups: undergraduates, .78 (n = 19); graduates (n = 27); .75, professional performers, .75 (n = 29); and university professors, .89 (n = 27). Descriptive statistics for each of the 18 items can be found in Table 3.

Research Question 3

A multivariate analysis of variance (MANOVA) was conducted in order to evaluate the psychometric qualities of the rating scale at multiple ensemble achievement levels. First, in order to group all evaluated performances into performance achievement subgroups, raw scores for each of the 102 Likert-scale responses were summed across the 18 items included in the factor structure and classified into three performance qualities based upon z-probability: upper group (top 25%), middle group (middle 50%), and low group (bottom 25%). The three criterion levels

Table 2. Pattern factor loading, communalities, variance, and factor correlations based on a principal axis factoring analysis with oblimin oblique rotations for 18 items of the Jazz Big Band Performance Rating Scale (JBBPRS) (*N* = 102).

Items	Factor				h^2
	BB	TF	IN	E	
14. Ensemble plays with a balanced sound in full passages	.88	.07	−.02	.00	.66
17. Good overall blend between brass and saxophones	.81	.17	−.12	.02	.74
9. Ensemble is balanced to the lead trumpet player during ensemble passages	.75	−.05	.10	.04	.59
16. Good overall balance between winds and rhythm section	.55	−.03	.06	−.03	.69
4. Dynamic extremes are controlled	.52	−.13	−.08	.18	.74
21. Rhythm section makes appropriate balance adjustments between ensemble and solo sections	.52	−.04	−.02	−.04	.76
15. Ensemble plays with a large, full sound	.47	−.12	−.25	.25	.55
3. Background figures are well-balanced to the soloist during solo sections	.45	−.13	−.09	.03	.61
10. Ensemble maintains a steady time feel	.08	−.89	.12	.20	.93
1. A steady tempo was kept throughout the performance	.22	−.75	.09	.02	.80
22. The rhythm section and winds share a common feel for the pulse	.11	−.70	−.01	.28	.59
8. Ensemble demonstrates a uniform feeling of pulse	.01	−.58	−.06	.31	.70
13. Ensemble performs with understanding of the swing eighth note concept	.09	.51	.27	.29	.28
12. Ensemble performs with a time feel appropriate to the composition	−.15	.04	.78	−.02	.33
11. Ensemble performs composition at an appropriate, idiomatic tempo	.06	−.07	.73	−.14	.27
5. Eighth note values are given appropriate duration	−.06	−.03	−.05	.88	.62
6. Ensemble accents figures in an appropriate manner	.04	.01	−.04	.83	.77
2. Articulations are consistent with a good concept of jazz phrasing	.07	.02	−.06	.71	.30
Percentage of total variance accounted for	41.00	8.80	7.01	6.50	
Factor intercorrelations					
Blend/Balance	–				
Time-feel	−.39	–			
Idiomatic nuance	−.33	.36	–		
Expression	.58	−.29	−.30	–	

were defined as high performance achievement (*n* = 26), moderate performance achievement (*n* = 54), and low performance achievement (*n* = 22).

To evaluate the appropriateness of utilizing a MANOVA as a statistical method in this study, three assumptions were checked. Tests of normality (i.e., Kolmogorov-Smirnov and Shapiro-Wilk) demonstrated an approximately normal distribution for the composite score data (i.e., all items were significant, p <.001). Levene's test of homogeneity indicated that 14 out of the 18 items indicated that the variances are not statistically significant ($p > .05$). It is assumed that not all items will satisfy the assumption (Stevens, 2002). Box's test of equality indicated that $p = .05$. Because multivariate normality was reasonable and p was equal to .05, Box's test was bypassed (Stevens, 2002).

Table 3. Descriptive statistics for the 18-item Jazz Big Band Performance Rating Scale (JBBPRS) data.

Items	Group means/(*SD*s)			Univ. *F** (*p*)
	High	Av	Low	
14. Ensemble plays with a balanced sound in full passages	3.69 (.48)	2.44 (.62)	1.69 (.70)	38.30 (< .001)
17. Good overall blend between brass and saxophones	3.38 (.65)	2.47 (.67)	1.88 (.62)	19.24 (< .001)
9. Ensemble is balanced to the lead trumpet player during ensemble passages	3.69 (.63)	2.53 (.76)	1.69 (.79)	26.02 (< .001)
16. Good overall balance between winds and rhythm section	3.62 (.65)	2.69 (.74)	2.56 (.73)	9.49 (< .001)
4. Dynamic extremes are controlled	2.92 (.76)	2.38 (.71)	1.13 (.34)	31.42 (< .001)
21. Rhythm section makes appropriate balance adjustments between ensemble and solo sections	3.31 (.95)	2.59 (.71)	2.13 (.89)	7.65 (.001)
15. Ensemble plays with a large, full sound	3.54 (.52)	2.31 (.86)	1.44 (.51)	30.63 (< .001)
3. Background figures are well-balanced to the soloist during solo sections	3.85 (.38)	3.06 (.62)	2.63 (.81)	13.53 (< .001)
10. Ensemble maintains a steady time feel	3.31 (.75)	2.94 (.67)	1.69 (.70)	23.75 (< .001)
1. A steady tempo was kept throughout the performance	3.31 (.86)	2.91 (.73)	1.88 (.89)	15.81 (< .001)
22. The rhythm section and winds share a common feel for the pulse	3.38 (.65)	2.97 (.74)	1.94 (.85)	15.16 (< .001)
8. Ensemble demonstrates a uniform feeling of pulse	3.31 (.86)	2.84 (.68)	1.88 (.86)	13.64 (< .001)
13. Ensemble performs with understanding of the swing eighth note concept	2.46 (.97)	2.00 (.84)	2.00 (1.21)	1.14 (.33)
12. Ensemble performs with a swing feel appropriate to the composition	1.92 (.86)	2.44 (.98)	2.69 (.95)	2.39 (.10)
11. Ensemble performs composition at an appropriate, idiomatic tempo	2.15 (.90)	2.72 (.89)	2.44 (1.09)	1.73 (.19)
5. Eighth note values are given appropriate duration	3.46 (.66)	2.38 (.71)	1.75 (.68)	22.29 (< .001)
6. Ensemble accents figures in an appropriate manner	3.54 (.88)	2.59 (.84)	1.69 (.79)	17.71 (< .001)
2. Articulations are consistent with a good concept of jazz phrasing	3.62 (.51)	2.75 (.80)	2.13 (.81)	14.10 (< .001)

Note. * $df_1 = 2$; $df_2 = 58$.

A one-way (performance achievement: high, moderate, low) between-subjects MANOVA was conducted in order to analyse the group mean differences of performance achievement on the sum scores of each of the 18 items. The analysis indicated a significant main effect of performance achievement grouping on item sums scores (Wilks λ =.074, $F(38,162) = 11.45$, $p < .001$). The partial eta squared value indicated that 72.9% of the variance was accounted for by performance achievement level. A series of one-way ANOVAs was conducted on each of the 18 items as a follow-up test. Out of the 18 items, 17 demonstrated significance between the

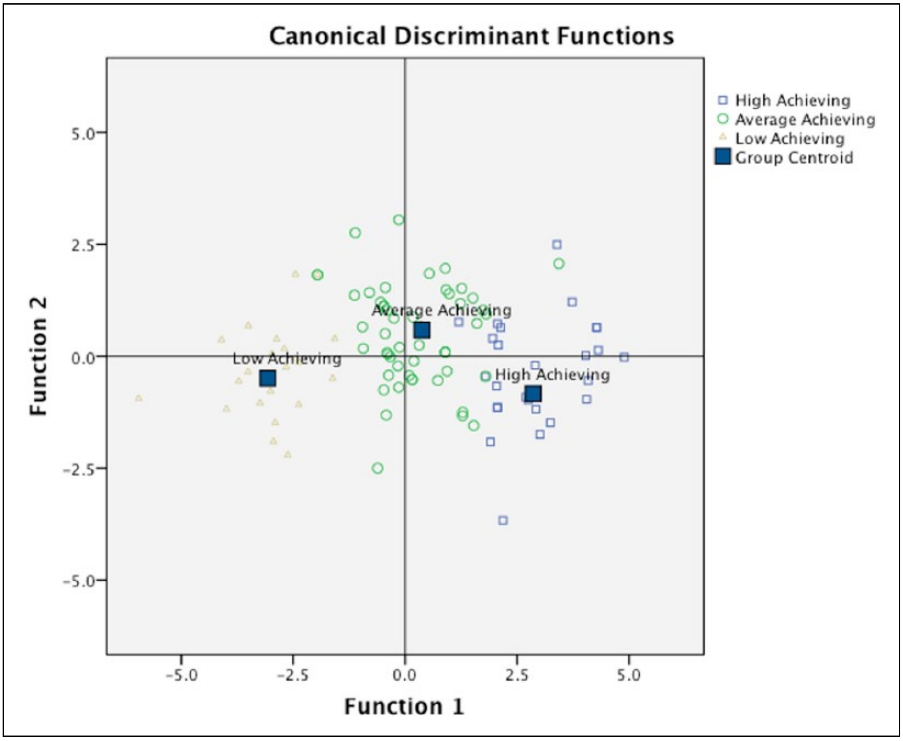

Figure 1. Canonical discriminant function of high, average, and low achievement levels, indicating two discriminant functions using group centroids

performance achievement levels ($p < .001$). Item 13, (Ensemble performs with understanding of the swing eighth note concept) ($p = .90$), did not demonstrate significance. A series of LSD adjusted post hoc tests were conducted in order to evaluate pairwise comparisons among the adjusted means. Significant mean differences occurred between all groups on all statistically significant items with the exception of four items: Item 16, (Good overall balance between winds and rhythm section), moderate/low ($p = .34$); Item 21 (Rhythm section makes appropriate balance adjustments between ensemble and solo sections), moderate/low ($p = .27$); Item 12, (Ensemble performs with a time feel appropriate to the swing style of the composition), moderate/low ($p = .19$); and Item 11 (Ensemble performs composition at an appropriate, idiomatic tempo), high/low ($p = .48$), moderate/low ($p = .21$).

Research Question 4

In order to detect what criteria best divide group membership into each of the three performance achievement levels, a discriminant function analysis (DFA) and predictive function analysis (PFA) were conducted. Two significant canonical correlates were extracted using a DFA (Dimension 1, $\lambda = .13$, $\chi^2(12) = 113.81$, $p < .001$); Dimension 2, $\lambda = .71$, $\chi^2(5) = 19.22$, $p = .002$). Dimension 1 isolated the low achieving ensembles and Dimension 2 isolated the moderate achieving ensembles (See Figure 1). High achieving ensembles were unable to be separated. Table 4 indicates the items that most heavily affect each of the dimensions. Scores

Table 4. Standardized canonical discriminant function coefficients and linear discriminant functions (LDFs) at group centroids.

Variables	D_1	D_2
9. Ensemble is balanced to the lead trumpet player during ensemble passages	.41	−.43
4. Dynamic extremes are controlled	.54	.61
15. Ensemble plays with a large, full sound	.45	−.48
10. Ensemble maintains a steady time feel	.36	.61
11. Ensemble performs composition at an appropriate, idiomatic tempo	.77	.50
5. Eighth note values are given appropriate duration	.57	−.32
LDFs at group centroids		
High-achieving	2.86	−.84
Average-achieving	.37	.59
Low-achieving	−3.07	−.49

Note. Table reflects standardized discriminant function coefficients.
Dimension 1, $\lambda = .13$, $\chi^2_{(12)} = 113.81$, $p < .001$; Dimension 2, $\lambda = .71$, $\chi^2_{(5)} = 19.22$, $p < .002$.

of the low achieving ensembles (i.e., dimension 1) were most affected by the following three items: Item 11, (Ensemble performs composition at an appropriate, idiomatic tempo); Item 5, (Eighth note values are given appropriate duration); and Item 4 (Dynamic extremes are controlled). Scores of the moderate achieving scores (i.e., dimension 2) were most affected by the following three items: Item 10 (Ensemble maintains a steady time feel); Item 4 (Dynamic extremes are controlled); and Item 11 (Ensemble performs composition at an appropriate, idiomatic tempo). In general, the items extracted from the DFA demonstrated relatively high factor loadings in the confirmatory factor analysis, indicting that they are more relevant in defining their respective factor's dimensionality.

Predictive discriminant function analyses were conducted in order to test the overall predictability of the 18 individual items contained within the measure (i.e., predictive validity). As recommended by Huberty and Olejnik (2006), the cross-validation (leave-one-out) rule was reported. As shown in Table 5, the predictive accuracy for the cross validation sample was 88.5%. The hold out accuracy rate of 88.5% indicates that this model surpasses the proportional chance criteria by 49.6% (Hair, Anderson, Tatham, & Black, 1998). Additionally, the Press Q statistic of 83.5 exceeded the critical value of 6.63 (Hair et al., 1998). The prediction model therefore exceeded all criteria of having accuracy beyond chance and supports the predictions of the dependent variable and provides evidence that the model is valid and can be generalized (Hair et al., 1998).

Conclusion and discussion

The purpose of this study was to develop a valid and reliable rating scale to assess jazz big band performance and to evaluate and compare the psychometric properties of jazz big bands at three performance achievement levels. In answer to research questions 1 and 2, the factor analysis indicated four distinct factors utilizing 18 items for assessing a jazz big band: Blend/Balance (*n* = 8), Time-feel (*n* = 5), Idiomatic Nuance (*n* = 2), and Expression (*n* = 3) (See Table 6 for a finalized rating scale). In answer to research question 3, low and moderate achieving ensembles demonstrated significant mean differences on items 11 (Ensemble performs composition at an appropriate, idiomatic tempo), 12 (Ensemble performs with a time feel appropriate to the

Table 5. Hit ratio for cross-validated (leave-one-out) cases selected in the analysis ($N = 61$).

Performance quality	Predicted group membership		
	High *n* (%)	Moderate *n* (%)	Low *n* (%)
High ($n = 23$)	11 (84.6)	2 (15.4)	0 (0)
Average ($n = 49$)	4 (12.5)	28 (87.5)	0 (0)
Low ($n = 30$)	0 (0)	1 (6.3)	15 (93.8)

Note. Percentage of 'grouped' cases correctly classified utilizing the cross-validation analysis for the 18 items was 88.5%.
Numbers in parenthesis indicate the row percentages.
Sample size ($N = 61$) indicates a 65% analysis sample.
Press Q calculated value of 83.6; Press Q table value of 6.63; $p < .001$.

Table 6. 18-item, 4-factor Jazz Big Band Performance Rating Scale.

Blend and Balance				
Ensemble plays with a balanced sound in full passages	SD	D	A	SA
Good overall blend between brass and saxophones	SD	D	A	SA
Ensemble is balanced to the lead trumpet player during ensemble passages	SD	D	A	SA
Good overall balance between winds and rhythm section	SD	D	A	SA
Dynamic extremes are controlled	SD	D	A	SA
Rhythm section makes appropriate balance adjustments between ensemble and solo sections	SD	D	A	SA
Ensemble plays with a large, full sound	SD	D	A	SA
Background figures are well-balanced to the soloist during solo sections	SD	D	A	SA
Time Feel				
Ensemble maintains a steady time feel	SD	D	A	SA
A steady tempo was kept throughout the performance	SD	D	A	SA
The rhythm section and winds share a common feel for the pulse	SD	D	A	SA
Ensemble demonstrates a uniform feeling of pulse	SD	D	A	SA
Ensemble performs with understanding of the swing eighth note concept	SD	D	A	SA
Idiomatic Nuance				
Ensemble performs with a time feel appropriate to the composition	SD	D	A	SA
Ensemble performs composition at an appropriate, idiomatic tempo	SD	D	A	SA
Expression				
Eighth note values are given appropriate duration	SD	D	A	SA
Ensemble accents figures in an appropriate manner	SD	D	A	SA
Articulations are consistent with a good concept of jazz phrasing	SD	D	A	SA

composition), 16 (Good overall balance between winds and rhythm section), and 21 (Rhythm section makes appropriate balance adjustments between ensemble and solo sections). Additionally, high and low achieving ensembles demonstrated a significant mean difference on item 11 (Ensemble performs composition at an appropriate, idiomatic tempo). Lastly, in answer to research question 4, ensembles' ratings on items 4 (Dynamic extremes are controlled), 5 (Eighth note values are given appropriate duration), 10 (Ensemble maintains a steady time feel), and 11 (Ensemble performs composition at an appropriate, idiomatic tempo) had the most impact on predicting membership into achievement level classification (with 88.5% accuracy).

The four factors resulting in this study (blend/balance, time-feel, idiomatic nuance, and expression) parallel previously existing factors from facet-factorial rating scale studies as outlined by Russell (2010b). Most notably, expressive elements (i.e., time-feel, nuance, and expression) tended to be the predominant domain criteria for assessing jazz big band performances. This supports the conclusions that similar musical domains bind various musical performances under diverse stylistic conditions. The argument can be made, then, as to the need for the development of a unique measurement tool for jazz performance if the assessment of all musical performances is bound by similar evaluation criteria. The argument, as maintained by Wilson (2005), is that the deliberate specification of a combination of items and content domains infer the measurement of a specific and particular latent construct. In this case, the latent construct was jazz big band performance achievement. As Wilson explains:

> . . . the measurer must think of some way that this theoretical construct could be manifested in a real-world situation. At first this will be not more than a hunch, a context that one believes the construct must play some determining role in the situation. Later, this hunch will become more crystallized and will settle into a certain pattern. The relationship between the construct and the items is not necessarily one way . . . often the items will be thought of first and the construct will be elucidated only later . . . the important thing is that the construct and items should be distinguished, and that *eventually* the items are seen as realizations of the construct. (p. 10)

The continuing refinement of the scale presented in this study is the only manner in which these realizations can be comprehended. Wilson continues:

> . . . in many, if not most cases, the construct is not clearly defined until a large set of items has been developed and tried out with respondents. Each new context brings about the possibility of developing new and different sorts of items or adapting existing ones. (p. 42)

Further testing of this scale with a different set of respondents analysed by applications of modern measurement theory (e.g., Rasch analysis) may be a fruitful endeavor in more clearly defining the specific construct being measured and improving the function and design of included items. Specifically, the application of the Rasch measurement model under new assessment and equating conditions may provide valuable insight into areas of item difficulty, rating scale structure, ensemble achievement level, and differential facet functioning.

In this study, the 18 items that describe jazz performance achievement may provide diagnostic value and pedagogical insight into jazz big band teaching and learning that could not be gleaned from previously existing literature. The decision to change the individual continuous data scores into categorical performance achievement groups was partly based upon pedagogical utility, as grouping students into ability levels provides an opportunity to differentiate instruction more aligned with student needs (Lou et al., 1996). The data derived from the MANOVA and discriminant function analyses proves to be insightful, as differences among high, moderate, and low performing ensembles existed. Significant canonical correlates were extracted, indicating two sets of explanatory variables (i.e., dimensions) among the low and moderate achieving ensembles. Both dimensions shared two common items: Item 4 (Dynamic extremes are controlled), and Item 11 (Ensemble performs composition at an appropriate, idiomatic tempo). This indicates that compared to high-achieving ensembles, low- and moderate-achieving ensembles consistently demonstrated lower scores on these two items. Low-achieving ensemble scores were more heavily affected by Item 16 (Eighth note values are given appropriate duration) and moderate-achieving ensembles were more heavily affected by Item 9 (Ensemble maintains a steady time feel). These dimensions and items may provide a pedagogical starting point for instructors and students of these performing ensembles. Low-achieving ensembles will likely most affect noticeable change in their performance achievement by focusing on the improvement of items 4, 11, and 16 in ensemble rehearsals. These items are representative of basic musical ensemble concepts: awareness and detail to dynamics (Item 4), steady tempo (Item 11), and ensemble balance (Item 16). Therefore, the instructor should continually remind young students that concepts developed in other large ensemble contexts that they may be more experienced in (e.g., band, orchestra) also warrant the same consideration in a jazz big band performance context. Moderate-achieving ensembles will most likely affect noticeable change in performance achievement by focusing on the improvement of items 4, 11, and 9 in ensemble rehearsals, as these items most heavily influenced moderate-achieving classification using the measurement tool proposed in this research scale. As with the low-achieving ensemble, moderate-achieving ensembles should be aware of and continually develop basic ensemble skills of dynamic contrast (Item 4) and steady tempo (Item 11). Additionally, with moderate-achieving ensembles, the focus of ensemble balance should be directed to the lead trumpet (Item 9), a new skill for the developing jazz musician specific to the jazz big band performance context. For both low- and moderate-achieving ensembles, working to develop performance skills represented in the factors of time-feel, idiomatic nuance, and expression should come after principles of basic ensemble playing (e.g., blend, balance, steady tempo) are achieved. This may be a predictor that young students with stronger fundamental ensemble performance skills are most suited to participate in jazz big band settings.

This study demonstrated that the prediction rate to accurately classify an ensemble's membership into the appropriate level group is 88.5%. The results of identifying the criteria of group membership can be most efficiently utilized in a pedagogical capacity. By highlighting the specific items that most contribute to group classification, educators can target specific performance elements that correlate to their current achievement level. This may provide a more concrete rehearsal plan tailored to individual ensemble needs. By isolating and developing the performance criteria related to each of these items, ensembles may demonstrate the most noticeable improvement in performance achievement.

Provided that the students participating in a jazz ensemble can perform with good overall technical facility on their instrument, the items contained within the Jazz Big Band Performance Rating Scale (JBBPRS) provide a foundation for attributes associated with jazz big band performance. The items may engage students in guided listening within the ensemble as well as foster classroom discussions through self-assessment and the assessment of other ensembles regarding the attributes of time-feel, nuance, expression, and ensemble sonority idiomatic to jazz big band performance. The items may also provide a pedagogical underpinning to pre-service and in-field teachers lacking confidence or experience in jazz pedagogy. According to Dunscomb and Hill (2002), listening to and assessing big bands in the areas of sound, balance, articulation, style, and nuance is the most accessible place to formulate critical listening skills. This measurement tool may provide a guidepost for such listening. However, further testing of the scale's reliability with non-jazz experts is warranted in order to fully state this claim.

In order to further develop the assessment process, it is suggested that a formative assessment tool be developed to supplement the JBBPRS. Although a summative assessment tool such as the JBBPRS can provide a benefit for the developing jazz ensemble, it may not provide the necessary feedback needed to improve the teaching and learning processes. The implementation of a valid and reliable formative assessment tool such as a rubric or criteria-specific measurement scale aligned with the JBBPRS may improve student learning as well as better shape the instructional process. The combination of the JBBPRS and a supplemental formative measure may improve the overall effectiveness of the educational process.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Abeles, H. F. (1971). *An application of the facet-factorial approach to scale construction in the development of a rating scale for clarinet music performance* (Unpublished doctoral dissertation). University of Maryland, College Park.

Asmus, E. P. (1981). The effect of altering the number of choices per item on test statistics: Is three better than five? *Bulletin of the Council for Research in Music Education*, *65*, 1–15.

Asmus, E. P. (1989). Factor analysis: A look at the technique through the data of Rainbow. *Bulletin of the Council for Research in Music Education*, *85*, 1–13.

Asmus, E. P. (1999). Music assessment concepts. *Music Educators Journal*, *86*(2), 19–24.

Asmus, E. P., & Radocy, R. E. (1992). Quantitative analysis. In R. Colwell (Ed.), *Handbook of research of music teaching and learning* (pp. 141–195). New York, NY: Schirmer Books.

Baker, D. (1989). *Jazz pedagogy: A comprehensive method of jazz education for teacher and student*. Van Nuys, CA: Alfred Publishing.

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

Berry, J. (1990). *The jazz ensemble director's handbook: Understanding the A-to-Z's of each section, including the rhythm section*. Milwaukee, WI: Jenson Publications.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Coker, J. (1978). *Listening to jazz*. Englewood Cliffs, NJ: Prentice-Hall.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.

Crowe, E. (2010). *Measuring what matters: A stronger accountability model for teacher education*. Washington, DC: Center for American Progress.

DCamp, C. B. (1980). An application of the facet-factorial approach to scale construction in the development of a rating scale for high school band performance. *Dissertation Abstracts International*, *41*, 1462A.

Dumas, J (1999) *Usability testing methods: Subjective measures, part II – measuring attitudes and opinions*. Washington, DC: American Institutes for Research.

Dunscomb, J. R., & Hill, W. (2002). *Jazz pedagogy: The jazz educator's handbook and resource guide*. Miami, FL: Warner Bros. Publications.

Ellis, M. C. (2007). An analysis of taped comments from a high school jazz band festival. *Contributions to Music Education*, *34*, 35–49.

Field, A. (2012). *Discovering statistics using IBM SPSS statistics* (4th ed.). London, UK: SAGE.

Fiske, H. E. (1979). Musical performance evaluation ability: Toward a model of specificity. *Bulletin of the Council for Research ion Music Education*, *59*, 27–31.

Goins, W. E. (2003). *The jazz band director's handbook: A guide to success*. Lewiston, NY: Edwin Mellen Press.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Harman, H. H. (1976). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Horowitz, R. A. (1994). The development of a rating scale for jazz guitar improvisation performance. *Dissertation Abstracts International*, *55* (11A), 3443.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Jarvis, J., Beach, D., & Wiest, S. (2002). *The jazz educator's handbook*. Delevan, NY: Kendor Music.

Jones, H. (1986). An application of the facet-factorial approach to scale construction in the development of a rating scale for high school vocal solo performance. *Dissertation Abstracts International*, *47*, 1230A.

Keith, T. Z. (2006). *Multiple regression and beyond*. Boston, MA: Pearson Education, Inc.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York, NY: Holt, Rinehart, and Winston.

Kernfeld, B. (1995). *What to listen for in jazz*. New Haven, CT: Yale University Press.

Kulik, C. L., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal*, *19*, 415–428.

Kulik, J. A., & Kulik, C. L. (1989). Effects of ability grouping on student achievement. *Equity and Excellence*, *23*(1–2), 22–30.

Kuzmich, J., & Bash, L. (1984). *Complete guide to instrumental jazz instruction: Techniques for developing a successful jazz program*. West Nyack, NY: Parker Publishing Co., Inc.

LaPorta, J. (1965). *Developing the school jazz ensemble*. Boston, MA: Berklee Press.

Lawn, R. (1981). *The jazz ensemble director's manual: A handbook of practical methods and materials for the educator*. Oskaloosa, IA: C.L. Barnhouse.

Lehman, P. R. (2007). Getting down to basics. In T. S. Brophy (Ed.), *Assessment in music education: Integrating curriculum, theory, and practice* (pp. 17–27). Chicago, IL: GIA Publications, Inc.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Appolonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, *66*(4), 423–458.

Lucas, S. R., & Beresford, L. (2010). Naming and classifying: Theory, evidence, and equity in education. *Review of Research in Education*, *34*, 25–84.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*(3), 215–237.

Miles, R., & Carter, R. (Eds.) (2008). *Teaching music through performance in jazz*. Chicago, IL: GIA Publications.

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological review*, *91*(3), 328–346.

Nichols, J. P. (2005). A factor analysis approach to the development of a rating scale for snare drum performance. *Dissertation Abstracts International*, *46*, 3282A.

Pazitka-Munroe, W. L. (2002). The construction and validation of an audition instrument to measure the vocal performance of college singers auditioning for choral ensembles. *Dissertation Abstracts International*, *63*(08), 2821.

Pfenninger, R. C. (1990). The development and validation of three rating scales for the objective measurement of jazz improvisation achievement. *Dissertation Abstracts International*, *51*(8), 2674A.

Pizer, R. A. (1990). *Evaluation programs for school bands and orchestras*. West Nyack, NY: Parker Pub. Co.

Prouty, K. (2012). *Knowing jazz: Community, pedagogy, and canon in the information age*. Jackson: University Press of Mississippi.

Rummell, R. J. (1970). *Applied factor analysis* . Chicago, IL: Northwestern University Press.

Russell, B. E. (2010a). The development of a guitar performance rating scale using a facet factorial approach. *Bulletin of the Council of Research in Music Education*, *184*, 21–34.

Russell, B. E. (2010b). The empirical testing of a musical performance assessment paradigm. *Open Access Dissertations*. Paper 387. Retrieved from http://scholarlyrepository.miami.edu/cgi/viewcontent.cgi?article=1386&context=oa_dissertations

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best evidence synthesis. *Review of Educational Research*, *60*(3), 471–499.

Smith, B. P., & Barnes, G. V. (2007). Development and validation of an orchestra performance rating scale. *Journal of Research in Music Education*, 55, 268–280.

Smith, D. T. (2009). Development and validation of a rating scale for wind jazz improvisation performance. *Journal of Research in Music Education*, *57*(3), 217–235.

Spector, P. E. (1992). *Summated rating scale construction: An introduction*. Newbury Park, CA: SAGE Publications.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston, MA: Allyn and Bacon.

Treinen, C. M. (2011). *Kansas high school band directors and college faculties' attitudes towards teacher preparation in jazz education* (Unpublished doctoral dissertation). Kansas State University, Manhattan.

Weiner, B., & Kukla, A. (1970). An attributional analysis of achievement motivation. *Journal of Personality and Social Psychology*, *15*(1), 1–20.

Wesolowski, B. C. (2014). Documenting student learning in music performance – a framework. *Music Educators Journal*, *101*, 77–85.

West, C. L. (2011). *Teaching middle school jazz: An exploratory sequential mixed methods study* (Doctoral dissertation, University of Michigan). Retrieved from http://deepblue.lib.umich.edu/bitstream/handle/2027.42/84484/cleewest_1.pdf

Wheaton, J. (1975). *How to organize and develop the stage band: Director's manual*. North Hollywood, CA: Maggio Music Press.

Whybrew, W. E. (1973). Research in evaluation in music education. *Bulletin of the Council for Research in Music Education*, *35*, 9–17.

Wilson, M. (2005). *Constructing measures*. New York, NY: Taylor & Francis Group.

Wiskirchen, G. (1966). *Developmental techniques for the jazz ensemble musician*. Boston, MA: Berklee Press.

Yorke, M. (2008). *Grading student achievement in higher education: Signals and shortcomings*. New York, NY: Routledge.

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.