ISSN: 1938-2065

Factors that Affect the Measurement of Rhythm Achievement

by
Christina Schneider
South Carolina Department of Education

Robert Johnson
University of South Carolina

Sameano Porchea
South Carolina Department of Education

The purpose of this study was to examine factors that affect the reliability and validity of an 8-beat rhythm improvisation performance task that was used during a pilot test of a large-scale music assessment. In addition to analyzing the holistic improvisation scores, two domains of rhythm achievement associated with improvisation performance task were investigated: student ability to begin the improvised "response" on the correct downbeat and student ability to maintain a steady beat during the 8-beat improvisation. Although reliability results were strong and positive for the holistic improvisation scores and entrance analysis, it was found that even with a 5-judge panel, sufficient reliability in measuring student ability to maintain a steady beat could not be established.

Schneider, C., Johnson, R., & Porchea, S. (2004). Factors that affect the measurement of rhythm achievement. *Visions of Research in Music Education, 5.* Retrieved from http://www-usr.rider.edu/~vrme/

Introduction

Improvisation is becoming a more commonly taught skill in music classes since its inclusion in the National Standards for Music Education (MENC, 1994). As a result, state and national assessments are beginning to measure student abilities in this area, and specific curriculums have been designed to teach this skill to young music students (e.g., Azzara, Grunow, & Gordon, 1996). The National Assessment of Educational Progress (NAEP) 1997 Music Assessment assessed students' abilities to vocally improvise over a blues background at the 8th-grade level (Persky, Sandene & Askew, 1998). Likely, similar improvisation skills will be measured when the NAEP Music Assessment is administered again in 2008.

Incorporating a music performance into a large-scale assessment is often difficult and costly (Schneider, 2003). In addition to performance tasks being less time effective measures of student achievement than paper and pencil measures, performance tasks, when separated from multiple-choice items in a large-scale assessment, are often less reliable measures of what students know and can do. Therefore, investigating the technical quality of a performance task (i.e., the rubric and scoring issues associated with the task) helps to ensure that student abilities in a performance area are measured precisely. More important, the inferences or decisions that educators make regarding student abilities that result from such situations will be defensible.

Traditionally, interrater reliability has been calculated using the Pearson-product moment correlation (PPMC) to determine how consistently raters evaluate a performance. Although, the PPMC indicates the degree to which two raters are ranking student performances consistently, it neither allows test developers or researchers to detect potential stringency level differences among judges (Brennan, 1992) nor does the PPMC allow for a single calculation of reliability

across a panel of judges. If one rater on a judging panel is consistently more severe or lenient than his or her peers, his or her scores will have more weight than a rater who is less severe or lenient. Test developers and researchers, then, must work to eliminate differences (i.e., error) in scoring that are attributable to differences in rater expectations. For that reason, test developers and researchers may use the coefficient alpha. That statistic while providing a single reliability calculation across a panel of judges does not allow for investigation of potential sources of measurement error.

Measurement error when rating music performances may be attributed to several different areas and is often labeled in educational research as a source of variability (Shavelson and Webb, 1991). Most sources of variability are unwanted, the exception being when the source of variability is derived from the objects of measurement (Brennan, 2001), that is, the students themselves (p). Differences in improvisation achievement scores should occur because of differences in the students' abilities to improvise. A source of unwanted variability is error variance among raters that was noted previously (R). In addition, an interaction among raters and students may occur when raters rank order students differently (pR, Brennan, 1992). Generalizability theory is often used to investigate the sources of unwanted variability that occur when a performance task is rated so that the measurement of student abilities may be improved.

Generalizability theory may be used to determine the overall reliability among a panel of raters who evaluate a particular performance task so that unwanted sources of variability may be investigated through the use of variance components. In addition, given that a degree of error occurs in any measurement situation, a decision (D) study may be used to determine the number

of raters that are necessary to provide reliable ratings to similar student performances using the same rubric and performance task in future assessments.

The main purpose of this study was to examine the factors that affect the overall reliability and validity of an 8-beat rhythm improvisation performance task during a pilot test of a large-scale music assessment. In addition to analyzing the holistic improvisation scores, two domains of rhythm achievement associated with improvisation performance task were investigated: student ability to begin the improvised "response" on the correct downbeat and student ability to maintain a steady beat during the 8-beat improvisation. The number of judges necessary to reliably measure these specific domain elements was also investigated.

Procedures

Two hundred and eleven fourth-grade students from intact but randomly selected classrooms in 11 elementary schools participated in the rhythm improvisation performance task. Each student heard an 8-beat rhythm pattern prompt in common time on rhythm sticks and was instructed to provide an improvised "answer" using rhythm sticks. Students were told that their answer should be 8 beats, enter on time, maintain a steady beat, and use a different arrangement of quarter notes and eighth notes than the rhythm pattern prompt. Task directions, the rhythm pattern prompt, and a model performance sequence were recorded onto a compact disc so that the task instructions and administration could be standardized for each student (Appendix A). Each student's performance was recorded using a high quality cassette tape recorder and microphone and then transferred onto compact discs for subsequent scoring by a panel of three raters.

In the first round of scoring, raters scored each improvisation using a four-point holistic rubric (Appendix B). Raters were trained to use the rubric through the use of anchor

performances and a sample rating session as is common in large-scale assessment situations. Raters also received instructions regarding the use of augmentation scoring. For augmentation scoring, raters were instructed to first identify the integer-level score (i.e., 0, 1, 2, or 3) and then to decide if the response reflected a performance level slightly higher or slightly lower than the typical response at that level. If the student response reflected a performance slightly higher than the typical response at that level, then .33 was added to the integer-level score. If the student response reflected a performance slightly lower than the typical response at that level, then .33 was subtracted from the integer-level score. The use of augmentation scoring has most notably been used in National Board for Professional Teaching Standards (NBPTS) portfolio assessment (Bond, 1998; Educational Testing Service, 1999; NBPTS, 2001).

After each rater had scored each performance holistically, raters were instructed to rescore each performance. In the second round of scoring, raters were asked to focus on whether the student entered on the correct downbeat and whether the student maintained a steady beat. Raters were instructed to give a 1-point rating if the student's entrance was on the downbeat and a 0-point rating if the student's entrance was not on the downbeat. Raters were also instructed to give a 1-point rating if the student maintained a steady beat throughout the 8-beat improvisation and 0-point rating if the student did not maintain a steady beat throughout the 8-beat improvisation.

Results

After checking the reliability of the raters, the scores of the three raters were averaged to provide a total score for each student. Table 1 provides descriptive statistics for overall improvisation achievement. A mean of 1.74 indicates that this task was moderately difficult as the p-value for the item was .58 (1.74/3=.57). That is, the task was one that differentiated among students'

abilities. The average student tended to improvise patterns that were less than or more than eight beats, tended not to enter on the phrase downbeat, and tended to be unable to maintain a steady beat.

Table 1

Summary Statistics for Improvisation

N	M	SD	SEM
211	1.74	0.71	0.05

Table 2 provides the estimated variance components from the generalizability study. As may be determined from Table 2, the students and the interaction among students and raters contribute as sources of variability in the study. The largest contributor of variability in scores is the student component (p). The interaction of student and rater (pR) contributes the next largest amount of variance. This indicates that judges did not always rank order students similarly. Examination of the variance associated with raters (R) also indicates that one rater was not more stringent or lenient than another leading the researchers to conclude that most of the improvisation score differences may be attributed to the students' differing ability levels.

Table 2

Model Variance Components Estimates

EFFECT	*df*	Estimated Variance Components for Single Observations	Estimated Variance Components for Mean Scores
p	210	.447	.447
R	2	.020	.007
PR	420	.185	.062

p=persons, R=raters, and pR=the interaction between persons and raters

In a D study of raters, the intent is to generalize from a single rating panel's scores on performance task to a performance task universe score given if all similar raters in the universe were rating the same task (Brennan, 2001). That is, the researcher may generalize to an average improvisation score using hypothetical randomly parallel raters. Using three raters, the index of dependability (F) coefficient for this performance task is .88. This coefficient indicates that we may expect another 3-judge panel to also obtain sufficient reliability in future situations with this performance task and rubric. Moreover, based on the results of Table 3, in future situations, the judging panel may be reduced to two persons while maintaining sufficient reliability (.81).

Table 3

Number of Raters	Index of Dependability (Φ)
1	.69
2	.81
3	.88

For the analysis of the domains of "entering on the downbeat" and "maintaining a steady beat" scores of the three raters were averaged to provide a total score for each student (a) entering on the correct down beat and (b) maintaining a steady beat. Reliability estimates for the downbeat and steady beat total scores were .90 and .62, respectively. The relationship between the students' abilities to begin the eight-beat answer on the correct downbeat and the students' improvisation scores was .58 meaning that roughly 34% of the variance in students' improvisation scores was related to beginning the improvisation on the correct downbeat. The relationship between the students' abilities to maintain a steady beat and the improvisation score was .40 meaning that roughly 16% of the variance in students' improvisation scores was related to maintaining a steady beat. One caveat should be noted. The 3-judge panel had difficulty rating students' abilities to maintain a steady beat as evidenced by the phi coefficient of .62.

To determine if more raters would be useful in measuring students' abilities to maintain a steady beat, phi coefficients were calculated. As may be determined from Table 4, even a panel of 5 raters would not produce sufficient reliability for precise measurement of this music skill.

Table 4

Phi Coefficients by Number of Raters for Steady Beat

Number of Raters	Φ
1	.36
2	.52
3	.62
4	.69
5	.73

Discussion

Students scoring around the mean in this study tended to have difficulty internalizing the eight-beat rhythm phrase structure even though an example of a correct response was modeled. Students tended not to enter on the phrase downbeat, and they tended to be unable to maintain a steady beat. Entering on the phrase downbeat and maintaining a steady beat jointly accounted for 50% of the rhythm improvisation scores. Given that accurately beginning the improvisation was one aspect of the holistic performance and accounted for 34% of the variance in improvisation scores, the influence of the domain on the improvisation scores appears appropriate. If one domain had accounted for a majority of the variance in the improvisation scores, perhaps 70%-80%, then it could be argued that judges were attending to a single domain at the expense of other aspects of the student's performance.

The reliability of the domain of "maintaining a steady beat" presents several interesting issues. As was shown in Table 4, even with a five-judge panel, sufficient reliability for isolating

a student's ability to maintain a steady tempo in future situations is not expected based on the results of this study. This brings an interesting question to research: Is it possible to reliably measure a student's ability to maintain a steady beat? In this performance task, a click track was present both during the first eight-beat phrase (the recorded prompt) and the second eight-beat phrase (the student response). Presumably, raters had a tempo reference through both the eight-beat call and the click track that remained present during the student's performance. The raters who seemed unable to measure students' abilities to maintain a steady beat were the same raters who were quite successful in measuring the overall improvisation ability and phrase entrance of students.

This steady beat finding may be an integral issue to music education research and measurement because maintaining time is of paramount importance and foundational to music making. It is possible that measuring a student's ability to maintain a steady beat through a binary variable is not appropriate due to the developmental nature of music achievement (Gordon, 2001). Students who are learning to maintain a steady beat often tend to "rush" when performing and may be heard, for example, rushing the end of a phrase. This may be particularly true in the fourth grade where students nationally receive music instruction merely one or two days per week.

This study found that the rhythm improvisation performance task used in the pilot phase of a large scale assessment had sufficient technical characteristics to be used in the future; however, some interesting findings regarding the measurement of students' abilities to maintain a steady tempo became apparent. Further research and development of a steady beat measure would be of

use not only to those interested in measuring student abilities for accountability purposes but also for those interested in music development research.

References

Azzara, C., Grunow, G., & Gordon, E. (1996).
Creativity in improvisation: Getting started. Chicago: G.I.A.

Bond, L. (1998).
"Culturally responsive pedagogy and the assessment of accomplished teaching." *Journal of Negro Education*, 67(3), 242-254.

Brennan, R. (1992).
"NCME instructional module: Generalizability theory." *Educational Measurement: Issues and Practices*, 11 (4), 27-34.

Educational Testing Service. (1999).
Technical analysis report. Princeton: Author.

Gordon, E. (2001).
Learning sequences in music: Skill, content, and patterns. Chicago: GIA.

Music Educators National Conference. 1994.
National Standards for Arts Education. Reston, VA: Music Educators National Conference.

National Board for Professional Teaching Standards. (2001).
Early and middle childhood art scoring guide. Washington, DC: Author.

Persky, H. R., Sandene, B. A., & Askew. J. (1998).
The NAEP 1997 arts report card. Washington, DC: National Center for Educational Statistics.

Schneider, M.C. (2003).
"Merging the ideal with reality: Reflecting on the NAEP 1997 arts assessment and looking toward 2008." *Arts Education Policy Review.*

Shavelson, R. & Webb, N. (1991).
Generalizability theory: A primer. Newbury Park, CA: Sage Publications.

Appendix A

Recorded Performance Task Directions

You are going to answer a rhythm pattern using your rhythm sticks. Use only quarter notes and eighth notes to answer the pattern. Do not play your rhythm sticks yet. First, **listen** to the rhythm pattern and then inside of your head think of an answer that is different from the pattern. Your pattern should be 8 beats long.

Students hear a sound clip of 8-beat pattern that includes a click track; the click track is 16 beats in length.

Now, you are going to hear the rhythm pattern again followed by an example of an answer that is different. Notice the person uses only quarter notes and eighth notes, plays for 8 beats, and keeps a steady tempo.

Students hear a sound clip of the same 8-beat pattern plus a model 8-beat answer that includes a click track.

Now, you answer the 8-beat pattern with your rhythm sticks using only quarter notes and eighth notes.

Students hear a sound clip of the same 8-beat pattern that includes a click track; the click track is 16 beats in length.

Appendix B

Improvisation Rubric

Rating	*The performance should be characterized by most of the following:*
3	The student performs an 8-beat improvisation. The improvisation begins on time, maintains a steady beat, and includes only quarter and eighth notes in an order different from the prompt.
2	The student performs an improvisation that may vary slightly from 8-beats. The improvisation may or may not enter on time, tends to maintain a steady beat, and includes only quarter and eighth notes in an order different from the prompt.
1	The student performs some form of an improvisation that may vary significantly from 8 beats. The improvisation may or may not enter on time, may not maintain a steady beat, and may not include only quarter and eighth notes in an order different from the prompt. OR, the performance imitates the prompt.
0	The student does not perform an improvisation