by Brian C. Wesolowski

"Classroometrics"

The Validity, Reliability, and Fairness of Classroom Music Assessments

Photo of Brian C. Wesolowski courtesy of the author

What makes a music assessment in the classroom fair and accurate? Here are some ideas to consider.

Abstract: Validity, reliability, and fairness are three prominent indicators for evaluating the quality of assessment processes. Each of the indicators is most often written about and applied in the context of large-scale assessment. As a result, the technical properties of these indicators make them limited in both their practicality and relevance for classroom assessments. The purpose of this article is to describe validity, reliability, and fairness in a way that is meaningful and applicable toward improving the quality of classroom music assessments.

Keywords: assessment, evaluation, inference, measurement, quality, reliability, test, validity

Assessment is a process that includes the collection, analysis, interpretation, and application of information about student-learning outcomes in order to make educational decisions about students, curricula, programs, schools, and/or educational policies.[1] For classroom teachers, the assessment process includes (1) developing

Brian C. Wesolowski is an associate professor of music education at the University of Georgia, Athens; he can be contacted at bwes@uga.edu.

NAfME is glad to offer one hour of professional development recognition to you for reading this article. Please follow the link below and complete a short quiz to receive your certificate of completion.
http://bit.ly/Classroometrics

DOI: 10.1177/0027432119894634
http://journals.sagepub.com/home/mej

tests that purport to measure a specified learning outcome, (2) using the tests to collect evidence of student behaviors that represent the learning outcome, (3) making inferences about student-learning outcomes from the collected evidence, and (4) using the inferences to make educational decisions.

Classroom assessments are inherently inferential in nature. In the context of classroom assessments, teachers often piece together varieties of student achievement evidence to draw conclusions, or make inferences, about the intended outcomes of instruction. Outcomes of instruction (i.e., student-learning outcomes) include the students' demonstration of specific knowledge, skills, abilities, and/or dispositions gained from their learning environment. Classroom assessment processes are integrated in all the instructional processes occurring in the classroom, and teachers' educational decisions are based primarily on their inferences of student-learning outcomes. In today's standards-based instructional environment, educational decisions made from classroom assessment processes are often intended to close the gap between a student's present level of ability and the minimum required level of ability as set forth by district, state, and/or national standards.[2] As a result, the quality of inferences teachers make about student-learning outcomes through classroom assessments not only influences future teaching processes but also directly affects their accuracy in judging the alignment between a student's learning outcomes and a student's achievement of curricular standards.

The quality of educational decisions in the classroom is predicated on teachers' abilities to make accurate inferences about student-learning outcomes. Therefore, it is important that teachers consider the overall quality of their assessment processes or, more specifically, the quality of inferences they make about the outcomes of their instruction to make the most reasonable and appropriate educational decisions possible. The concepts of validity, reliability, and fairness each provide quality-assurance indicators of assessment processes as well as unique evidence of the quality of the inferences a teacher makes about student-learning outcomes. Each of the concepts, however, is deeply rooted in the evaluation of large-scale assessment processes[3] that are vastly different from classroom assessment processes and offer little relevance to classroom assessment processes.

From a large-scale assessment perspective, validity, reliability, and fairness are used to ensure the adequacy of the testing instrumentation, data-gathering procedures, and the interpretation of student achievement data. In other words, the inferences made about student achievement in large-scale assessment contexts are directly associated with the testing context itself and are external to the student's day-to-day learning environment. Any actions taken by teachers to improve instruction are indirectly made from the results of a large-scale assessment.

Additionally, the concepts of validity, reliability, and fairness in large-scale assessment contexts are strongly connected to the concept of *measuring* student achievement. In these cases, inferences about student achievement are made through the implementation of measurement models. Measurement models provide a wide variety of statistical evidence that lays the groundwork for various validity, reliability, and fairness arguments.

Seeing the disparity between large-scale assessment and classroom assessment processes, educational consultant Susan Brookhart coined the term *classroometrics,* or *classroometric theory,* as a conceptual method for considering how these quality-control indicators could be meaningfully applied to classroom teaching and assessment.[4] Specifically, Brookhart called for the development of improved methodologies to evaluate the quality of classroom assessments across all educational contexts. In particular, she highlighted three important needs toward improving classroom assessment processes: (1) highlighting the role of the classroom teacher in developing quality assessments for classroom use, (2) elevating the value of classroom assessments via a unique conceptual approach appropriate for evaluating their quality, and (3) using the conceptual approach to clearly delineate the evaluation of the quality of assessments in classroom contexts from the quality of assessments in large-scale assessment contexts.[5]

According to Brookhart, classroom assessments warrant the same theoretical considerations as large-scale assessments. However, classroometrics moves beyond *measuring* student ability via a onetime large-scale assessment condition and focuses more broadly on the *evaluation* of students' learning processes and their learning advancement over the course of an instructional cycle. Classroometrics therefore represents an important consideration for evaluating the quality of classroom assessment processes and quality of inferences regarding student-learning outcomes in the area of music education.

The purpose of this article is to describe validity, reliability, and fairness as it pertains to the improvement of classroom music assessment processes and provide considerations for evaluating the quality of assessment processes in the context of music teaching and learning (see Sidebar 1 for definitions of some assessment terms).

Tests, Inferences, and Use

The definition of a test, according to psychometrician and educational measurement leader Gregory Cizek,[6] "is simply a data collection procedure; more precisely, . . . a sample of behavior(s) taken and interpreted under specified, systematic, and uniform conditions." Cizek continues, "regrettably, the concept of a test is . . . frequently interpreted too narrowly . . . it is mistaken as referring to a specific format for data collection instead of more broadly as any structured processes for doing so."[7]

Traditional notions of a test in music may include a formal evaluation of a performance, a self-evaluation of a performance, a multiple-choice examination containing musical content, or an

SIDEBAR 1

Definitions of Some Assessment Terms
Content Representativeness
The alignment of content taught in the classroom to national and/or state standards, learning objectives, and related classroom assessments.
Fairness (Classroom Assessment)
The opportunities for a student to best demonstrate student-learning outcomes.
Fairness (Large-Scale Assessment)
Responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses.
Formal Assessment
An assessment that uses collected quantitative or qualitative data to support an inference of a student's learning outcome.
Formative Assessment
An activity within the instruction cycle that measures a student's progress toward an intended learning outcome in the classroom or a program.
Inference
A conclusion drawn about a student's learning outcome.
Informal Assessment
An assessment that does not use collected quantitative or qualitative data to support an inference of a student's learning outcome.
Reliability (Classroom Assessment)
The dependability of assessments to adequately support inferences made about student-learning outcomes.
Reliability (Large-Scale Assessment)
Consistency over replications of the testing procedure.
Student-Learning Outcome
The knowledge, skills, abilities, and/or dispositions intended to be assessed.
Summative Assessment
An assessment at the end of an instructional cycle to measure student growth and learning.
Test
A sample of behaviors taken and interpreted under specified, systematic, and uniform conditions.
Validity (Classroom Assessment)
The teacher's confidence in the quality of the inferences made about student-learning outcomes.
Validity (Large-Scale Assessment)
The degree to which evidence and theory support the interpretations of the test scores for proposed uses of tests.

essay related to any musical concept. Using Cizek's definition, however, a test in the music classroom can be considered any circumstance in which a music teacher makes a systematic observation of a student's musical behavior. These observations can include all the formal, informal, formative, and summative assessments used within an instructional cycle. For example, a test may include an informal observation of a student's ability to follow a mental checklist in putting an instrument together, an informal observation of a group of students working together to create a musical map, the ability of a student to be called on to perform his or her individual part of a musical work, a survey of a class's affective response to a piece of music, an informal observation of a select group of students clapping together in rhythmic unison, students' aural identification of a musical motive in a piece of music by raising their hands, an interview for a leadership position, an audition for an ensemble seating position, a performance pass-off, or a call-and-response classroom exercise. In each of these examples, different assessment methods can be used to collect varying types of student learning outcome information. Together, all this student-learning outcome information provides a holistic picture of student achievement.

In each of these contexts, the term *test* does not specifically refer to a

formal assessment context where a testing instrument is used for collecting data, such as a written exam, a rubric, or a rating scale, for example.[8] Furthermore, a test does not refer to a grade or score.[9] Empirical data do not need to be collected and students do not need to be given a grade for authentic and meaningful assessment to occur in the music classroom. A test, in each of these examples, refers to the notion that some type of information about a student is being collected in a specified, systematic, and uniform way. In the context of classroom music teaching, a broadening of the term, *test*, can provide music teachers with a clearer picture of the variety of evidence that can be collected to make inferences about student learning as well as provide a better contextualization for all the opportunities in which evidence of student learning can be collected.

Classroom assessment encompasses *all* activities undertaken by teachers and students that provide information to be used as feedback to modify the teaching and learning activities occurring in the classroom. The multitude of assessments embedded within classroom instruction move beyond the purpose of assessment *of* learning (a summative approach to assessment) as often used for the purpose of large-scale assessment processes and serve as foundation of assessment *for* learning (a formative approach to assessment), where they underscore the planning, guiding, monitoring, and facilitation of teaching and learning in the classroom. In other words, classroom assessments are nested within the instructional environment, inferences made about student learning are internal to the learning environment, and the inferences teachers make about student learning outcomes directly affect the actions taken by teachers and students to improve instruction.

Traditional Theories

To apply the concepts of validity, reliability, and fairness in a manner appropriate for classroom assessment, it is important to first consider their broad considerations in the context of large-scale assessment. Traditional theories of validity, reliability, and fairness are most often considered, discussed, and written about from a large-scale assessment context. Therefore, each of the quality indicators is considered as it relates to the assessment context itself or more specifically, the inferences gleaned from the individual testing condition.[10] As a result, each quality indicator concerns itself with the accuracy of inferences regarding student achievement but with a specific focus on how the testing instrument itself performed under the testing conditions. Under large-scale assessment conditions, only empirical data gathered from the implementation of a measurement model can provide the validity, reliability, and fairness evidence needed to objectively establish the assessment context as a quality assessment. Quality, under large-scale assessment conditions, broadly means that there is confidence in the inferences made from the assessment context with regard to the way the measurement instrument performed and the way the testing scores represent the student's overall achievement.

According to the *Standards for Educational and Psychological Testing*,[11] validity is defined as "the degree to which evidence and theory support the interpretations of the testing scores for proposed uses of tests." Validity asks the question, "How strong of an argument can be made that the inferences drawn from the testing scores are truly representative of the student taking the test?" The validation process traditionally includes the gathering of varied types of evidence that support a test's outcomes in relation to the context in which it was used. This can include evidence of how well the test measures what is intended to be measured in relation to a similar type of test (e.g., criterion validity), how well the content included in the test represents what is intended to be measured (e.g., content validity), how well the test items work together to measure what is intended to be measured (e.g., construct validity), or the considerations for the intended use of the test once the assessment data are collected and inferences are made (e.g., consequential validity). In each of these validity arguments, there is a clear relationship between the student's testing score, the testing instrument itself, and the inferences made.

Reliability, according to the *Standards for Educational and Psychological Testing*, "is defined in terms of consistency over replications of the testing procedure. Reliability/precision is high if the testing scores for each person are consistent over replications of the testing procedure and is low if the testing scores are not consistent over replications."[12] In large-scale assessment, replication refers to consistency of the testing scores across equivalent tests or across test items, for example. Various types of empirical reliability coefficients have been proposed that reflect different underlying conceptualizations of the notion of replication in measuring a test's consistency. Similar to the various validity arguments, there is a clear relationship between the student's testing score, the testing instrument itself, and the proposed inferences when reliability arguments are made.

In the context of large-scale assessment, fairness is a validity issue concerned with the degree to which the measurement procedures and testing scores result in accurate estimates of student achievement.[13] According to the *Standards for Educational and Psychological Testing*,[14] fairness is the "responsiveness to individual characteristics and testing contexts so that testing scores will yield valid interpretations for intended uses." In other words, arguments for fairness include the statistical investigation into differences in students' testing scores based on any subgroup affiliation of interest to the test constructor or stakeholders. Examples of these affiliations can include socioeconomic status, parental involvement, access to technology, gender, race, and parents' educational background, for example. Fairness, therefore, is often associated with considerations related to opportunity to learn, inclusion, and social justice.

The traditional definitions and uses of validity, reliability, and fairness are

not necessarily meaningful or useful in their application or interpretation as they relate to classroom assessments. First, large-scale assessments include the measurement of student ability through a single assessment context *outside* of the classroom, whereas classroom assessments include the evaluation of student learning throughout the instructional process using multiple assessments *within* the classroom. Second, large-scale assessments use testing scores derived from a *measurement model* to infer student ability, whereas classroom assessments use *formal and informal evaluations* to infer student learning.

The statistical indices that support validity arguments in large-scale assessment are not of day-to-day interest to the music teacher. The use of measurement models is inaccessible or even inappropriate based on the context of classroom assessments, and classroom assessments are too embedded within and affected by the instructional process. Therefore, from a traditional validity perspective, there is not enough information to support a traditional validity argument from a large-scale assessment perspective.

Because replication is essential to reliability arguments, the student would need to be either judged by multiple people or respond more than once to the same prompt or the teacher would have to collect and investigate large amounts of assessment data to statistically simulate replication of the assessment. Therefore, from a traditional reliability perspective, there are little to no data to support a reliability argument from a large-scale assessment perspective.

The small number of students in any classroom, the embedded nature of performance assessments, and the lack of statistical tools to support the analysis of fairness make its testing virtually impossible for the classroom teacher. Therefore, there are no data or opportunity to support a fairness argument from a large-scale assessment perspective.

Although this is an oversimplification of validity, reliability, and fairness procedures, it is clear that there is an incompatibility of paradigms in considering the quality of large-scale assessments versus the quality of classroom assessments. However, the broad principles of validity, reliability, and fairness, from a qualitative perspective, may be loosely applied to classroom assessment contexts as a method for evaluating and improving their quality. Sidebar 2 presents some questions to consider when creating classroom assessments.

"Classroometric" Theory

From a classroom assessment perspective, validity is the confidence that a teacher has in the quality of the inferences he or she makes about student-learning outcomes. In particular, teachers should consider three broad areas when considering the validity of any classroom assessment: (1) relevance, (2) level of thinking processes, and (3) congruency.

Relevance refers to the alignment between the national and/or state standards relevant to the instructional cycle, the learning objectives, the content taught, and the content of the

SIDEBAR 2

Validity, Reliability, and Fairness in Classroom Assessments: Questions to Consider	
Validity ***The confidence in the quality of the inferences made about student-learning outcomes.***	
Relevance ***The alignment between the national/state standards, learning objectives, and content on the assessment.***	
	Is the content on the assessment properly aligned with the learning outcomes of the instructional unit?
	Is the content on the assessment properly aligned with the content taught throughout the instructional unit?
	Is the content on the assessment properly aligned with district, state, or national standards?
Level of Thinking Processes ***Considerations of the cognitive rigor of the assessment in relation to the cognitive rigor of the class content.***	
	Is the difficulty of the assessment adequately matched to the student's ability level?
	Is the difficulty of the assessment adequately matched to the classroom content?
	Does difficulty of the assessment adequately represent the student's day-to-day behaviors?
	Does the range of difficulty on the assessment adequately represent the knowledge, skills, abilities, and/or dispositions being assessed?
Congruency ***Relationship of the outcome of an assessment with previous patterns of student performance.***	
	Does the result of the assessment generally match the expected result based on prior student behavior?
	Are the majority of the students unexpectedly overachieving or underachieving on the assessment?
	Is the student bringing with him or her prior experiences that can affect the outcome of the assessment?

SIDEBAR 2 (CONTINUED)

Reliability ***The dependability of assessments to adequately support inferences made about student-learning outcomes.***	
Differentiation of Assessment Types ***Use of multiple assessment types to ensure a student's opportunity to demonstrate student-learning outcomes.***	
	Is there enough information to make an accurate judgment about the student's knowledge, skill, ability, or disposition being assessed?
	If the student were to be assessed again, is there confidence that he or she would be evaluated or respond to the questions the same way?
	What varying types of information is the assessment providing to make a judgment of what the student knows or is able to do?
	Are there other types of assessments that can be used to elicit the knowledge, skills, abilities, and/or dispositions being assessed in a different way?
Clear Communication of Expectations ***Ensuring student understanding of teacher's learning outcome expectations.***	
	For performance tasks, is an assessment instrument, such as a scoring rubric or rating scale, being used to clearly communicate expectations of the assessment?
	Is there a set of illustrative student work that serves as exemplars for expectations at all achievement levels?
Systematic Assessment Procedures ***Ensuring student understanding, familiarity, and engagement with assessment procedures.***	
	Is the student comfortable with the assessment process?
	Is the assessment procedure itself affecting the ability of the student to demonstrate the knowledge, skills, abilities, and/or dispositions being assessed?
	If a formal assessment procedure is used, is the student aware and/or prepared for it?
	Does the student understand how to engage with the assessment?
Fairness ***The opportunities for a student to best demonstrate student-learning outcomes.***	
Transparency ***Clear communication between teacher and student with regard to assessment content, content, and use.***	
	Does the student know what the assessment is going to be used for?
	Is the student aware of any positive and negative consequences of this assessment?
	Are the consequences of the assessment procedure itself affecting the ability of the student to demonstrate the knowledge, skills, abilities, and/or dispositions being assessed?
Student Opportunities ***The ability of the student to adequately demonstrate student-learning outcomes in varied ways.***	
	Are students being provided multiple and varied opportunities to demonstrate what they know and what they are able to do?
	Are accommodations necessary to allow some students to best demonstrate their knowledge, skills, abilities, and/or dispositions?
	Are unnecessary accommodations being made for the student?
	Is the student actively engaged in the learning process being assessed?
	Does the assessment authentically evaluate the day-to-day knowledge, skills, abilities, and/or dispositions of the student?
Teachers' Critical Reflection ***Considerations of personal bias or stereotyping that may impede the assessment process.***	
	Are assumptions of prior knowledge being made about the student that can affect the outcome of the assessment?
	Is there flexibility between teacher expectations of the level of knowledge, skills, abilities, and/or dispositions being assessed and the actual level of knowledge, skills, abilities, and/or dispositions?
	Are any teacher's stereotypes of the student affecting the assessment process?
	Are any group affiliations of the student (e.g., gender, race, ethnicity, ability level, instrument, etc.) affecting the assessment outcome?
	Are any personal interactions with the student affecting the assessment outcome?

assessment.[15] Assessments occurring in the music classroom should be aligned to national and/or state standards and should directly reflect the content taught in the classroom (i.e., content representativeness). Content representativeness, therefore, is a critical consideration for the validity of any music assessment.[16]

Level of thinking processes refers to the considerations of the cognitive rigor of the assessment in relation to the cognitive rigor of the class content. The level of cognitive rigor, as often defined in educational taxonomies such as Bloom's Taxonomy or Webb's Depth of Knowledge, for example, should be considered for what is being asked of the student in the assessment. In particular, teachers should strategically use, make reference to, and align assessment criteria to an appropriate educational taxonomy.[17] This not only ensures age appropriateness and skill level appropriateness of the assessment but also helps align the cognitive rigor of the assessment to the cognitive rigor of typical classroom and/or rehearsal experiences.

Congruency refers to the relationship of the outcome of an assessment with the previous patterns of student performance. In general, high-achieving students in the context of classroom activities should typically demonstrate high achievement on most assessments. Average and low-achieving students in the context of classroom activities should generally demonstrate average and low achievement on most assessments, respectively. This is not to say that some students will inevitably demonstrate some fluctuation in achievement across various assessments or content areas; however, drastic changes in patterns to these relationships may indicate cause to investigate the validity of the inferences the teacher is making.

Reliability, in the context of classroom assessment, refers to the dependability of assessments to adequately support inferences made about student-learning outcomes. Consistency, in the case of classroom assessments, refers to the uniformity between what a student demonstrates on a day-to-day basis in the classroom and what a student demonstrates during an assessment. Teachers considering the reliability of their assessments should consider three areas: (1) differentiation of assessment types, (2) clear communication of expectations, and (3) systematic assessment procedures.

Particularly in music, the demonstration of student-learning outcomes can occur under multiple contexts. The ability of a student to perform a rhythm, identify the same rhythm orally, or identify the same rhythm in written notation may all address a similar learning outcome but vary slightly in the context and skill. By providing students with varied opportunities to demonstrate their abilities through the *differentiation of assessment types*, a more thorough and consistent picture of student-learning outcomes can be obtained.

A *clear communication of the teacher's expectations* is essential for reliable assessment. For performance evaluations, the use of rubrics or rating scales can provide the student with clearly outlined and defined criteria for what specifically is being evaluated.[18] Furthermore, having exemplar performances aligned to varying levels of achievement across the criteria of the rubric or rating scale can provide students with tangible evidence of performance expectations.

Having a *systematic set of assessment procedures* can help improve the consistency with which students are assessed. Students' comfort with what, when, where, and why they are being assessed can improve his or her engagement in the assessment process and provide a more realistic depiction of his or her knowledge, skills, abilities, or dispositions.

Fairness, in the context of classroom assessment, refers to an assessment context whereby the opportunity is available for a student to best demonstrate student-learning outcomes. Teachers considering the fairness of their assessments should consider three areas: (1) transparency, (2) student opportunities, and (3) teachers' critical reflection.[19]

Transparency refers to the communication between teacher and student regarding the assessment context, content, and use. Teachers should focus on explicitly communicating with students what specific learning outcome is being evaluated, how the student will be evaluated, what the result of the assessment means about the student's level of achievement, and what the consequences of the assessment are.

Student opportunities refer to the ability for students to adequately demonstrate accurate student-learning outcomes in varied ways. Just as teaching should be differentiated to accommodate students' learning processes in a classroom, assessment should be differentiated to provide the best opportunity for a student to demonstrate his or her ability.

Teacher's critical reflection refers to the consideration of the teacher's knowledge and perceptions of his or her students. In particular, teachers should be considerate of any personal biases or stereotypes that may affect the outcomes of the assessment process.

Different validity, reliability, and fairness considerations can be considered at the various stages throughout an assessment cycle. Figure 1 provides considerations for each of these indicators at four different stages of the assessment cycle: (1) designing the assessment, (2) administering the assessment, (3) evaluating the assessment results, and (4) redesigning the assessment/future teaching considerations.

A New Conception

Arguments for the validity, reliability, and fairness of assessments are traditionally in the context of large-scale assessment contexts. As a result, the properties for evaluating their quality make them limited in both their practicality and relevance for classroom assessments. Their consideration in a context more appropriate for music classrooms may lend itself to improving the quality of classroom assessments. Now, more than ever, it is important for music educators to strongly embrace the potential for assessment in providing a more clear, accurate, and precise picture of student-learning outcomes in the music classroom. Continued dialogues

FIGURE 1

Validity, Reliability, and Fairness Considerations in Action throughout a Classroom Assessment Cycle

[1] Designing the Assessment

Validity Considerations

- Is the content on the assessment properly aligned with the learning outcomes of the curricular unit?
- Is the content on the assessment properly aligned with the content taught throughout the curricular unit?
- Is the content on the assessment properly aligned with district, state, or national standards?
- Is the difficulty of the assessment adequately matched to the student's ability level?
- Is the difficulty of the assessment adequately matched to the classroom content?
- Does difficulty of the assessment adequately represent the student's day-to-day behaviors?
- Does the range of difficulty on the assessment adequately represent the knowledge, skills, abilities, and/or dispositions being assessed?

Reliability Considerations

- Is there enough information to make an accurate judgment about the student's knowledge, skill, ability or disposition being assessing?
- What different types of information is the assessment providing to make a judgment of what the student knows or is able to do?
- Is there a set of illustrative student work that serves as exemplars for expectations at all achievement levels?

Fairness Considerations

- Are students being provided multiple and varied opportunities to demonstrate what they know and what they are able to do?
- Does the assessment authentically evaluate the day-to-day knowledge, skills, abilities, and/or dispositions of the student?

[2] Administering the Assessment

Validity Considerations

- Is the student bringing with him/her prior experiences that can affect the outcome of the assessment?

Reliability Considerations

- Is the student comfortable with the assessment process?
- Is the assessment procedure itself affecting the ability of the student to demonstrate the knowledge, skills, abilities, and/or dispositions being assessed?
- Is the student aware and/or prepared for this assessment?
- Does the student understand how to engage with the assessment?

Fairness Considerations

- Does the student know what the assessment is going to be used for?
- Is the student aware of any positive and negative consequences of this assessment?
- Are the consequences of the assessment procedure itself affecting the ability of the student to demonstrate the knowledge, skills, abilities, and/or dispositions being assessed?
- Are accommodations necessary to allow some students to best demonstrate their knowledge, skills, abilities, and/or dispositions?
- Are unnecessary accommodations being made for the student?
- Are any personal interactions with the student affecting the assessment outcome?
- Are students being provided multiple and varied opportunities to demonstrate what they know and what they are able to do?
- Does the assessment authentically evaluate the day-to-day knowledge, skills, abilities, and/or dispositions of the student?

[3] Evaluating the Assessment Results

Validity Considerations

- Does the result of the assessment generally match the expected result based upon prior student behavior?
- Were the majority of the students unexpectedly overachieving or underachieving on the assessment?

Reliability Considerations

- Did the assessment instrument clearly communicate the expectations of the assessment?
- If the student were to be assessed again, is there confidence that he/she would be evaluated or respond to the questions the same way?
- Are there other types of assessments that can be used to better elicit the knowledge, skills, abilities, and/or dispositions being assessed?

Fairness Considerations

- Was the student actively engaged in the learning process being assessed?
- Were assumptions of prior knowledge made about the student that can affect the outcome of the assessment?
- Did any stereotypes of the student affect the assessment process?
- Did any group affiliations of the student (e.g., gender, ethnicity, ability level, instrument, etc.) affect the assessment outcome?

[4] Redesigning the Assessment/Future Teaching Considerations

Validity Considerations

- Was the content on the assessment properly aligned with the learning outcomes of the curricular unit?

FIGURE 1 (CONTINUED)

- Was the content on the assessment properly aligned with the content taught throughout the curricular unit?
- Was the content on the assessment properly aligned with district, state, or national standards?
- Was the difficulty of the assessment adequately matched to the student's ability level?
- Was the difficulty of the assessment adequately matched to the classroom content?
- Did the difficulty of the assessment adequately represent the student's day-to-day behaviors?
- Did the range of difficulty on the assessment adequately represent the knowledge, skills, abilities, and/or dispositions being assessed?

Reliability Considerations

- Was there enough information to make an accurate judgment about the student's knowledge, skill, ability, or disposition being assessing?
- What different types of information did the assessment provide to make a judgment of what the student knew or was able to do?
- Was there an appropriate set of illustrative student work that served as exemplars for expectations at all achievement levels?

Fairness Considerations

- Were students being provided multiple and varied opportunities to demonstrate what they know and what they are able to do?
- Did the assessment authentically evaluate the day-to-day knowledge, skills, abilities, and/or dispositions of the student?

on classroom music assessment are critical toward improving teaching and learning in the music classroom. Taken together, each of these three quality indicators can help improve assessment practice in the music classroom.

Notes

1. Edwards P. Asmus, "Considerations for Teaching Music Education Assessment," in *Proceedings of the Tenth International Symposium of the Research Alliance of Institutes for Music Education*, ed. Sven-Erik Holgersen and Frede V. Nielsen (Copenhagen, Denmark: RAIME, 2010), 27; Susan M. Brookhart and Anthony J. Nitko, *Educational Assessment of Students Seventh Edition* (Boston, MA: Pearson), 2016.
2. The National Coalition for Core Arts Standards (NCCAS) 2014 Music Standards can be found at https://nafme.org/my-classroom/standards/.
3. For the purpose of this article, *large-scale assessment* is defined as a test administered to a large number of students, usually at the state or district level, that implements measurement models as a method for inferring student achievement.
4. Susan M. Brookhart, "Developing Measurement Theory for Classroom Assessment Purposes and Uses," *Educational Measurement: Issues and Practice* 22, no. 4 (2003): 5–12.
5. Ibid.
6. Gregory J. Cizek, "An Introduction to Contemporary Standard Setting," in *Setting Performance Standards: Foundations Methods, and Innovations*. 2nd ed., ed., Gregory J. Cizek (New York: Routledge, 2012), 3.
7. Ibid.
8. See Brian C. Wesolowski, "Understanding and Creating Rubrics for the Assessment of Music Performance," *Music Educators Journal* 98, no. 3 (2012): 36–42.
9. A distinction must be made between grading and assessment. The purpose of a grade is to empirically label student learning, most often for summative or reporting purposes. The purpose of assessment is to improve student learning.
10. A thorough evaluation of validity, reliability, and fairness from a large-scale music assessment context can be found in Brian C. Wesolowski and Stefanie A. Wind, "Validity, Reliability, and Fairness in Music Testing," in *The Oxford Handbook of Assessment Policy and Practice in Music Education*, ed. Timothy S. Brophy (New York: Oxford University Press, 2019), 437–60.
11. American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME), *Standards for Educational and Psychological Testing* (Washington, DC: American Psychological Association, 2014).
12. Ibid., 35.
13. Brian C. Wesolowski, Stefanie A. Wind, and George Engelhard Jr., "Rater Fairness in Music Performance Assessment: Evaluating Model-Data Fit and Differential Rater Functioning," *Musicae Scientiae* 19, no. 2 (2015): 147–70.
14. AERA et al., *Standards*, 50.
15. Brian C. Wesolowski, "Tracking Student Achievement in Music Performance: Developing Student Learning Outcomes for Growth Model Assessments," *Music Educators Journal* 102, no. 1 (2015): 39–47.
16. Susan M. Brookhart and Anthony J. Nitko, *Educational Assessment of Students*, 8th ed. (Boston, MA: Pearson, 2019).
17. See Wesolowski, "Tracking Student Achievement in Music Performance," for developing student-learning objectives and aligning assessment criteria to educational taxonomies in the context of music performance.
18. See Wesolowski, "Understanding and Creating Rubrics," for information on developing and implementing rubrics and rating scales in the context of music performance.
19. Robin D. Tierney, "Fairness in Classroom Assessment," in *SAGE Handbook of Research on Classroom Assessment*, ed. James H. McMillian (Thousand Oaks, CA: SAGE, 2013), 125–44.