5 Validity, Reliability, and Fairness in Classroom Tests

Brian C. Wesolowski

Chapter Overview

This chapter addresses the evaluation of music classroom testing quality using three key indicators: **validity**, **reliability**, and **fairness**.

Learning Expectations for the Chapter

- Define and describe a latent construct.
- Define and describe validity, reliability, and fairness in the context of large-scale testing.
- Define and describe validity, reliability, and fairness in the context of music classroom testing.
- Describe processes to evaluate the validity, reliability, and fairness of classroom testing outcomes.

Essential Questions for the Chapter

- What makes the testing of musical constructs so difficult?
- Why are validity, reliability, and fairness important considerations when making inferences about student achievement in the music classroom?
- How can we ensure that we truly capture student learning through the process of testing?

When we hear the word *test*, we may automatically consider a multiple-choice, fill-in-the-blank, or a true-and-false exam. We may think of the process of sitting down at a desk with a pencil and paper or sitting in front of a computer answering systematically crafted questions that reflect some type of content knowledge. As music students and music performers, the word test, in the context of music, specifically, might automatically bring about the idea of a performance or playing test,

as the field does not usually engage in standardized tests or written classroom exams. According to Cizek (2012), however, a test can be more broadly conceptualized as "simply a data collection procedure; more precisely, . . . a sample of behavior(s) taken and interpreted under specified, systematic, and uniform conditions" (p. 3). As Cizek suggests, the concept of a test is frequently interpreted too narrowly, and it is often mistaken as referring to a specific format for data collection instead of the broad, structured process used to collect the information. Using Cizek's definition, a test, in the context of a music classroom, can be considered *any circumstance in which a teacher makes a systematic observation of students' musical behavior*. As music teachers, we test our students every day, whether the information is collected for diagnostic, formative, or summative reasons, formally or informally, in the context of a cognitive task or a performance task, through general, day-to-day observations or systematic, scheduled examinations. A test, broadly considered, is not necessarily the mechanism for how a music teacher evaluates students' musical behaviors or the particular method of scoring or grading, but the notion that some type of information about a student's knowledge, skills, abilities, or dispositions is being systematically collected.

The testing and **measurement** of any musical behavior is an abstract concept. If we were to measure the height of a person, we could use a physical ruler with inches as our unit of measurement. If we were to measure the weight of a person, we could use a physical scale with pounds as our unit of measurement. In both cases, the measurement is made with a pre-established physical instrument and is conducted with some unit of measurement that is specifically being used to make comparisons or more or less of the attribute being measured. Does one person have more or less height than another person? Does one person have more or less weight than another person? If we were to measure the performance achievement of a clarinetist, however, the process becomes much more complex. How would we define performance achievement? What musical behaviors would we evaluate about the clarinetist that reflect the idea of performance achievement? How would we collect evidence of performance achievement? What type of measurement instrument would we use to evaluate performance achievement? How would we score the student based upon the behaviors we observe? Unlike the physical sciences, where testing and measurement are concrete, physical, and utilize a specified unit of measurement, testing and measurement in the behavioral sciences, including music, are abstract, psychological, and do not often utilize a specified unit of measurement.

From a theoretical testing perspective, tests of musical behavior measure some type of unobservable, or *latent*, construct (Loevinger, 1957). Latent constructs (sometimes referred to as latent traits, behaviors, or attitudes) can be defined as "*any construct that cannot be directly measured*

but rather inferred through the measurement of secondary behaviors" (Wesolowski, 2019). More specifically, a latent construct:

> is an idea developed or "constructed" as a work of informed, scientific imagination; that is, it is a theoretical idea developed to explain and organize some aspects of existing knowledge . . . the construct is much more than a label; it is a dimension understood or inferred from its network of interrelationships[.]
>
> (APA, AERA, & NCME, 1974, p. 29)

Musical constructs such as musical aptitude, music performance achievement, musical preference, ear-training ability, or any other music content knowledge or music performance expectation, for example, cannot be measured with physical instruments and there are no specified units of measurement associated with the constructs. In most of the educational, behavioral, and psychological sciences, such as music, these constructs cannot be directly measured; rather, they are *inferred* using secondary, observable behaviors that represent the latent construct. The test itself acts as an operational definition of the latent construct. The items (for a cognitive task) or criteria (for a performance task) included within a test are the secondary behaviors that are observable. The interactions between the items/criteria and the student are observable in the sense that the student engages with the items/criteria in order to demonstrate some level of achievement. If it is a cognitive task, the student directly engages with the item by responding to it. If it is a performance task, the student indirectly engages with the criteria (mediated by the teacher scoring the performance). **Inferences** are the conclusions that are made by the teacher about the adequacy of the test in regard to the latent construct being measured, the data-gathering procedures, the level of achievement of the student, and their interpretation of the data gathered from the test. Inferences are the link between the secondary, observable behaviors and the primary, unobservable construct (see Figure 5.1).

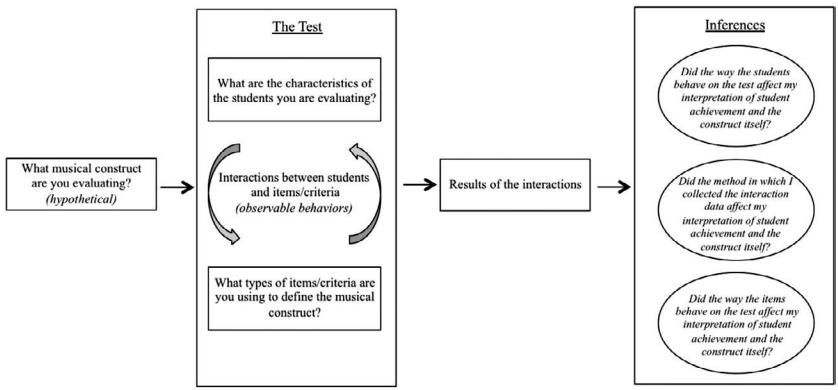

Figure 5.1 Making inferences from hypothetical constructs

As an example, the latent construct of music aptitude is an unobservable behavior that cannot be directly measured. As music teachers, we cannot simply hold a theoretical ruler up to a student and indicate that the student has certain amount of musical aptitude. Music aptitude, such as defined by Edwin Gordon's (1995) *Musical Aptitude Profile*, for example, is inferred based upon secondary, observable behaviors including students' interactions with test items associated with (a) tonal imagery (melody and harmony), (b) rhythm imagery (tempo and meter), and (c) musical sensitivity (phrasing, balance, and style). In Arnold Bentley's (1966) *Measures of Music Abilities*, music aptitude is inferred based upon secondary, observable behaviors including students' interactions with test items associated with (a) pitch discrimination, (b) tonal memory, (c) chord analysis, and (d) rhythmic memory. In these two examples, different measures developed by both Gordon and Bentley include different observable behaviors (i.e., domains and test items), but both purport to measure the same latent construct of musical aptitude.

As another example, music performance achievement is an unobservable behavior that cannot be directly measured. In the National Association for Music Education's (NAfME) (2018) *Concert Band Assessment Form*, music performance achievement is defined based upon secondary, observable behaviors including an ensemble's ability to perform a task associated with (a) sound quality (tone quality and pitch), (b) technical accuracy (technique and rhythm), (c) musicality (interpretation, musicianship, dynamics, and breath), and (d) stage deportment. In the Florida Bandmasters Association (FBA) (2018) *Concert Band Assessment Form*, music performance achievement is inferred based upon secondary, observable behaviors including an ensemble's ability to perform a task associated with (a) performance fundamentals (tone quality, intonation, balance, blend, sonority, and physical articulation), (b) technical preparation (note accuracy, rhythmic accuracy, precision, entrances, releases, interpretive articulation, clarity of articulation, technique, stability of pulse, dynamics observed, and transitions), and (c) musical effect (expression, shaping of line, style, interpretation, phrasing, tempo, and dynamic expression). In this example, different measures developed by both NAfME and FBA include different observable behaviors (i.e., domains and evaluative criteria), but both purport to measure the same latent construct of music performance achievement.

Learning Experience: *Locate the performance assessment for large groups in your state. Then, locate the performance assessment for large groups in another state. Identify and list the observable behaviors in the assessment. Then, in groups, compare and contrast both state performance assessments and discuss the importance placed on each through the development of the given assessment.*

In both the music aptitude and music performance achievement examples, the test developers each have a unique, but similar operational definition of the construct. So who is right, which measure is better, and what does this mean in the context of the music classroom? For a classroom music teacher, it is the teacher's responsibility to define the construct, outline the specific observable behaviors that define the construct, and ensure that instruction is aligned to the observable behaviors. For example, if a music teacher is collecting information about his or her students' rhythm identification achievement, rhythm identification achievement is the latent construct. Rhythm identification achievement, theoretically, is an unobservable construct in itself. It becomes the teacher's job to operationally define the latent construct through some type of test. Whether the data collection process is a cognitive task (e.g., multiple-choice items, dictation items) or a performance task (e.g., performance criteria via a rating scale, rubric), the items/criteria that the students directly or indirectly interact with are the observable behaviors that represent the teacher's operational definition of rhythm identification achievement. Each student interacts with each item or criterion, either directly providing answers for each question for a cognitive task or being judged by the teacher for a performance task, resulting in multiple observable behaviors that represent the construct of rhythm identification achievement. It is then the teacher's job to make inferences of the student achievement based upon the collection of behaviors made. The most important question to ask in the music aptitude example, the music performance achievement example, and the rhythm identification achievement example is, *How good are the inferences about the student's achievement based upon the data collected from the observations?* In order to answer questions pertaining to the quality of inferences made from a test, three important indicators should be considered: validity, reliability, and fairness.

> ***Learning Experience:*** *Choose a performance task of your choice that you might use in your classroom. Design criteria that identify observable behaviors and your expectations as they relate to those given behaviors.*

Traditional Perspectives of Validity, Reliability, and Fairness

Traditional views of validity, reliability, and fairness are most often written about in the context of large-scale, high-stakes testing (See Wesolowski & Wind, 2019 for detailed overview). Examples of these large-scale tests may include SATs, GREs, or other state-based competency exams. The

strict, operational definitions that stem from these writings, however, are not entirely relevant for the application in classroom settings. In order to evaluate the quality (i.e., validity, reliability, and fairness) of the inferences gleaned from large-scale tests, psychometricians typically apply measurement models and statistical analyses to the collected observations that provide both quantitative and qualitative evidence of quality, each distinctly related to validity, reliability, and fairness considerations. In these contexts, validity, reliability, and fairness investigations are test-centered, external to the classroom, standardized, and student achievement–based. Classroom tests, on the other hand, are student-centered, internal to the classroom, non-standardized, and student-learning-based.

Classroom teachers are not particularly focused on how a specific classroom test performs; rather, they are more concerned with the how multiple classroom tests, together, can holistically provide information about student achievement to help understand, develop, foster, and improve student learning. Results of classroom testing are much more embedded into the instructional environment and are used to guide and facilitate student learning through cyclical teaching and learning processes. In these contexts, validity, reliability, and fairness investigations should be student-centered, internal to the classroom, non-standardized, and student-learning-based (see Table 5.1). Therefore, principles of validity, reliability, and fairness in the context of classroom teaching and learning should include different applications and considerations compared to those typically described from a large-scale testing perspective.

Table 5.1 Differences between large-scale tests and classroom tests

	Large-Scale Tests	*Classroom Tests*
Reporting	Standardized	Non-standardized
Purpose	Measurement of learning (summative)	Measurement for learning (formative and summative)
Context	Linear	Cyclical
Relevance	Achievement-based	Learning-based
Process	Mastery	Developmental
Situatedness	Norm-referenced/ criterion-referenced	Individual learning–based
Consequences	Often high-stakes/formal	Often low-stakes/ informal
Data-type	Quantitative	Quantitative and qualitative
Construct Complexity	Unidimensional	Multidimensional

(Continued)

Table 5.1 (Continued)

	Large-Scale Tests	*Classroom Tests*
Responsibility of Validity	Test publisher, test user	Teacher, student
Content Representativeness	Standards-based	Curriculum-based
Decision-making	Policy-based, program-based	Instructional-based, curriculum-based
Changes	External (outside classroom)	Internal (inside classroom)
Objectivity	Evaluated by statistical error	Evaluated by systematic procedures
Consequences	Accountability	Improvement of student learning

In order to consider validity, reliability, and fairness in the context of classroom assessment, it is important to first conceptualize and understand validity, reliability, and fairness in the context of large-scale testing contexts. Using these definitions and applications as a foundation, we can then build upon them to make them more meaningful and relevant for classroom use.

Validity

Validity, according to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), is defined as follows:

> Validity refers to the degree to which evidence and theory support the interpretations of the test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations.
>
> (p. 11)

The validation process in large-scale assessment contexts includes gathering a combination of qualitative and quantitative evidence to support the inferences about student achievement in relation to the test itself. This evidence can include multiple types of validity, including three of the broadest validity types: criterion validity, content validity, and construct validity.

Criterion validity refers to the extent to which a test matches related outcomes of a similar test measuring the same construct (Cronbach & Gleser, 1965). From our earlier music aptitude example, we could test

our students' music aptitude ability using Edwin Gordon's (1995) *Musical Aptitude Profile* and Arnold Bentley's (1966) *Measures of Music Abilities* and see how similarly the students perform on both measures. If there is a high correlation, meaning that the ordering of the students (from high-achieving to low-achieving) in terms of their test performance is similar, one could provide an argument for the criterion validity of the measures.

Content validity refers to how adequately the content of the test covers the construct being measured (Messick, 1989). From our earlier music performance achievement example, we would want to ensure that there is an adequate use of domains and items in a way that (a) satisfactorily describes and defines music performance achievement and (b) effectively separates students based upon varying music performance achievement levels. For example, if a music performance achievement measure altogether omits items related to intonation, an arguably important criterion for high-quality performance achievement, there is an argument for underrepresentation of the latent construct and a source of content invalidity. A result of this omission may also play a role in inadequately separating out students who satisfactorily perform their musical selections in tune from students who do not satisfactorily perform their musical selections in tune. Other considerations for content validity arguments include clarity in the writing and overall appropriateness for the items/criteria in relation to the construct.

Construct validity refers to how well the items function together to represent the construct being measured (Cronbach & Meehl, 1955). As Wesolowski (2018) notes:

> the development of criteria within the context of an assessment is, more broadly, the process of developing a hypothetical, latent construct. . . . Therefore, the development of assessments is dualistic: (a) to provide a means for measuring students, and (b) to develop a hypothetical latent construct . . . therefore, the [testing] results . . . not only provide information about student performance; more importantly, they provide diagnostic information about each of the latent constructs represented by the measurement instruments[.]
>
> (p. 151)

Because the latent construct being development is a hypothetical abstraction, evidence is needed for how well the items and/or criteria function together to define the construct. From our earlier music performance achievement example in the context of content validation, we would want to ensure that there is an adequate use of domains and items in a way that (a) satisfactorily describes and defines music performance achievement and (b) effectively separates students based upon varying music performance achievement levels. Construct validity is

closely related to content validity in that the goals of the validation process are the same: to satisfactorily define the latent construct and to effectively separate students based upon their achievement. However, in the case of construct validity, the validation process becomes more concrete. We are interested in the difficulty range of the items/criteria (i.e., how easy the least difficult item is to how difficult the most difficult item is) in relation to the achievement range of the students (i.e., how low the least-achieving student is to how high the most-achieving is) as well as how the items and students interact throughout both ranges.

Most broadly, validity answers 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?* It is worth noting that tests themselves are not validated. It is the inferences, or meaning one derives from testing outcomes, that are validated.

> ***Learning Experience:*** *As an in-class discussion, describe a testing moment in your life that may have been invalid. Why did you feel this way? How would you describe the infringement upon validity? Discuss with your class.*

Reliability

According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), reliability is defined as:

> the 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.

Evidence of reliability is demonstrated through various types of statistical estimations, including stability reliability coefficients (*Will the ordering of the students from high-scoring to low-scoring be the same across tests if repeated?*), equivalence coefficients (*How strong of a relationship is there between student scores using two or more tests of the same difficulty?*), internal consistency coefficients (*How consistent are the students' scores across the items within the same test?*), and rater reliability coefficients for performance tasks (*How consistent are raters with themselves and/or with each other?*), for example. Reliability answers the question, *How stable are the measures?*

> ***Learning Experience:*** *As an in-class discussion, describe a testing moment in your life that may have been unreliable. Why did you feel this way? How would you describe the infringement upon reliability? Discuss with your class.*

Fairness

Fairness, according to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), is defined as the *responsiveness to individual characteristics and testing contexts so that testing scores will yield valid interpretations for intended uses*. Investigations into fairness often include considerations that may adversely affect student outcome scores, including but not limited to (a) test content (e.g., item content that may systematically favor or disadvantage some groups of students over others based on prior knowledge, experiences, level of interest or motivation, or other variables), (b) test context (e.g., aspects of the testing environment that systematically affect student outcome scores, such as clarity used in the test instructions, the complexity of vocabulary within the test items or tasks, or the language in which the test is administered), (c) test response (e.g., writing or speaking tasks may result in differences in responses that are unrelated to the construct due to cultural views related to wordiness or rate of speech, and survey items may result in differences in responses due to perceptions of social desirability), or (d) opportunity-to-learn (e.g., the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test). Fairness is the investigation of biases (i.e., systematic lower or higher scoring outcomes) against subgroups of students related to four considerations described above, including socioeconomic status, parental involvement, access to technology, gender, race, and parent's educational background, for example. Fairness answers the question, "Do all students receive equitable treatment during the testing process?"

> ***Learning Experience:*** *As an in-class discussion, describe a testing moment in your life that may have been unfair. Why did you feel this way? How would you describe the infringement upon fairness? Discuss with your class.*

Music Classroom Perspectives on Validity, Reliability, and Fairness

In the case of large-scale assessment, the three quality indicators of validity, reliability, and fairness are largely connected to the measurement of student achievement based upon scoring outcomes of standardized tests, external to the classroom. In order to provide grounded arguments for each of these indicators, an inference must be made about student achievement using a specified measurement model stemming from some theory of measurement. These theories can include Classical Test Theory, Generalizabilty Theory, or Item Response Theory (IRT), for example. From a large-scale testing perspective, measurement of student achievement for any educational, psychological, or behavioral construct, including music, can only be achieved via a measurement model. In making inferences about student achievement from their engagement with a large-scale test, the implementation of measurement models provides objective, empirical evidence that supports validity, reliability, and fairness arguments.

Applications of measurement models are inaccessible or even inappropriate for classroom music teachers. The statistical indices and other qualitative considerations that support item and construct analyses are not necessarily of day-to-day interest to the music teacher, and the time, energy, and effort to conduct such analyses would draw the teacher away from instructional time. Therefore, from a traditional validity perspective, there are no data to support a validity argument. Because replication is essential to reliability arguments, the student would need to be either judged by multiple people, the student would need to respond more than once to a prompt in order to evaluate the reliability of the assessment, or statistical reliability analyses would need to be conducted by the teacher to investigate the reliability of a test. Therefore, from a traditional reliability perspective, there are no data to support a reliability argument. The small number of students in any classroom, the embedded nature of the performance assessment, and the lack of statistical tools to support the analysis of bias makes fairness testing virtually impossible for the classroom music teacher. Therefore, from a traditional fairness perspective, there are no data to support a fairness argument. 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 the assessment contexts themselves or, more importantly, quality of the inferences gleaned from the assessment contexts (see Table 5.2).

Seeing this inconsistency, Brookhart (2003) called for the development of improved methodologies for evaluating the qualities of classroom

Table 5.2 Validity, reliability, and fairness in the context of large-scale assessments and classroom assessments

		Large-Scale Assessment	*Classroom Assessment*
Validity	*Purpose*	Purpose is meaningful inference of student achievement.	Purpose is meaningful inference of student learning.
	Measurement	Objective measurement of the student using a measurement model.	Implied measurement of the student using raw scores and qualitative interactions.
	Alignment	Knowledge/skill/ability/disposition content is set forth by content standards.	Knowledge/skill/ability/disposition content is set forth by learning objectives and instructional activities.
	Performance standards	Level of achievement is specified by performance standards using norm-referenced reporting (student classification) or criterion-referenced reporting (students on a continuum).	Level of achievement is specified by illustrative exemplars of student work at multiple achievement levels.
	Testing	Formal test is conducted externally and validity argument is made by test constructor and test user.	Formal and informal tests conducted internally and validity argument is made by teacher and student.
	Evidence for arguments	Statistical evidence supports validity argument.	Instructional decisions support validity argument.
Reliability	*Purpose*	Purpose is statistical evidence of stability of scores across multiple testing occasions.	Purpose is to attain sufficient information that demonstrates learning across instructional cycle.
	Interpretation	Statistical property of the score itself.	Qualitative property of teacher's diagnosis of student learning.
	Evidence	Small standard errors, high internal consistency.	Systematic classroom procedures aligned with test content.

(Continued)

Table 5.2 (Continued)

		Large-Scale Assessment	*Classroom Assessment*
Fairness	*Purpose*	Purpose is statistical evidence of construct-irrelevant variability stemming from non-assessment characteristics.	Purpose is to assure demonstration of learning is not limited by non-classroom characteristics.
	Test response	Statistical evidence of construct-irrelevant variability stemming from student characteristics.	Teacher provides varied and differentiated opportunities for students to demonstrate learning.
	Test content	Statistical evidence of construct-irrelevant variability stemming from test content.	Teacher demonstrates that the test content is aligned with learning objectives (appropriateness).
	Test context	Statistical evidence of construct-irrelevant variability stemming from testing context.	Students clearly understand what is being assessed and how to engage with the test.
	Test construct	Statistical evidence of construct-irrelevant variability stemming from response type.	Teacher uses varied types of assessments, such as cognitive tests, performance evaluations, portfolios, self-reflections.
	Opportunity-to-learn	Statistical evidence of construct-irrelevant variability stemming from opportunity-to-learn factors.	Teacher assures demonstration of learning is not affected by individual student characteristics.

assessments across all educational contexts through her notion of classroometrics. Classroometrics, according to Brookhart,

> should take into account that classroom assessments provide information about students that immediately becomes part of their learning environment and their own psychology . . . . the main actions of interest are relatively immediate, internal changes in the students who are measured.
>
> (pp. 8–9)

Drawing on Brookhart's suggestions for improved quality control of classroom testing, Wesolowski (in press) provided suggestions for validity,

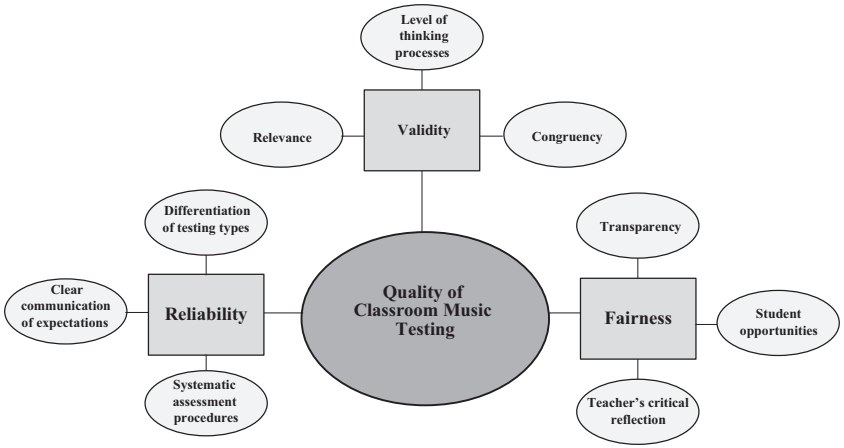

Figure 5.2 Considerations toward the quality of classroom music testing

reliability, and fairness considerations in classroom music testing that are more accessible, applicable, manageable, and relevant for music teaching and learning in the classroom. These considerations are more qualitative in nature and are based upon important questions to ask while crafting, implementing, and reflecting upon any classroom music testing circumstances. Figure 5.2 is an overview of these considerations.

Validity

In the context of classroom testing, validity can be defined as the confidence that a teacher has in the quality of the inferences they make about student-learning outcomes. There are three important considerations toward the validity of classrooms tests: (a) relevance, (b) levels of thinking processes, and (c) congruency.

Relevance

Relevance refers to the alignment between the content of the test and any related national/state standards and learning objectives that underscore the related teaching. There are three important questions to consider when evaluating validity with respect to relevance:

1. Is the content of the test properly aligned with the learning outcomes of the curricular unit?
2. Is the content of the test properly aligned with the content taught throughout the curricular unit?
3. Is the content of the test properly aligned with district, state, or national standards?

Level of thinking processes refers to the considerations of the cognitive rigor of the test in relation to the cognitive rigor of the course content and student abilities. There are four important questions to consider when evaluating validity with respect to level of thinking processes:

1. Is the difficulty of the test adequately matched to the student's ability level?
2. Is the difficulty of the test adequately matched to the instructional content taught in the classroom?
3. Does the difficulty of the test adequately match the ability represented by the student during class instruction?
4. Is the range of difficulty within the test adequately distributed?

Congruency refers to the relationship of the outcome of the test with previous patterns of student achievement. There are three important questions to consider when evaluating validity with respect to congruency:

1. Does the student's outcome of the test generally match the teacher's expected outcome?
2. Are there large groups of students who are unexpectedly overachieving or underachieving on the test?
3. Is the student bringing with him or her any prior experiences or unrelated knowledge that can affect the outcome of the test?

Reliability

In the context of classroom testing, **reliability** can be defined as the dependability of the test to adequately support the inferences made about the student learning outcomes. There are three important considerations toward the reliability of classroom tests: (a) differentiation of assessment types, (b) clear communication of expectations, and (c) systematic assessment procedures.

Differentiation of assessment types refers to the use of multiple assessment types to ensure a student's opportunity to demonstrate student-learning outcomes. There are four important questions to consider when evaluating reliability with respect to differentiation of assessment types:

1. Is there enough information to make an accurate judgment about the student's knowledge, skills, abilities, or dispositions being assessed?
2. If the student were to be tested again, is there confidence that he or she would be evaluated or respond to the questions in the same way?
3. What different types of information is the test soliciting in order to make a judgment of what the student knows or is able to do?
4. Are there other types of tests that can be used to elicit the knowledge, skills, abilities, or dispositions being tested?

Clear **communication of expectations** refers to the teacher ensuring that student has a comprehensive understanding of the teacher's learning outcome expectations. There are two important questions to consider when evaluating reliability with respect to communication of expectations:

1. Do the items (for cognitive tasks) or criteria (for performance tasks) clearly represent the learning outcomes and clearly communicate the expectations of the assessment?
2. Is there a set of illustrative student work that serves as exemplars for expectations across all achievement levels?

Systematic assessment procedures refer to the teacher ensuring student understanding, familiarity, and engagement with assessment procedures. There are four important questions to consider when evaluating reliability with respect to systematic assessment procedures:

1. Is the student comfortable with the testing process?
2. Is the testing procedure itself affecting the ability of the student to demonstrate the knowledge, skills, abilities, and/or dispositions being tested?
3. If a formal testing procedure is used, is the student aware of it and prepared for it?
4. Does the student understand how to engage with the test?

Fairness

In the context of classroom testing, fairness can be defined as the opportunities for a student to best demonstrate student-learning outcomes. There are three important considerations toward the fairness of classroom tests: (a) transparency, (b) student opportunities, and (c) teachers' critical reflection.

Transparency refers to the clear communication between teacher and student in regard to the testing context, testing content, and testing use. There are three important questions to consider when evaluating fairness with respect to fairness:

1. Does the student know what the test is going to be used for?
2. Is the student aware of any positive and/or possible negative consequences of the test?
3. Are the consequences of the testing procedure itself affecting the ability of the student to optimally demonstrate the knowledge, skills, abilities, and/or dispositions being tested?

Student opportunities refers to the ability of the student to adequately and accurately demonstrate their ability to meet student-learning outcomes in

varied ways. There are five important questions to consider when evaluating fairness with respect to student opportunities:

1. Are students being provided multiple and varied opportunities to demonstrate what they know and what they are able to do?
2. Are accommodations necessary to allow for some students to best demonstrate their knowledge, skills, abilities, and/or dispositions being tested?
3. Are unnecessary accommodations being made for the student?
4. Is the student actively engaged day-to-day in the learning process and manner in which the learning process is being assessed?
5. Does the test authentically evaluate the day-to-day knowledge, skills, abilities, dispositions, and ways of being engaged with what is being tested?

Teachers' critical reflection refers to the teacher's considerations toward personal biases or stereotyping that may impede the testing process. There are five important questions to consider when evaluating fairness with respect to teachers' critical reflections:

1. Are assumptions of prior knowledge being made about the student that can affect the outcome of the test?
2. Is there flexibility between teacher expectations of the level of knowledge, skills, abilities, and/or dispositions being tested and the actual level of knowledge, skills, abilities, and/or dispositions being tested?
3. Are any teacher's stereotypes of the student affecting the testing process?
4. Are any group affiliations of the student, such as gender, ethnicity, ability level, instrument, affecting the testing outcome?
5. Are any personal interactions between the teacher and the student affecting the testing outcome?

Learning Experience: *Select either a music or non-music course that you are currently enrolled in. Consider an individual testing context, specifically, and all of the testing contexts, broadly, that demonstrate your learning of the course content. What evidence of validity, reliability, and fairness have you witnessed? What is your evidence? Prepare a short presentation to share your findings with the class.*

Summary

Musical constructs are difficult to measure because they are hypothetical in nature. More specifically, they are based upon secondary, observable behaviors that are used to infer the abstraction of the construct. As such, validity, reliability, and fairness considerations are necessary in order to ensure the quality of the inferences a teacher makes from testing outcomes. In the context of classroom testing, validity is defined as *the confidence that a teacher has in the quality of the inferences they make about student-learning outcomes* and is underscored by three important considerations: (a) relevance, (b) level of thinking processes, and (c) congruency. In the context of classroom testing, reliability is defined as *the dependability of the test to adequately support the inferences made about the student-learning outcomes* and is underscored by three important considerations: (a) differentiation of assessment types, (b) clear communication of expectations, and (c) systematic assessment procedures. In the context of classroom testing, fairness is defined as *the opportunities for a student to best demonstrate student-learning outcomes* and is underscored by three important considerations: (a) transparency, (b) student opportunities, and (c) teachers' critical reflection.

Activities and Worksheets

1. Reflect upon an instance when you were assessed in a music classroom setting in the context of a cognitive test and a performance test. Answer the following. Upon completion, discuss with the class.

Cognitive Test
What latent construct did this test measure?
_____
What were the testing conditions?
_____ _____
Were the inferences the teacher concluded about your achievement of high quality? (yes or no)
_____

Were the inferences the teacher concluded about your achievement valid? Explain.
_____ _____ _____
Were the inferences the teacher concluded about your achievement reliable? Explain.
_____ _____ _____
Were the inferences the teacher concluded about your achievement fair? Explain.
_____ _____ _____
Performance Test
What latent construct did this test measure?
_____
What were the testing conditions?
_____ _____
Were the inferences the teacher concluded about your achievement of high quality? (yes or no)
_____
Were the inferences the teacher concluded about your achievement valid? Explain.
_____ _____ _____
Were the inferences the teacher concluded about your achievement reliable? Explain.
_____ _____ _____
Were the inferences the teacher concluded about your achievement fair? Explain.
_____ _____ _____

2. Pair up with a classmate and construct a hypothetical scenario in which you are going to evaluate a student's achievement on a musical (a) cognitive task and (b) performance task. In both instances, describe any safeguards you would put in place to ensure the quality of the inferences you would make with regard to (a) validity, (b) reliability, and (c) fairness.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association (APA), American Educational Research Association (AERA), & National Council on Measurement in Education (NCME). (1974). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Bentley, A. (1966). *Musical ability in children and its measurement*. New York: House Inc.

Brookhart, S. M. (2003, Winter). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, 5–12.

Cizek, G. J. (2012). An introduction to contemporary standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 3–14). New York, NY: Routledge.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Florida Bandmasters Association. (2018). *Florida Bandmasters Association adjudicator's comment sheet: Concert band*. Retrieved January 3, 2019, from http://fba.flmusiced.org/media/1483/judgesheet-concertmus-with-rubric-rev2013-bullets.pdf

Gordon, E. E. (1995). *Musical aptitude profile (Grades 5–12)*. Chicago, IL: GIA Publishers.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

National Association for Music Education. *National music adjudication coalition concert band or orchestra music assessment form*. Retrieved January 3, 2019 from https://nafme.org/wp-content/files/2016/04/Ensemble-Adjudication-Form-PDF.pdf

Wesolowski, B. C. (2018). Examination of the psychometric properties of the Model Cornerstone Assessments. In F. Burrack & K. A. Parkes (Eds.), *Applying Model Cornerstone Assessments in K-12 music: A research-supported approach* (pp. 151–180). Lanham, MD: Roman and Littlefield.

Wesolowski, B. C. (2019). Item response theory and music testing. In T. S. Brophy (Ed.), *The Oxford handbook of assessment, policy, and practice in music education* (pp. 479–503). New York: Oxford University Press.

Wesolowski, B. C. (in press). "Classroometrics": The validity, reliability, and fairness of classroom music assessments. *Music Educators Journal*.

Wesolowski, B. C., & Wind, S. A. (2019). Validity, reliability, and fairness in music testing. In T. S. Brophy (Ed.), *The Oxford handbook of assessment, policy, and practice in music education* (pp. 437–460). New York: Oxford University Press.