CHAPTER 20

VALIDITY, RELIABILITY, AND FAIRNESS IN MUSIC TESTING

BRIAN C. WESOLOWSKI AND
STEFANIE A. WIND

The focus of this chapter is on validity, reliability, and fairness in music testing. A test can be defined simply as any procedure whereby data is collected. In the context of music, a test can be described more specifically as the collection and interpretation of data representing a particular musical behavior using a systematic and uniform procedure. As such, a test in music can encompass the evaluation of any number of musical behaviors and the collection of data using any number of techniques. Examples may include a multiple choice test to measure musical aptitude, a checklist to measure a musical procedure, a performance assessment to measure music performance achievement, an essay or oral exam to measure a student's ability to discuss steps they took to solve a musical problem, or a systematic observation to measure a student's ability on a particular musical task. With any test, the collection of data needs to be valid, reliable, and fair in order to ensure quality and provide defensibility of the assessment. With any assessment, test users need to ensure that quality inferences are made from a test (i.e., validity), be confident that students with similar knowledge, skills, and abilities will receive similar scores across repeated administrations (i.e., reliability), and certify that the test is accessible and appropriate for all students, regardless of any subgroup affiliation such as race, ethnicity, gender, socioeconomic status, and so forth (i.e., fairness). This chapter is designed to help music educators better understand validity, reliability, and fairness within the framework of measurement and evaluation and to help contextualize it in the field of music education.

Validity

General Considerations and Definitions

The recent revision of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) defines validity 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)

Tests are developed with specific objectives in mind. These objectives may include the measurement of how well a student performs at a given point in time in a particular content area, to forecast or estimate a student's future standing based on a related criterion, or to infer the degree to which a student possesses a particular latent quality that can be reflected by a testing outcome. Each of these objectives can be supported by specific validation methods. These may include the statistical demonstration that a particular set of items supports a particular content area, the correlation between a test and a criterion variable, or statistical support of how well particular items fit a latent construct. According to the *Standards*, multiple findings must be taken into account in order to assess the validity of a test for a particular interpretation and use.

The validation process includes the gathering of specific evidence that supports a test's outcomes in relation to the context in which it was used. Therefore, it is not the test itself that is validated, but the inferences one makes from the measure based on the context of its use. As Vernon (1960) argues, a test is always valid for some purpose, and therefore can be considered on a continuum of more valid or less valid depending on the circumstance for which it is being implemented.

History and Definitions of Validity Theory

Historically, "validation" is a catchall term for the procedures that determine the usefulness of a measure. Based on classic definitions of validity, a test is considered to be valid if it measures what it purports to measure. Over the years, however, philosophical shifts have occurred in the paradigms and frameworks of validity theory. Most notably, this shift is evidenced through the focus of the classification of multiple types of validity of the test itself to a unified approach of validity evidence in the context in which the test is being used. Broadly, the first half of the 20th century was led by the singular concept of a criterion-based (and closely associated content-based) model of validity; in the 1950s and 1960s, the concept shifted toward the construct-based model of validity. Starting in

the 1980s, the concept began to shift toward a more unified, consequential-based model of validity that included expanded discussions of the moral foundations to testing and measurement (see Angoff, 1988; Hubley & Zumbo, 1996; Jonson & Plake, 1998; Kane, 2001, for detailed histories of validity theory).

The Criterion-Based Model of Validity

The earliest writings on validity include sophisticated summaries of how to calibrate values of an attribute of interest and correlate them to a preexisting, established criterion value (Cronbach & Gleser, 1965). Validity, from this perspective, is the accuracy in which the estimated values match the criterion values. As Cureton (1950) notes:

> A more direct method of investigation, which is always preferred wherever feasible, is to give the test to a representative sample of the group with whom it is to be used, observe and score performances of the actual task by the members of this sample, and see how well the test performances agree with the task performances. (p. 623)

As an example, if a new music aptitude measure were to be constructed, the data gathered from the new measure could be correlated with data gathered from a previously existing music aptitude measure (i.e., criterion scores) to gather validity evidence of the new measure. In this case, each measure purports to measure musical aptitude and therefore, if the correlations are high and positive, evidence of criterion validity exists. According to the criterion-based model, there are four particular types of validity: (1) *predictive validity* (how well a score on a test predicts a measurable criterion); (2) *concurrent validity* (how well a test score corresponds to previously established measures of the same construct); (3) *convergent validity* (how well a test score converges, or is similar to tests that should theoretically have high correlations), and (4) *discriminant validity* (how well a test score diverges, or is dissimilar to tests that should theoretically have low correlations).

Criterion-based models are useful in contexts where an appropriate criterion exists. However, a disadvantage to the criterion-based model is the availability and/or validity of the criterion measure itself.

The Content-Based Model of Validity

The disadvantage of criterion validity is accessibility to a high-quality criterion measure in the criterion-based model. This led to a content-based model of validity. The content-based model posits the establishment of a measure based on a desired outcome with content expert interpretation of the scores. As an example, intrinsic validity of a measure is accepted if a music performance achievement measure can appropriately discriminate between high and low achieving performances based on music content experts' opinions. Here, the measure itself is the criterion for performance achievement, as opposed to

correlating it to another criterion measures. Messick (1989) describes content validity as "the domain of relevance and representativeness of the test instrument" (p. 17). Content validity arguments include two key components. First, items must adequately represent the domain of interest. The items ideally represent an adequate sample of a particular dimension of interest. Second, the measure should be constructed sensibly. Specifically, detail should be given to features such as clarity in instruction, appropriateness of language, and overall appearance of the measure itself. Although this model is acceptable in contexts such as auditions, placements, or other achievement-based assessments, it often comes with a confirmatory bias and subjectivity of content expert opinion.

The Construct-Based Model of Validity

The previous descriptions make clear that criterion- and content-based validity models are of limited value for the evaluation of constructs common to the field of music (as well as many other psychological and behavioral fields). For example, the constructs of musical aptitude, performance achievement, and musical intelligence are theoretical constructs that are not directly observable. Therefore, inferences must be made based on operationalizations set forth in the testing context and measurement instruments.

Cronbach and Meehl's (1955) construct-based model of validity was based on an adaptation of a hypotheticodeductive (HD) model of theory (Popper, 1959, 1962). The HD model is a scientific method of inquiry where hypotheses can be validated or falsified based on tests of observable data. The model is based on a set of axiomatic theorems connected by sets of empirical laws used to validate the observable data in order to infer evidence of an unobserved, latent trait. According to the 1966 *Standards* (APA, AERA, & NCME, 1966), construct validity "is ordinarily studied when the tester wishes to increase his understanding of the psychological qualities being measured in by the test. . . . Construct validity is relevant when the tester accepts no existing measure as a definitive criterion" (p. 13).

Construct validation is based on theories of unobservable (i.e., latent) traits that are inferred through secondary, observable variables:

> A psychological 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 to 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)

The validation process is underscored by four distinct processes: (1) the development of a theory, (2) the development of a measure to both directly and indirectly reflect the theory, (3) the development of a hypothesis to reflect the theory, and (4) the testing of the hypothesis against the theory. As a result, the content validation process comprises three methodological principles: (1) a broader theoretical conceptualization of validation principles, (2) the explicit statement of a proposed hypothesis and related interpretation, and (3) the explicit statement of alternative hypotheses.

A Unified Model of Validity

Construct validation processes provided the foundation for broader and more current perspectives in the concept of validation. Messick (1989, 1996a, 1996b) provided early discussions of a unified view to the validity argument under the umbrella of construct validity, where heavier emphasis is placed on how the test scores are used and what context they are used in. According to Messick (1989), "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13, emphasis in original).

Messick (1996b) provides six key types of validity criteria that should function as interdependent and complementary forms of evidence for all educational and psychological measurement: (1) *content* (i.e., the knowledge, skills, and other attributes to be revealed by the assessment task), (2) *substantive* (i.e., the verification of the domain processes to be revealed in assessment tasks), (3) *structure* (i.e., scoring consistency with what is known about the internal structure of the construct domain), (4) *generalizability* (i.e., representative coverage of the content and processes of the construct domain), (5) *external factors* (i.e., the extent that the assessment scores' relationship with other measures and nonassessment behaviors reflect the expected high, low, and interactive relations implicit in the specified construct), and (6) *consequential* (i.e., evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short term and long term).

Most notable of Messick's (1989) validity argument is the notion of consequence. If a music performance achievement measure is to be constructed, there may be several different meanings and interpretations of the resulting scores. These may include the result of a particular instructional intervention for a student, placement in an ensemble, acceptance into a program of study, or an ensemble adjudication (e.g., in the United States, an all-state ranking). Lane and Stone (2002) outline several examples of both intended and unintended consequences of validity arguments. Examples of intended consequences may include:

Motivation of the measure;
Content and strategies underscoring the assessment protocol;
Format of the assessment;
Improved learning;
Professional development support;
Use and nature of test preparation activities; and/or
Beliefs and awareness of the assessment, assessment process, criteria for judging; and use of the results.

Examples of unintended consequences may include:

The narrowing of the instructional process or curriculum to focus on specific learning outcomes assessed;

Use of preparation materials closely associated with the assessment without direct changes to curriculum or instruction;

Use of unethical assessment preparation or related materials; and/or Inappropriate test scores by administrators.

In these cases, the validation process can be supported by the processes of constructing, investigating, and evaluating arguments for and against the intended interpretation and application of the resulting test scores. Questions that guide the development of these arguments may include:

What is the purpose of this assessment?
How will the results of the assessment be used?
Who are the stakeholders that require an explanation of the results?
What are the stakeholders' perceptions of the assessment?
What are the intended and unintended consequences of the assessment?
What benefits have the participants gained in participating in the assessment?
What is the direct influence of student achievement through participation in the assessment?
To what extent do particular variables influence and/or predict assessment outcomes?
What is the intended impact of assessment outcomes on future teaching and learning?
To what extent are students and/or teachers included in the accountability process of the assessment?

Messick (1989) argues, "Inferences are hypotheses and the validation of inferences is hypothesis testing" (pp. 13–14). Kane (2001) further elaborates that these inferences are based on a network, which leads "from the scores to the conclusions and decisions based on the scores, as well as the assumptions supporting these inferences" (p. 56). Validity, therefore, can be conceptualized as an evaluative *argument* based on scientific theory. Because validity evidence is contextual, no single hypothesis can be confirmed or disconfirmed, only supported through appropriate evidence.

Sources of Invalidity

Two plausible rival hypotheses that challenge the proposed interpretation of the results of an assessment may be generated when encountered with unintended consequences: *construct underrepresentation* (i.e., *construct deficiency*) and *construct-irrelevant variance* (i.e., *construct contamination*). *Construct underrepresentation* "refers to the degree to which a test fails to capture important aspects of the construct" (AERA et al., 2014, p. 12). In other words, the task being measured fails to account for important dimensions of the construct, yielding an outcome that fails to provide evidence of a student's true ability. As an example, an assessment developed to measure individual student performance achievement may underrepresent the achievement construct if the difficulty level of the student's instrument is not accounted for; or, the assessment instrument only accounts for auditory facets of the performance when psychomotor facets, such as body carriage or instrument positioning, may influence the outcome.

Construct-irrelevant variance "refers to the degree to which test scores are affected by processes that are extraneous to the test's intended purpose (AERA et al., 2014, p. 12). In other words, the test measures too many variables that are not relevant to the assessment context, leading to a systematic over- or underestimation of a student's ability level. As an example, in a large group music performance assessment, a sight-reading piece may be overly difficult for an ensemble. This may lead to a case of *construct-irrelevant difficulty*. Another ensemble may have familiarity with the sight-reading piece. This may lead to a case of *construct-irrelevant easiness*. Studies of internal and external assessment structures can provide empirical evidence of systematic patterning (Wesolowski, Wind, & Engelhard, 2015). These studies include analyses of *differential item functioning* (DIF) (i.e., subgroups of test takers with similar status on the assessment criterion have systematically different responses to a particular item), *differential rater functioning* (DRF) (i.e., subgroups of test takers with similar status on a performance-based assessment criterion show have systematically different results due to differential severity/leniency of a rater), and *differential facet functioning* (DFF) (i.e., subgroups of test takers with similar status on the assessment criterion have systematically different responses to any facet of interest).

The *Standards* (AERA, APA, & NCME, 1999) deliberately state:

> Validation is the joint responsibility of the test developer and the test user. The test developer is responsible for furnishing relevant evidence and a rationale in support of the intended use. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used. (p. 11)

Therefore, the joint endeavor of both test developer and test user toward the validation of an assessment context is fundamental to the unified model of validity.

The validity of a musical test is one of the most important considerations in musical assessments. As Asmus (2010) notes, "Gathering evidence of validity should be standard operating procedure for anyone developing a music assessment" (p. 143). If the validity of a test is not established, the resulting data and any inferences drawn from the testing context are meaningless. Therefore, with the administration of any musical test, it is paramount for the test user, whether a policymaker, music administrator, or music educator, to be clear in what the test purports to be measuring and whether the test results will specifically answer what is being asked.

Reliability

General Considerations and Definitions

The recent revision of the *Standards for Educational and Psychological Testing* (AERA et al., 2014) defines reliability as follows:

> The general notion of reliability/precision is defined in terms of *consistency over replications* of the testing procedure. Reliability/precision is high if the scores for

> each person are consistent over replications of the testing procedure and is low if the scores are not consistent over replications. (p. 35, emphasis added)

This definition highlights an emphasis within the *Standards* on the notion of replications as central to the conceptualization of reliability, where replications are defined to reflect a particular interpretation and use of test scores. Brennan (2001) points out that replications can be specified that reflect the various sources of measurement error of interest or concern for a given testing procedure. For example, in a rater-mediated music performance assessment, replications could be defined as raters, such that consistency over raters provides evidence of reliability. Likewise, replications of a standardized college admissions examination could be defined as administrations, such that consistency of student scores across administrations provides evidence of reliability.

A major theme emphasized in the recent revision of the *Standards* is the need for reliability evidence to support each intended interpretation and use of test scores. A variety of reliability coefficients have been proposed that reflect different underlying conceptualizations of replications and the role of measurement error. In the next section, we describe reliability estimates based on classical test theory (CTT; i.e., true score theory), followed by a discussion of reliability estimation methods based on modern measurement theories.

True Score Reliability Estimates

The first set of reliability coefficients presented in this chapter are based on the CTT model. As noted by Wesolowski, Wind, and Engelhard (2015, 2016a, 2016b), much of the empirical work done in the field of music teaching and learning related to measurement is based on the CTT framework (see Boyle & Radocy, 1986; Asmus & Radocy, 1992). Developed in relation to early work on correlation coefficients (e.g., Spearman, 1904, 1910, 1913), the CTT model is defined as shown in Figure 20.1.

In Figure 20.1, X is the observed score, T is the true score, and E represents error. Within the framework of CTT, true scores are a theoretical concept defined as the mean observed score (i.e., expected value) across infinite replications of a measurement procedure. Error scores are the difference between the observed score and the true score. The CTT model is based on three major underlying assumptions that facilitate the calculation of reliability coefficients. First, the average value of error scores for a population of test takers is zero. Second, the correlation between true scores and error scores within a population of test takers is zero. Finally, the correlation between error scores for repeated testing procedures (unique tests or unique occasions) is expected to be zero (Crocker & Algina, 1986; Lord & Novick, 1968).

$$X = T + E$$

FIGURE 20.1 Classical test theory model.

$$\rho_{XT} = \frac{\sigma_T}{\sigma_X}$$

FIGURE 20.2 Classical test theory reliability index formula.

Based on these assumptions, reliability is defined as the correlation between true scores and observed scores, which can be expressed as the ratio of the standard deviation of true scores to the standard deviation of observed scores using the *reliability index*. This is shown in Figure 20.2.

Because true scores are unknown, it is necessary to approximate the reliability index using observed scores. Accordingly, the *reliability coefficient* (ρ_{xx}) is defined as the correlation between observed scores across replications. The reliability coefficient ranges between .00 and 1.00. A reliability of 1.00 indicates that the differences between observed scores are perfectly consistent with differences in true scores. A reliability of .00 indicates that the differences between observed scores are perfectly inconsistent with differences in true scores.

Differences in the definition of replications lead to different specifications of the reliability coefficient. In this section, we describe reliability coefficients that reflect three major specifications of replications: (1) Replications across administrations, or stability coefficients; (2) Replications across alternate forms, called equivalence coefficients; and (3) Replications across items, known as internal consistency coefficients.

Replications Across Administrations: Stability Reliability Coefficients

The first category of CTT reliability coefficients describes the consistency of observed scores across administrations. Coefficients in this category can be considered *stability reliability coefficients*, because they describe the degree to which the ordering of test takers in terms of total scores is stable across administrations. First, the *test-retest coefficient* is the correlation between observed scores obtained from separate administrations of the same test form, where there is some time interval between administrations. Values of the test-retest coefficient describe the degree to which differences in observed scores can be attributed to differences across administrations, such as administration procedures, scoring errors, guessing, or other changes in behavior. No single rule exists regarding the appropriate length of time between administrations when test-retest reliability is of interest. However, Crocker and Algina (1986) offer the following advice:

> The time period should be long enough to allow effects of memory or practice to fade but not so long as to allow maturational or historical changes to occur in the examinees' true scores. The purpose for which test scores are to be used should be taken into account when designating the waiting time. (p. 133)

In addition to the repeated administration of a single test form, it is also possible to estimate reliability using two administrations of similar test forms. Specifically, the *test-retest with alternative forms* coefficient is calculated as the correlation between observed scores on two similar test forms. Values of this reliability coefficient describe the degree to which differences can be attributed to administrations *as well as* differences in the content included in each form. Various criteria exist for establishing the degree of acceptable similarity between test forms, with the strictest definition based on the concept of *parallel forms*. Parallel forms exist when students have equivalent true scores on each form, and error variance for each form is equal—such that the two test forms have equivalent means and variances of observed scores.

Replications Across Alternate Forms: Equivalence Coefficients

The second category of CTT reliability coefficients includes coefficients that describe the consistency or equivalence of observed scores on alternative forms of a test. The *alternative forms reliability* or *equivalence coefficient* is calculated as the correlation between observed scores on parallel or similar test forms. Unlike the test-retest coefficient, data collected for alternative forms reliability studies should only include a short time period between administrations to prevent test-taker fatigue. Further, in order to prevent ordering effects, the presentation of the two forms should be reversed in half the sample, such that one half of test takers receives Form A first, followed by Form B, while the second half receives Form B first, followed by Form A. Values of alternative forms reliability coefficients describe the degree to which observed scores are affected by differences related to the two forms.

Replications Across Items: Internal Consistency Coefficients

The third category of CTT reliability coefficients includes coefficients that describe the consistency of observed scores across items within a single test form. These coefficients are desirable in situations in which only one test form will be administered on only one occasion. Internal consistency coefficients reflect the correlation between observed scores across items or subsets of items within a single form. Values of *internal consistency coefficients* describe the degree to which individual items or item subsets reflect the same content domain. A variety of CTT internal consistency coefficients have been proposed and are commonly applied in practice, including split-half methods and methods based on item covariances. This section describes two commonly used internal consistency reliability coefficients that reflect these two major categories: Spearman-Brown reliability and Cronbach's alpha coefficient.

Split-Half Methods

The first major category of internal consistency reliability coefficients is based on creating two subsets of items within a test form (i.e., two halves of a test) that are parallel or approximately parallel forms. A variety of methods are used in practice to create test halves, including random assignment of items to halves, splitting items based on odd and even numbers, rank-ordering items by difficulty and assigning items with odd and even ranks to different halves, and splitting the items based on content. After test halves are established, the correlation between observed scores on each half is used as an estimate of reliability.

Because reliability coefficients calculated from split-halves do not reflect the full length of test forms, several correction procedures have been proposed that provide an estimate of the reliability coefficient that would have been obtained using the entire test form. The *Spearman-Brown Correction* (i.e., the Spearman-Brown prophecy formula) is one such correction that is calculated using two halves of a single test form (Brown, 1910; Spearman, 1910). Stated mathematically, the correction is shown in Figure 20.3, where ρ_{AB} is the observed correlation between half-tests A and B. When applying the Spearman-Brown correction, it is important to note that the degree to which the halves used to calculate ρ_{AB} are parallel will affect the corrected value of the reliability coefficient. Deviations from parallel forms will result in less-accurate reliability coefficients.

Methods Based on Item Covariances

The second major type of internal consistency reliability coefficient is calculated using observed scores on individual items. These reliability estimates reflect the consistency of test taker performance across individual items on a test, thus overcoming potential problems associated with different reliability coefficients resulting from different definitions of half-tests. Several reliability coefficients based on item covariances are commonly used in practice, including the Kuder-Richardson formulas (Kuder & Richardson, 1937), and Cronbach's alpha coefficient (Cronbach, 1951). Because the two formulas yield equivalent results, and because Cronbach's alpha coefficient can be applied to both dichotomous and polytomous items, we focus here on Cronbach's alpha.

Cronbach (1951) presented a reliability coefficient for a single test administration that is based on the view of the *k* individual items within a test form as *k* individual subtests that act as a type of replication of the measurement procedure. Using the covariances

$$\hat{\rho}_{XX'n} = \frac{2\rho_{AB}}{1 + \rho_{AB}}$$

FIGURE 20.3 Spearman-Brown prophecy formula.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_X^2}\right)$$

FIGURE 20.4 Cronbach *Coefficient Alpha* formula.

among the *k* individual items, Cronbach presented *coefficient alpha* (see Figure 20.4) as an internal consistency reliability coefficient that is calculated as follows for dichotomous or polytomous items:

In Figure 20.4, *k* is the number of items, $\hat{\sigma}_i^2$ is the variance of item *i*, and $\hat{\sigma}_x^2$ is the variance of the total test score. Values of coefficient alpha reflect an estimate of the proportion of variance that can be attributed to the true score, where higher values reflect higher levels of consistency across items.

Alpha-if-Deleted Statistics

Item analysis procedures based on CTT often include the use of item-level statistics that describe the change in internal consistency reliability that would occur if the item of interest was not included in the reliability estimate. For individual items, these *alpha-if-deleted statistics* are calculated using the equation in Figure 20.4 on the remaining *k* items. These statistics provide a valuable diagnostic tool for identifying items that are not homogeneous with the other items in a measurement procedure by comparing the observed alpha-if-deleted statistic to the value of coefficient alpha calculated from the entire set of items.

Modern Measurement Theory reliability coefficients

In addition to reliability coefficients based on CTT, additional methods for evaluating reliability have been proposed based on modern measurement theory frameworks. This section provides an overview of generalizability theory as a measurement framework that focuses on issues related to reliability, followed by a brief description of reliability analyses within the context of item response theory (IRT).

Reliability Estimates Based on Generalizability Theory

Generalizability theory (G theory) is a framework for considering a variety of reliability issues based on a combination of principles from CTT and analysis of variance (ANOVA; Fisher, 1925). The advent of G theory is marked by Cronbach and his colleagues' (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) recognition that typical reliability coefficients are limited in that they do not provide a method for distinguishing among multiple sources of error in measurement procedures. Drawing on the statistical tools provided by ANOVA, Cronbach et al. (1972) applied ANOVA methods to the context of educational and psychological testing in order to systematically explore sources

of error variance. In particular, they observed that researchers can "learn far more by allocating variation to facets than by carrying the conventional reliability analysis" (p. 2).

Specifically, G theory allows researchers to define reliability coefficients that reflect the unique combination of variables in a given measurement procedure, such as items, raters, or occasions, in order to identify the most efficient sample sizes and combinations of facets for future administrations. G theory analyses include two major steps. First, *generalizability studies* are conducted in order to estimate reliability using the observed set of facets. Second, *decision studies* are conducted in order to calculate the predicted changes in reliability coefficients that would occur with different sample sizes and subsets of the facets included in the generalizability study. While it is beyond the scope of the current chapter to explore methods for calculating reliability coefficients based on G theory in depth, there are useful introductory texts on G theory by Shavelson and Webb (1991) and Brennan (2010).

Reliability Estimates Based on Item Response Theory

Within the context of IRT (see Volume 1, Chapter 22, of this handbook), reliability analyses focus primarily on the concept of *targeting* between items, test takers, and other facets in a measurement procedure, where the magnitude of measurement error varies depending on the "appropriateness" of items for persons. Because IRT models allow for the estimation of items, test takers, and other facets on a common scale, it is possible to make direct comparisons between the two in order to examine targeting. Targeting can be maximized in order to yield optimal levels of "information" or precision in the estimation of achievement levels (Embretson & Reise, 2000). Additional details regarding the consideration of reliability issues within the context of IRT can be found in Hambleton and Swaminathan (1985) and Embretson and Reise (2000).

High reliability of a test is essential for a test to be valid. Therefore, it is important to ensure high reliability of all tests in music. Testing, however, is never without error. A misconception often demonstrated by practitioners is that a test itself has high reliability and/or high validity. Regardless of how well a test has been developed or how high of a reliability is cited for a test or measure, practitioners should understand that reliability changes from testing context to testing context, sample to sample. Therefore, estimates of reliability should be considered properties of the scores of the examinees of within the testing context, not the properties of the test itself. Because the estimates of reliability are sample-dependent, how can a practitioner ensure reliability of a measure within their classroom? According to Brookhart (2003), the goal of having a reliable assessment in the classroom is achieved by collecting "stable information about the gap between students' work and 'ideal' work (as defined in students' and teachers' learning objectives" (p. 9). Therefore, in the context of music assessment, teachers need to communicate clear learning objectives for their students, specifically how the students will be measured for each objective, and have an open communication between where the students' current ability is and what the standard for achievement is.

Fairness

This section describes the concept of fairness as it applies to educational and psychological testing in general.

General Considerations and Definitions

Essentially, fairness is a validity issue concerned with the degree to which measurement procedures result in accurate estimates of student achievement in terms of a construct. The 2014 revision of the *Standards* (AERA et al.) is the first version to include a standalone chapter on fairness as a separate foundational area for evaluating the psychometric quality of educational and psychological tests. In the standards, fairness is defined as "responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses" (p. 50).

The discussion of fairness in the *Standards* emphasizes the need to maximize test takers' opportunity to demonstrate their standing on the construct of interest in order to ensure fair testing procedures. Accordingly, test developers and test users are called to consider potential obstacles to these opportunities for individual and groups of test takers. These threats to fairness include a variety of aspects of the testing procedure that result in systematic patterns of observed scores within subgroups that lead to inappropriate interpretations and uses of test scores.

Threats to Fairness in Testing

A variety of factors can act as threats to fairness in a testing situation. The *Standards* (AERA et al., 2014) identify four major categories of threats to fairness: (1) test content; (2) test context; (3) test response; and (4) opportunity to learn. Minimizing these threats is key to ensuring fair testing procedures.

Threats to fairness related to *test content* result in systematic variation in test scores that are not related to the construct of interest (i.e., construct-irrelevant variance) and can be attributed to item content that systematically favors or disadvantages some groups over others based on prior knowledge, experiences, level of interest or motivation, or other variables. Such test content confounds the interpretation of test scores. Next, threats to fairness related to *test context* include aspects of the testing environment that systematically affect performance in a construct-irrelevant way. For example, a testing environment might introduce construct-irrelevant variance through the clarity or degree of specificity used in the test instructions, the complexity of vocabulary within the test items or tasks, and the language in which the test is administered.

The third major threat to fairness discussed in the *Standards* is related to the *response types* elicited by assessment tasks. Specifically, it is possible for fairness issues to arise

when the type of response required for an assessment task results in different score interpretations across individual and groups of test takers. For example, 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. Finally, threats to fairness can result from differences among test takers related to *opportunity-to-learn*. The *Standards* define opportunity-to-learn as "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" (AERA et al., 2014, p. 56). Different opportunities to learn across test takers can result in construct-irrelevant differences in test scores that threaten the validity of test score interpretations, particularly for high-stakes decisions.

Clarity in Operational Definitions

When considering issues related to fairness in testing, it is important to note the distinction between operational definitions of several key terms. Although the terms "differential facet functioning," "bias," and "fairness" are often used interchangeably, these terms have distinct definitions within the field of psychometrics. A brief description of these concepts is provided below.

Differential Facet Functioning

The terms "differential item functioning" (DIF), "differential person functioning" (DPF), and other forms of "differential facet functioning" (DFF) are frequently used in discussions related to testing to describe a lack of fairness that can be attributed to various facets in a measurement procedure, including items and persons. Broadly, the concept of DFF refers to differences in the probability for a correct or positive response between individual or subgroups of test takers with equivalent levels on the construct (i.e., the same ability level). A variety of methods exist for detecting DFF related to both CTT and IRT (see IRT, Volume 1, Chapter 22, of this handbook); across both approaches, DFF analyses focus on identifying interactions between response probabilities (i.e., item difficulties) and subgroup membership (e.g., race/ethnicity, gender, or socioeconomic status). It is important to note that the finding of DFF does not necessarily imply a threat to fairness. Rather, DFF analyses are used to identify subgroup differences for which qualitative evidence is needed to determine whether threats to fairness are observed (Millsap, 2011).

Bias

Whereas DFF refers to statistical evidence of interactions between a facet of interest and characteristics of test takers, *bias* in testing refers to qualitative evidence that these interactions result in meaningful patterns that suggest threats to fairness. Accordingly, statistical evidence of DFF is not sufficient to identify bias in a testing procedure; these patterns must be explored qualitatively in order to detect potential threats to fairness.

It is important to note that, although DFF provides a starting point for mixed-methods investigations of bias, statistical evidence of DFF does not always suggest the presence of bias.

Maximizing Fairness Through Changes to the Testing Procedure

As discussed previously, fairness is a broad concept that describes the degree to which test score interpretations are free from construct-irrelevant variance that results from differences in test takers' opportunities to demonstrate their standing on a construct (AERA et al., 2014). A variety of sources of evidence are needed in order to ensure fair testing practices for all test takers. The *Standards* summarize the requirements for ensuring fairness as follows:

> Fairness is a fundamental issue for valid test score interpretation, and it should therefore be the goal for all testing applications. Fairness is the responsibility of all parties involved in test development, administration, and score interpretation for the intended purposes of the test. (AERA et al., 2014, p. 62)

According to the *Standards* (AERA et al., 2014), test design should be approached from the perspective of *universal design*, such that tests are designed to be as "usable as possible for all test takers" (AERA et al., 2014, p. 56). However, design principles are not always sufficient to ensure fair testing practices for all test takers. As a result, fairness efforts often result in adaptations to testing procedures in order to overcome threats to fairness and maximize access to the construct being measured for individuals or groups of test takers. These adaptations vary in terms of the degree to which the resulting test scores remain comparable with those obtained from the original procedure. Specifically, *accommodations* are relatively minimal changes to the testing procedure that maintain the intended construct, such as changes to the presentation, format, or response procedures. On the other hand, *modifications* are more substantial in terms of the construct; test scores obtained from modified testing procedures do not maintain the same interpretation in terms of the construct.

On the other hand, changes to testing procedures may result from test security considerations that result in the need for alternative forms of a test for use across administrations. When different forms of the same test are used, it is necessary to perform statistical procedures to arrive at a common scale for the two (or more) sets of scores obtained from different forms. A range of *scale linking* methods have been proposed based on CTT and IRT that aim to provide a common score scale across test forms. One commonly used form of scale linking is *test equating*, which is a statistical process used to identify exchangeable scores across different tests or test forms that measure the same construct and that are built to reflect the same content specifications.

Wolfe (2004) identified three major requirements for equating: (1) symmetry, (2) group invariance, and (3) equity. First, *symmetry* requires that student achievement measures

be comparable in both directions (Form A can be equated to Form B and vice versa). *Group invariance* requires that the results from equating be independent of the particular group of students used to conduct the equating. Finally, *equity* requires that test takers will be indifferent to which of the equated test forms is used to obtain measures of their achievement (Cook & Eignor, 1991; Hambleton & Swaminathan, 1985; Lord, 1980).[1]

Fairness Considerations in Rater-Mediated Assessments

In the context of rater-mediated assessments, it is essential that fairness efforts include explicit consideration of sources of construct-irrelevant variance that can be attributed to the raters whose judgmental processes are used to evaluate test-taker performances. Concerns with the quality of ratings are prevalent across research on rater-mediated assessments in a variety of domains, including music assessment. As a result, numerous methods have been proposed for detecting patterns of ratings that may suggest threats to fairness, including rater errors and differential rater functioning. Framed from a fairness perspective, threats to fairness that can be attributed to raters suggest construct-irrelevant differences in rater interpretation of performances across individual or groups of test takers.

Several authors have proposed classification schemes for patterns of ratings that suggest threats to fairness. These patterns are often described as *rater errors*, and classifications of rater errors attempt to describe systematic scoring patterns that can be attributed to construct-irrelevant influences on a rater's judgmental process. For example, Saal, Downey, and Lahey (1980) conducted a review of research on methods for evaluating rating quality in research on educational and psychological testing and concluded that there is a general lack of consistency in the terms, definitions, and methods used to detect rater errors. However, these authors noted that the most common rater errors could be classified in four broad categories: (1) leniency/severity; (2) halo; (3) response sets; and (4) score range restriction. In general, these classifications are used to describe systematic deviations from expected rating patterns that are used to "flag" raters during training or operational scoring procedures for remediation in order to improve the quality of ratings.

Leniency/severity errors include the tendency for a rater to systematically assign lower or higher ratings than are warranted by the quality of performances. These errors are identified by comparing rater severity estimates based on observed average ratings or rater calibrations from an IRT model to the overall average ratings or calibrations for a group of raters. On the other hand, *halo* errors describe the tendency for a rater to judge performances holistically, rather than distinguishing between distinct aspects of performances. Murphy and Cleveland (1991) and Saal et al. (1980) identified several methods for identifying halo error in practice. Specifically, evidence for a halo effect may be obtained through examination of correlations among ratings assigned to distinct aspects of a performance (e.g., domain ratings on an analytic rubric), where high correlations suggest a lack of discrimination among domains. Other methods for

detecting halo focus on standard deviations across domains and interaction effects in a rater-by-student-by-domain ANOVA. Methods based on IRT focus on patterns of residuals that suggest less variation than expected across domains (see IRT, Volume 1, Chapter 22, in this handbook).

The third category of rater errors is *response sets*. Response set errors occur when raters interpret and use rating scale categories in an idiosyncratic or unintended fashion. Several classifications have been proposed for rater response sets. For example, within the context of IRT, Engelhard (2013) distinguishes between "noisy" raters, who provide extreme unexpected responses, and "muted" raters, who demonstrate less variation in their rating patterns than expected. A type of response set, the final category of rater errors is *range restriction*. This class of rater errors refers to a rater's tendency to assign ratings that cluster around a particular rating scale category; this category may be located anywhere on the rating scale. Essentially, the definition of this rater error reflects a view that the true scores in a population are distributed across the score range, such that a uniform or tightly clustered rating distribution would be incorrect. Indices of range restriction that are used in practice include small standard deviations for individual raters across students within domains, kurtosis of a rating distribution, and the lack of a significant student main effect in a rater-by-student-by-domain ANOVA (Murphy & Cleveland, 1991; Saal et al., 1980).

When considering fairness issues for rater-mediated assessments, it is important to note that the differences between DFF and bias in testing in general also apply to raters. Specifically, analyses used to detect DFF can be extended to raters, where interactions between rater severity and construct-irrelevant characteristics of test takers suggest the presence of differential rater functioning (Engelhard, 2008). Further investigation using qualitative methods is needed in order to establish evidence for rater bias that may compromise the interpretation of ratings for certain test takers.

Thinking Holistically About Validity, Reliability, and Fairness in Music Assessment

Data-driven educational reform plans have required all stakeholders to standardize, centralize, and provide test-based accountability procedures in our school systems. As a result, assessment as a means for academic (i.e., teaching- and learning-based) improvement and assessment as a means for accountability (i.e., policy- and program-based) evidence have become increasingly intertwined for music educators (Wesolowski, 2014). Decision-making processes underscore assessment practices in both the accountability and academic assessment paradigms. As demonstrated in Table 20.1, decisions based on music assessment results can be made at local, district, state, and national levels by a variety of individuals, including music teachers, music specialists, and administrators.

Table 20.1 Summary of Music Assessment Decisions

Type of decision	Who makes the decision?	Type of assessment used to make the decision	Example
Instructional Decision	Music Educator	Educator -constructed	Music educator uses the assessment results to determine the pace, content, and focus of their class.
Grading Decision	Music Educator	Educator -constructed	Music educator uses the assessment results to assign a grade in their class.
Diagnostic Decision	Music Educator	Educator -constructed	Music educator uses the assessment results to understand student's strengths and weaknesses.
Placement Decisions	Music Educator	Educator -constructed or Standardized	Music educator or music specialist uses the assessment results to place students in chairs or ensembles within the school program or outside of the school program (e.g., district-, state-, or national-level opportunities).
Selection Decisions	Music Educator	Standardized	Music educator uses the assessment results to select students for program-, or institutional-level admissions decisions.
Guidance Decisions	Music Educator	Standardized	Music educator uses the assessment results to guide students toward specific majors and careers that best match the student's achievement level.
Program and Curriculum Decisions	Music Educator or Administrator	Standardized	Music educator or administrator use assessment results to determine success of a program or curriculum and to determine whether a program should be implemented or eliminated.
Administrative Policy Decisions	Administrator	Standardized	Administrator use assessment results to determine how money should be spent and what programs should receive money.

Source: Adapted from Thorndike, Cunningham, Thorndike, and Hagan (1991).

Each of these individuals is a part of the assessment process and therefore can be considered a test user, as they can be involved in the development, administration, scoring, interpretation, and/or application of the assessment measure. As a result, each individual needs to be trained and informed on the appropriate use and application of such assessments. If those involved in the assessment process are not properly trained, unintended consequences to test takers can be substantial.

Kane (2013) (as cited in Caines, Bridglall, & Chatterji, 2014) provides three perspectives on assessment that should be considered by all test users in order to support the validity of measures, reliability of scoring, and fairness of the testing context to all test takers:

> A *measurement perspective*, reflecting the typical mindset of test developers and measurement researchers in standardized testing programs, describing how this group tends to approach the tasks of test development and validation;
>
> A *contest perspective*, reflecting the typical mindset of test takers, who view passing a high stakes test as being similar to winning a contest to help them reach their goals.
>
> A *pragmatic perspective*, or the mindset of typical decision makers, who count on a test to be cost-efficient, objective, fair and dependable for taking necessary actions from an institutional perspective, without adverse or untoward repercussions. (p. 9)

The *measurement perspective* concerns itself with the interpretation, precision, and accuracy of the assessment outcomes. In music assessment contexts, it is of great concern that assessments estimate the "true" measure of the individual student, ensemble, or program's achievement as consistently and reliably as possible. The *contest perspective* concerns itself with a student, ensemble, or program receiving the highest possible outcome of an assessment. In music assessment contexts, it is of great concern that the assessment is fair in providing student, ensemble, or program the same opportunity to succeed. The *pragmatic perspective* concerns itself with the concrete and public perception of validity, reliability, and fairness from a policy-driven perspective. In music assessment contexts, inferences drawn from music assessment contexts can provide grounds for important decisions regarding program longevity and related funding. The current data-driven educational climate has left music students, music educators, music specialists, and administrators in need of valid, reliable, and fair assessment measures. Most importantly, the various contexts in which the assessment of musical experiences is conducted must meet the demands of all three of Kane's (2013) perspectives in order for the measurement and evaluation processes to be considered effective.

Final Thoughts

The psychometric theories of validity, reliability, and fairness discussed in this chapter are most often considered in the context of high-stakes, large-scale assessment. So what do the considerations brought forth in this chapter mean for the music classroom and

music educator? The most clear and direct information students, teachers, and parents receive about achievement in the music classroom come from classroom assessments. The way in which classroom assessments are conducted differs from the way large-scale assessments are conducted. If the psychometric principles of validity, reliability, and fairness and the related development of development and implementation of classical test theory (CTT; see Volume 1, Chapter 21, in this handbook) and item response theory (see IRT, Volume 1, Chapter 22, in this handbook) are rooted in this tradition, how can a connection be made? We propose that in the context of classroom assessments, rather than considering the individual applications of validity, reliability, and fairness in terms of psychometric theory, these concepts should be considered holistically in order to develop and implement a "high quality music classroom assessment system." By considering all types of validity evidence in the assessment building process, a music educator may gain more accurate insight into the congruency between the intentions of the assessment and what is actually being measured. Furthermore, ensuring that there is a clear and open communication between teacher and student in terms of expectations set forth by the measurement instrument and student awareness of how their performance aligns with the teacher's expectations is foundational for the validity of the assessment process. Reliability within the assessment system can be achieved by ensuring consistency in the assessment process. More specifically, regularly evaluating student performances and communicating the level of performance achievement for all students can provide stability, and more importantly, reliability, in the assessment process. Lastly, providing multiple opportunities for students to demonstrate their knowledge, skills, and abilities across different assessment-types can provide more equitability of student assessment.

Validity, reliability, and fairness are all complex issues engrained within the music assessment paradigm. As conversations continue with regard to these constructs in the context of music assessment, it is important that all test users deeply consider the philosophical aspects of music assessment and continue to refine their technical knowledge and abilities related to measurement and evaluation.

Note

1. Additional details about scale linking and equating within the context of CTT are provided by Kolen and Brennan (2014), and within the context of IRT are provided by von Davier (2011).

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

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: AERA.

American Psychological Association (APA), American Educational Research Association (AERA), & National Council on Measurement in Education (NCME) (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological 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.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H.nI. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Asmus, E. P. (2010). Assuring the validity of teacher-made assessments. In T. S. Brophy (Ed.), *The practice of assessment in music education: Frameworks, models, and designs* (pp. 131–144). Chicago: GIA Publications.

Asmus, E. P., & Radocy, R. E. (1992). Quantitative analysis. In R. Colwell (Ed.), *Handbook of research on music teaching and learning* (pp. 141–183). New York, NY: Schirmer Books.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295–317. doi: 10.1111/j.1745-3984.2001.tb01129.x

Brennan, R. L. (2010). *Generalizability theory*. New York, NY: Springer-Verlag.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practices, 22*(4), 5–12.

Boyle, J. D., & Radocy, E. E. (1986). *Measurement and evaluation of musical experiences*. New York: Schirmer Books.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322. doi: 10.1111/j.2044-8295.1910.tb00207.x

Caines, J., Bridglall, B. L., & Chatterji, M. (2014). Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education, 22*(1), 5–18.

Cook, L. L., & Eignor, D. L. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37–45.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi: 10.1007/BF02310555

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: John Wiley.

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

Cureton, E. E. (1950). Validity. In E. F. Lingquist (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Engelhard, G. (2008). Differential rater functioning. *Rasch Measurement Transactions, 21*, 1124.
Engelhard, G. (2013). *Invariant measurement: Using Rasch Models in the social, behavioral, and health sciences*. New York, NY: Routledge.
Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.
Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.
Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123*, 207–215. doi: 10.1080/00221309.1996.9921273
Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*, 736–753. doi: 10.1177/0013164498058005002
Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342. doi: 10.1111/j.1745-3984.2001.tb01130.x
Kane, M. (2013). Validity and fairness in the testing of individuals. In Chatterji, M. (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity*. Bingley, UK: Emerald Group.
Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.
Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160. doi: 10.1007/BF02288391
Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 23–30. doi: 10.1111/j.1745-3992.2002.tb00082.x
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
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.
Messick, S. (1996a). Standards-based score interpretation: Establishing valid grounds for valid inferences. *Proceedings of the joint conference on standard setting for large scale assessments*, sponsored by National Assessment Governing Board and The National Center for Education Statistics. Washington, DC: Government Printing Office.
Messick, S. (1996b). Validity of performance assessment. In Philips, G. (1996). *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Educational Statistics.
Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.
Murphy, K. R., & Cleveland, J. M. (1991). *Performance appraisal: An organizational perspective*. Boston, MA: Allyn and Bacon.
Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.
Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge*. New York, NY: Basic Books.
Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413–428. doi: 10.1037/0033-2909.88.2.413
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295. doi: 10.1111/j.2044-8295.1910.tb00206.x

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417–426. doi: 10.1111/j.2044-8295.1913.tb00072.x

Thorndike, R. M., Cunningham, G., Thorndike, R. L., & Hagen, G. (1991). *Measurement and evaluation in psychology and education*. New York, NY: Macmillan.

Vernon, P. E. (1960). *Intelligence and attainment tests*. London, UK: University of London Press.

von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking*. New York, NY: Springer.

Wesolowski, B. C. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal, 101*(1), 77–85.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae, 19*, 147–170.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016a). Rater analyses in music performance assessment: Application of the Many Facet Rasch Model. In *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education* (pp. 335–356). Chicago, IL: GIA Publications.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016b). Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted Rasch partial credit model. *Music Perception, 5*, 662–678.

Wolfe, E. W. (2004). Equating and item banking with the Rasch model. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.