A Crisis of Measurement in Music: Is a Revolution Needed for Improved Inferences in Latent Construct Development?

Brian C. Wesolowski, University of Georgia, USA

Abstract

The purpose of this paper is to discuss the state of measurement in music research, and the need to implement a new paradigm for the improved treatment of data used to measure latent constructs of musical experiences. From a philosophical and conceptual perspective, this paper discusses how the use of additive conjoint measurement models based on the requirements of invariant measurement bare similar properties to scientific measurement, therefore properly preparing observed data for the application of statistical analyses. The use of models containing such properties can improve the validity of inferences related to musical experiences by providing additive, conjoint, sample- and scale-free, interval-scale measures with equal, meaningful units of measurement. Specifically, this paper discusses the epistemology of psychological measurement and its related problems, problems with measurement in music research, foundations of fundamental measurement in psychometrics using additive conjoint measurement models, criteria for fundamental measurement and related properties of invariance, and the calling for a paradigm shift in the measurement of musical experiences. Implications for the field of music and related psychological research are also discussed.

Introduction

Scientific measurement can be defined as "the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute" (Michell, 1997, p. 358). Scientific measurement, then, requires two essential functions (Michell, 1997). The first function is an instrumental function, which includes the systematic development of a set of operations to identify measurable characteristics (i.e., magnitudes) of a single attribute (i.e., quantity). In the context of measurement, magnitudes are expressed as consequential, genuine numbers through the use of a set of meaningful, equal units of measurement in relation to one another (i.e., ratios). The second function is scientific/epistemic, whereby the attribute of interest is derived specifically from a quantitative structure. The focus of true scientific measurement is on the

discovery of concrete, physical properties through the specific and intentional invention of operations. In order for observed data to properly form the foundation toward the deduction of inferences and knowledge claims through mathematical and statistical analyses, the discovered properties are required to be multiplicative; specifically, they must contain true zeros with clearly defined and equal units of measurement (Kline, 2000).

In music and related behavioral sciences, however, psychological measurement is used as a method for inference. One fundamental distinction of psychological measurement from scientific measurement is that it is concerned with measuring abstract, latent properties that cannot be physically demonstrated. Arguably, latent measurement lacks a true zero and does not establish clearly defined equal units of measurement (Michell, 1999). Thorndike (1904) notes:

> If one attempts to measure even so simple a thing as spelling, one is hampered by the fact that there exist no units in which to measure. One may arbitrarily make up a list of words and observe ability by the number spelled correctly. But if one examines such a list one is struck by the inequality of the units. All results based on the equality of any one word with any other are necessarily inaccurate. (p. 7)

As a physicist and philosopher of science, Campbell (1928) claimed that psychometric measurement is too inconsistent to form the basis of true, fundamental measurement as demonstrated in scientific measurement. More current criticisms against the strength of psychometric measurement is that where all scientific measurement contains multiplicative properties, psychometric operations, at best, are based upon additive properties where equality of the units is often assumed based upon a set of pragmatic procedures that conveniently discriminate individual differences (Michell, 1990, 1997). Most notable of the models subject to this criticism are those of the true score test theory variety. According to Lazarsfeld and Barton (1951):

> The idea that 'social science must develop measurements' has sometimes led to misunderstandings. Some optimists want to start measuring social phenomena immediately with all the precision of the most advanced sectors of physical science; some pessimists deny that man and his works can ever be measured at all, and recommend an entirely intuitive approach to the understanding of society. The false assumption underlying both positions is that science can be carried on only with one particular kind of device—the quantitative scale with equal intervals and a zero point—and that aside from this device there is nothing but a chaos of guesswork and intuition. (p. 155)

The purpose of this paper is to discuss the state of measurement in music and the need to implement a new paradigm for the improved treatment of data that measures latent constructs in music. It is not the intent of this paper to argue against the use of various statistical methods often employed in music research such as general linear modeling, factor analysis, structural equation modeling, or hierarchical linear modeling, for example. These methods have benefits and specific purposes beneficial to the research process. This paper, from a philosophical and conceptual perspective, discusses how the use of additive conjoint measurement models based on the requirements of invariant measurement bare similar properties to scientific measurement. Therefore, the applications of conjoint measurement models to raw score, observed data properly prepare observed music data for the application of statistical analyses. The use of models containing such properties can improve the validity of inferences related to musical experiences by providing additive, conjoint, sample- and scale-free, interval-scale measures with equal, meaningful units of measurement. Specifically, this paper will discuss the epistemology of psychological measurement and its related problems, problems with measurement in music research, foundations of fundamental measurement in psychometrics using additive conjoint measurement models, criteria for fundamental measurement and related properties of invariance, and the calling for a paradigm shift in the measurement of musical experiences. Implications for the field of music education and psychological research are discussed.

Epistemology and Problems of Psychological Measurement

The epistemology of measurement refers to the study of knowledge deduced from the relationship between the inferences that tie together raw score observations and the latent construct being defined. In order to deduce a knowledge claim based upon raw score data, dual processes need to occur: (a) the use of a valid, reliable, and fair measurement instrument that allows for the collection of concrete and observable data; and (b) an abstract inference as to what latent construct can be defined by the collected observations. According to Mari (2003):

> Measurement is a specific kind of evaluation, i.e., it is an operation aimed at associating an information entity, the result of measurement, with the state of the system under measurement in reference to a given quantity, the measureand. (pp. 17-18)

Historically, much debate exists over the nature of how inferences are drawn from observed data in psychological measurement. Between 1932 to 1940, a committee formed by the British Association for the Advancement of Science debated the nature of measurement in science and psychology. According to its

final report, the committee noted, "… any law purporting to express a quantitative relation is not merely false but is in fact meaningless unless and until a meaning can be given to the concept of addition…" (Ferguson, Myers, & Bartlett, 1940, p. 245).

The committee was hopeful, however, for another alternative:

> Some members, perhaps all, admit that their opinion might change if new facts were established; but the facts that would be necessary for this purpose are not of the kind that can be established by any experimental method at present in general use. (Ferguson, Myers, & Bartlett, 1940, p. 14)

Stevens (1946), identifying a disparity between the use of measurement in the physical sciences and the use of measurement in the behavioral sciences, then adopted an operational definition of representation for psychological measurement. He stated:

> Perhaps agreement can better be achieved if we recognize that measurement exists in a variety of forms and that scales of measurement fall into certain definite classes. These classes are determined both by the empirical operations invoked in the process of "measuring" and by the formal (mathematical) properties of the scales. Furthermore-and this is of great concern to several of the sciences-the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered. (Stevens, 1946, p. 677)

At the time, Stevens' representational approach to measurement was forward thinking and distinctive in its shift of focus from the development of scientific operations to a systematic classification of operations. Stevens (1946) defined measurement as "the assignment of numerals to events or objects according to rule" (p. 677). This now famous hierarchical taxonomy of operational classifications consists of four broad categories: nominal, ordinal, interval, and ratio. This definition is the most widely adopted definition of measurement in the behavioral sciences (Michell, 1997). However, in modern psychometrics, Steven's definition of measurement and descriptions of operational categories has led him to be considered the proverbial "whipping boy," (Bond & Fox, 2007, p. 2) as it often sanctions mathematical mistreatment, particularly due to his underlying contention that particular types of statistical analysis were permitted based upon the level of measurement. One significant weakness of Steven's writing, "was its apparent implication that the nature of a scale is somehow defined by the investigator" (Cliff, 1992, p. 186). As a result, Wright (1997c) argues that although this definition is heavily cited and taught, the cause of such misuse and misrepresentation is due to the that fact that it is seldom read critically, yielding a blatant misunderstanding. This implication has

often led to the misappropriation or even pure abandonment of concepts in measurement in the behavioral sciences. Measurement in music education is often subject to such misappropriation.

The Calling for a Kuhnian Revolution in the Measurement of Musical Experiences

The positivist/Enlightenment perspective of normal science is one of linearity, where the accumulation of knowledge toward truth is a persistent and gradual process. However, Kuhn (1963) argues that science is non-cumulative and moves in cycles of paradigms, a term he defines as "universally recognized scientific achievements that for a time provide model problems and solutions to a community of scholars and practitioners" (p. viii). The measurement of musical experiences in music is most often shaped by the true score test theory model. According to Kuhn, paradigms govern normal science through a framework of assumptions that are driven by the paradigm of question, having a direct influence on the questions being asked, methodology, and interpretation of results. Furthermore, the theory of normal science and its assumptions are not even questioned. He states:

> Though they may begin to lose faith and then to consider alternatives, they do not renounce the paradigm that has led them into crisis. They do not, that is, treat anomalies as counterinstances, though in the vocabulary of philosophy of science that is what they are. The decision to reject one paradigm is always simultaneously the decision to accept another, and the judgment leading to that decision involves the comparison of both paradigms with nature *and* with each other. (pp. 76, 79)

As normal science progresses, anomalies can develop that can constrain the progress of scientific endeavors. When anomalies cannot be explained they begin to compound, their combined strength pushes science out of normal science and into crisis science. In crisis science, current paradigms are not thrown out, but brought to light as being inadequate. As conceptually argued in this paper, current paradigms of measurement in music are limited by weak theoretical assumptions and lack of properties related to fundamental measurement and invariance. When new paradigms prove old paradigms inadequate, a paradigm shift occurs. A paradigm shift involves a revolution, whereby scientific ideas and methodology change. A paradigm can encounter a crisis that calls for change due to new problems. Has music education faced a crisis in need of a paradigm revolution?

The Crisis of Measurement in Music

Based upon Kuhn's model, anomalies constrain the development of scientific endeavors. Here, I argue three important concepts in measurement in the field of music education that are constraining the development of science and blurring the validity of measures: (a) use of true score test theory; (b) circular dependency of item discrimination and person ability; and (c) use of raw scores and the false assumption of linearity.

Use of True Score Test Theory. In music education and related psychological research, inference of psychological responses is most often deduced upon the classical true-score test theory (CTT) measurement model and its related extensions (i.e., generalizability theory, factor analysis, structural equation modeling, and hierarchical linear modeling). These models stretch across all areas of music research. CTT measurement models are based upon two weak theoretical assumptions: (a) an observed score obtained from a measure is comprised of a true score and error; and (b) random errors are normally distributed and uncorrelated to each other or the true score. Test responses and item statistics grounded in CTT can meet these assumptions rather easily. As Heene (2013) argues, models with easily met assumptions are preferred in the behavioral sciences to those with strict sets of requirements in order to avoid falsifiability and inconvenient truths. Michell (1990) notes:

> In general psychologists have found refuge in quantitative methods that, because they assume more, demand less foundational research as the basis for their application. Methods that always yield a scaling solution, like the method of summated ratings, are almost universally preferred to methods which ... do not produce a scaling solution when they are falsified by the data. Surprisingly, vulnerability to falsification is commonly deemed by psychologists to be a fault rather than a virtue. (p. 130)

Circular Dependency. A striking limitation of the CTT framework is circular dependency: parameters that characterize person ability (e.g., true scores) are dependent on the items; and (b) the parameters that characterize the items (e.g., item discrimination and item difficulty) are dependent on the sample (Fan & Sun, 2013). Under the CTT framework, item discrimination, or the ability of an item to discriminate between examinees of varying ability levels, is indicated statistically in dichotomous scoring using Pearson product-moment correlation coefficients. In the case of polytomously scored items such as Likert rating scales, for example, adjusted proportion-correct values (*p*-values) and correlation coefficients are used in order to indicate an index item difficulty and overall ability level of an examinee. Indices of item difficulty are sample dependent and calculated using a point biserial correlation coefficient based upon the total success rate of the pool of examinees. These relationships create an inherent circular dependency, which limits the ability to develop valid and reliable measures and to make informed inferences related to person ability, item

difficulty, and item discrimination that extend beyond the context of a single assessment situation. Because of these limitations, any changes in the sample or scale affect the unit of the measure. According to Thurstone's (1928) crucial experimental test:

> One crucial experimental test must be applied to our method of measuring attitudes before it can be accepted as valid. A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement. The scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. This may tum out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale. (p. 228)

Raw Scores and False Assumption of Linearity. In the field of music, raw scores are most often used for drawing inferences under the CTT paradigm. In instances where raw score data are used, linear magnitudes are all too often assumed, laying the "foundations for misinference" (Merbitz, Morris, & Grip, 1989). According to Fraenkel and Wallen (2009), "Researchers rarely know for certain whether their data justify the assumption that interval scales have actually been used" (p. 229).

Raw scores are in fact ordinal counts of observations that indicate "more" or "less" (Wright & Masters, 1982). Bond and Fox (2007) state:

> In terms of Stevens' (1946) levels… nominal and ordinal levels are NOT any form of measurement in and of themselves. Admittedly, we concur that his interval and ratio levels actually would constitute genuine measurement, but the scales to which we routinely ascribe that measurement status in the human sciences are merely *presumed* to have measurement properties; those measurement properties are almost never tested empirically. It is not good enough to allocate numbers to human behaviours and then, merely to *assert* that this is measurement in the social sciences. (p. 4)

> According to Wright and Linacre (1989):

> They cannot be measures because a measure implies the previous construction and maintenance of a calibrated measuring system with a well-defined origin and unit which has been shown to work well enough to be useful. (p. 2)

The problem with ordinal-level scales is that they:

> … fail to indicate the absolute position of subjects on the variable and, in addition, there is no indication of how far apart subjects are from those next to them. These obvious flaws of measurement make it obvious that there is little quantitative information that can be obtained from ordinal scales. Ordinal scales are regarded as primitive or lower forms of measurement than other scales. (Kline, 2000, p. 45)

This lack of indication evokes a raw score bias in the magnitudes of person ability, because magnitudes of the units in a monotonically increasing ogive, increments of one additional correct answer under probabilistic conditions, change based upon the person's relative location on the latent construct. As (Wright, 1997a) notes:

> Any statistical method like linear regression, analysis of variance, generalizability, or factor analysis that uses raw scores or Likert scales as though they were linear measures will have its output hopelessly distorted by this bias. That is why so much social "science" has turned out to be no more than transient description of never to be reencountered situations, easy to contradict with almost any replication (p. 34).

Furthermore, "All results from raw score analyses are spoiled by their non-linearity, their extreme score bias and their sample dependence" (Wright, 1997a, p. 40). Wright (1992) points out several additional limitations of raw scores compared to model-controlled linearization (see Table 1). Wright and Masters (1982) state:

> For observations to be combined into measures they must be brought together and connected to the idea of measurement which they are intended to imply. The recipe for bringing them together is a mathematical formulation or measurement model in which observations and our ideas about the relative strengths of persons and items are connected to one another. (p. 4)

Wright (1999) laments the fact that:

> … this simple point is sometimes belittled. It is not negotiable. It is vital and decisive! We will never build a useful, let along moral, social science until we stop deluding ourselves by analyzing local concrete ordinal raw scores as though they were general abstract linear measures. (p. 100)

Table 1. Comparing the Use of Raw Scores and Linear Measures (based upon Wright, 1992)

Standing	**Raw Scores**	**Linear Measures**
Psychometrics	Classical Test Theory CTT	Rasch Model
Data	must be "complete" ordinal	robust against missing data and non-administered items
Meaning	ordinal ranking on hoped-for latent variable	linear positioning on latent variable explicitly-defined by item content
Status	mistaken for truth	known as estimates
Additivity	non-linear, bent	linear, straight
Continuity	discrete, lumpy	continuous, smooth
Precision	unknown, except on average	quantified by standard errors
Accuracy	unknown	quantified by local fit statistics
Linear Analysis	unsuited to usual statistics	ideal for usual statistics
Validity	sample-dependent test "reliability"	item-dependent construct validity
Diagnosis	sample-dependent item point biserial	individual item and person fit statistics
Conceptualization	concrete	abstract
Occurrence	accidental	deliberate
Construction	irreproducible	reproducible

Cliff (1992) states that "Stevens' enormous contribution was his successful argument that there are different kinds of scales, defined in terms of their degree of resemblance to the real number line." However, as Torgerson (1958) notes, a weakness was that Stevens was concerned with "the systematic classification of various limited sets of [concrete] objects, rather than methods of measurement of [an abstract] property" (p. 18).

Fundamental Measurement and the Use of Additive Conjoint Models for Inference of Latent Constructs

Fundamental Measurement. Fundamental measurement is a theoretical property that can be defined by two salient principles: (a) measurement which is not derived from other measurements; and (b) measurement which is produced by an additive (or equivalent) measurement operation (Wright, 1997a). Fundamental measurement is built upon three axiomatic theorems of natural concatenation (See Campbell, 1920, 1928; Krantz, Luce, Suppes, & Tversky, 2006 for theorems and formal proofs). One elementary interpretation of fundamental measurement includes a physical principle deduced from a physical interpretation. Examples of this include piling bricks to concatenate weight or the joining of rigid straight rods end-to-end to concatenate length. A second interpretation is a principle demonstrating associative and additive representation (Campbell, 1957). An example includes the orthogonal concatenation for length measurement, where the length of the hypotenuse of a right triangle is deduced by applying the algebraic rules of right angles with knowledge of the lengths of the other two sides. Physical concatenation, however, is not possible in latent construct measurement. Luce and Tukey (1964) note that "when no natural concatenation operation exists, one should try to discover a way to measure factors and responses such that the 'effects' of different factors are additive" (p. 4).

In latent construct measurement, according to Bond and Fox (2007), additive representation comes in the form of abstraction of equal and meaningful units. They contend: "Abstractions of equal units must be created and calibrated over sufficiently large samples so we are confident in their utility. Then these abstractions can be used to measure attributes of our human subjects'' (p. 3). According to Wright (1997a) in order to infer meaningful data from abstraction, a two-step process needs to occur. First is the identification of the stochastic process by which sufficient approximations based upon inverse probability can be defined. A stochastic process is one where item difficulty and person ability are separable but conjointly considered. This is a direct contrast of the CTT framework, where the measure of a person's ability is assumed to be equal to a person's observed score. Sufficiency, according to (Wright, 1997a), is the estimation requirement that replaces a parameter in a model with a sufficient statistic, forming the basis for fundamental measurement. Through the use of sufficient statistics, as discovered by Fisher (1920), independent parameter estimation of the model is allowed. According to (Wright, 1997a):

> When a psychometric model employs parameters for which there are no sufficient statistics, that model cannot construct useful measurement because it cannot estimate its parameters independently of one another… By 1960 Rasch had proven that formulations in the compound Poisson

> family, such as Bernoulli's binomial, were both sufficient and, more surprising, necessary for the construction of stable measurement. (p. 36)

Inverse probability is the mechanism by which inference is drawn:

> Uncertainty is the motivation for inference. The future is uncertain by definition. We have only the past by which to foresee. Our solution is to capture uncertainty in a construction of imaginary probability distributions which regularize the irregularities that disrupt connections between what seems certain now but is uncertain later. The solution to uncertainty is Bernoulli's inverse probability. (Wright, 1997a, p. 38)

Second is to discover what mathematical models can govern the stochastic process in a way that enables a stable, ambiguity resilient estimation of the model's parameters from the limited data in hand (e.g., the Rasch measurement model). According to Wright (1997a):

> Rasch had found that the 'multiplicative Poisson' was the only mathematical solution to the second step in inference, the formulation of an objective, sample [free] and test free measurement model… This is solved by formulating the mathematical function which governs the inferential stochastic process so that its parameters are either infinitely divisible or conjointly additive i.e. separable. (p. 37)

> The ability of the Rasch Model to compare persons and items directly means that when adequate model data fit occurs, person-free measures and item-free calibrations are possible, as what is expected in the physical sciences (Bond & Fox, 2007).

Additivity and Conjoint Measurement. Additive conjoint measurement models are an algebraic measurement theory that allows for the monotonic transformation of ordinal, raw scores into interval-leveled data from which an additive representation can be constructed. Specifically, it provides a quantitative (i.e, additive and stochastic) and representative structure (i.e., abstraction) in instances when operations cannot be deduced from a physical concatenation. As (Heene, 2013) notes:

> By avoiding tests of the assumption of a quantitative structure of psychological attributes, psychologists have yet failed to make progress on the basis of the fundamental scientific principle of falsification and in regard to their most fundamental assumptions of quantitative psychological attributes. (p. 3)

Additive conjoint measurement models are concerned with the ordering of dependent variables (i.e., score) as a function of the joint effects of two or more independent variables. In instances of educational and psychological testing, the probability of a person answering a question correctly is a function of that person's ability and the item difficulty. Such models work under the requirement that interactions between independent factors and ordinal outcome variables (i.e, scores) be removed through a monotonic transformation of the outcome variable. When the set of conjoint axioms as set forth by Luce and Tukey (1964) hold true, concomitantly constructed independent variables (i.e., item difficulty and person ability) and transformed dependent variables (i.e., person response) are conjointly represented on a common, equal unit scale.

Wright and Mok (2004) note five important requirements for constructing inference from observation: (a) produce linear measures, (b) overcome missing data, (c) give estimates of precision, (d) have devices for detecting misfit, and (e) the parameters of the object being measured and of the measurement instrument must be separable (p. 4). The Rasch measurement model is a practical realization of additive conjoint measurement with an underlying stochastic structure (Perline, Wright, & Wainer, 1979) and the only family of measurement models to make proper inference under these requirements.

Rasch Measurement: Ideal-Type Models and Invariance

The theory behind the Rasch measurement model is one of fundamental measurement, ideal model-type, and invariance. The work of Campbell's (1920) requirement of concatenation for fundamental measurement, Fisher's (1920) sufficiency statistic, Thurstone's (1925) absolute scaling and Law of Comparative Judgment (1927), Guttman's (1950) joint ordering, Kolmogorov's (1950) divisibility of independent parameter estimates, and Luce and Tukey's (1964) conjoint additivity has contributed to what Engelhard (2013) has aptly called the "quest for invariance" (See Engelhard, 2013 and Wright, 1997a for historical and theoretical details of measurement lineage). Each of these principles is embodied in the Rasch family of measurement models.

Rasch measurement theory (Rasch, 1960/1980) is often preferred in scale development as well as the measurement of latent traits in the behavioral, social, and health sciences for the reason that it is an ideal-model type (Engelhard, 2013). An ideal-type model is one that provides a strong theoretical basis based upon an a priori set of requirements. Utilizing ideal-type models with strict requirements provides a framework that can offer detailed, diagnostic information on whether the collected data fit the model. This is a philosophical shift in thought and approach for music research. The field's current paradigm is one of a statistical perspective, where the model fits the data based upon *assumptions*. Under this traditional paradigm, the argument for choosing one model over another is that "it accounts better for the data" (Andrich, 2004, p. 8).

However, the Rasch perspective is one where the data must fit the model based upon sets of *requirements*. According to Andrich (1989):

> The view that the model should be fitted to the data, rather than the other way around, where the model is not chosen capriciously, has profound consequences for the psychometric research agenda. In the traditional approach, the agenda is to search for models that best account for the data. That tends to be carried out by statisticians… [The Rasch perspective of measurement] leads to a search for qualitative understandings of why some responses *do not* [emphasis added] accord with the model. That task needs to be carried out by researchers who understand the substance of the variables. (p. 15).

This implies a shift from a nomothetic approach to an ideographic approach of data analysis and interpretation, where the focus is on the individual items, persons, or even raters that do *not* fit the model. Outliers and misfits are not ignored or thrown out of the model, but further investigated and given attention for a substantive story or meaning in their behavior.

Engelhard and Perkins (2011) provided a set of five requirements that can be used to determine the degree to which invariant measurement is obtained for persons and items. These requirements include (a) item-invariant measurement of persons where the measurement of persons must be independent of the particular items that happen to be used for the measurement; (b) non-crossing person response functions (i.e., a more able person must always have a better chance of success on an item than a less able person); (c) person-invariant calibration of test items (i.e., the calibration of the items must be independent of the particular persons used for calibration); (d) non-crossing item response functions (i.e., any person must have a better chance of success on an easy item than on a more difficult item); and (e) variable map (i.e., items and person must be simultaneously located on a single underlying latent variable).

The major benefit of the Rasch model is that, when adequate fit to the model is observed, invariant measurement is achieved. In the context of assessments, invariant measurement implies that the measurement of persons is not influenced by the particular items that they happen to take, and the measurement of items is not influenced by the particular persons by whom they are measured. The property of invariant measurement that characterizes the Rasch model must be evaluated in empirical data. Invariant measurement is a hypothesis that must be confirmed or disconfirmed by evidence in a data set (Engelhard, 1994). Based upon the difference in item and person locations on the variable map, items can be evaluated for their usefulness in providing information about persons' varying achievement levels.

Rasch models use probabilistic distributions of responses as a logistic function of person and item parameters in order to define a latent trait. In

contrast to CTT where raw scores are directly used in the analyses, Rasch measurement theory converts raw scores to a log-odds scale using a logistic transformation. The transformed test score data can then be conceptualized as a dependent variable with multiple independent variables (i.e., facets) of interest, including measures of, item difficulty, task difficulty, person ability, rater severity and leniency, or any other facets of interest. Hierarchies of difficulty for each relevant item, and each examinee's discrete item responses are mapped onto a single logit (log-odds units) scale. As a result of the mapping of facets onto a single continuous latent variable scale, it is possible to construct a variable map to use as a visual display for illustrating relative differences in locations among facets.

The benefit of Rasch approaches to measurement and construct modeling is the strong requirement that a set of items being used can measure a single, unidimensional construct or latent trait. Thurstone (1931) notes that unidimensionality is a universal characteristic of all measurement. Lazarsfeld and Barton (1951) state, "…we must decide what attributes of the concrete items we wish to observe and measure: do we want to study 'this-ness' or 'that-ness' or some other '-ness'" (p. 155). In instances when thinking in multidimensional scaling occurs, notes that confusion can caused by interdependencies between items and/or persons. Wright (1997a) states:

> The method we use to control confusion is to enforce our ideas of unidimensionality. We define and measure one invented dimension at a time. The necessary mathematics is parameter separability. Models which introduce putative "causes" as separately estimable parameters are our laws of quantification. These models define measurement, determine what is measurable, decide which data are useful and expose data which are not. (p. 38)

As summarized by (Cavanagh, 2009), Wright and Masters (1981) identified seven criteria for evidence of fundamental measurement:

1. Each item should be evaluated to see whether it functions as intended;
2. the relative position (difficulty) of each valid item along the scale that is the same for all persons should be estimated;
3. Each person's responses should be evaluated to check that they form a valid response pattern;
4. Each person's relative score (attitude or achievement) on the scale should be estimated;
5. The person scores and the item scores must fit together on a common scale defined by the items and they must share a constant interval from one end of the scale to the other so that their numerical values mark off the scale in a linear way;

6. The numerical values should be accompanied by standard errors which indicate the precision of the measurements on the scale; and
7. The items should remain similar in their function and meaning from person to person and group to group so that they are seen as stable and useful measures.

Conclusion

The purpose of this paper was to discuss the state of measurement in music and the need to implement a new paradigm for the improved treatment of data proposing to measure latent constructs in music. This paper serves as a call for improved standards of good science in music research. With improved standards comes improved inference. The suggested paradigm shift of implementing Rasch measurement in the field of music includes the shift in perspective from the traditional ideology of descriptive models for which the model fit the data to the use of prescriptive models for which the data fit the model. Embedded in this paradigm shift is a new idiographic approach to knowledge with the intent of garnering a detailed understanding of items and persons that do not fit the model, versus the traditional nomothetic approach to knowledge with the intent of generalizing items and persons that do fit the model. Rasch (1960/1980) himself understood the implications of such a paradigm shift:

> It is tempting, therefore, in the case with deviations of one sort or other to ask whether it is the model or the test that has gone wrong. In one sense, this of course turns the question upside down, but in another sense the question is meaningful. For one thing, it is not easy to believe that several cases of accordance between model and observations should be isolated occurrences. Furthermore, the applicability of the model must have something to do with the construction of the test; at least, if a pair of tests showed results in accordance with our theory, this relationship could easily be destroyed by adding alien items to the tests. Anyhow, it may be worthwhile to know of conditions for the applicability of such relatively simple principles for evaluating test results. (p. 51)

Kuhn enlightens us to such an incommensurability of paradigms. Paradigms cannot be logically chosen between due to the way the data is conceived. Accepting a new paradigm requires a kind of faith, based upon critical examinations and value judgments of functionality and problem-solving power.

This paper has brought to light limitations of the current framework for inference in music education and related music psychology research and the need to adopt a probabilistic framework utilizing additive conjoint models based

on the requirements of invariance. The field of music education research suffers from the problem with theory-laden observations; what scholars in the field see is due in large part to what they have come to expect or believe (Kuhn, 1963). What the field of music psychology observes and interprets is predicated and mediated by our paradigm, creating an inherent resistance to change. The lens through which we, as a field, interpretively construct views of the world is shaped by what we know and have come to expect (Bruner, 1974). The dogmatic disposition of methodology is a direct result of how the field is primed. Consider the status of graduate training in music education and psychology. The professional background of the majority of music education graduate students is practitioner based. More specifically, initial undergraduate training and the career focus of many music educators is on performance-related teaching. In both the fields of music education and psychology, the first academic introduction to theoretical measurement comes for most at the doctoral level. For few others, this introduction may come at the master's level. A traditional one-semester introduction to measurement and evaluation is traditionally a brief introduction to classical test theory, with little or no mention of probabilistic models. In lieu of courses in psychometrics and measurement, graduate students in music are often exposed to if not required to participate in statistics courses mostly focused on a sequence general linear models including analysis of variance and regression. For a rare few, additional electives may include a course in multivariate analysis, structural equation modeling, or time series analysis. However, detailed instruction in measurement, the often missed and necessary step before the application of statistical analysis, is overlooked. Consideration in the field for priming graduate students for higher standards for inference is necessary toward a revolution in measurement; specifically, deeper thought and inquiry toward the general understanding of inference through the teaching of multiple paradigms. Babbie (2014) notes:

> When we recognize that we are operating within a paradigm, two benefits accrue. First, we are better able to understand that seemingly bizarre views and actions of others who are operating from a different paradigm. Second, at times we can profit from stepping outside our paradigm. We can see new ways of seeing and explaining things. We can't do that as long as we mistake our paradigm for reality. (p. 33)

From a social objectivity perspective, scientific inquiry is interest-relative, bound by the consensus of the field and dependent on individual researchers (Weber, 1949). In the pursuit of scientific knowledge, scholars aim to disclose truths that are based upon a commitment to a set of unifying and underscoring standards that are field-dependent. Music's truths and standards seem to be in perpetual motion. However, are they perpetual because they are proven to be explicitly valid, reliable, and accurate or are they perpetual because they have yet

to be disputed? What if music's' "standards" are limited and "truths" are an artifact of convenience? Previous work and efforts in the field should never be discredited; however, questioning convention is essential in the quest for scientific knowledge and truth.

References

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath, & S. H. Lovibond (Eds.), *Mathematical and theoretical systems* (pp. 7–16). North Holland, The Netherlands: Elsevier Science.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(1), I7–I16. doi: 10.1097/01.mlr.0000103528.48582.7c

Babbie, E. (2014). *The basics of social research*. Belmont, CA: Wadsworth.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.

Bruner, J. (1974). *Toward a theory of instruction*. Boston: Harvard University Press.

Campbell, N. R. (1920). *The elements*. London, UK: Cambridge University Press.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London, UK: Longmans, Green & Co.

Campbell, N. R. (1957). *Foundations of science: The philosophy of theory and experiment*. New York: Dover Publications, Inc.

Cavanagh, R. (2009). Measurement issues in the use of rating scale instruments in learning environment research. In *Australian Association for Research in Education AARE*.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*(3), 186–190. doi: 10.1111/j.1467-9280.1992.tb00024.x

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Engelhard, G., & Perkins, A. F. (2011). Person response functions and the definition of units in the social sciences. *Measurement: Interdisciplinary Research & Perspective*, *9*(1), 40–45.

Engelhard Jr., G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. doi: 10.1111/j.1745-3984.1994.tb00436.x

Fan, X., & Sun, S. (2013). Item response theory. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 45–67). Rotterdam, The Netherlands: Sense Publishers.

Ferguson, A., Myers, C. S., & Bartlett, R. J. (1940). *Quantitative estimates of sensory events, final report*.

Fisher, R. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean square error. *Notices of the Royal Astronomical Society*, *53*, 758–770.
Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed). New York: McGraw-Hill.
Guttman, L. (1950). The basis for scalogram analysis. In *Measurement and prediction, volume 4* (pp. 60–90). Princeton, NJ: Princeton University Press.
Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology*, *4*, 1–4.
Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). New York: Routledge.
Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York: Chelsea Publishing.
Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (2006). *Foundations of measurement volume 1: Addititive and polynomial representations*. Mineola, NY: Dover Publications, Inc.
Kuhn, T. S. (1963). *The structure of scientific revolutions*. Chicago: University of of Chicago Press.
Lazarsfeld, P. F., & Barton, A. H. (1951). Qualitative measurement in the social sciences: Classification, typologies, and indices. In D. Lerner & H. D. Lasswell (Eds.), *The policy sciences: Recent developments in scope and methods*. Stanford, CA: Stanford University Press.
Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.
Mari, L. (2003). Epistemology of measurement. *Measurement*, *34*, 17–30.
Merbitz, C., Morris, J., & Grip, J. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*,*70*, 308–312.
Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Erlbaum.
Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. doi:10.1111/j.2044-8295.1997.tb02641.x
Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.
Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237–255. doi: 10.1177/014662167900300213
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. doi: 10.1126/science.103.2684.677

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York: Teacher's College.
Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433–451.
Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273–286.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554.
Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology*, *26*, 249–269.
Torgerson, T. S. (1958). *Theory and methods of scaling*. New York: Wiley & Sons, Inc. Publication.
Weber, M. (1949). *The methodology of the social sciences*. (E. A. Shils & H. A. Finch, Eds.). Glencoe, IL: The Free Press.
Wright, B. D. (1992). Raw acores are not linear measures: Rasch vs. classical test theory CTT comparison. *Rasch Measurement Transactions*, *6*(1), 208.
Wright, B. D. (1997a). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33-45.
Wright, B. D. (1997b). S. S. Stevens revisited. *Rasch Measurement Transactions*, *11*(1), 552–553.
Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857–860.
Wright, B. D., & Masters, G. N. (1981). *The measurement of knowledge and attitude*. University of Chicago, Department of Education.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 1–24). Maple Grove, MN: JAM Press.