Phenomenography: Bringing Together Theory and Practice through the Process of National Standards Development and Measure Construction

Brian C. Wesolowski, University of Georgia, USA
Fredrick Burrack, Kansas State University, USA
Kelly A. Parkes, Teachers College, Columbia University, USA

Abstract

In music teaching and learning, all individuals experience musical phenomena in unique ways as a result of experiences, perceptions, cognitive engagement with, and awareness of musical events. Phenomonography is a research method for mapping the qualitatively different ways in which people experience, conceptualize, perceive, and understand various aspects of, and phenomena in, the world around them. While developing measures for assessment tasks in a music classroom, phenomenographic analysis can be conceptualized as "outcome space," where the validity of measurement instruments is explored through a qualitative, theoretical framework to better understand how teachers use measurement instruments to instruct toward curricular goals and how both teachers and students interact with measurement instruments from a research-based (e.g., construct validity) and practitioner-based (e.g., content validity) perspective. The unique feature of phenomenography is that outcome space processes focus on the variation in differences of responses rather than the similarities of responses. The purpose of this paper is to describe the phenomenographic framework used for the development of the Model Cornerstone Assessments within the context of a national music assessment study in the United States.

Introduction

Marton (1986) defines phenomenography as "a research method for mapping the qualitatively different ways in which people experience, conceptualize, perceive, and understand various aspects of, and phenomena in, the world around them" (p. 31). Wilson (2005) brings to light the application of phenomenography as an important component of the measure construction

process that is often overlooked in the field of education. In considering measure construction processes in the context of student learning and achievement, Wilson conceptualizes phenomenographic analysis as an "outcome space," where the validity of measurement instrument(s) can be explored through a qualitative, theoretical framework to better understand how measures engage student responses and teacher use both from a research-based (e.g., construct validity) and practitioner-based (e.g., content validity) perspective. The unique feature of phenomenography is that outcome space processes focus on the variation in *differences* of responses rather than the *similarities* of responses.

In the context of music teaching and learning, music practitioners and researchers experience musical phenomenon in unique ways, as their interaction with music may be quite different from day-to-day. More broadly, all *individuals* experience musical phenomena in unique ways as a result of experiences, perceptions, cognitive engagement with, and awareness of musical events. In the United States, the recent revision of the National Core Arts Standards in music by the National Coalition for Core Arts Standards (2014) and development of the paralleling Model Cornerstone Assessments (MCAs) (2014-2016) have embodied the phenomenographic approach to measure construction. The development, piloting, scoring, and validation processes underscoring the development of the MCAs specifically implemented an approach that aimed to bridge the gap between the theoretical and practical, qualitative and quantitative. The purpose of this paper is to discuss the framework of phenomenographic processes embedded in the development of the MCAs.

Validity and the Measurement of Musical Processes

The procedure for measuring psychological and behavioral phenomena such as musical outcomes and processes, for example, is inherently subjective (Engelhard, 2013). Therefore, the broad considerations of validity in these contexts, according to Messick (1989), should be thought of as "… and integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriates* of *inferences* and *actions* based on test scores and other modes of assessment" (p. 13). The thoughtful development of a system of valid measurement instruments, then, should take into consideration the mutual understanding between test developers, users, test takers, and the public at-large. In this large-scale study, we propose two types of methodological discourses that are addressed in this paper. First, as described under the *Construct Development and Item Design* and *Scoring Procedures* headings, is a methodological outline of a presumptive argument. Here, we first describe the process of developing the hypothetical, theoretical constructs, and the specific items written to define the intended constructs. Second, we describe data analysis procedures used to provide empirical evidence of the hypothetical constructs. It is important to note that the empirical evidence does not

necessarily prove the hypothetical constructs, rather, they provide presumptive evidence of the hypothetical constructs. In the second discourse, under the heading *Post-Scoring Data*, were describe the types of data collected to form our validity argument. Ultimately, as Kane (2006) notes, "… the need for validation derives from the scientific and social requirement that public claims and decisions can be justified" (p. 17). Here, the use of phenomenological processes to investigate differences in classroom use, teaching contexts, and students' results provide the context of the inferences gleaned from the scoring procedures. The variability in responses offers a varied scope of experiences, linked to Kane's notion of "social requirement" in his description of validity as stated above. The evaluation of the post-scoring data from a phenomenological perspective, therefore, provides contextual evidence of both the plausibility and cohesiveness of the outcome scores and overall assessment process. In this paper, we describe the methodologies used, not the arguments themselves.

Construct Development and Item Design

The Model Cornerstone Assessments (MCAs) are built upon the principle of measuring artistic processes educatively (Wiggins & McTighe, 2005). Because students engage in the artistic process in a variety of ways, the structure of the assessments is designed as a model to enable students to demonstrate individual learning and for teachers to integrate the assessment task as it relates to the unique instructional context and varied forms of musical literacy. A team of researchers in collaboration with K-12 music teachers and the National Coalition for Core Arts Standards (NCCAS) music standards writing team leaders developed the MCAs in music to reflect the learning expectations defined in the performance standards. The Research Advisors (RAs) were selected from higher education institutions with specialized knowledge in the research regarding music development and assessment practices. Their work focused on creating MCAs for music learning in second, third, and fifth grades, ensembles, harmonizing instruments, theory/composition, and technology. Researchers guided the development with consideration of current understandings of developmental readiness, psychological principals of learning, and research-supported pedagogical assessment methods.

An assessment framework designed by NCCAS was adapted for the music MCAS, supported by focused and consistent interaction with practitioners and current learning theory resulting in measures of process rather than performance outcome. Designing the framework for assessment of the artistic process defined in National Standards for Music required that the RAs and the sub-committee members to carefully consider how students at each developmental level broadly and uniquely demonstrate expectations defined in the performance standards. Knowledge and skills relevant to the process components were defined by the writing committees and considered in the

development of the tasks and measures. Structure of the tasks and characteristics of preparatory instruction were designed to integrate into current practice in school music programs *with consideration of multiple classroom contexts inherent in schools across the United States.* The developmental process illustrates a shift of focus from a prescribed performance outcome to recognition of varied student demonstrations of achievement. The assessment framework provides a basis for teachers to guide and facilitate students to authentically demonstrate artistic processes and to assess student work that illustrates the nature and quality of student achievement envisioned in the standards.

Between 2013 and 2016 researchers oversaw multiple rounds of assessment task and scoring-device development. Careful consideration of traits to be demonstrated that illustrate the standards required revision by teachers, researchers, and standard writers. The expected traits were defined by the performance standard and categorized into rubric criteria. Individual rubrics categories were used as an assessment item as they pertain to individual process components. Corresponding levels of the prescribed taxonomy (e.g., below standard, approaching standard, meets standard, exceeds standard) and the defining characteristics of each level were agreed upon to reflect the task framework. The scoring rubrics were specifically designed to allow for task flexibility while measuring the student learning defined by the process components and performance standards.

Throughout the measure construction process, data documenting the variations in measures were collected from the researchers and sub-committee members through meeting notes and narrative feedback. Furthermore, piloting classroom teachers provided feedback through multiple stages of the pilot process to help reveal and authenticate immediate and developing perceptions of the MCAs, as well as to provide qualitative reflections concerning administrations of the assessments and student responses to the tasks.

As described in the MCAs, classroom teachers have the most informed grasp on student learning needs and contextual factors that influence student achievement. Students in one setting may need more time and additional instruction to guide successful achievement, while other school settings may not. The content of an MCA may not directly match the content already in a given school's curriculum, so the music educator would replace the content of a task with something that reflects the curriculum and the students' needs. Decisions of adaptation were collected throughout the process and analyzed along with specific demographic factors for disaggregation during analysis.

Scoring Procedures

In traditional performance assessment scoring procedures, indices of inter-rater reliability, intra-rater reliability, and correlation coefficients are used to assess the degree in which raters agree in overall scoring and use of the

rubrics. The goal in such procedures is for the raters to use the rubrics with machinelike consistency, and any divergence in scoring is considered as a source of error. For the scoring of the MCAs, Rasch Measurement Theory was used to evaluate rater consistency, consensus, and internal reliability of the measures.

Application of the Rasch measurement model was specifically chosen due to its strict requirements of invariance. In assessment contexts when raters are used to mediate the process, five fundamental requirements of rater-invariant measurement apply to the model:

a) scorer-invariant measurement of persons (i.e., the measurement of persons must be independent of the particular scorers that happen to be used for the measuring);
b) non-crossing person response functions (i.e., a more able person must always have a better chance of obtaining higher ratings from scorers than a less able person;
c) person-invariant calibration of scorers (i.e., the calibration of the scorers must be independent of the particular persons used for calibration);
d) non-crossing scorer response functions (i.e., any person must have a better chance of obtaining a higher rating from lenient scorers than from more severe scorers; and
e) variable map (i.e., persons and scorers must be simultaneously located on a single underlying latent variable).

Invariant measurement is a hypothesis that must be confirmed or disconfirmed by evidence in a data set (Engelhard, 1994). When the data fit the requirements of the Rasch model, then it becomes possible to support rater-invariant measurement of performances (Engelhard, 2013). With the use of Rasch Measurement Theory, the focus is not on raters as machines, rather, raters as independently acting *experts* who will sometimes disagree in their overall evaluations of the performances. Rater variation, from this perspective, is embraced as marked differences from experience, background, differing expertise, and vantage point. From this perspective, it was more advantageous to qualitatively engage with the expert raters in terms of their shared understanding of the construct while quantitatively controlling for rater errors (e.g., severity/leniency) versus expecting and insisting upon machine-like consistency in their scoring.

Andrich (1989) discusses the notion of gleaning qualitative understandings of measurement phenomena in the context of the Rasch measurement model:

> 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).

In instances when raters, items, or performances did not adequately fit the measurement model (evaluation of model-data fit), their scores were not discarded to provide a better fitting model to the data. Oppositely, signs on inadequate fit provided a mechanism for qualitatively investigating the cause of the misfit. As an example, a item demonstrating inadequate fit to the model may be evaluated for word structure or location in the rubric. A rater demonstrating inadequate fit to the model may be questioned as to why particular scores were provided. A performance demonstrating inadequate fit to the model may be investigated as to why it would have received specific scores. The data analysis was conducted using the program *Facets* (Linacre, 2014). For complete details of the data analysis results, please see Wesolowski (in press).

Post-Scoring Data

After the teachers scored and peer-scored the student work, they were contacted and asked to fill out a post-pilot follow-up survey. Data collected through the post-pilot survey included: (a) reflection on clarity and ease of administration of pilot protocol; (b) connection and usefulness within current curriculum; (c) ways that the MCA was adapted to fit context; (d) observed student response; (e) curricular impact; (f) instruction administered to prepare students for the assessment; (g) changes in teacher perception of the MCAs and student learning, and (h) suggestions for enhancements to the made to the MCAs, rubrics, and/or protocols as appropriate. Information attained through selected response and open-ended questions were sorted and analyzed by grade level and specific MCA administered using Microsoft's PowerBI. Of particular interest beyond the constructs in the post-pilot questions was uniqueness within or among demographic groups. The demographic filters obtained through a pre-pilot survey included: (a) region in the United States as designated at the National Association for Music Education (NAfME); (b) school district size; (c) socio-economic status of students in district as designated by percent of free and reduced lunches; (d) student opportunity-to-learn as designated by number of days per week the teacher meets with the students and number of minutes allocated per class period.

The Research Advisors (RAs) additionally collected qualitative information from the piloters throughout and following the pilot with interview and email interaction. Discussion notes were maintained through the consistent development and administration of the MCAs concerning the assessment tasks, scoring devices, administration protocol, and interaction with piloters. Analysis

of qualitative and quantitative data disaggregated by demographic characteristic provided the opportunity to find unique, as well as common characteristics of the MCAs among various assessment contexts.

Discussion

As early as the 1992 annual meeting of the American Education Research Association (AERA), psychometrician Wim van der Linden urged measure constructors in the field of education to consult context experts in all decision-making processes, including the carefully crafted definition of the construct intended for measurement, a clear definition of the descriptive components of the item design, strategic development of the item pool, and post-hoc refinement of the items. He urged test constructors to, "... start applying [measurement] models in discussion with content experts...[and] to start pretesting items… we should be looking for the intersection between what content experts agree on and what fits the [measurement] model" (van der Linden, 1992, 5:33-6:18).

The MCA developers decided to involve higher education researchers to work directly with teachers as content experts, as the first step in the construction process. We pretested items in classrooms very early in the process, garnering feedback not only about the items (components in the rubric scoring devices) themselves but also about the administration and instructional merit of the assessment tasks. We asked teachers as content experts to examine the overview and purpose of the MCAs and to determine whether the expected student competencies were reasonable, along with the instructional expectations. By sharing detailed descriptions of the tasks and items with the piloting teachers in an iterative process, we were able to work cyclically, improving the MCAs over the course of 3 years. Intending the MCAs to be embedded in curriculum (rather than positioning the MCAs as external assessments for the Music Standards), we wanted to ensure that piloting teaching recognized that they were to authentically represent skills and knowledge within their school and classroom context. Clearly, they needed to assess progressive development of artistic processes cross grades and achievement levels but had to be developed in a way that users could adapt with their own curriculum (repertoire, etc.).

In our review of acceptable evidence of reliability, validity, and fairness through Rasch analyses, we were pleased to report that teachers are the most appropriate expert to use and score the MCAs to measure student learning. Through our application of a somewhat new theoretical analysis to the data and by going back to teachers to identify the cause of outlier items and/or scores, we were able to determine the overall psychometric quality of the scales (items, student performances and raters) and determine how well the traits and tasks fit the measurement model. We also determined how each MCA varied across individual tasks. The Rasch model provided analysis to more tightly focus the scoring devices, as well as evidence supporting their rigor. The discoveries from

the qualitative analysis provided a mutual understanding of instructional usefulness, contextual practicality, applicability to determining student progress, and consistency in measuring student learning.

As evidenced in the MCA development process, the phenomenographic approach provided both measure developers and users the ability to better understand the psychological constructs and content underscoring the National Core Arts Standards in music. Furthermore, it brought together the music practitioner, music researcher, and music policy communities to construct and apply measures in ways that uses language and content true to what music educators are actively doing in their classrooms while establishing validity, reliability, and fairness in the scale construction process.

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: Vol. 4* (pp. 7-16). North-Holland, Netherlands: Elsevier Science Publishers.

Engelhard, 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.tb00436x

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

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th Ed.). Wesport, CT: American Council on Education.

Linacre, J. M. (2014). *Facets*. Chicago, IL: MESA Press.

Marton, F. (1986). Phenomenography: A research approach to investigating different understandings of reality. *Journal of Thought*, *21*, 29–49.

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.

Model Cornerstone Assessments (2015). National Association for Music Education. Retrieved from http://www.nafme.org/my-classroom/standards/mcas-information-on-taking-part-in-the-field-testing/

National Coalition for Core Arts Standards (2014). National Coalition for Core Arts Standards. Retrieved from http://www.nafme.org/my-classroom/standards/.

van der Linden, W. J. (1992). IRT in the 1990s- which models work best. Retried from Rasch Research Papers, Explorations & Explanations. Retrieved from http://www.rasch.org/audio/IRT-van-der-Linden-2.mp3

Wiggins, G. P. & McTighe, J. (2005). *Understanding by Design*. Alexandria, VA: Association for Curriculum and Development.

Wesolowski, B. C. (in press). Examination of the psychometric qualities of the Model Cornerstone Assessments using Rasch measurement theory: A technical report. In K. A. Parkes and F. Burrack (Eds.), *Authentic assessments for music: The Model Cornerstone Assessment research project*.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York, NY: Taylor & Francis Group, LLC.