ASSESSMENT

Special Research Interest Group

Fall Issue 2016

Assessment, in particular, measurement of student achievement and teacher evaluation, is an important component of music teaching and music teacher education. This issue features several research project underway across the country, an invitation to work collaboratively with your colleagues on relevant issues, and professional updates and conference information.

CALL FOR CONTRIBUTIONS and INVITATION TO JOIN:

- The NAfME Assessment SRIG publishes a newsletter annually. We accept articles, book reviews, and news items.

- Invite your colleagues to join the Assessment Special Research Group. To join, they simply go to the NAfME SRIG website and follow the directions: http://musiced.nafme.org/resources/special-research-interest-groups-srigs

- Our NAfME SRME SRIG Liaison is Steven M. Demorest, Professor, Music Education, Henry and Leigh Bienen School of Music, Northwestern University, 711 Elgin Road, Evanston, IL 60208-1200. He can be reached at sdemorest@northwestern.edu.

Website:
http://assessmentsrig.weebly.com/

VOLUME 30

Contents

2016 Music Research and Teacher Education National Conference Recap

In March 2016, the Assessment SRIG met as a part of the NAfME Music Research and Teacher Education National Conference. During the Assessment SRIG meeting, the following presentations were made:

The Ethics of Assessment: A Theoretical Model for Defining Fair Assessment - Ann Clements, Pennsylvania State University

Concerns of Validity and Inference in Music Assessments with Multiple Raters - Brian Wesolowski, University of Georgia

Interactivity of Rater-Schema in the Context of Large Scale Ensemble Performance Assessment - Brian Wesolowski, University of Georgia

Model Cornerstone Assessment - National research project: Overview. Kelly Parkes, Teachers College, Columbia University, and Fred Burrack, Kansas State University.

During the meeting, Brian Wesolowski was announced as the Chair-Elect for the Assessment SRIG. Dr. Wesolowski is an Assistant Professor of Music Education at the University of Georgia, Hugh Hodgson School of Music. He earned his Ph.D. in music education from the University of Miami in Coral Gables, Florida. He holds a Bachelor of Music in music education and jazz performance as well as a Master of Music Education from Oberlin College Conservatory of Music. In addition, he holds a Master of Music in jazz studies from the University of North Texas.

Dr. Wesolowski's research interests include rater behavior, scale development, policy of educational assessment, and the cognition, action, perception, and expressiveness in improvised jazz and groove-based performances. He has worked closely with the Georgia Department of Education, the National Coalition for Core Arts Standards (NCCAS), the State Education Agency Directors of Arts Education (SEADAE), NAfME's Model Cornerstone Assessment Pilot Study, and the National Board of Professional Teaching Standards on music assessment-related matters. He continues to present research papers internationally and nationally and has articles/chapters forthcoming or published in The Journal of Research in Music Education, Bulletin of the Council for Research in Music Education, The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, Educational Assessment, The Oxford Handbook of Assessment Policy and Practice in Music Education, Music Perception, Musicae Scientiae, International Journal of Music Education, Psychology of Music, Psychomusicology: Music, Mind, and Brain, Research Perspectives in Music Education, Music Educators Journal, Saxophone Symposium, Florida Music Director, and Georgia Music News.

You will find some of Dr. Wesolowski's recent work on page 7.

At the end of the meeting, Jeffrey Ward, Chair for the Assessment SRIG, shared a vision for the SRIG that included the SRIG's role of facilitating collaborative research through research subgroups. These groups met and discussed research needs in their respective areas and planned collaborative projects. If you are interested in collaborative research in one of the area subgroups (see below), please send an email expressing your interest to the area facilitator. I think this is an exciting time for our SRIG and I hope that you will engage with your colleagues!

Teacher Licensure and Evaluation: Doug Orzolek (dcorzolek@stthomas.edu)

Model Cornerstone Assessments: Phillip Payne (ppayne@ksu.edu)

Higher Education Music Admissions: Jeff Marlatt (jmarlatt@su.edu)

Program and Institutional Assessment: Jeffrey Ward (jeff98@ksu.edu)

ISME Assessment, Measurement, and Evaluation SIG
Timothy S. Brophy, University of Florida

The 32nd biennial ISME Conference was held in Glasgow, Scotland, from July 24-29, 2016. The Assessment, Measurement, and Evaluation (AME) SIG was well represented at the conference with a total of 49 sessions – 46 spoken papers, two workshops, and one symposium. Paper presenters represented many countries, including Estonia, Hungary, New Zealand, the United Kingdom, Australia, and the United States, to name a few. Paper topics ranged widely from classroom practices and assessment methodology to national initiatives that gather standardized test data to drive national educational policy. The AME-SIG leadership met during the meeting and made plans to move the SIG forward through the next biennium with Professor Dr. Andreas Lehmann-Wermser, University of Hannover, Germany, as the Chair (replacing founding Chair Timothy S. Brophy). The leadership decided to establish a formal SIG meeting time during the biennial International Symposia on Assessment in Music Education, and outgoing chair Timothy S. Brophy has established a SIG listserv at the University of Florida. Interested NAfME members who wish to join the listserv should send their contact information to Tim at tbrophy@aa.ufl.edu.

Development and Piloting of Model Cornerstone Assessments
Frederick Burrack and Kelly Parkes

Beginning in 2012, a team of 12 researcher advisors (RAs) have been involved with developing frameworks to help school music teachers integrate the updated National Standards for Music Education (finalized June, 2014) in their classrooms. This team worked with music teachers from across the country to develop tasks reflective of current practice in school music programs that result in authentic student demonstrations of the artistic processes. These tasks can independently exemplify the process components. Focused and consistent interaction of practitioners with the researchers, along with consideration of current learning theory, led toward Model Cornerstone Assessments work as a framework to measure the 2014 National Standards. The purpose of the model tasks were to provide examples of student learning assessments that could be used as is or adapted to the unique context and curriculum.

Developing Scoring Devices

Rubrics were created within a series of task processes that defined the traits expected to be demonstrated by students. 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. The expected traits were defined by the performance standard and categorized into rubric criteria. Careful consideration of traits required multiple phases of revision by teachers, researchers, and standard writers. Beginning 2014, the research team began to pilot the tasks and scoring devices in schools across the country. There were multiple rounds of revisions made to the assessment tasks and scoring-devices resulting from teacher feedback and analysis of student achievement data. 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.

MCA Description

Each Model Cornerstone Assessment (MCA) begins with a narrative description of the intent of the task and an estimation of the time required to prepare and administer the assessment in a music class. Prerequisite skills and knowledge required of the students are described to assist the music educator in appropriate preparation. Then a detailed outline of the assessment task clearly defining alignment with the performance standard enables the music educator to fully comprehend the task and indicators of achievement for each process component.

Testing the MCA

The resulting MCAs have been administered in schools Fall 2015, Spring 2016, and Fall 2016 to test the tasks and scoring devices for content and construct validity, consistency of administration and

scoring, usefulness in the classroom throughout multiple contexts, and effectiveness to guide student learning. Each assessment task is broken down in detail to guide: (a) instructional preparation; (b) classroom environment for the task; and (c) administration protocol and data to collect. To foster commonality and consistency within the assessment task administrations, all teachers piloting the MCAs were trained on the construct of the assessment task, provided "open-door" guidance on integrating the task into their curriculum, appropriate use of the scoring devices and procedures for uploading and scoring student work. Following administration of the MCAs in their classroom, all student work was uploaded to MeasureSuccess.org and the teachers scored student work using the online rubric. All student work was also peer-scored by two or three piloting teachers who were also using the same MCA.

Teachers were provided opportunities to reflect on the relevance of the task and scoring process as compared to their current conception of student achievement. With encouragement, teachers shared how they made relevant the MCAs in their students' learning experience. Data collected allowed for inference to instructional implications as to the varieties of student work that can authentically reflect the construct being measured by reviewing the illustrative examples in each achievement level.

Analysis of Data

In traditional performance assessment scoring procedures (classical test theory), indices of inter-rater reliability, intra-rater reliability, and correlation coefficients are used to assess the degree in which raters agree in overall scoring. The goal in these procedures is for the raters to use the rubrics with machinelike consistency, and any divergence in scoring is considered a source of error. For testing of the tasks and scoring rubrics in MCAs, the use of Rasch Measurement Theory was used to evaluate rater consistency, consensus, and internal reliability of the measures. With the use of Rasch Measurement Theory, the focus was 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).

In instances when raters, items, or performances did not adequately fit the measurement model (evaluation of model-data fit), their scores were not discarded as a means to provide a better fitting measurement 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 evaluation of shared understanding of the construct and improvement of the measures was found to be better accomplished in this context through the examination of multiple, divergent perspectives.

Understanding the Usefulness of the MCAs

The information collected from the pre-pilot was used to understand and compare the data collected post-pilot. The pre-pilot data included: (a) contact information; (b) number of years teaching; (c) NAfME region in which their school was located; (d) size and socio-economic status of their school district; (e) grade levels taught; (f) opportunity to learn; (g) prior experience in administering externally designed assessments; and (h) perceptions of the MCAs in relation to current student learning expectations. Data collected through the post-pilot survey included: (a) reflection on clarity and ease of administration of pilot protocol; (b) connection to current curriculum; (c) ways that the MCA was adapted to fit context; (d) observed student response; (e) curricular impact; and (f) adaptations made to the MCAs to fit curriculum, changes in teacher perception of the MCAs and student learning, and suggestions for enhancements to the made to the MCAs, rubrics, and/or protocols as appropriate.

Current Status of the Project

The RAs are collecting qualitative information from the piloters throughout and following the pilot through interview and email interaction. Discussion notes are being maintained through the consistent development and administration of the MCAs concerning the assessment tasks, scoring devices, administration protocol, and interaction with piloter teachers. Analysis of qualitative and quantitative data will be disaggregated by demographic characteristic provides the opportunity to find unique, as well as common characteristics of the MCAs among various assessment contexts. The final semester of student work will be teacher and peer-teacher scored by the end of January 2017. All data from the three testing semesters will be analyzed to determine construct validity, appropriate fit of measures to process components, consistency of scoring, and selection of student work that illustrates the variety of student work representative of the achievement levels.

The content of an MCA may not directly match the content in a given school's curriculum, so the study will also expose the extent that music educator can replace the content of a task with something more appropriate to reflect the local curriculum, as well as their students' needs, while maintaining the integrity of the performance standards. Decisions of adaptations were collected throughout the process and will be analyzed with consideration of specific demographic factors.

The Model Cornerstone Assessments are intended to be a basis for teachers to guide and facilitate students to authentically demonstrate artistic processes. The study allows us to collect student work that illustrates the nature and quality of student achievement envisioned in the standards, however at the conclusion of our research, student work will no longer be collected. The description of the development process and methodology of the pilot study, along with the findings, will be published in a book with Roman and Littlefield in such as way as to be appropriate and useful for school music teachers, as well as the research community.

Featured Research

Wesolowski, B.C., Wind, S.A., & Engelhard, Jr., G. (in press). Evaluating Differential Rater Functioning Over Time in the Context of Solo Music Performance Assessment. *Bulletin of the Council for Research in Music Education.*

Rater variability studies in the context of music performance assessment treat rater effects as static characteristics of raters, where the effects occur similarly across each assessed performance. The purpose of this study was to investigate expert raters' (N = 13) differential severity/leniency as dynamic processes, where the rater effects occur over time. In particular, we sought to examine the manifestation of group and individual variability using a class of rater effects referred to as Differential Rater Functioning Over Time (DRIFT). DRIFT refers to the changes in rater performance in relation to a parameter of time. Three classes of Multifaceted Rasch (MFR) models were specified in order to explore differences in raters' systematic changes in their interpretation of a four-point rating scale structure across a five-day rating session: (a) time-static model, (b) rater- by-time interaction model, and (c) partial credit model for time points. Results indicated a significant difference in severity/leniency across time for both the group of raters as a whole and some individual raters. Overall, raters demonstrated a general trend of decreasing severity over the five-day rating session. Interaction analyses suggested that differential severity/leniency existed for both the raters as a group and for nine out of the thirteen individual raters. Of the total 65 potential pairwise interaction terms examined between raters and days, 21 (33.31%) were found to be statistically significant. Ten interactions systematically underestimated the performances and 11 interactions systematically overestimated the performances. Implications for the improved fairness of ratings in music assessment contexts are discussed.

Wesolowski, B.C. (in press). Exploring Rater Cognition: A Typology of Raters in the Context of Music Performance Assessment. *Psychology of Music.*

This manuscript sought to investigate rater cognition by exploring rater types based upon differential severity and leniency associated with rating scale items, rating scale category functioning, and dimensions of music performance assessment. The purpose of this study was to empirically identify typologies of operational raters based upon systematic differential severity indices in the context of large ensemble music performance assessment. A rater cognition information-processing model is explored based upon two frameworks: a *framework for scoring* and a *framework for audition*. Rater scoring behavior was examined using a *framework for scoring*, where raters' mental processes compare auditory images to the scoring criteria used to generate a scoring decision. The scoring decisions were evaluated using the Multifaceted Rasch Partial Credit Measurement Model. A rater typology was then examined under the *framework of audition*, where similar schemata were defined through raters' clustering of differential severity indices related to items and compared across performance dimensions. The results provided three distinct rater-types: (a) the syntactical rater; (b) the expressive rater; and (c) the mental representation rater. Implications for equitability and precision in the assessment process are discussed as well as considerations for validity of scoring processes.

Context Matters: The 6th International Symposium on Assessment in Music Education

April 18-21, 2017
Birmingham City University
Birmingham, England, UK

For conference details, see http://reg.conferences.dce.ufl.edu/isame

Email: isame6@aa.ufl.edu

The University of Florida and Birmingham City University will host the 6th International Symposium on Assessment in Music Education in April 2017 at Birmingham City University in the UK. The purpose of the symposium is to bring together music education and measurement professionals worldwide to share the latest research, thought, and practice in music education assessment. We invite primary and secondary school music educators, higher education professionals and music education researchers, national, state and local education officials from across the world to join us in Birmingham.

Key Questions. The Topics and Key Questions for this symposium are:

- **Diverse methodologies.** What models, designs, and practices are most successful in assessing student achievement and skill in music in diverse educational systems?
- **Effective data use.** In what ways are music educators using assessment data effectively to improve music teaching and learning?
- **Assessing music learning.** How is music learning assessed in local, regional, state, national, and international contexts?
- **Data analysis.** In what ways are music assessment data analyzed to yield meaningful information in different contexts?
- **Evaluation.** What forms, tools, or processes are used to assess and/or evaluate the practices of music teaching and learning?
- **Validity and reliability.** In what ways are reliability and validity influenced by music assessment contexts?
- **Unintended consequences.** What unforeseen and unplanned outcomes of assessment and evaluation have been observed?
- **Ethical obligations.** How does the music education profession ensure that the ethical obligations of assessment in music education (professional competency, integrity, honesty, confidentiality, objectivity, and fairness) are met?

Past ISAME symposium published by GIA can be ordered from
http://www.giamusic.com/music_education/

For comments, suggestions, or future submissions to this newsletter, please contact the Chair or Chair Elect of the SRME Assessment SRIG

JEFFREY WARD, CHAIR

Director
School of Music, Theatre, and Dance
Kansas State University
785-532-5740
jeff98@ksu.edu

BRIAN WESOLOWSKI, CHAIR ELECT

Assistant Professor of Music Education
The Hugh Hodgson School of Music
The University of Georgia
706-542-3737
bwes@uga.edu