Applying Model Cornerstone Assessments in K–12 Music

A Research-Supported Approach

Edited by Frederick Burrack
and Kelly A. Parkes

Published in Partnership with the
National Association for Music Education

ROWMAN & LITTLEFIELD
Lanham • Boulder • New York • London

Published in partnership with the National Association for Music Education,
1806 Robert Fulton Drive, Reston, Virginia 20191; nafme.org

Published by Rowman & Littlefield
A wholly owned subsidiary of The Rowman & Littlefield Publishing Group, Inc.
4501 Forbes Boulevard, Suite 200, Lanham, Maryland 20706
www.rowman.com

Unit A, Whitacre Mews, 26-34 Stannary Street, London SE11 4AB, United Kingdom

British Library Cataloguing in Publication Information Available

Library of Congress Cataloging-in-Publication Data

Names: Burrack, Frederick. | Parkes, Kelly A.
Title: Applying model cornerstone assessments in K–12 music : a research-supported approach / [edited by] Frederick Burrack and Kelly A. Parkes.
Description: Lanham : Rowman & Littlefield, [2018] | Includes bibliographical references.
Identifiers: LCCN 2018017989 (print) | LCCN 2018018381 (ebook) | ISBN 9781475837407 (Electronic) | ISBN 9781475837384 (cloth : alk. paper) | ISBN 9781475837391 (pbk. : alk. paper)
Subjects: LCSH: Music—Instruction and study—Evaluation. | Music—Instruction and study—Research.
Classification: LCC MT1 (ebook) | LCC MT1 .M75 2018 (print) | DDC 780.71—dc23 LC record available at https://lccn.loc.gov/2018017989

™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI/NISO Z39.48-1992.

Printed in the United States of America

Eleven

Methodology for Examining the Psychometric Qualities of the Model Cornerstone Assessments

Brian C. Wesolowski

Classroom assessments and, more specifically, the measurement instruments used in classrooms, undergo little if any psychometric (e.g., validity, reliability, and fairness) evaluation. In the rare event that classroom assessments in the arts are evaluated for psychometric quality, the most used measurement model is based on classical test theory (CTT). CTT, or "true-score theory," uses raw scores gleaned from a pool of examinees to test their relative success or failure on individual items. Known for its relatively weak theoretical assumptions, the CTT measurement model assumes that the observed scores obtained from a measure comprise two parts: true score and measurement error. CTT analysis is usually limited to item-difficulty indices based on proportion-correct and reliability coefficients that summarize proportions of variance. In the case of polytomously scored items, such as those found in the Music Model Cornerstone Assessments (herein referred to as MCAs), adjusted proportion correct values (p-values) and correlation coefficients are used to indicate item difficulty and overall ability level of an examinee. The major disadvantage of using CTT as a means for analyzing performance assessment data is the sample and test dependency of estimated person parameter estimates (e.g., true scores) and item parameters (e.g., item discrimination and item difficulty). This limits the ability to develop valid and reliable measures and to make informed inferences related to examinee ability and item difficulty that extend beyond the context of the sample of student work or performances used in the assessment context.

In contrast, the Rasch family of measurement models offers a more grounded theory compared to CTT (Cavanagh & Waugh, 2011). Rasch measurement theory (Rasch, 1960/1980) is often preferred in scale development as well as in the measurement of latent traits in the behavioral, social, and health sciences (Engelhard, 2013). 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 measuring. Rasch measurement 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 scorer severity and leniency, criterion difficulty, and student performance achievement level. Hierarchies of difficulty for each relevant criterion, 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.

It is important to note that the property of invariant measurement that characterizes the Rasch model must be evaluated using empirical data. Invariant measurement is a hypothesis that must be confirmed or disconfirmed by evidence in a data set (Engelhard, 1994). 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 (i.e., 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 does 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). Based on 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. 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 construct (i.e., latent trait), the local independence of items, and sample-independent estimations of person and item parameters (i.e., invariant measurement).

On the occasion that scorers facilitate the assessment process, as is the case with the MCA pilot study, the Many Facet Rasch (MFR) model can be used to simultaneously define student ability, criterion difficulty, and scorer severity (Linacre, 1989/1994). The MFR model stems from the family of Rasch measurement models and can be used for both dichotomous or polytomous items (Wright & Mok, 2004). As pointed out by Engelhard (2013), the five requirements for invariant measurement can be extended to the context in which assessments are mediated by scorers: (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 does 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). When the data fit the requirements of the Rasch model, then it becomes possible to support invariant measurement that also implies scorer-invariant measurement of performances (Engelhard, 2013).

The purpose of the technical report in the following chapter is to investigate the psychometric properties (e.g., validity and reliability) of the National Association for Music Education's (NAfME) Music Model Cornerstone Assessment 2015–2016 Pilot Study. The MCAs investigated in this study include (a) Grade 2 create, (b) Grade 2 perform, (c) Grade 2 respond, (d) Grade 5 perform, (e) Grade 5 respond, (f) Grade 8 create, (g) composition/theory, (h) ensemble perform (intermediate), (i) ensemble perform (proficient), (j) harmonizing instruments, and (k) harmonizing Instruments (revised). MCAs not investigated in this study involved challenges for the analysis for three possible reasons: (a) no data collected; (b) not enough data collected to warrant analysis; or (c) no cross-scoring occurred, resulting in student work confounding with the scorer. The specific research questions that guide this study include:

1. What is the overall psychometric quality (e.g., validity and reliability) of each of the model cornerstone assessments?
2. How well do the criteria fit the measurement model and how do they vary in difficulty?[1]
3. How does the rating-scale structure (i.e., levels) of each Model Cornerstone Assessment vary across individual criteria?

$$\ln\left[\frac{P_{nijmk}}{P_{nijmk-1}}\right] = \theta_n - \lambda_i - \delta_j - \gamma_m - \tau_{ik} \quad , \qquad (1)$$

where

$\ln[P_{nijmk}/P_{nijmk-1}]$ = the probability that Student work n rated by Scorer i on Criterion j receives a rating in level k rather than level k-1,

θ_n = the logit-scale location (e.g., achievement) of Student Work n,

λ_i = the logit-scale location (e.g., severity) of Scorer i,

δ_j = the logit-scale location (e.g., severity) of Scoring Type (e.g., peer- or self-scored) j,

γ_m = the logit-scale location (e.g., achievement) of Criterion m,

t_{ik} = the location on the logit scale where scale levels k and k-1 are equally probable for Scorer i.

Figure 11.1.

MEASUREMENT MODEL

The measurement model used in this study was the Multifaceted Rasch Partial Credit (MFR-PC) measurement model (Linacre, 1989/1994) (see Figure 11.1). The Partial Credit (PC) version of the model (Masters, 1982) adds an additional parameter to the model that allows for the investigation of the rating-scale structure across each of the criteria, thereby making it possible to test the null hypothesis of equidistant rating-scale levels across each criterion. The addition of this parameter provides construct evidence of the measure through the verification of an increasingly monotonic relationship between adjacent levels (i.e., the preservation of increasingly positive ordering that establishes an intended direction of "more achievement"), acceptable discrimination between performances, appropriate distribution of frequency use by scorers (i.e., multimodal use of all available rating-scale levels), and levels of acceptable randomness for the stochastic process of probabilistic modeling (i.e., acceptable levels of unsystematic variability for probabilistic processes; Linacre, 2002a). The PC model is specified as follows:

DATA ANALYSIS PROCEDURES

Examining Psychometric Quality

Table 11.1 presents a set of statistics and displays based on the MFR-PC model that can be used to examine the psychometric quality of musical performance assessments. Indices based on three distinct levels were examined: (a) logit-scale locations, (b) separation, and (c) model-data fit. Logit-scale locations provide a method for summarizing student work,

Table 11.1. Statistics and Displays Based on the MFR-PC Model

Level	*Indicators and Displays based on the MFR-PC Model*	*Substantive Interpretation (Question)*			
		Student Work Facet	*Scorer Facet*	*Scoring-Type Facet*	*Criteria Facet*
A. Logit-Scale Locations	1. Variable map	Where is the student work located on the construct being measured (criterion achievement)?	Where are the scorers located on the construct being measured (criterion achievement)?	Where are the scoring types located on the construct being measured (criterion achievement)?	Where are the criteria located on the construct being measured (criterion achievement)?
	2. Location of elements within the facet	What is the location of each student work (criterion achievement)?	What is the location of each scorer (severity/leniency)?	What is the location of each scoring type (severity/leniency)?	What is the location of each criterion (difficulty)?
	3. Standard error	How precisely has the location of each student work been estimated?	How precisely has the location of each scorer been estimated?	How precisely has the location of each scoring type been estimated?	How precisely has the location of each criterion been estimated?

(continued)

Table 11.1. ***(Continued)***

Level	*Indicators and Displays based on the MFR-PC Model*	*Substantive Interpretation (Question)*			
		Student Work Facet	*Scorer Facet*	*Scoring-Type Facet*	*Criteria Facet*
B. Separation	4. Reliability of separation statistic	How spread out are the scored performance locations on the logit scale?	How spread out are the scorer locations on the logit scale?	How spread out are the scoring-type locations on the logit scale?	How spread out are the criterion locations on the logit scale?
	5. Chi-square statistic	Are the overall differences between student work locations significant?	Are the overall differences between scorer locations significant?	Are the overall differences between scoring-type locations significant?	Are the overall differences between criterion locations significant?
C. Model-Data Fit	6. Mean Square Error (MSE) and standardized fit statistics	How consistently has each student work been interpreted by the scorers?	How consistently has each scorer interpreted the items and rating scale levels across the student work?	How consistently has each scoring type interpreted the items and rating scale levels across the student work?	How consistently has each criterion been interpreted by the scorers?

scorer severity, scoring-type severity, and criterion difficulty on a single linear scale that represents the latent construct. Separation indicates the degree to which the elements within a facet can be reliably differentiated from one another. Elements refer to each piece of individual student work (within the student work facet), each individual scorer (within the scorer facet), each scoring type (within the scoring type facet), and each individual criterion (within the criterion facet). The separation statistic for student work can be interpreted similarly to Cronbach's alpha, indicating high reproducibility of relative measure locations. Separation statistics for scorers, scoring type, and criteria can be interpreted as the separation verification of the hierarchy for the elements within each facet (i.e., construct validity). Model-data fit indices describe how closely the raw score observations provided by the scorers approximate the useful, invariant properties of the Rasch model. Mean square error (MSE) fit statistics demonstrate the overall randomness within the model (Linacre, 2002b). Perfect predictability is represented by the value of 1.00. Values less than 1.00 indicate too much predictability/redundancy in the data (i.e., muted data). Values above 1.00 indicate too little predictability in the data (i.e., sporadic data). In particular, infit MSE fit statistics represent inlier-sensitive fit, where over- or under-fit for Guttman probabilistic patterns are detected. Outfit MSE fit statistics represent outlier-sensitive fit, where over-fit for observations of model variance are detected. When reasonable model-data fit to the model is observed, invariant measurement is achieved. At the parameter level, the reasonable mean-square range for infit and outfit is 0.50–1.50. The reasonable mean-square range for infit and outfit at the element level for *high stakes assessments* is 0.80–1.20. In the case of the MCAs, the data will be treated as *survey/rating-scale data* where the scorers are untrained. Therefore, the reasonable mean-square range for infit and outfit for the element level is 0.60–1.40. Standardized fit statistics (*Zstd*) are t-tests (reported as z-scores) that test the hypothesis of perfect model data fit for predictability of data. Less than the expected score of 0.00 indicates predictability, and values above 0.00 indicate lack of predictability. All data fitting in the range of –1.90 to 1.90 indicates reasonable predictability and good model data fit (Linacre & Wright, 2004). Fit statistics within the threshold indicate sufficient accuracy and predictability of model data fit, validity evidence for the construct, and good productivity for measurement (Linacre, 2002b).

Because the Rasch model is unidimensional, it is possible to display the location estimates for each facet on a single linear scale. The *variable map* is a useful method for visually displaying descriptions of student work, criteria, scorers, and other facets of interest in terms of a single, unidimensional latent variable. The usefulness of the variable map is a major factor in the adoption of Rasch modeling by many national and international

assessments, including but not limited to the National Assessment of Educational Progress and the Program for International Student Assessment, for example. In the context of this study, the variable maps provide a graphical representation of the student work, scorer, score type, and criterion facets on a common "ruler."

Examining Rating-Scale Structure

Application of the Partial Credit Model (Wright & Masters, 1982) to Linacre's (1989/1994) MFR model extends the analysis to allow the distance between rating-scale level thresholds to vary across each criterion. Statistically, the process of freeing each criterion from a rating-scale grouping and allowing it to define its own partial credit scale allows for each ordered rating-scale level to be estimated. Substantively, the process indicates that in addition to each criterion having its own unique difficulty level as identified by its location on the logit scale, each rating-scale level within each item, too, has its own difficulty level as indicated by its own unique location on the logit scale. For polytomous items, such as those found in the MCAs, inter-adjacent-level discrimination indices (i.e., Rasch-Andrich thresholds) provide the location on the latent continuum where adjacent rating-scale levels are discriminated. In the instance where four levels are used for each criterion (e.g., Level 1 = "Emerging"; Level 2 = "Approaching Criterion"; Level 3 = "Meets Criterion"; Level 4 = "Exceeding Criterion"), the levels are modeled by a set of three discrimination indices that describe the location where (1) Level 1 and Level 2 are discriminated, (2) where Level 2 and Level 3 are discriminated, and (3) where Level 3 and Level 4 are discriminated. The substantive interpretation of the Rasch-Andrich thresholds, however, is based on adequate functioning and proper optimization of the rating-scale levels.

Linacre (2002b) indicates that methodological steps can be taken in order to optimize rating-scale level structures. Modification of the structure based on this empirical methodology provides more rigorous examination and precise estimation of performances, ultimately addressing and improving validity issues surrounding construct validity of the measurement instrument. Additionally, this post hoc investigation can clarify the meaning of the collected data and improve subsequent use of the scale. First, frequency counts for each of the four levels were examined. Uniformly distributed frequency counts across each of the rating-scale levels are optimal for the calibration of rating-scale difficulties. Any frequency count demonstrating less than 10% of the total level usage provides incentive to collapse the level into an adjacent level. Second, outfit MSEs were examined for values ≥2.0. Values greater than 2.0 indicate excessive noise in the ratings. More

specifically, levels exhibiting MSE values ≥2.0 indicate that they have been used by scorers in unexpected contexts and warrant their collapse into an adjacent level. Third, average observed logit measures were examined for violations of monotonicity. Monotonicity can be described as the continuous advancement of threshold calibrations (Andrich, 1996). This is a requirement for inferential interpretability of the rating scale. In instances when incrementally higher measures were not observed, it is suggested that violating levels are to be collapsed into adjacent levels.

NOTE

1. Throughout this report, each component of the MCA (e.g., interpret, evaluate, etc.) is referred to as a *trait*. Each row of the MCA is referred to as a *criterion*, and the columns (emerging, approaching standard, meets standard, exceeds standard) are referred to as *levels*.

REFERENCES

Andrich, D. A. (1996). Measurement criteria for choosing among models for graded responses. In A. von Eye & C. C. Clogg (Eds.), *Analysis of categorical variables in developmental research* (pp. 3–35). Orlando FL: Academic Press.

Cavanagh, R. F., & Waugh, R. F. (2011). *Applications of Rasch measurement in learning environment research*. Rotterdam, The Netherlands: Sense 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.

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

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

Linacre, J. M. (1989/1994). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2002a). Judge ratings with forced agreement. *Rasch Measurement Transactions, 16*(1), 857–858.

Linacre, J. M. (2002b). Optimizing rating scale level effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

Linacre, J. M., & Wright, B. D. (2004). Construction of measures from many-facet data. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theories, models, and applications* (pp. 296–321). Maple Grove, MN: JAM Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. Expanded edition. (1980). Chicago, IL: University of Chicago Press.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Wright, B. D., & Mok, M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith and R. M. Smith (Eds.), *Introduction to Rasch measurement: Theories, models, and applications* (1–24). Maple Grove, MN: JAM Press.