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

TWELVE

Examination of the Psychometric Qualities of the Model Cornerstone Assessments

Brian C. Wesolowski

In music and related behavioral sciences, 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. Accordingly, latent properties must be defined by secondary, observable behaviors. The purpose of the Model Cornerstone Assessment (MCA) Pilot Study was to develop criteria that provide formative and summative means to measure student achievement of performance standards in the National Core Music Standards. However, it is important to be aware that the development of criteria within the context of an assessment is, more broadly, the process of developing a hypothetical, latent construct. In the broadest sense, the constructs are the processes directly associated with the National Core Music Standards. Therefore, the development of assessments is dualistic: (a) to provide a means for measuring students, and (b) to develop a hypothetical latent construct. Therefore, the results reported in this chapter not only provide information about student performance; more importantly, they provide diagnostic information about each of the latent constructs represented by the measurement instruments. Therefore, it is recommended that the suggestions provided here be strongly considered as a mechanism to redevelop and refine the hypothetical construct intended to be developed.

The MCAs investigated in this study include Grade 2: create, perform, respond; Grade 5: perform, respond; Grade 8: create; Composition/Theory create, respond, perform (proficient level); Ensemble: perform (intermediate and proficient); and Harmonizing Instruments: create (proficient). Responses to research question 1, "*What is the overall psychometric quality (e.g., validity and reliability) of each of the Model Cornerstone Assessments?*" are provided under the summary statistics subheadings and depicted through the related variable maps. Responses to research

question 2, "*How well do the criteria fit the measurement model, and how do they vary in difficulty?*" are provided under the calibrations of traits and calibration of criteria subheadings. Responses to research question 3, "*How does the rating-scale structure of each Model Cornerstone Assessment vary across individual criteria?*" are provided under the rating-scale-level diagnostics and inter-adjacent-level discrimination indices subheadings. Each MCA report includes *summary statistics*, *calibration of student work*, *calibration of scorer findings*, *calibration of scoring type*, *calibration of traits*, *calibration of criteria*, *rating-scale-level diagnostics*, and *inter-adjacent-level discrimination indices* and a *Findings* section.

PRECURSORY INFORMATION

In all reports that follow, the *calibration of student work* assumes that the student work facet was allowed to float (i.e., noncentered) relative to all other facets in the model. Common practice indicates that objects of measurement be allowed to float. All student work outside the reasonable mean-square range for infit and outfit (0.60–1.40) did not adequately fit the measurement model. It is suggested that first, these cases not be used as exemplars of student work, and second, that they be investigated qualitatively for indicators as to why they may not have demonstrated adequate model-data fit.

The *calibration of scorer findings* assumes that the scorer facet was centered on the logit scale (mean of 0.00 logits) to provide a frame of reference for the interpretation of the student work locations (i.e., objects of measurement), and as a result, means are reported in logits.

The *calibration of scoring type* assumes that student work was crossed-scored, providing two types of scoring: (a) peer scoring and (b) self-scoring. The scoring-type facet was centered on the logit scale (mean of 0.00 logits) to provide a frame of reference for the interpretation of the student work locations (i.e., objects of measurement).

The *calibration of traits* assumes that the trait facet was centered on the logit scale (mean of 0.00 logits) to provide a frame of reference for the interpretation of the student work locations (i.e., objects of measurement). Similarly, the *calibration of criteria* assumes that the student work facet was centered on the logit scale (mean of 0.00 logits) to provide a frame of reference for the interpretation of the student work locations (i.e., objects of measurement).

The *inter-adjacent-level discrimination* indices all assume that evidence exists to reject the null hypothesis of equidistant rating-scale levels. Prior to evaluating the inter-adjacent-level discrimination indices provided, any recommendations of collapsing levels under the *Rating-Scale-Level Di-*

agnostics subheading should be considered. Then, assuming acceptability of the related rating-scale-level diagnostics, the range of rating-scale Level 1 (emerging) is –∞ to the logit score indicated under the *Level 2 Threshold* subheading, minus the standard error. The range of rating-scale Level 2 (approaching) is the logit score listed under the *Level 2 Threshold* subheading, minus the standard error and the logit score listed under the *Level 3 Threshold* subheading, minus the standard error. The range of rating-scale Level 3 (meets) is the logit score listed under the *Level 3 Threshold* subheading, minus the standard error and the logit score listed under the *Level 4 Threshold* subheading, minus the standard error. The range of rating-scale Level 4 (i.e., exceeding) is the logit score listed under the *Level 4 Threshold* subheading, minus the standard error to ∞.

Construct validity is addressed in each report in the *Findings* section as demonstrated by the reasonable parameter separation for each of the considered parameters. The parameters considered for each MCA were (a) student work, (b) scorers, (c) scoring type, and (d) criteria. Reasonable parameter separation indicates that one set of parameters (the performance achievement of student work, for example) can be estimated without any interference from any of the other parameters (the severity of the raters or difficulties of the criteria, for example). As a result, one can directly and meaningfully interpret the characteristics of each parameter (e.g., the overall performance achievement of a student, the overall severity of the scorers, the overall severity of scoring types, and the overall difficulties of the level) as a unique characteristic to those parameters. This is a characteristic of fundamental measurement that is specifically underscored by properties of invariance that is expected in physical measurement, and a unique characteristic of the Rasch family of measurement models in the context of psychological measurement as applied here. As with the previous chapter and for the purposes of this chapter, each process 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*. Tables and figures are found at http://nafme.org/wp-content/files/2017/10/Chapter-12-Data-Tables-and-Figures.pdf.

GRADE 2 CREATE MCA

Summary Statistics

Table 12.A1 provides the summary statistics for the MFR-PC model analysis of student work ($n = 333$), scorers ($n = 8$), scoring type ($n = 2$) and criteria ($n = 1$). The analysis indicated an overall good model data fit with

significant differences between student work ($\chi^2_{(332)} = 841.90$, $p < .01$), scorers ($\chi^2_{(7)} = 123.80$, $p < .01$), scoring type ($\chi^2_{(1)} = 61.60$, $p < .01$), and criteria ($\chi^2_{(10)} = 3821.20$, $p < .01$). Reliability of separation statistics for all four facets are as follows: student work (*Rel* = .63), scorers (*Rel* = .71), scoring type (*Rel* = .97), and criteria (*Rel* = .99). See figure 12.I1 for the variable map.

Calibration of Student Work

Table 12.B1 provides student work calibration information. The mean of the student work was 0.89 logits with a range of 2.95 logits for the highest fit achieving student work (Student ID 4778) to –1.39 logits for the lowest fit achieving (Student ID 4802).

Calibration of Scorers

Table 12.C1 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.39 for the most severe scorer (Scorer 8) to –0.43 for the most lenient scorer (Scorer 1). All scorers demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Scoring Type

Table 12.D1 provides scoring-type calibration information. The mean of the scoring-type facet was 0.00 logits with a range of 0.15 for the most severe scoring type (peer score) to –0.15 for the most lenient scoring type (self score). Both scoring types demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Traits

Table 12.E1 provides trait calibration information. The mean of the traits was 0.00 logits with a range of 0.60 for the most difficult trait (scoring device 1, part 2) to –0.25 for the least difficult trait (scoring device 1, part 1). All traits demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F1 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 2.76 for the most difficult criterion (rhythmic complexity) to –1.05 for the easiest criterion (uniqueness of response). All criteria demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G1. Three steps were taken in order to evaluate the rating-scale-level structure. First, frequency counts for each level were evaluated for usage under 10%. Results indicated that four levels did not reach the prescribed 10% usage: (a) Level 1 of criterion 2, (b) Level 1 of criterion 7, (c) Level 1 of criterion 8, and (d) Level 2 of criterion 9. It is therefore recommended that all first levels be collapsed into their respective adjacent Level 2 and be rewritten, and Level 2 of criterion 9 be collapsed into its adjacent Level 1 and be rewritten. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that no levels were found to have an MSE value ≥2.0. Third, average observed logit measures were examined for violations of monotonicity. Results indicated no violations of monotonicity.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H1 for each criterion within the respective trait.

Findings

The Grade 2 Create MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the considered parameters.

Results indicated that the rank order of traits by difficulty (from most difficult to least difficult) was (1) scoring device 1, part 2, (2) scoring device 2, and (3) scoring device 1, part 1. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Create* construct for second grade. An analysis of the usability and meaningfulness of the levels across each criterion indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G1. Based on the analysis, the following is recommended: (a) Level 1 of criterion 2 should be collapsed with Level 2 of criterion 2 and be rewritten; (b) Level 1 of criterion 7 should be collapsed with Level 2 of criterion 7 and be rewritten; (c) Level 1 of criterion 8 should be collapsed with Level 2 of criterion 8 and be rewritten; and (d) Level 2 of criterion 9 should be collapsed with Level 1 of criterion 9 and be rewritten. The tables displaying all psychometric analyses for the Grade 2 Create MCA include tables 12.A1, 12.B1, 12.C1, 12.D1, 12.E1, 12.F1, 12.G1, and 12.H1. The variable map can be found in figure 12.I1 (http://nafme.org/wp-content/files/2017/10/Chapter-12-Data-Tables-and-Figures.pdf).

GRADE 2 PERFORM

Summary Statistics

Table 12.A2 provides the summary statistics for the MFR-PC model analysis of student work (n = 222), scorers (n = 9), scoring type (n = 2) and criteria (n = 7). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(221)}$ = 1649.30, p< .01), scorers ($\chi^2_{(8)}$ = 184.80, p< .01), scoring type ($\chi^2_{(1)}$ = 5.00, p< .01), and criteria ($\chi^2_{(6)}$ = 1115.30, p< .01). Reliability of separation statistics for all four facets are as follows: student work (*Rel* = .86), scorers (*Rel* = .94), scoring type (*Rel*< .00), and criteria (*Rel* = .99). See figure 12.I2 for the variable map.

Calibration of Student Work

Table 12.B2 provides student work calibration information. The mean of the student work was 0.70 logits with a range of 3.86 logits for the highest fit achieving student work (Student ID 4934) to –3.53 logits for the lowest fit achieving (Student ID 4343).

Calibration of Scorers

Table 12.C2 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.68 for the most severe scorer (Scorer 5) to –0.49 for the most lenient scorer (Scorer 6). All scorers demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Scoring Type

Table 12.D2 provides scoring-type calibration information. The mean of the scoring-type facet was 0.00 logits. Because separation of reliability was low (0.00), both scoring types demonstrated logit measures 0.00, indicating that student performances were not separated based on scorers' affiliation to a scoring type.

Calibration of Criteria

Table 12.F2 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 1.46 for the most difficult criterion (expressive quality) to –1.55 for the easiest criterion (rhythm). All criteria

demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G2. Three steps were taken in order to evaluate the rating-scale-level structure. First frequency counts for each level were evaluated for usage under 10%. Results indicated that two levels did not reach the prescribed 10% usage: (a) Level 1 of criterion 2, and (b) Level 4 of criterion 7. It is therefore recommended that Level 1 of criterion 2 be collapsed into its respective adjacent Level 2 and be rewritten, and Level 4 of criterion 7 be collapsed into its adjacent Level 3 and be rewritten. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that one level was found to have an MSE value ≥2.0: Level 3 of criterion 1. It is therefore recommended that Level 3 be examined and rewritten. Third, average observed logit measures were examined for violations of monotonicity. Results indicated no violations of monotonicity.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H2 for each criterion within the respective trait.

Findings

The Grade 2 Perform MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the considered parameters.

Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) expressive quality, (2) tonal center, (3) intonation/pitch, (4) starting pitch, (5) singing voice, (6) tempo, and (7) rhythm. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Perform* construct for second grade. An analysis of the usability and meaningfulness of the levels across each criterion indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G2. Based on the analysis, the following is recommended: Level 1 of criterion 2 be collapsed into its respective adjacent Level 2 and be rewritten, and Level 4 of criterion 7 be collapsed into its adjacent Level 3 and be rewritten, Level 3 be rewritten with more specificity. The tables displaying all psychometric analyses for the Grade 2 Perform MCA in-

clude tables 12.A2, 12.B2, 12.C2, 12.D2, 12.E2, 12.F2, 12.G2, and 12.H2. The variable map can be found in figure 12.I2.

GRADE 2 RESPOND

Summary Statistics

Table 12.A3 provides the summary statistics for the MFR-PC model analysis of student work ($n = 261$), scorers ($n = 8$), scoring type ($n = 2$) and criteria ($n = 3$). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(260)} = 714.60$, $p < .01$), scorers ($\chi^2_{(7)} = 124.40$, $p < .01$), scoring type ($\chi^2_{(1)} = 10.40$, $p < .01$), and criteria ($\chi^2_{(2)} = 33.20$, $p < .01$). Reliability of separation statistics for all four facets are as follows: student work (*Rel* = .65), scorers (*Rel* = .94), scoring type (*Rel* = .81), and criteria (*Rel* = .91). See figure 12.I3 for the variable map.

Calibration of Student Work

Table 12.B3 provides student work calibration information. The mean of the student work was –1.12 logits with a range of 4.69 logits for the highest fit achieving student work (Student ID 2596) to –6.06 logits for the lowest fit achieving (Student ID 3808).

Calibration of Scorers

Table 12.C3 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 1.73 for the most severe scorer (Scorer 5) to –0.29 for the most lenient scorer (Scorer 6). Scorer 3 demonstrated muted rating patterns as evidenced by an infit MSE less than 0.60.

Calibration of Scoring Type

Table 12.D3 provides scoring-type calibration information. The mean of the scoring-type facet was 0.00 logits with a range of 0.26 for the most severe scoring type (peer score) to –0.26 for the most lenient scoring type (self score). Both scoring types demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Traits

Table 12.E2 provides trait calibration information. The mean of the traits was 0.00 logits with a range of 0.16 for the most difficult trait (evalu-

ate) to –0.31 for the least difficult trait (interpret). All traits demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F3 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 0.51 for the most difficult criterion (cite musical reasons) to –0.31 for the easiest criterion (movement). All criteria demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G3. Three steps were taken in order to evaluate the rating-scale level structure. First frequency counts for each level were evaluated for usage under 10%. Results indicated that three levels did not reach the prescribed 10% usage: (a) Level 4 of criterion 1, and (b) Level 4 of criterion 2, and (c) Level 4 of criterion 3. It is therefore recommended that each of these levels be collapsed into their respective Level 3s and be rewritten. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that one level was found to have an MSE value ≥2.0: Level 4 of criterion 3. Because Level 4 of criterion 3 additionally did not reach the minimum level usage, it is recommended that this level be collapsed into Level 3. Third, average observed logit measures were examined for violations of monotonicity. Results indicated one level that violated the rule of monotonicity: Level 4 of criterion 4. Because Level 4 of criterion 3 did not reach the minimum frequency use and was found to have an MSE value ≥2.0, it is recommended to collapse Level 4 into the adjacent Level 3.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H3 for each criterion within the respective trait.

Findings

The Grade 2 Respond MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the considered parameters.

Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) cite musical reasons, (2) verbal response,

and (3) movement. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Respond* construct for second grade. An analysis of the usability and meaningfulness of the levels across each criterion (i.e., emerging, approaches standard, meets standard, exceeds standard) indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G3. Based on the analysis, the following is recommended: (a) Level 4 of criterion 1 should be collapsed with Level 3 of criterion 1 and be rewritten, and (b) Level 4 of criterion 2 should be collapsed with Level 3 of criterion 2 and be rewritten, and (c) Level 4 of criterion 3 should be collapsed with Level 3 of criterion 3 and be rewritten. The tables displaying all psychometric analyses for the Grade 2 Respond MCA include tables 12.A3, 12.B3, 12.C3, 12.D3, 12.E3, 12.F3, 12.G3, and 12.H3. The variable map can be found in figure 12.I3.

GRADE 5 PERFORM

Summary Statistics

Table 12.A4 provides the summary statistics for the MFR-PC model analysis of student work (n = 43), scorers (n = 3), scoring type (n = 2) and criteria (n = 12). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(42)}$ = 356.80, p< .01), scorers ($\chi^2_{(2)}$ = 99.00, p< .01), and criteria ($\chi^2_{(11)}$ = 172.40, p< .01). Non-significance was found for score type ($\chi^2_{(1)}$ = 2.20, p = .55). Reliability of separation statistics for the three significant facets are as follows: student work (*Rel* = .85), scorers (*Rel* = .97), and criteria (*Rel* = .94). See figure 12.I4 for the variable map.

Calibration of Student Work

Table 12.B4 provides student work calibration information. The mean of the student work was –0.12 logits with a range of 3.46 logits for the highest fit achieving student work (Student ID 3058) to –3.16 logits for the lowest fit achieving (Student ID 5132).

Calibration of Scorers

Table 12.C4 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.71 for the most severe scorer (Scorer 1) to –0.97 for the most lenient scorer (Scorer 3). All scorers demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Traits

Table 12.E3 provides trait calibration information. The mean of the traits was 0.00 logits with a range of 0.59 for the most difficult trait (performance) to –0.91 for the least difficult trait (music selection). All traits demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F4 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 1.05 for the most difficult criterion (expressive quality) to –2.36 for the easiest criterion (P interest). All criteria demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G4. Three steps were taken to evaluate the rating-scale-level structure. First frequency counts for each level were evaluated for usage under 10%. Results indicated that eight levels did not reach the prescribed 10% usage: (a) Level 1 of criterion 1, (b) Level 1 of criterion 2, (c) Level 1 of criterion 10, (d) Level 1 of criterion 11, (e) Level 1 of criterion 12, (f) Level 4 of criterion 3, (g) Level 4 of criterion 5, and (h) Level 4 of criterion 6. It is therefore recommended that each Level 1 be collapsed into its respective adjacent Level 2 and be rewritten, and each Level 4 of criterion be collapsed into its adjacent Level 3 and be rewritten. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that one level was found to have an MSE value ≥2.0: Level 2 of criterion 4. It is therefore recommended that this level be examined, collapsed into Level 1, and rewritten. Third, average observed logit measures were examined for violations of monotonicity. Results indicated no violations of monotonicity.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H4 for each criterion within the respective trait.

Findings

The Grade 5 Perform MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the

considered parameters. Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) expressive quality, (2) accuracy, (3) technical ability, (4) quality of interpretation, (5) appropriateness, (6) context, (7) considerations of personal performance, (8) analyze, (9) consideration of feedback, (10) performance decorum, (11) consideration of teacher-provided criteria and peer feedback, and (12) personal interest. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Perform* construct for fifth grade. An analysis of the usability and meaningfulness of the levels across each criterion (i.e., emerging, approaches standard, meets standard, exceeds standard) indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G4. Based on the analysis, the following is recommended: (a) Level 1 of criterion 1 should be collapsed with Level 2 of criterion 1 and be rewritten, (b) Level 1 of criterion 2 should be collapsed with Level 2 of criterion 2 and be rewritten, (c) Level 1 of criterion 10 should be collapsed with Level 2 of criterion 10 and be rewritten, (d) Level 1 of criterion 11 should be collapsed with Level 2 of criterion 11 and be rewritten, (e) Level 1 of criterion 12 should be collapsed with Level 2 of criterion 12 and be rewritten, (f) Level 4 of criterion 3 should be collapsed with Level 3 of criterion 3 and be rewritten, (g) Level 4 of criterion 5 should be collapsed with Level 3 of criterion 5 and be rewritten, and (h) Level 4 of criterion 6 should be collapsed with Level 3 of criterion 6 and be rewritten. The tables displaying all psychometric analyses for the Grade 5 Respond MCA include tables 12.A4, 12.B4, 12.C4, 12.D4, 12.E4, 12.F4, 12.G4, and 12.H4. The variable map can be found in figure I4.

GRADE 5 RESPOND MCA

Summary Statistics

Table 12.A5 provides the summary statistics for the MFR-PC model analysis of student work (n = 110), scorers (n = 6), scoring type (n = 2) and criteria (n = 5). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(109)} = 587.10$, $p < .01$), scorers ($\chi^2_{(5)} = 30.90$, $p < .01$), scoring type ($\chi^2_{(1)} = 48.10$, $p < .01$), and criteria ($\chi^2_{(4)} = 64.80$, $p < .01$). Reliability of separation statistics for all four facets are as follows: student work (*Rel* = .83), scorers (*Rel* = .78), scoring type (*Rel* = .96), and criteria (*Rel* = .92). See figure 12.I5 for the variable map.

Calibration of Student Work

Table 12.B5 provides student work calibration information. The mean of the student work was –0.98 logits with a range of 3.41 logits for the

highest fit achieving student work (Student ID 3229) to –5.36 logits for the lowest fit achieving (Student ID 4331).

Calibration of Scorers

Table 12.C5 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.26 for the most severe scorer (Scorer 3) to –0.63 for the most lenient scorer (Scorer 4). All scorers demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Scoring Type

Table 12.D5 provides scoring-type calibration information. The mean of the scoring-type facet was 0.00 logits with a range of 0.38 for the most severe scoring type (peer score) to –0.38 for the most lenient scoring type (self score). Both scoring types demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F5 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 0.64 for the most difficult criterion (selecting best representation Q2) to –0.56 for the easiest criterion (interpreting qualities Q1). All criteria demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G5. Three steps were taken in order to evaluate the rating-scale-level structure: First, frequency counts for each level were evaluated for usage under 10%. Results indicated that one level did not reach the prescribed 10% usage: Level 4 of criterion 3. It is therefore recommended that Level 4 be collapsed into its respective adjacent Level 3 and be rewritten. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that no levels were found to have an MSE value ≥2.0. Third, average observed logit measures were examined for violations of monotonicity. Results indicated no violations of monotonicity.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H5 for each criterion within the respective trait.

Findings

The Grade 5 Respond MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the considered parameters. Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) selecting best representation Q2, (2) reflection Q2, (3) interpreting qualities Q3, (4) connections Q4, and (5) interpreting qualities Q1. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Respond* construct for fifth grade. An analysis of the usability and meaningfulness of the levels across each criterion (i.e., emerging, approaches standard, meets standard, exceeds standard) indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G5. Based on the analysis, the following is recommended: Level 4 should be collapsed with Level 3 of criterion 3 and be rewritten. The tables displaying all psychometric analyses for the Grade 5 Respond MCA include tables 12.A5, 12.B5, 12.C5, 12.D5, 12.E5, 12.F5, 12.G5, and 12.H5. The variable map can be found in figure 12.I5.

GRADE 8 CREATE

Summary Statistics

Table 12.A6 provides the summary statistics for the MFR-PC model analysis of student work ($n = 40$), scorers ($n = 2$), scoring type ($n = 2$) and criteria ($n = 8$). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(39)} = 204.90$, $p < .01$), scorers ($\chi^2_{(1)} = 30.60$, $p < .01$), scoring type ($\chi^2_{(1)} = 4.50$, $p < .01$), and criteria ($\chi^2_{(7)} = 111.20$, $p < .01$). Reliability of separation statistics for all four facets are as follows: student work (*Rel* = .83), scorers (*Rel* = .93), scoring type (*Rel* = .55), and criteria (*Rel* = 93). See figure 12.I6 for the variable map.

Calibration of Student Work

Table 12.B6 provides student work calibration information. The mean of the student work was –0.71 logits with a range of 1.74 logits for the highest fit achieving student work (Student ID 4518) to –3.77 logits for the lowest fit achieving (Student ID 3530).

Calibration of Scorers

Table 12.C6 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.61 for the most severe scorer (Scorer 1) to –0.61 for the most lenient scorer (Scorer 2). All scorers

demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Scoring Type

Table 12.D6 provides scoring-type calibration information. The mean of the scoring-type facet was 0.00 logits with a range of 0.41 for the most severe scoring type (peer score) to –0.41 for the most lenient scoring type (self score). Both scoring types demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F6 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 1.14 for the most difficult criterion (applying criteria) to –1.48 for the easiest criterion (craftsmanship). All criteria demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G6. Three steps were taken in order to evaluate the rating-scale-level structure: First, frequency counts for each level were evaluated for usage under 10%. Results indicated that eight levels did not reach the prescribed 10% usage: (a) Level 1 of criterion 1, (b) Level 1 of criterion 2, (c) Level 1 of criterion 8, (d) Level 4 of criterion 3, (e) Level 4 of criterion 4, (f) Level 4 of criterion 5, (g) Level 4 of criterion 6, and (h) Level 4 of criterion 7. It is therefore recommended that each Level 1 be collapsed into its respective adjacent Level 2 and be rewritten, and each Level 4 of criterion be collapsed into its adjacent Level 3 and be rewritten. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that one level was found to have an MSE value ≥2.0: Level 1 of criterion 1. Because Level 1 of criterion 1 additionally did not reach the minimum level usage, it is recommended this level be collapsed into Level 2. Third, average observed logit measures were examined for violations of monotonicity. Results indicated one level that violated the rule of monotonicity: Level 2 of criterion 1. It is recommended to collapse Level 2 into the adjacent Level 1.

Inter-Adjacent-Level Discrimination Indices

Andrich thresholds can be found in table 12.H6 for each criterion within the respective trait.

Findings

The Grade 8 Create MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the considered parameters. Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) applying criteria, (2) rational for refinement, (3) evaluation, (4) music ideas, (5) expressive intent from the Imagine Plan and Make Worksheet, (6) effective crafting, (7) expressive intent from the composition scoring device, and (8) craftsmanship. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Create* construct for eighth grade. An analysis of the usability and meaningfulness of the levels across each criterion (i.e., emerging, approaches standard, meets standard, exceeds standard) indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G6. Based on the analysis, the following is recommended: (a) Level 1 of criterion 1 should be collapsed with Level 2 of criterion 1 and be rewritten, (b) Level 1 of criterion 2 should be collapsed with Level 2 of criterion 2 and be rewritten, (c) Level 1 of criterion 8 should be collapsed with Level 2 of criterion 8 and be rewritten, (d) Level 4 of criterion 3 should be collapsed with Level 3 of criterion 3 and be rewritten, (e) Level 4 of criterion 4 should be collapsed with Level 3 of criterion 4 and be rewritten, (f) Level 4 of criterion 5 should be collapsed with Level 3 of criterion 5 and be rewritten, (g) Level 4 of criterion 6 should be collapsed with Level 3 of criterion 6 and be rewritten, and (h) Level 4 of criterion 7 should be collapsed with Level 3 of criterion 7 and be rewritten, (i) Level 2 of criterion 1 should be collapsed with Level 1 of criterion 1 and be rewritten. The tables displaying all psychometric analyses for the Grade 8 Create MCA include tables 12.A6, 12.B6, 12.C6, 12.D6, 12.E6, 12.F6, 12.G6, and 12.H6. The variable map can be found in figure 12.I6.

COMPOSITION/THEORY: CREATE, RESPOND, PERFORM—PROFICIENT

Summary Statistics

Table 12.A7 provides the summary statistics for the MFR-PC model analysis of student work ($n = 40$), scorers ($n = 3$), scoring type ($n = 2$) and criteria ($n = 12$). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(86)} = 225.20$, $p < .01$), scorers ($\chi^2_{(2)} = 72.80$, $p < .01$), scoring type ($\chi^2_{(1)} = 1.70$, $p < .01$), and criteria ($\chi^2_{(11)} = 153.50$, $p < .01$). Reliability of separation statistics for all four facets are as follows: student work (*Rel* = .86), scorers (*Rel* = .77), scoring type (*Rel*<.01), and criteria (*Rel* = .93). See figure 12.I7 for the variable map.

Calibration of Student Work

Table 12.B7 provides student work calibration information. The mean of the student work was 0.20 logits with a range of 3.86 logits for the highest fit achieving student work (Student ID 5075) to –2.42 logits for the lowest fit achieving (Student ID 5096).

Calibration of Scorers

Table 12.C7 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.87 for the most severe scorer (Scorer 2) to –0.95 for the most lenient scorer (Scorer 1). All scorers demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Scoring Type

Table 12.D7 provides scoring-type calibration information. The mean of the scoring-type facet was 0.00 logits with a range of 0.14 for the most severe scoring type (peer score) to –0.14 for the most lenient scoring type (self score). Both scoring types demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Traits

Table 12.E4 provides trait calibration information. The mean of the traits was 0.00 logits with a range of 1.38 for the most difficult trait (imagine) to –0.27 for the least difficult trait (process). All traits demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F7 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 1.38 for the most difficult criterion (imagine) to –1.47 for the easiest criterion (strategies for improvement). The recognizability criterion and the feedback for refinement criterion both demonstrated sporadic rating patterns as evidenced by an infit MSE greater than 1.40. Therefore, it is recommended that both criteria should be either rewritten and further tested or removed.

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G7. Three steps were taken in order to evaluate the rating-scale-level structure: First, frequency

counts for each level were evaluated for usage under 10%. Results indicated that 14 levels did not reach the prescribed 10% usage: (a) Level 1 of criterion 2, (b) Level 1 of criterion 4, (c) Level 1 of criterion 5, (d) Level 1 of criterion 7, (e) Level 1 of criterion 8, (f) Level 1 of criterion 10, (g) Level 1 of criterion 11, (h) Level 1 of criterion 12, (i) Level 2 of criterion 11, (j) Level 4 of criterion 1, (k) Level 4 of criterion 6, (l) Level 4 of criterion 7, (m) Level 4 of criterion 8, and (n) Level 4 of criterion 9. It is therefore recommended that each Level 1 be collapsed into its respective adjacent Level 2 and be rewritten, and each Level 4 of criterion be collapsed into its adjacent Level 3 and be rewritten. In the case of criterion 11 where both Levels 1 and 2 demonstrated insufficient usage, the collapsing of both levels into each other would render sufficient usage. However, considering the large percentages of their respective Levels 3 and 4, it is recommended to consider removing both Levels 1 and 2. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that one level was found to have an MSE value ≥2.0: Level 1 of criterion 4. Because Level 1 of criterion 4 additionally did not reach the minimum level usage, it is recommended this level be collapsed into Level 2 and be rewritten. Third, average observed logit measures were examined for violations of monotonicity. Results indicated two levels that violated the rule of monotonicity: (a) Level 2 of criterion 6, and (b) Level 2 of criterion 11. It is recommended that both levels be collapsed into their adjacent Level 1 and be rewritten.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H7 for each criterion within the respective trait.

Findings

The composition/theory MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the considered parameters. Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) recognizability, (2) imagine, (3) evaluation of T&E, (4) analysis from the responding scoring device, (5) interpretation, (6) analysis from the plan, make, and analyze scoring device, (7) craftsmanship, (8) organization, (9) selection, (10) verbal, (11) strategies for improvement, and (12) feedback for refinement. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *composition/theory construct*. An analysis of the usability and meaningfulness of the levels across each criterion (i.e., emerging, approaches standard, meets standard, exceeds standard) indicated some evidence for suggested revisions. The

labels associated with each criterion are found in table 12.G7. Based on the analysis, the following is recommended: (a) Level 1 of criterion 2 should be collapsed with Level 2 of criterion 2 and be rewritten, (b) Level 1 of criterion 4 should be collapsed with Level 2 of criterion 4 and be rewritten, (c) Level 1 of criterion 5 should be collapsed with Level 2 of criterion 5 and be rewritten, (d) Level 1 of criterion 7 should be collapsed with Level 2 of criterion 7 and be rewritten, (e) Level 1 of criterion 8 should be collapsed with Level 2 of criterion 8 and be rewritten, (f) Level 1 of criterion 10 should be collapsed with Level 2 of criterion 10 and be rewritten, (g) Level 1 of criterion 11 should be collapsed with Level 2 of criterion 11 and be rewritten, (h) Level 1 of criterion 12 should be collapsed with Level 2 of criterion 12 and be rewritten, (i) Level 2 of criterion 11 should be collapsed with Level 1 of criterion 11 and be rewritten, (j) Level 4 of criterion 1 should be collapsed with Level 3 of criterion 1 and be rewritten, (k) Level 4 of criterion 6 should be collapsed with Level 3 of criterion 6 and be rewritten, (l) Level 4 of criterion 7 should be collapsed with Level 3 of criterion 7 and be rewritten, (m) Level 4 of criterion 8 should be collapsed with Level 3 of criterion 8 and be rewritten, (n) Level 4 of criterion 9 should be collapsed with Level 3 of criterion 9 and be rewritten, and (o) Level 2 of criterion 6 should be collapsed with Level 1 of criterion 6 and be rewritten. The tables displaying all psychometric analyses for the composition/theory MCA include tables 12.A7, 12.B7, 12.C7, 12.D7, 12.E4, 12.F7, 12.G7, and 12.H7. The variable map can be found in figure 12.I7.

ENSEMBLE PERFORM (INTERMEDIATE)

Summary Statistics

Table 12.A8 provides the summary statistics for the MFR-PC model analysis of student work (n = 137), scorers (n = 6), scoring type (n = 2) and criteria (n = 11). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(136)} = 685.50$, $p < .01$), scorers ($\chi^2_{(5)} = 55.40$, $p < .01$), scoring type ($\chi^2_{(1)} = 7.90$, $p < .01$), and criteria ($\chi^2_{(10)} = 203.40$, $p < .01$). Reliability of separation statistics for all four facets are as follows: student work (*Rel* = .82), scorers (*Rel* = .88), scoring type (*Rel* = .75), and criteria (*Rel* = .95). See figure 12.I8 for the variable map.

Calibration of Student Work

Table 12.B8 provides student work calibration information. The mean of the student work was –0.89 logits with a range of 1.44 logits for the highest fit achieving student work (Student ID 2416) to –4.84 logits for the lowest fit achieving (Student ID 2571).

Calibration of Scorers

Table 12.C8 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.50 for the most severe scorer (Scorer 4) to –0.46 for the most lenient scorer (Scorer 1). All scorers demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Scoring Type

Table 12.D8 provides scoring-type calibration information. The mean of the scoring-type facet was 0.00 logits with a range of 0.12 for the most severe scoring type (peer score) to –0.12 for the most lenient scoring type (self score). Both scoring types demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Traits

Table 12.E5 provides trait calibration information. The mean of the traits was 0.00 logits, with a range of 0.27 for the most difficult trait (rehearsal) to –0.31 for the least difficult trait (perform). All traits demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F8 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 0.79 for the most difficult criterion (selection of varied program) to –0.98 for the easiest criterion (awareness of Exp Qual). All criteria demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G8. Three steps were taken in order to evaluate the rating-scale-level structure: First, frequency counts for each level were evaluated for usage under 10%. Results indicated that 10 levels did not reach the prescribed 10% usage: (a) Level 4 of criterion 1, (b) Level 4 of criterion 2, (c) Level 4 of criterion 3, (d) Level 4 of criterion 4, (e) Level 4 of criterion 5, (f) Level 4 of criterion 6, (g) Level 4 of criterion 7, (h) Level 4 of criterion 9, (i) Level 4 of criterion 10, and (j) Level 4 of criterion 11. It is therefore recommended that each Level 4 be collapsed into its respective adjacent Level 3 and be

rewritten. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that no levels were found to have an MSE value ≥2.0. Third, average observed logit measures were examined for violations of monotonicity. Results indicated one level that violated the rule of monotonicity: it is recommended that Level 4 criterion be collapsed into its adjacent Level 3 and be rewritten.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H8 for each criterion within the respective trait.

Findings

The ensemble perform (intermediate) MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the considered parameters. Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) selection of varied program, (2) awareness of technical challenges, (3) evaluate/refine, (4) interpretation, (5) expressive qualities, (6) analysis, (7) rehearsal plan, (8) rhythm/pulse accuracy, (9) tone production, (10) pitch/intonation accuracy, and (11) awareness of expressive qualities. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Intermediate Ensemble Perform* construct. An analysis of the usability and meaningfulness of the levels across each criterion (i.e., emerging, approaches standard, meets standard, exceeds standard) indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G8. Based on the analysis, the following is recommended: (a) Level 4 of criterion 1 should be collapsed with Level 3 of criterion 1 and be rewritten, (b) Level 4 of criterion 2 should be collapsed with Level 3 of criterion 2 and be rewritten, (c) Level 4 of criterion 3 should be collapsed with Level 3 of criterion 3 and be rewritten, (d) Level 4 of criterion 4 should be collapsed with Level 3 of criterion 4 and be rewritten, (e) Level 4 of criterion 5 should be collapsed with Level 3 of criterion 5 and be rewritten, (f) Level 4 of criterion 6 should be collapsed with Level 3 of criterion 6 and be rewritten, (g) Level 4 of criterion 7 should be collapsed with Level 3 of criterion 7 and be rewritten, (h) Level 4 of criterion 9 should be collapsed with Level 3 of criterion 9 and be rewritten, (i) Level 4 of criterion 10 should be collapsed with Level 3 of criterion 10 and be rewritten, and (j) Level 4 of criterion 11 should be collapsed with Level 3 of criterion 11 and be rewritten. The tables displaying all psychometric analyses for the ensemble perform (intermediate) MCA include tables 12.A8, 12.B8, 12.C8,

12.D8, 12.E5, 12.F8, 12.G8, and 12.H8. The variable map can be found in figure 12.I8.

ENSEMBLE PERFORM (PROFICIENT)

Summary Statistics

Table 12.A9 provides the summary statistics for the MFR-PC model analysis of student work (n = 84), scorers (n = 3), scoring type (n = 2) and criteria (n = 11). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(83)}$ = 586.20, p< .01), scorers ($\chi^2_{(5)}$ = 298.40, p< .01), scoring type ($\chi^2_{(1)}$ = 115.10, p< .01), and criteria ($\chi^2_{(10)}$ = 210.10, p< .01). Reliability of separation statistics for all four facets are as follows: student work (*Rel* = .87), scorers (*Rel* = .99), scoring type (*Rel* = .90), and criteria (*Rel* = .95). See figure 12.I9 for the variable map.

Calibration of Student Work

Table 12.B9 provides student work calibration information. The mean of the student work was –1.42 logits with a range of 1.22 logits for the highest fit achieving student work (Student ID 2635) to –3.68 logits for the lowest fit achieving (Student ID 2650).

Calibration of Scorers

Table 12.C9 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.75 for the most severe scorer (Scorer 3) to –1.04 for the most lenient scorer (Scorer 1). All scorers demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Scoring Type

Table 12.D9 provides scoring-type calibration information. The mean of the scoring-type facet was 0.00 logits with a range of 0.48 for the most severe scoring type (peer score) to –0.48 for the most lenient scoring type (self score). Both scoring types demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Traits

Table 12.E6 provides trait calibration information. The mean of the traits was 0.00 logits with a range of 0.18 for the most difficult trait (re-

hearsal) to –0.19 for the least difficult trait (perform). All traits demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F9 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 0.92 for the most difficult criterion (analysis) to –1.43 for the easiest criterion (tone production). The selection of a varied program criterion demonstrated sporadic rating patterns as evidenced by an infit MSE greater than 1.40. Therefore, it is recommended that this criterion be either rewritten and further tested or removed.

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G9. Three steps were taken in order to evaluate the rating-scale-level structure: First, frequency counts for each level were evaluated for usage under 10%. Results indicated that 11 levels did not reach the prescribed 10% usage: (a) Level 4 of criterion 1, (b) Level 4 of criterion 2, (c) Level 4 of criterion 3, (d) Level 4 of criterion 4, (e) Level 4 of criterion 5, (f) Level 4 of criterion 6, (g) Level 4 of criterion 7, (h) Level 4 of criterion 8, (i) Level 4 of criterion 9, and (j) Level 4 of criterion 10, and (k) Level 4 of criterion 11. It is therefore recommended that each Level 4 be collapsed into its respective adjacent Level 3 and be rewritten. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that one level was found to have an MSE value ≥2.0: Level 4 of criterion 1. Because Level 4 of criterion 1 additionally did not reach the minimum level usage, it is recommended this level be collapsed into Level 3 and be rewritten. Third, average observed logit measures were examined for violations of monotonicity. Results indicated one level that violated the rule of monotonicity: Level 4 of criterion 10. It is therefore recommended this level be collapsed into its adjacent Level 3 and be rewritten.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H9 for each criterion within the respective trait.

Findings

The ensemble perform (proficient level) MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter

separation for each of the considered parameters. Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) analysis, (2) pitch/intonation accuracy, (3) evaluate/refine, (4) awareness of expressive qualities, (5) expressive qualities, (6) rehearsal plan, (7) awareness of technical challenges, (8) selection of varied program, (9) rhythm/pulse accuracy, (10) interpretation, and (11) tone production. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Proficient Ensemble Perform* construct. An analysis of the usability and meaningfulness of the levels across each criterion (i.e., emerging, approaches standard, meets standard, Exceeds Standard) indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G9. Based on the analysis, the following is recommended: (a) Level 4 of criterion 1 should be collapsed with Level 3 of criterion 1 and be rewritten, (b) Level 4 of criterion 2 should be collapsed with Level 3 of criterion 2 and be rewritten, (c) Level 4 of criterion 3 should be collapsed with Level 3 of criterion 3 and be rewritten, (d) Level 4 of criterion 4 should be collapsed with Level 3 of criterion 4 and be rewritten, (e) Level 4 of criterion 5 should be collapsed with Level 3 of criterion 5 and be rewritten, (f) Level 4 of criterion 6 should be collapsed with Level 3 of criterion 6 and be rewritten, (g) Level 4 of criterion 7 should be collapsed with Level 3 of criterion 7 and be rewritten, (h) Level 4 of criterion 8 should be collapsed with Level 3 of criterion 8 and be rewritten, (i) Level 4 of criterion 9 should be collapsed with Level 3 of criterion 9 and be rewritten, (j) Level 4 of criterion 10 should be collapsed with Level 3 of criterion 10 and be rewritten, and (l) Level 4 of criterion 11 should be collapsed with Level 3 of criterion 11 and be rewritten. The tables displaying all psychometric analyses for the ensemble perform (proficient) MCA include tables 12.A9, 12.B9, 12.C9, 12.D9, 12.E6, 12.F9, 12.G9, and 12.H9. The variable map can be found in figure 12.I9.

HARMONIZING INSTRUMENTS: CREATE (PROFICIENT LEVEL)

Summary Statistics

Table 12.A10 provides the summary statistics for the MFR-PC model analysis of student work ($n = 20$), scorers ($n = 2$), scoring type ($n = 2$) and criteria ($n = 7$). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(19)} = 46.60$, $p < .01$), scorers ($\chi^2_{(1)} = 48.30$, $p < .01$), scoring type ($\chi^2_{(1)} = 19.20$, $p < .01$), and criteria ($\chi^2_{(6)} = 69.90$, $p < .01$). Reliability of separation statistics for all four facets are as follows: student work (*Rel* = .79), scorers (*Rel* = .96), scoring type (*Rel* = .90), and criteria (*Rel* = .91). See figure 12.I10 for the variable map.

Calibration of Student Work

Table 12.B10 provides student work calibration information. The mean of the student work was 0.07 logits with a range of 4.86 logits for the highest fit achieving student work (Student ID 3517) to –1.66 logits for the lowest fit achieving (Student ID 3479).

Calibration of Scorers

Table 12.C10 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.83 for the most severe scorer (Scorer 2) to –0.83 for the most lenient scorer (Scorer 1). All scorers demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Scoring Type

Table 12.D10 provides scoring-type calibration information. The mean of the scoring-type facet was 0.00 logits with a range of 0.47 for the most severe scoring type (self score) to –0.47 for the most lenient scoring type (peer score). Both scoring types demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Traits

Table 12.E7 provides trait calibration information. The mean of the traits was 0.00 logits with a range of 0.42 for the most difficult trait (plan/make) to –0.99 for the least difficult trait (imagine). All traits demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F10 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 2.06 for the most difficult criterion (development of harmonization) to –1.21 for the easiest criterion (analysis). All criteria demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G10. Three steps were taken in order to evaluate the rating-scale-level structure: First, frequency

counts for each level were evaluated for usage under 10%. Results indicated that three levels did not reach the prescribed 10% usage: (a) Level 1 of criterion 3, (a) Level 4 of criterion 1, and (b) Level 4 of criterion 2. It is therefore recommended that each Level 1 be collapsed into its respective adjacent Level 2 and be rewritten and Level 4 be collapsed into its respective adjacent Level 3 and be rewritten. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that one level was found to have an MSE value ≥2.0: Level 4 of criterion 2. Because Level 4 of criterion 2 additionally did not reach the minimum level usage, it is recommended this level be collapsed into Level 3 and be rewritten. Third, average observed logit measures were examined for violations of monotonicity. Results indicated one level that violated the rule of monotonicity: Level 4 of criterion 2. It is therefore recommended that Level 4 be collapsed into its adjacent Level 3 and be rewritten.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H10 for each criterion within the respective trait.

Findings

The harmonizing instruments MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the considered parameters. Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) development of harmonies, (2) interpretation, (3) craftsmanship, (4) recognition of notation, (5) verbal presentation, 6) imagine, and (7) analysis. All criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Harmonizing Instruments* construct. An analysis of the usability and meaningfulness of the levels across each criterion (i.e., emerging, approaches standard, meets standard, exceeds standard) indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G10. Based on the analysis, the following is recommended: (a) Level 1 of criterion 3 should be collapsed with Level 2 of criterion 3 and be rewritten, (b) Level 4 of criterion 1 should be collapsed with Level 3 of criterion 1 and be rewritten, and (c) Level 4 of criterion 2 should be collapsed with Level 3 of criterion 2 and be rewritten. The tables displaying all psychometric analyses for the harmonizing MCA include tables 12.A10, 12.B10, 12.C10, 12.D10, 12.E7, 12.F10, 12.G10, and 12.H10. The variable map can be found in figure 12.I10.

HARMONIZING INSTRUMENTS: CREATE—PROFICIENT (REVISED)

Summary Statistics

Table 12.A11 provides the summary statistics for the MFR-PC model analysis of student work ($n = 49$), scorers ($n = 4$), scoring type ($n = 2$) and criteria ($n = 6$). The analysis indicated an overall good model data fit with significant differences between student work ($\chi^2_{(48)} = 536.10$, $p < .01$), scorers ($\chi^2_{(3)} = 31.90$, $p < .01$), and criteria ($\chi^2_{(5)} = 154.00$, $p < .01$). Nonsignificance was found for score type ($\chi^2_{(1)} = .30$, $p = .58$). Reliability of separation statistics for the three significant facets are as follows: student work (*Rel* = .89), scorers (*Rel* = .91), and criteria (*Rel* = .95). See figure 12.I11 for the variable map.

Calibration of Student Work

Table 12.B11 provides student work calibration information. The mean of the student work was 0.60 logits with a range of 4.96 logits for the highest fit achieving student work (Student ID 4588) to –6.51 logits for the lowest fit achieving (Student ID 5136).

Calibration of Scorers

Table 12.C11 provides scorer calibration information. The mean of the scorers was 0.00 logits with a range of 0.54 for the most severe scorer (Scorer 3) to –1.30 for the most lenient scorer (Scorer 1). Scorer 4 demonstrated sporadic rating patterns as evidenced by an infit MSE greater than 1.40.

Calibration of Traits

Table 12.E8 provides trait calibration information. The mean of the traits was 0.00 logits with a range of 1.20 for the most difficult trait (plan/make) to –0.80 for the least difficult trait (imagine). All traits demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40).

Calibration of Criteria

Table 12.F11 provides criterion calibration information. The mean of the criteria was 0.00 logits with a range of 2.42 for the most difficult criterion (documentation of harmonization) to –0.99 for the easiest criterion

(recognition of notation). All criteria demonstrated adequate fit to the model based on the reasonable mean-square range for infit and outfit (0.60–1.40). The development of harmonization criterion demonstrated muted rating patterns as evidenced by an infit MSE less than 0.50. Therefore, it is recommended that this criterion be either rewritten and further tested or removed.

Rating-Scale-Level Diagnostics

Rating-scale diagnostics can be found in table 12.G11. Three steps were taken in order to evaluate the rating-scale-level structure. First, frequency counts for each level were evaluated for usage under 10%. Results indicated that eight levels did not reach the prescribed 10% usage: (a) Level 1 of criterion 1, (b) Level 1 of criterion 4, (c) Level 1 of criterion 5, (d) Level 2 of criterion 2, (e) Level 2 of criterion 4, (f) Level 4 of criterion 1, (g) Level 4 of criterion 2, and (h) Level 4 of criterion 3. It is therefore recommended that each Level 1 be collapsed into its respective adjacent Level 2 and be rewritten, each Level 2 be collapsed into its respective adjacent Level 1 and be rewritten, and each Level 4 be collapsed into its respective adjacent Level 3 and be rewritten. In the case of criterion 4 where both Levels 1 and 2 demonstrated insufficient usage, the collapsing of both levels into each other would still not render sufficient usage. Therefore, it is recommended to consider removing both Levels 1 and 2 and rewriting Levels 3 and 4. Second, outfit mean squares (MSE) were examined for values ≥2.0. Results indicated that one level was found to have an MSE value ≥2.0: Level 1 of criterion 3. It is examined that this level be examined and rewritten. Third, average observed logit measures were examined for violations of monotonicity. Results indicated no violations of monotonicity.

Inter-Adjacent-Level Discrimination Indices

Rasch-Andrich thresholds can be found in table 12.H11 for each criterion within the respective trait.

Findings

The harmonizing instruments (revised) MCA demonstrated overall strong construct validity as demonstrated by the reasonable parameter separation for each of the considered parameters. Results indicated that the rank order of criteria by difficulty (from most difficult to least difficult) was (1) documentation of harmonization, (2) development of harmonization, (3) melodic interpretation, (4) feedback, (5) imagine, and (6) recognition of notation. The Development of Harmonization criterion

was found to be too predictable, therefore not contributing meaningful information toward the measurement of student work. Therefore, it is recommended that this criterion be either rewritten and further tested or removed. All other criteria were found to be appropriate and meaningful in their overall functioning within the context of measuring the *Harmonizing Instruments* (revised) construct. An analysis of the usability and meaningfulness of the levels across each criterion (i.e., emerging, approaches standard, meets standard, exceeds standard) indicated some evidence for suggested revisions. The labels associated with each criterion are found in table 12.G11. Based on the analysis, the following is recommended: (a) Level 1 of criterion 1 should be collapsed with Level 2 of criterion 1 and be rewritten, (b) Level 1 of criterion 4 should be collapsed with Level 2 of criterion 4 and be rewritten, (c) Level 1 of criterion 5 should be collapsed with Level 2 of criterion 5 and be rewritten, (d) Level 2 of criterion 2 should be collapsed with Level 1 of criterion 2 and be rewritten, (e) Level 2 of criterion 4 should be collapsed with Level 1 of criterion 4 and be rewritten, (f) Level 4 of criterion 1 should be collapsed with Level 3 of criterion 1 and be rewritten, (g) Level 4 of criterion 2 should be collapsed with Level 3 of criterion 2 and be rewritten, and (h) Level 4 of criterion 3 should be collapsed with Level 3 of criterion 3 and be rewritten. The tables displaying all psychometric analyses for the Harmonizing (revised) MCA include tables 12.A11, 12.B11, 12.C11, 12.D11, 12.E8, 12.F11, 12.G11, and 12.H11. The variable map can be found in figure 12.I11.

REFERENCE

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.