Old Wine in New Bottles: A Methodology for Developing and Validating Performance Measures Using Modern Measurement Theory

Dorothy J. Musselwhite, The University of Georgia, USA
Brian C. Wesolowski, The University of Georgia, USA

Abstract

In the development of a measure to assess music performance in the United States, the method of data analysis is most often factor analysis. However, Rasch Measurement Theory is a branch of item response theory that is underscored by properties of invariance using a fixed model across independent items, persons, and raters. It is because of the properties of invariance that Rasch Measurement Theory is the preferred method for the development of measures in the context of performance assessment. The purpose of this paper is to provide a clearly defined, thirteen-step methodology for developing and validating music performance measures using Rasch Measurement Theory.

Introduction

In the development of a measure to assess music performance, the traditional method of data analysis is factor analysis (Miksza, 2012; Russell, 2010; Ten Holt, et al., 2010; Smith, 2009; Smith & Barnes, 2007; Zdzinski & Barnes, 2002; Nichols, 1991; Brand, 1985, for example). Factor analysis is a statistical method rooted in the Classical Test Theory tradition with the purpose of describing the variability among correlated variables, using raw scores and covariance matrices. Raw scores are not indicative of measurement because they are not linear, additive, or unidimensional (Wright & Stone, 1999). The use of factor analysis is an acceptable data analysis method for very specific purposes (for example, when interest is in reducing data while also defining latent variable).

Rasch Measurement Theory (RMT) is a method of analysis that offers valid measures that, when developed, are independent from the sample used (Granger, 2008). RMT is a branch of item response theory that is underscored by properties of invariance using a fixed model across independent items, persons, and raters. Due to the requirements of invariance, students' level of achievement, items' level of difficulty, and in the context of performance evaluation, rater severity, will not affect the overall model. It is because of the properties of

invariance that RMT is the preferred method for the development of measures in the context of performance assessment. In particular, there are five requirements for invariant measurement (Engelhard & Perkins, 2011): (a) the calibration of the items must be independent of the particular persons used for calibration; (b) any person must have a better chance of success on an easy item than on a more difficult item; (c) the measurement of persons must be independent of the particular items that happen to be used for the measuring; (d) a more able person must always have a better chance of success on any item than a less able person; and (e) items must be measuring a single underlying latent variable. Specifically, this paper addresses the development of performance assessments where raters are used to gather data. Invariance is a property that is defined by empirical data, specifically model-data fit. With the inclusion of raters, rater-invariant measurement must also be determined. This property implies that persons and raters are independent (Wind & Engelhard, 2013).

The process of constructing a measure for music performance should be guided by two underlying questions:

1. How can raw score data be collected from raters in a valid and meaningful way?
2. How can test construction and development be handled in order to make inferences that are valid, reliable, and fair?

The thirteen-step methodology described in this paper provides a framework for developing measures in the context of music performance. Due to the limitations of the length of this paper, this methodology should be considered a basic framework. Throughout the paper, aspects of decision making will be addressed in relation to the process of test construction and development. The purpose of this paper is to provide a clearly defined methodology for developing and validating music performance measures using Rasch Measurement Theory.

Method

Step 1: Observational Design

The observational design refers to the content and design of the items. The researcher must envision the construct he/she wants to build, then think about the items that would best describe that construct. After consulting subject matter experts and various pedagogical and methodological resources, item construction can begin, and qualitatively grouped into a priori domains. These domains and related items become the framework for the measure.

Step 2: Decide between using a Rating Scale or Rubric

There are two types of preferred response formats: rating scales or rubrics. The researcher may choose to use ether either a rating scale response format or rubric-based response format based upon needs and requirements of the assessment context (time, detail, test requirements, requirements of stakeholders, for example). By choosing a rubric-based response from the start, more work is required up front. In a rubric, categories of performance are listed (i.e., tone, articulation, posture) with accompanying levels of performance. Also important is the terminology used across the categories within the rubric. The language used in each category must be consistent. For example, the type of language to address tone could be level of desirability (e.g., very undesirable, undesirable, desirable, and very desirable). The type of language to address appropriate use of articulation could be level of acceptability (e.g., totally unacceptable, slightly unacceptable, slightly acceptable, and perfectly acceptable) (see Vagias, 2006). It is preferred for there to be between three and five levels at most (Dumas, 1999; Wright, 1977).

A rating scale is different in that a statement is given to the rater, then the rater must decide the level of agreement based on the performance (i.e., strongly agree, agree, disagree, and strongly disagree). The previously discussed domains and subsequent items can be paired with a Likert-type scale in order to meaningfully design a rating scale structure. While a five-category Likert scale is most common, it does not provide meaningful feedback. By removing a middle category (i.e. undecided, neutral), the rater is forced to make a choice, and the researcher is provided with a more accurate picture of the performance (Cox, 1980).

Step 3: Design a Judging Plan

Raters may be organized in a variety of ways. The type of linking design chosen will have an effect on the amount of information and related standard error of the assessment context (Wind, Engelhard, & Wesolowski, 2016). Rater variability is a necessary component in the development of a measure, as multiple perspectives only serve to improve the validity of the measurement instrument (Wilson, 2005). In a complete linking design, every judge or rater will evaluate every performance. While this is the most reliable of the linking designs, a complete system has drawbacks. There may be an increased cost due to the workload of every rater having to evaluate every performance. Potentially, raters may drop out due to time and energy constraints. In addition, the time requirement could impact consistency among raters. In an incomplete design, there are more raters, more performances to evaluate, but more information will be provided by the design. Here, all raters are involved, but they will not evaluate every performance. There are multiple incomplete designs. For

example, Rater 1 will evaluate performances 1, 2, 3, and 4. Rater 2 will evaluate performances 3, 4, 5, and 6. Every performance will be judged by at least two raters, but raters are not having to spend a large amount of time rating.

Step 4: Collect Rater Data

Using the predetermined items and specified rating design, the researcher must develop a pilot measure to conduct with a sample group. This is the first time the raters are interacting with the measure. The raters must be instructed as to word choice, meaning, and the overall operational procedure of the performance assessment. Data must then be collected in a systematic way.

Step 5: Analyze the Data

Two models may be considered based on the qualitative decision-making of the researcher: The Rating-Scale Model (RSM) (Wright & Masters, 1982) or the Partial-Credit Model (PCM) (Masters, 1982). Linacre (2000) outlines the decision-making process between the Rating-Scale Model and the Partial-Credit (Linacre, 2000):

1. Design of the items: If the items are clearly intended to use the same rating scale throughout (e.g., a Likert-type scale), then the RSM should be used. If each item is intended to have a different rating scale, then the PCM should be used.
2. Communication: Each item should match the response-options. A question/item that merits a yes/no response should not be followed by four Likert-scale responses.
3. Size of the dataset: There should be at least 10 observations in each category. This will prevent accidents in the data. However, if the sample size does not allow for 10 observations per category, the RSM should be considered over the PCM.
4. Construct and Predictive Validity: If there is a meaningful difference between the item abilities and between the person abilities, the PCM should be used.
5. Fit Considerations: Underfit is a greater threat to validity than overfit. It is imperative to not only examine parameter-level fit statistics in addition to the fit statistics for each element. If the fit is poor, then better data is needed for the intended purposes. This is not an indication of the need for a better model (see Step 6).
6. Category Thresholds: In the PCM, category thresholds (i.e., step difficulties between rating scale categories) are unknown before data collection. In the RSM, the thresholds are set in advance.

7. Unobserved categories: In the PCM, unused categories will distort the structure of the rating scale. When there is an unobserved category in the RSM, its function is inferred from other items that employ the same category.
8. Statistical information: Both the PCM and RSM provide the same statistical information, therefore there is no benefit of choosing one over the other in this regard.
9. Optimization: Optimization refers to a process where careful examination of the items will lead to more effective use of the rating scale structure (see Step 10). Specifically, categories may need to be collapsed in order to achieve a lower standard error. In order to optimize the rating scale structure, the PCM should be used (Linacre, 2000).

Step 6: Evaluate Parameter-Level Fit Statistics

Parameter-level fit statistics will help determine overall how the components are working. A parameter refers to a measurable factor that is essential to understanding a set of data. Parameters may include items, persons, and raters. Specifically, in the context of music performance assessment, parameter-level statistics look at the student performances, items, and raters within the music performance assessment to see how these components are performing in the model. The range of reasonable mean-square fit values can change depending on the context of assessment (Wright & Linacre, 1994). As an example, there are five contexts: (a) high stakes, (b) run of the mill, (c) survey, (d) clinical observation, and (e) judged test, where agreement is encouraged. The choice of fit statistic thresholds is a qualitative decision.

Fit statistics (e.g., infit and outfit) describe the degree to which invariant measurement is achieved. Infit Mean Squares refers to data fit that is sensitive to inliers (Linacre, 2002). This statistic focuses on the configuration of responses to items aimed on the person. Outfit Mean Squares refers to data fit being outlier-sensitive. This statistic looks at any data that may lie far from the person and looks at what may affect the patterning of responses. Mean squares show how much randomness occurs in the specified data set. The expected mean square error statistics should be close to 1.00 with very little variation within the linear scale, usually a standard deviation of 0.20 at the most. Infit problems can be seen as a bigger threat to measurement, and therefore should be evaluated first (Linacre, 2002). For example, if the infit and outfit statistics fall within the range of 0.80-1.20, it can be concluded that the data demonstrates acceptable levels of invariance for that context. If the infit and outfit statistics fall outside the range of 0.80-1.20, it can be concluded that the data demonstrates unacceptable levels of invariance for that context should be qualitatively evaluated as to how the parameter can be improved (Wright & Linacre, 1994).

Step 7: Evaluate Fit Statistics for Elements

After fit statistics for the entire measure have been obtained, each facet can be examined to determine individual fit statistics for each element. An element is the individual component of the parameter. For example, in the item parameter, a specific item would be an element. The same thresholds from the parameter level will hold true within each element. The focus here, however, is on individual items, individual performances, and/or individual raters. The evaluation of fit statistics at the element level provides important diagnostic and qualitative information on how individual performers, items, and raters performed within the model.

Step 8: Manage Misfit

Misfit, quantitatively, means the item, student, or rater lies outside of the specified threshold described in Step 6 and Step 7. Misfit analysis of items should be viewed not as "bad items." Rather, misfit should be valued as an opportunity to learn and investigate. The same is true for misfitting raters and misfitting performances. Misfit should fuel the rewriting of items and draw attention to content and construct validity concerns.

Step 9: Refine the Measure

Misfitting items should either be removed from the measure or rewritten based upon qualitative decision making. Once the items have been removed or rewritten, the items should go through a follow-up pilot test. In the follow-up study, the same considerations for fit should be applied.

Step 10: Evaluate and Optimize the Rating Scale Structure

The structure of the rating scale can be evaluated when specifically using the Partial Credit Model (PCM). Linacre (2002) provides a set of nine guidelines for optimizing this rating scale structure.

1. All items within the rating scale should align with one latent variable.
2. There should be at least 10 observations per rating scale category.
3. There should be a uniform distribution of observations across categories.
4. Average measures advance monotonically with each category.
5. Outfit Mean Squares are less than 2.00.
6. Step calibrations should advance (shows that category usage is regular).
7. The ratings imply measures, and the measures imply ratings.
8. Step difficulties advance by at least 1.4 logits.
9. Step difficulties advance by at most 5.0.

In this step, each item is examined individually to find out how the categories were used. Each item must meet all guidelines listed above in order to justify the use of each individual category within the context of the rating scale structure.

Step 11: Refine the Learning Outcomes

If a rating scale has been used, and the developer wants to transition into a rubric, rewriting of the items is necessary. Item stems should be rephrased without directionality as to resemble outcome criteria. For example, an item worded as, "Student performs excerpt with desired tone," could be rephrased as simply "Tone." This step aids in the process of transitioning from rating scale to rubric.

Step 12: Write Qualitative Descriptors

Each of the four levels of performance should now be written to describe a specific outcome related to the item stem. Using the aforementioned item stems, all scale categories must be represented with similar tone and language (see Vagias, 2006).

Step 13: Begin the Revalidation Process

The rubric should be revalidated using all previous steps. Once the rating scale items have transformed into a rubric, the rubric must once again be tested for reliability and validity in the same manner described above. Assessment contexts can have an effect on performance of a measurement instrument, therefore it is important to consistently be evaluating items, performances, and raters in the context of a performance assessment. A rubric is a living breathing organism that can change based upon objects of measurement, raters, context, standards change.

In a performance-based assessment in psychological sciences (i.e. music), constructs must be defined and inferred through secondary behaviors (i.e., tone, articulation, posture). Music performance can be adequately assessed through the inferences from these secondary behaviors. It is important that in order to make inferences that are valid, reliable and fair, that researchers and educators are using tools that have been well-developed and maintained. The use of a measurement instrument in any context should be closely monitored and evaluated for its properties of invariance.

References

Brand, M. (1985). Development and validation of the home musical environment scale for use at the early elementary level. *Psychology of Music*, *13*(1), 40-48.

Cox, III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*, 407-422.

Dumas, J. (1999). *Usability testing methods: Subjective measures, part II – Measuring attitudes and opinions.* Washington, DC: American Institutes for Research.

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

Granger, C. V. (2008). Rasch analysis is important to understand and use for measurement. *Rasch Measurement Transactions, 21*(3), 1122-1123.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J.M. (2000). Comparing "partial credit models" (PCM) and "rating scale models" (RSM). *Rasch Measurement Transactions, 14*(3), 768.

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

Miksza, P. (2012). The development of a measure of self-regulated practice behavior for beginning and intermediate instrumental music students. *Journal of Research in Music Education, 59*(4), 321-338.

Nichols, J. P. (1991). A factor-analysis approach to the development of a rating scale for snare drum performance. *Dialogue in Instrumental Music Education, 15*, 11-31.

Russell, B. (2010). The development of a guitar performance rating scale using a facet-factorial approach. *Bulletin of the Council for Research in Music Education,* (184), 21-34.

Smith, B. P. & Barnes, G. V. (2007). Development and validation of an orchestra performance rating scale. *Journal of Research in Music Education*, (3), 268.

Smith, D. T. (2009). Development and validation of a rating scale for wind jazz improvisation performance. *Journal of Research in Music Education*, (3), 217.

Ten Holt, J. C., van Duijn, M. J., & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling*, *52*(3), 272-297.

Vagias, W. M. (2006). *Likert-type scale response anchors.* Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson Management.

Wind, S. A. & Engelhard, Jr., G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18*, 278-299.

Wind, S. A., Engelhard, Jr., G. & Wesolowski, B. C. (2016). Exploring the effects of rater linking designs and rater fit on achievement estimates within the context of music performance assessments. *Educational Assessment, 21*(4), 278-299.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97-116.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago, IL: MESA Press.

Wright, B. D. & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE.

Zdzinski, S. F., & Barnes, G. V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, (3), 245.