Article

MUSICÆ SCIENTIÆ

An Examination of Differential Item Functioning in a Rubric to Assess Solo Music Performance

Musicae Scientiae
1–15

DOI: 10.1177/1029864919859928
journals.sagepub.com/home/msx
SAGE

Brian C. Wesolowski
The University of Georgia, USA

Abstract

The purpose of this study was to examine differential item functioning (DIF) in a rubric used to assess middle-school solo and ensemble performances. This study was guided by the following research questions: (a) does measurement equivalence for all items exist when used to measure subgroups of students based on their musical instrument? (b) what patterns of differential item functioning effects exist for items when used to measure subgroups of students based on their musical instrument? and (c) what size of differential item functioning effects exists for items when used to measure subgroups of students based on their musical instrument? In total, 17 adjudicators evaluated 138 middle-school instrumental students (ages 11–13) in the context of a live, formal solo and ensemble performance assessment. Using the Many Facets Rasch Partial Credit measurement model, measurement equivalence for all items did not exist when used to measure subgroups of students based on their musical instrument ($\chi^2_{(252)} = 634.00$, $p < .01$). Of the 252 total pairwise interactions examined between items and instruments, 57 (22.62%) significant interaction terms were identified. Overall, 26 (10.3%) significant interaction terms demonstrated a moderate to large effect ($|\text{DIF}| >/= .63$ logits), nine (3.6%) significant interaction terms demonstrated a slight to moderate effect ($|\text{DIF}| = (.43, .63$ logits)), and 22 (8.7%) significant interaction terms demonstrated a negligible effect ($|\text{DIF}| < .43$ logits). Implications for the fairness of music performance assessments and improved assessment protocols are discussed.

Keywords

Assessment, fairness, measurement error, Rasch, variability

Introduction

Solo and ensemble festivals are an important part of the music education experience for many secondary-level music students across the USA. From a student perspective, participation in adjudicated events, such as solo and ensemble festivals, can increase quality of musicianship, improve motivation, and increase self-efficacy (Austin, 1988; Banister, 1992; Franklin, 1979;

Corresponding author:
Brian C. Wesolowski, Hugh Hodgson School of Music, The University of Georgia, 250 River Road, Athens, GA 30602, USA.
Email: bwes@uga.edu

Howard, 1994; Hurst, 1994). From a music program perspective, the results of student participation in adjudicated events can impact teaching decisions related to curricular content, learning objectives, and educational goals (Abeles, Hoffer, & Klottman, 1994; Crochet, 2006; Howard, 2002). From a community perspective, student achievement in adjudicated events not only affects perceptions of teacher and program quality, but these perceptions are often a catalyst for a program's continued participation with a particular focus on achieving high-level results and related accolades (Boyle, 1992; Kirchhoff, 1988).

The significance and importance of these adjudicated events in the landscape of secondary-level music education yields a broad and significant amount of research investigating the psychometric properties of their scoring systems as well as extraneous variables that may affect scoring outcomes. Historically, these psychometric inquiries most often include reliability considerations using some form of inter- or intra-rater reliability coefficients in the context of true score (i.e., Classical Test Theory) testing traditions (Bergee, 2003; Bergee, 2007; Bergee & McWhirter, 2005; Bergee & Westfall, 2005; Ciorba & Smith, 2009; Conrad, 2003; Fiske, 1975; Fiske, 1983; Hash, 2012; King & Burnsed, 2009; Latimer, Bergee, & Cohen, 2010; Norris & Borst, 2007; Saunders & Holahan, 1997; Silvey, 2009, for example). Investigations into extraneous variables that may have an overall effect on rating results historically fall into two broad categories: performer-centered variables and contextual variables (Wesolowski, Wind, & Engelhard, Jr., 2015). Examples of performer-centered variables include expressive communication via body movement (Davidson, 1993, 1994, 1995; Davidson & Coimbra, 2001); physical attractiveness (Bermingham, 2000; Davidson & Coimbra, 2001; Wapnick, Darrow, Kovacs, & Dalrymple, 1997; Wapnick, Mazza, & Darrow, 1998, 2000); musical expression (Schubert, 2002); stage entrance and facial expressions (Waddell & Williamon, 2017); and performing from memory (Kopiez, Wolf, & Platz, 2017). Examples of contextual variables include the environmental effect on acoustics (Ando, 1988), and social factors (McPherson & Thompson, 1998). Additionally, more recent investigations into rater behavior have suggested that, from a cognitive perspective, variations in audio and visual presentation cues can affect performance assessment outcome scores (Griffiths & Reay, 2018; Krath, Hahn, & Whitney, 2015; Platz & Kopiez, 2012; Tsay, 2015).

Fairness and Differential Item Functioning

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 50) (herein referred to as the *Standards*), fairness is defined as "responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses." More broadly, fairness is a validity issue concerned with the degree to which measurement procedures result in accurate estimates of student ability. Notably, measurement equivalence, or "lack of measurement bias," is one of the four general views of fairness highlighted by the *Standards* (2014, pp. 51–52):

> Characteristics of the test itself that are not related to the construct being measured, or the manner in which the test is used, may sometimes result in different meanings for scores earned by members of different identifiable groups. For example, *differential item functioning* (DIF) is said to occur when equally able test takers differ in their probabilities of answering a test item correctly as a function of group membership.

Differential item functioning (DIF) is a statistical framework that aids in the detection of systematic patterning of responses based on students' subgroup affiliations (Holland & Thayer,

1988). More specifically, DIF is said to exist if students with the same estimated ability level from different group affiliations demonstrate different probabilities of success on a specific item.

In the case of music performance assessment, the type of musical instrument a student plays is an important grouping characteristic that may potentially account for systematic differences in outcome scores and violate the principle of test fairness as prescribed by the *Standards*. Each musical instrument comes with its own unique set of foundational skills, problems, and challenges. As Colwell and Hewitt (2011, p. 1) note:

> Mixed-instrument classrooms present unique challenges . . . many of which are unique to each instrument. Poor hand position on the clarinet is not poor hand position on the violin. Carefully monitoring each student on all of the foundational skills . . . requires knowledge of each instrument.

Duke and Byo (2011, p. 9–10) acknowledge that "Double reeds, horns, tubas . . . all present unique problems . . . double reeds are overly finicky for a variety of reasons (e.g., reeds, instrument adjustments) . . . horn . . . is problematic because of its register . . . Tubas simply require a great deal of physical capacity . . ."

In many solo and ensemble performance assessment contexts, the same evaluation/scoring form is used for all students, regardless of the instrument they are playing. Furthermore, students of heterogeneous instrument types are evaluated on the same performance standard of superior, excellent, good, fair, or poor (National Association for Music Education, 2016). From a feedback and evaluation perspective, Duke and Byo (2011, p. 4) noted, "Consider that, in all likelihood . . . feedback about a group's performance is almost always the wrong feedback for one or more members of the group." With these considerations in mind, systematic differences in estimates of student ability, due to the interaction between item difficulties and musical instrument, may exist specifically because of the unique performance characteristics and challenges associated with specified musical instruments. Therefore, an investigation into possible DIF effects related to musical instrument is warranted.

In the case of music performance assessment research, DIF related to students' musical instrument grouping has not been examined in the music assessment research literature. The purpose of this study was to examine DIF in a rubric used to assess middle-school solo and ensemble performances. In particular, this study was guided by the following research questions:

1. Does measurement equivalence for all items exist when used to measure subgroups of students based upon their musical instrument?
2. What patterns of DIF effects exist for items when used to measure subgroups of students based on their musical instrument?
3. What size of DIF effect exist for items when used to measure subgroups of students based on their musical instrument?

Method

Assessment Context, Participants, and Raters

The assessment context in this study was a district-level middle-school solo and ensemble festival in the southern part of the USA. A total of 138 middle school instrumental students (ranging from grades 6–8, ages 11–13) were evaluated in this study (flute, $n = 29$; oboe, $n = 5$; clarinet, $n = 28$; saxophone, $n = 23$; French horn, $n = 8$; trumpet, $n = 19$; trombone, $n = 15$; euphonium, $n = 8$; tuba, $n = 3$). The sample of students was representative of four suburban

middle-school instrumental music programs from the same district participating in the assessment. The sample size meets the requirement for productive measurement (Linacre, 1994, 1996b) and sample sizes of the groupings are satisfactory for identifying potentially problematic areas of DIF considerations (Linacre, 2018).

A total of 17 adjudicators (i.e., raters) voluntarily participated in this study. The assessment took place over one full day. All raters were qualified by the district (and/or closely located outside districts) as having met the appropriate prerequisites to be approved as an official district solo and ensemble adjudicator. As Bergee (2007) noted, a one-rater assessment is often the norm in solo and ensemble-type adjudicated events. The problem with this design, however, is with a one-to-one ratio of rater-to-student performances, the student's estimated performance ability is confounded with the rater's severity and each item's difficulty, making it impossible to separate estimations of person ability, item difficulty, and rater severity. The rater assessment design used in this study was an unbalanced incomplete assessment network, consisting of an consisting of a incomplete, three-facet design three-facet design (person *x* item *x* rater) where a minimum of two raters evaluated one performance (Engelhard, 1997; Linacre & Wright, 2004; Wright & Stone, 1979). The benefit of an incomplete design is that fewer raters can evaluate more performances. The drawback of an incomplete design, as opposed to a complete design where all raters evaluate all performances, is the expectation of slightly larger estimates of standard error variances (Wind, Ooi, & Engelhard, in press; Wind, Engelhard, Jr., & Wesolowski, 2016). However, the differences in error are often negligible in substantive interpretation (less than .30 logits) (Engelhard & Myford, 2003). In the case of an authentic solo and ensemble festival where raters evaluate performances in real time, the implementation of a complete assessment network is impossible. In this case, all raters were connected through the incomplete design, thereby allowing for the sound estimation of person ability, item difficulty, rater severity, and instrument-type difficulty. At least two raters evaluated each performance and all sequentially assigned raters were connected across at least one musical performance. Raters did not confer during the assessment contexts and evaluated the performances in a live setting while simultaneously listening to the student perform.

Measurement Instrument

A 28-item rubric consisting of rating categories ranging from two to four performance criteria was used as the measurement tool for this study (see Supplementary Figure 1; Wesolowski, et al., 2017). The 28 item rubric consisted of the following eight domains: technique ($n = 2$ items), tone ($n = 2$ items), articulation ($n = 1$ item), intonation ($n = 1$ item), visual ($n = 9$ items), air support ($n = 3$ items), melody ($n = 4$ items), and expressive devices ($n = 6$ items). The rubric was developed and the psychometric qualities (i.e., validity and reliability) were tested using the MFR-PC model. The rubric was further tested in the context of a judgmental standard setting procedure (Wesolowski, et al., 2018). The rubric was uploaded into electronic format (Google Forms) and each rater completed one form per student evaluated in real time of the performance.

Psychometric Considerations for the MFR-PC Model

The MFR-PC measurement model (Linacre, 1989; Masters, 1982) was used in this study to convert the raters' unadjusted raw scores to estimated linear measures. More specifically, the probability of raw score responses was modeled as a function of student ability, item difficulty, rater severity, and music instrument classification parameters (Linacre, 1989). The Rasch

model was used for two specific purposes. First, the properties of the Rasch family of measurement models are fixed in their invariant measurement requirements for estimating data-to-model fit, making it particularly useful in contexts where raters mediate the assessment (Engelhard & Wind, 2019). The five requirements for rater-mediated invariance include: (a) rater invariant measurement of persons (i.e., the measurement of students must be independent of the particular raters that happen to be used for the measuring); (b) non-crossing person response functions (i.e., a higher-ability student must always have a better chance of obtaining higher ratings from raters than a lower-ability student); (c) person-invariant calibration of raters (i.e., the calibration of the raters must be independent of the particular students used for calibration); (d) non-crossing rater response functions (i.e., any student must have a better chance of obtaining a higher rating from lenient raters than from more severe raters); and (e) variable map (i.e., students and raters must be simultaneously located on a single underlying latent variable) (Engelhard, 2013). Second, as Penfield and Camilli (2007) note, approaches to DIF detection all share common theoretical interpretations rooted in the modeling capabilities of Item Response Theory (IRT). The Rasch family of measurement models, considered part of the same modeling (i.e., scaling) tradition as IRT, was thus a suitable measurement model for this study (Wesolowski, 2019).

The partial credit (PC) version of the measurement model (Masters, 1982) was used to allow the rating scales to vary by each item. Because the rubric was built in a manner where the rating scale categories are empirically unequal in the number of categories and substantively unequal in the difficulty of their content, the PC version of the model was a preferred choice. The PC version of the model allows for the empirical investigation of item differences across each of the item's individual rating scale structures (i.e., differences in step difficulty across rating scale categories), thereby providing a more detailed true score estimation of model-data fit.

The MFR-PC measurement model used in this study included facets for students, items, raters, and instrument classification. The model is specified as follows:

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

where

$\ln[P_{nijmk}/P_{nijmk}\text{-}1]$ = natural log of the probability that Student n rated by Rater i on Item j for Musical Instrument m receives a rating in category k rather than category k-1,

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

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

δ_j = the logit-scale location (e.g., difficulty) of Item j,

γ_m = the logit-scale location (e.g., difficulty) of Musical Instrument m, and

τ_{ik} = the logit-scale location where rating categories k and k-1 are equally probable for Student n.

Psychometric Considerations for DIF Interaction Model

A post hoc DIF analysis was carried out by adding an additional interaction parameter to the MFR-PC measurement model as specified in Equation 1. In the case of this study, DIF analysis identifies any potential significant differences in student scores across students' instrument classification at the same estimated ability level. In other words, DIF tested the null hypothesis

that students of the same ability did not differ significantly according to the music instrument they happen to play. In the case of this study, the interaction parameter included crossing the item facet with the instrument classification facet. The interaction term ($\delta_j\gamma_m$) was added to Equation 1 as follows:

$$ln\left[\frac{P_{nijmk}}{P_{nijmk}-1}\right]=(\theta_n-\lambda_i-\delta_j-\gamma_m-\tau_{ik})-\delta_j\gamma_m, \quad (2)$$

where $\delta_j\gamma_m$ is the interaction term between item difficulty and music instrument difficulty.

DIF analyses consist of two parts. The first is an omnibus test of the null hypothesis that the overall set of interaction terms between the item facet and the instrument classification facet do not differ significantly from zero. The test statistic that confirms or disconfirms the null hypothesis is a chi-square. If the omnibus test yields a significant chi-square, post hoc testing of individual interactions is conducted to explore meaningful patterns of interactions between individual elements (i.e., pairwise interactions between every item and every instrument grouping to identify where the significant interactions occur). The interaction terms are calculated as *t*-statistics with a mean of zero and standard deviation of 1 (i.e., a Z-score that approximates a normal distribution). In this study, significant interactions ($|Z| \geqslant 2.00$) indicate that an item is statistically significantly more difficult or less difficult for a student based on the instrument they play. All data analyses in this study were conducted using the *FACETS* computer program (Linacre, 2014).

Results

MFR-PC Model Results

The following results include the summary statistics of the MFR-PC model as well as the calibration results for each of the facets (student, rater, item, and instrument classification) in the model.

Summary statistics for the MFR-PC model. The full results of the summary statistics from the MFR-PC can be found in Supplementary Table 1. Overall, results indicated adequate fit of the data to the measurement model. In particular, the student facet ($\chi^2_{(138)} = 3{,}197.50$, $p < .01$, $Rel = .96$), rater facet ($\chi^2_{(16)} = 2{,}420.40$, $p < .01$, $Rel = .99$), item facet ($\chi^2_{(27)} = 16{,}951.50$, $p < .01$, $Rel = .96$), and instrument classification facet ($\chi^2_{(8)} = 791.70$, $p < .01$, $Rel = .99$) all demonstrated a statistically significant and high reliability of separation. Significant reliability of separation for the student facet indicates that students (i.e., the objects of measurement in the model) were able to be separated adequately across all ability levels based on varying difficulty levels of the items (i.e., the agents of measurement in the model), the varying severity levels of the raters, and the varying difficulty levels of the instrument groupings. Significant reliability of separation for raters, instruments, and items indicates that as a whole, each facet adequately separated out students at their varying levels of ability. Empirically, reliability of separation is interpreted in a similar manner as Cronbach's Alpha, ranging from .00 to 1.00. The rater, instrument, and item facets were each centered at .00 logits to act as an anchor to interpret the variability in the student facet appropriately. Significant and high reliabilities of separation for all facets included in the measurement model provide empirical support for the strong construct validity of the measure.

Student and rater calibrations. Supplementary Table 2 provides the calibration information for each of the 138 students. The highest-ability student was 67 (2.19 logits) and the lowest-ability student was 102 (-1.56 logits). Supplementary Table 3 provides the calibration information for each of the 17 raters. Overall, all raters demonstrated acceptable data-to-model fit based upon Wright and Linacre's (1994) acceptable range of infit and outfit mean squared error (MSE) statistics (.08 to 1.20) for high stakes testing contexts. The most severe rater was rater 3 (1.22 logits) and the least severe rater was rater 17 (-1.41 logits).

Item and instrument calibrations. Supplementary Table 4 provides the calibration information for each of the 28 items and instrument groupings. Overall, all items demonstrated acceptable data-to-model fit based on Wright and Linacre's (1994) acceptable range of infit and outfit MSE statistics (.08 to 1.20) for high stakes testing contexts. The most difficult item was item 23, "*Stylistically-related dynamics*" (2.07 logits) and the least difficult item was item 14, "*Cheeks*" (-2.30 logits).

Overall, all instrument classifications demonstrated an acceptable data-to-model fit based on Wright and Linacre's (1994) acceptable range of infit and outfit MSE statistics (.08 to 1.20) for high stakes testing contexts. The highest-difficulty instrument group was flute (0.38 logits) and the lowest-difficulty instrument group was trombone (-0.50 logits).

DIF Interaction Model Results

To determine whether measurement equivalence for all items exists when used to measure subgroups of students based upon their musical instrument (research question 1), a DIF omnibus analysis was conducted. The DIF analysis was conducted by crossing the item and instrument classification facets to test the null hypothesis that the overall set of interaction terms between the item facet and instrument classification facet do not significantly differ from zero. The analysis indicated an overall statistically significant item performance based on instrument classification ($\chi^2_{(252)} = 634.00$, $p < .01$), enabling the null hypothesis to be rejected. This suggests that measurement equivalence for all items does not exist when used to measure subgroups of students based upon their musical instrument. Table 1 illustrates the mean pairwise differences in average ability scores between instrument groupings.

To find out if any patterns of DIF effects exist for items when used to measure subgroups of students based on their musical instrument (research question 2), a post hoc pairwise interaction analysis between all items ($n = 28$) and all instrument classifications ($n = 9$) was conducted. Of the 252 possible interaction terms, a total of 57 (22.62%) interactions were found to be statistically significant ($|Z| \geqslant 2.00$). Supplementary Table 5 provides the individual significant pairwise interactions between items and instrument classification, as indicated by $|Z| \geqslant 2.00$, presented in order of positive to negative bias index values. The infit mean square (Infit MSQ) column shows how much misfit to the model remains after accounting for the bias. Infit MSQ is expected to be less than 1.00. Any value greater than 1.00 can be interpreted as having further unknown causes of misfit beyond that of the identified DIF effect. The observed minus expected values column (Obs.-Exp.) represents the total observed raw scores minus the total expected raw scores divided by the total observed count of raw scores. Positive values indicate the item was systematically less difficult for the identified instrument classification than expected. Negative values indicate the item was systematically more difficult for the identified instrument classification than expected. The bias index values (Bias) indicate the size of the DIF effect in logit units. Directionality and interpretation of bias index values correspond directly to the observed minus expected values. The expected value of bias indices is 0.00 logits, indicating there was no systematic differential item behavior exhibited. Positive values indicate the item was systematically less difficult for the identified instrument

Table 1. Mean Pairwise Differences Between Instrument Achievement Scores (In Logit Units).

Instrument	Average measure	Mean differences in achievement								
		Flute	Oboe	Clarinet	Saxophone	Trumpet	French horn	Trombone	Euphonium	Tuba
Flute	0.38	–	0.38	0.33	0.38	0.61	0.03	0.88	0.51	0.29
Oboe	0.00		–	–0.05	0.00	0.23	–0.35	0.50	0.13	–0.09
Clarinet	0.05			–	0.05	0.28	–0.30	0.55	0.18	–0.04
Saxophone	0.00				–	0.23	–0.35	0.50	0.13	–0.09
Trumpet	−0.23					–	–0.58	0.27	–0.10	–0.32
French horn	0.35						–	0.85	0.48	0.26
Trombone	−0.50							–	–0.37	–0.59
Euphonium	−0.13								–	–0.22
Tuba	0.09									–

All differences are statistically significant ($\chi^2_{(8)} = 791.70$, $p < .01$). Substantive differences $\geqslant .30$ logits (Engelhard & Myford, 2003).

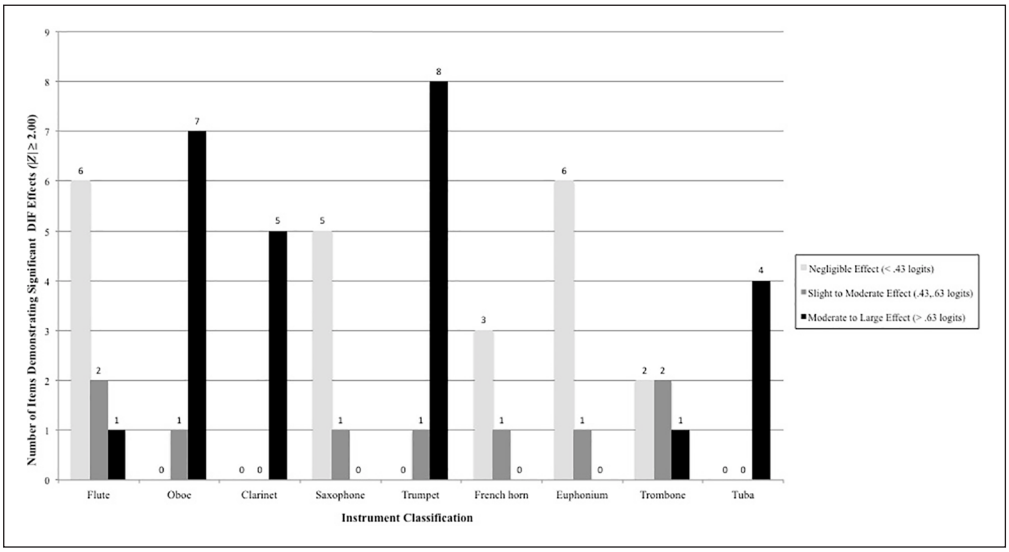

Figure 1. Frequency of significant pairwise interactions between items and musical instruments by differential item functioning (DIF) effect size.

classification than expected. Negative values indicate the item was systematically more difficult for the identified instrument classification than expected. The standard error (*SE*) represents the standard model error of the bias estimate. The *Z*-score (*Z*) is the statistic that tests the null hypothesis that there is no DIF effect other than the standard measurement error. Any value above 2.00 or below -2.00 indicates a significant DIF interaction effect. Note that only the 57 (of a possible 252) significant interactions ($|Z| \geqslant 2.00$) are reported in the table.

In the context of assessments involving comparatively small sample sizes, it is more advantageous and relevant to evaluate bias index values for items demonstrating statistically significant DIF effects (Linacre, 2018). To identify the size of DIF effects for items used to measure subgroups of students based on their musical instrument (research question 3), bias index values were examined to make substantive interpretations of their effect. Based on the guidelines of Zwick, Thayer, and Lewis (1999), bias indices where |DIF| >/= .64 logits are considered to represent a moderate to large effect, bias indices where |DIF| is in the range of .43 to .63 logits are considered to represent a slight to moderate effect, and bias indices where |DIF| < .43 logits are considered to represent negligible effects. A total of 26 (10.3%) significant interaction terms demonstrated a moderate to large effect where |DIF| >/= .63 logits; nine (3.6%) significant interaction terms demonstrated a slight to moderate effect, where |DIF| = (.43, .63 logits); and 22 (8.7%) significant interaction terms demonstrated a negligible effect, where |DIF| < .43 logits. Figure 1 provides descriptive statistics for the significant pairwise interactions between items and instruments by DIF effect size. Effects represented by |DIF| >/= .43 are highly relevant to considerations of the fairness of assessment outcomes.

Discussion

The purpose of this study was to examine DIF in a rubric used to assess middle-school solo and ensemble performances. To answer the first research question (*does measurement equivalence for all items exist when used to measure subgroups of students based on their musical instrument?*), an

omnibus test of DIF was used to test the null hypothesis that the overall set of interaction terms between all items and all instruments do not differ significantly from zero. An overall statistically significant interaction between item and instrument facets ($\chi^2_{(252)} = 634.00$, $p < .01$) was found, indicating that measurement equivalence did not exist for all items when used to measure subgroups of students based on the particular instrument they happened to play, so the null hypothesis was rejected.

To answer the second research question (*what patterns of DIF effects exist for items when used to measure subgroups of students based on their musical instrument?*), a post hoc analysis of pairwise interactions between all item elements ($n = 28$) and all instrument classification elements ($n = 9$) was conducted. Of the 252 possible interaction terms, a total of 57 (22.62%) of the interactions were found to be statistically significant ($|Z| \geqslant 2.00$).

To answer the third research question (*what size of DIF effects exist for items when used to measure subgroups of students based on their musical instrument?*), bias index values were examined to make substantive interpretations of the DIF effect. A total of 26 (10.3%) significant interaction terms demonstrated a moderate to large effect where |DIF| >/= .63 logits, nine (3.6%) significant interaction terms demonstrated a slight to moderate effect, where |DIF| = (.43, .63 logits), and 22 (8.7%) significant interaction terms demonstrated a negligible effect, where |DIF| < .43 logits.

It is important to consider assessment practices in music not just as a mechanism for justifying scoring and validating inferences but, more importantly, as a method for building theories and better understanding psychological constructs. As noted eloquently by Bond & Fox (2015, p. 44):

> . . . [We] can take the conventional approach and see this as a test of whether our data correspond with our substantive theory. However, it is more useful if we complement this use with the idea of whether our construct, as expressed in developmental or other theory, fits with our data. Ideally, this is just one part of an epistemologically iterative research program in which theory informs practice and practice informs theory dialectically.

It is clear the unique challenges and pedagogies for both teaching and learning a musical instrument affect the fairness of scoring outcomes when attempting to measure music performances multilaterally. Interestingly, the results demonstrated adequate data-to-model fit as evidenced by the acceptable fit statistics of the MFR-PC model. However, post hoc DIF analyses indicated a multitude of DIF effects. An attempt to understand the significant instrument-by-item interactions substantively is far beyond the scope of this study, and even the broad-based considerations mentioned here are speculative, at best. Follow-up phenomenological investigations into exemplar performances, related raw scores, and rater insights coupled with pedagogical considerations by experts of instrument-specific content would be one step in the right direction for explaining the particular results from both scholarly and pedagogical perspectives.

Solo and ensemble festivals, such as the one examined in this study, take place regularly and frequently each year throughout the USA. On the evidence of the challenges to fairness revealed by the findings of this study, it may well be that similar challenges pervade other similar assessment contexts. According to current frameworks whereby raw scores are used instead of estimated measures, performance standard classifications are assigned on the basis of summed raw scores equilaterally across all instrument types, judges mediate the assessment environment without controlling for judge severity, and DIF is not a consideration, fairness cannot be guaranteed. Such frameworks do not currently permit investigations into the functionality of measurement tools. These could be improved by examining data-to-model fit and conducting

DIF analyses, important methods for monitoring and maintaining the overall stability of measurement tools. Wright and Masters (1982, p. 114) note that "[when] a variable is used with different groups of persons or to measure the same persons on different occasions, it is essential that the identity of the variable be maintained from one occasion to the next" and, as Wright (1999, p. 704) argues, "It is up to us [the test developers] to construct and maintain [tests]."It is hoped that this vision may be realized in future.

A major consideration for the field of music assessment is how best to move forward in providing high-quality formal assessments for students. It is clear, from a pedagogical perspective, that different instruments make different demands on those who play them and, from a psychometric perspective, performances on different instruments are scored differently. These differences need to be acknowledged and accounted for in assessment practice. It is suggested that researchers and assessors work together to examine and reflect on current assessment paradigms and consider implementing unique measurement tools designed to evaluate individual instruments. A more focused approach whereby the challenges and idiosyncrasies of assessing performance on specific instruments are considered empirically may produce not only fairer assessment procedures but also better, more relevant formative feedback that is directly transferable to improved teaching and learning in relation to each instrument.

When considering improved approaches to formative assessment, one psychometric solution might be to implement a diagnostic classification modeling (DCM) approach to large-scale music assessment processes (Rupp, Templin, & Henson, 2010). With the use of Rasch measurement theory or other families of measurement models under the umbrella of IRT, students are assigned a single score along a 'continuum' representative of the particular latent construct being measured (e.g., from less to more music performance ability). This approach is arguably summative in nature. DCMs, by contrast, are a class of psychometric models that can characterize the traits of student performances as categorical latent constructs (Bradshaw, 2016). DCM approaches are thus more diagnostic than summative, as students are provided with a profile of proficiency attributes representing a meaningful pathway toward mastery of skill content rather than a single score. Such approaches may be better suited to the expectations of students participating in solo and ensemble festivals, in relation to formative feedback.

Lastly, it is important that fairness guidelines should be taken into consideration when conducting formal music assessments. The Educational Testing Service's (ETS) *Guidelines for Fairness Review of Assessments* (2009) and *International Principles for the Fairness of Assessments* (2016) are just two of many such sets of guidelines that can be helpful in promoting fairness when designing and implementing formal assessment measures. Most fairness guidelines are written from a high-stakes, cognitive-exam perspective, where cognitive considerations (i.e., related to topics and content), affective considerations (i.e., related to eliciting strong and distracting emotions), and physical considerations (i.e., related to physical barriers in items and stimulus materials) are needed when an examinee takes an exam. Many music assessments are performance based, therefore considerations of the rater's interaction with the items/criteria may be more relevant when considering fairness in the design of an assessment. Therefore, some ETS guideline considerations, specifically for writing assessment items, may be most relevant and appropriate to the context of music performance assessment. Some examples include:

- training test constructors to follow fairness guidelines when constructing and using measures;
- retraining test constructors periodically;
- asking expert peer reviewers to review measurement tools using a double-blind process;
- requiring content-expert reviewers to probe items for fairness considerations;

- checking that test constructors do not have conflicts of interest;
- not allowing items to be used until fairness challenges, if raised, have been resolved;
- allowing test constructors to dispute fairness challenges; and
- clarifying fairness considerations regularly throughout the assessment process.

Considering fairness, both qualitatively and quantitatively, and taking action to overcome fairness challenges in protocols for assessing music performance are important ways of enhancing the quality of programs and overall validity of inferences regarding student performances. Indeed they may be considered essential for improving current music assessment research and practice. From a scholarly perspective, continued investigations and improved training in graduate programs is a fundamental step towards improving the science of psychometrics as it relates to the field of music assessment. Graduate programs regularly train students in statistics, but often gloss over or completely ignore training in measurement theory. From the perspective of a practitioner, it is vital to hold discussions with teachers and music administrators about the impact of validity, reliability, and fairness on inferences of student ability if current practice is to change for the better.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

ORCID iD

Brian C. Wesolowski https://orcid.org/0000-0003-0615-9277

References

Abeles, H. F., Hoffer, C. R., & Klottman, R. H. (1994). *Foundations of music education (2nd edition)*. New York, NY: Schirmer Books.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association

Ando, Y. (1988). *Architectural acoustics: Blending sound sources, sound fields, and listeners*. New York, NY: Springer.

Austin, J. R. (1988). The effect of music contest format on self-concept, motivation, achievement, and attitude of elementary band students. *Journal of Research in Music Education, 36*, 95–107.

Banister, S. (1992). Attitudes of high school band directors toward the value of marching band and concert band contests and selected aspects of the overall band program. *Missouri Journal of Research in Music Education, 29*, 49–57.

Bergee, M. J., & McWhirter, J. L. (2005). Selected influences on solo and small-ensemble festival ratings: Replication and extension. *Journal of Research in Music Education, 53*, 177–190.

Bergee, M. J., & Platt, M. C. (2003). Influence of selected variables on solo and small-ensemble festival ratings. *Journal of Research in Music Education, 51*, 342–353.

Bergee, M. J., & Westfall, C. R. (2005). Stability of a model explaining selected extramusical influences on solo and small-ensemble festival ratings. *Journal of Research in Music Education, 53*, 253–271.

Bergee, M. J. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education, 55*, 344–358.

Bermingham, G. A. (2000). Effects of performers' external characteristics on performance evaluations. *Update: Applications of Research in Music Education, 18*, 3–7.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd edition)*. New York, NY: Routledge.

Boyle, D. J. (1992). Program evaluation for secondary school music programs. *NASSAP Bulletin, 76*, 63–68.

Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (Ed.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 297–326). West Sussex, UK: John Wiley & Sons, Inc.

Burnsed, V., Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert festivals. *Journal of Band Research, 21*, 22–29.

Ciorba, C. R., & Smith, N. Y. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *Journal of Research in Music Education, 57*, 5–15.

Colwell, R. J., & Hewitt, M. P. (2011). The teaching of instrumental music (4th Edition). New York, NY: Taylor and Francis.

Conrad, D. (2003). Judging the judges: Improving rater reliability at music contests. *NFHS Music Association Journal, 20*, 27–31.

Crochet, L. S. (2006). *Repertoire selection practices of band directors as a function of teaching experience, training, instructional level, and degree of success.* (Unpublished doctoral dissertation). University of Miami, Coral Gables, FL.

Davidson, J. W., & Coimbra, D. D. C. (2001). Investigating performance evaluation by assessors of singers in a music college setting. *Musicae Scientiae, 5*, 33–53.

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music, 21*, 103–113.

Davidson, J. W. (1994). Which areas of a pianist's body convey information about expressive intention to an audience? *Journal of Human Movement Studies, 26*, 279–301.

Davidson, J. W. (1995). What does the visual information contained in musical performances offer the observer? Some preliminary thoughts. In R. Steinberg (Ed.), *The music machine: Psychophysiology and psychopathology of the sense of music* (pp. 105–113). Berlin, Germany: Springer-Verlag.

Duke, R. A., & Byo, J. L. (2011). *The habits of musicianship: A Radical approach to beginning band*. Austin, TX: University of Texas Center for Music Learning.

Educational Testing Service (2009). *ETS guidelines for fairness review of assessments*. Retrieved from https://www.ets.org/Media/About_ETS/pdf/overview.pdf

Educational Testing Service (2016). *ETS international principles for the fairness of assessments: A manual for developing locally appropriate fairness guidelines for various countries*. Retrieved from https://www.ets.org/s/about/pdf/fairness_review_international.pdf

Engelhard, G. E. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement, 1*, 19–33.

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

Engelhard, G. E., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the advanced placement English literature and composition program with a Many-faceted Rasch Model* (ETS Research Report).

Engelhard, G. E., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York, NY: Routledge.

Fiske, H. E. (1975). Judge-group differences in the rating of secondary school trumpet performers. *Journal of Research in Music Education, 23*, 186–196.

Fiske, H. E. (1983). *The effect of a training procedure in music performance evaluation on judge reliability*. Ontario, Canada.

Franklin, J. O. (1979). *Attitudes of school administrators, band directors, and band students towards selected activities of the public school band program*. (Unpublished doctoral dissertation). Northwestern State University of Louisiana, Natchitoches, LA.

Griffiths, N. K., & Reay, J. L. (2018). The relative importance of aural and visual information in the evaluation of Western canon music performance by musicians and nonmusicians. *Music Perception: An Interdisciplinary Journal*, *35*, 364–375.

Hash, P. M. (2012). An analysis of the ratings and interrater reliability of high school band contests. *Journal of Research in Music Education*, *60*, 81–100.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.129–146). Engelwood Cliffs, NJ: Erlbaum.

Howard, K. K. (1994). *A survey of Iowa high school band students' self-perceptions and attitudes toward types of music contests*. (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Howard, R. L. (2002). *Repertoire selection practices and the development of a core repertoire for the middle school concert band*. (Unpublished doctoral dissertation). University of Florida, Gainsville, FL.

Hurst, C. W. (1994). *A nationwide investigation of high school band directors' reasons for participating in music competitions*. (Unpublished doctoral dissertation). The University of North Texas, Denton, TX.

King, S. E., & Burnsed, V. (2007). A study of the reliability of adjudicator ratings at the 2005 Virginia Band and Orchestra Directors Association state marching band festivals. *Journal of Band Research*, 27–33.

Kirchhoff, C. (1988). The school and college band: wind band pedagogy in the United States. In J. T. Gates (Ed.), *Music education in the United States: Contemporary issues* (pp. 259–276). Tuscaloosa, AL: The University of Alabama Press.

Kopiez, R., Wolf, A., & Platz, F. (2017). Small influence of performing from memory on audience evaluation. *Empirical Musicology Review*, *12*(1–2), 2–14.

Krahe, C., Hahn, U., & Whitney, K. (2015). Is seeing (musical) believing? The eye versus the ear in emotional responses to music. *Psychology of Music*, *43*(1), 140–148.

Latimer, M. E., Bergee, M. J., & Cohen, M. L. (2010). Reliability and perceived pedagogical utility of a weighted music performance assessment rubric. *Journal of Research in Music Education*, *58*, 168–183.

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.

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

Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (1996b). Sample size again. *Rasch Measurement Transactions*, *9*, 468.

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

Linacre, J. M. (2018). *DIF – DPF – bias – interaction concepts*. Retrieved from https://www.winsteps.com/winman/difconcepts.htm

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

McPherson, G. E., & Thompson, W. F. (1998). Assessing music performance: Issues and influences. *Research Studies in Music Education*, *10*, 12–24.

National Association for Music Education. (2016) Ensemble adjudication forms. Retrieved from http://www.nafme.org/my-classroom/ensemble-adjudication-forms/

Norris, C. E., & Borst, J. D. (2007). An examination of the reliabilities of two choral festival adjudication forms. *Journal of Research in Music Education*, 55, 237–251.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26: Psychometrics* (pp. 125–167). Amsterdam, The Netherlands: Elsevier.

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception: An Interdisciplinary Journal*, *30*, 71–83.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford.

Saunders, T. C., & Holahan, J. M. (1997). Criteria-specific rating scales in the evaluation of high school instrumental performance. *Journal of Research in Music Education*, *45*, 259–272.

Schubert, E. (2002). Continuous response methodology applied to expressive performance. In C. Stevens, D. Burnham, G. McPherson, E. Schubert, & J. Renswick (Eds.), *Proceedings of the Seventh International Conference on Music Perception and Cognition* (pp. 83–86). Adelaide, Australia: Causal Productions.

Silvey, B. A. (2009). The Effects of band labels on evaluators' judgments of musical performance. *Update: Applications of Research in Music Education*, *28*, 47–52.

Tsay, C. J. (2014). The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes*, *124*, 24–33.

Waddell, G., & Williamon, A. (2017). Eye of the beholder: Stage entrance behavior and facial expression affect continuous quality ratings in music performance. *Frontiers in Psychology- Cognitive Science*, *8*.

Wapnick, J., Darrow, A. A., Kovacs, J., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education*, *45*, 470–479.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (1998). Effects of performer attractiveness, stage behavior, and dress on evaluation of violin performance evaluation. *Journal of Research in Music Education*, *46*, 510–521.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (2000). Effects of performer attractiveness, stage behavior, and dress on evaluation of children's piano performances. *Journal of Research in Music Education*, *48*, 323–335.

Wesolowski, B. C. (2019). Item response theory and music testing. In T. S. Brophy (Ed.), *The Oxford handbook of assessment policy and practice in music education* (pp. 437–460). New York, NY: Oxford University Press.

Wesolowski, B. C., Amend, R. M., Barnstead, T. S., Edwards, A. S., Everhart, M., Goins, Q., Grogan III, R. J., Herceg, A. M., Jenkins, S. I., Johns, P. M., McCarver, C. J., Schaps, R E., Sorrell, G. W., & Williams, J. D. (2017). The development of a secondary-level solo wind instrument performance rubric using the Multifaceted Rasch Partial Credit Measurement Model. *Journal of Research in Music Education*, *65*, 95–119.

Wesolowski, B. C., Athanas, M., Burton, J., Edwards, A. S., Edwards, K. E., Goins, Q., Irby, A., Johns, P., Musselwhite, D. J., Parido, B., Sorrell, G., & Thompson, J. (2018). Judgmental standard setting: The development of objective content and performance standards for secondary –level solo instrumental music assessment. *Journal of Research in Music Education*, *66*, 224–245.

Wesolowski, B. C., Wind, S. A., & Engelhard, Jr., G. E. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*, 39–47.

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

Wind, S. A., Ooi, P. S., & Engelhard, Jr., G. E. (in press). Exploring decision consistency and decision accuracy across rating designs in rater-mediated music performance assessments. *Musicae Scientiae*. doi: 10.1177/1029864918761184

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

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

Wright, B. D. (1999). Common sense for measurement. *Rasch Measurement Transactions*, *13*, 704.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28.