sempre:
Society for Education, Music and Psychology Research

Article

Psychology of Music
1–25

DOI: 10.1177/0305735616665004
pom.sagepub.com
SAGE

Exploring rater cognition: A typology of raters in the context of music performance assessment

Brian C. Wesolowski

Abstract

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

Keywords

assessment, cognition, differential rater functioning, evaluation, partial-credit, Rasch Measurement

Music performance assessments are based upon a constructed-response format, where observed scores result from raters' interpretations of a set of evaluation cues (e.g. items and rating scale categories) provided within a measurement instrument. The construct of music performance achievement is latent. In latent construct assessment contexts, the cues set forth within

The University of Georgia, Athens, USA

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

the measurement instrument serve as the operational definition of the construct. Therefore, the functioning of cues needs to be valid, reliable, and fair in order to properly define the latent construct of intended measurement. Unlike traditional cognitive-type tests where psychometric concerns stem from item and person parameter variance, constructed-response assessments have an additional layer of complexity calling for psychometric concern – rater behavior (Engelhard, 2002). In both hermeneutic music performance scoring systems (i.e. systems that define performance achievement based upon multiple-content expert rater perspectives stemming from differences in rater schemata) and psychometric music performance assessment scoring systems (i.e. systems that emphasize and value a high level of quantitative consistency between trained raters), empirical results can serve as a foundation for important administrative and political decisions with impactful consequences (Wesolowski, 2014, 2015). It is therefore necessary for decision makers to verify that rater functioning is valid, reliable, and fair. The inherent problem with the validation of rater-mediated assessments and rater quality management is that: (a) raters' schemata vary in the use of evaluation cues and the cognitive processes by which the scoring is based (Wolfe, 1997); and (b) observed rater scores are more often associated with characteristics of the raters and less with the performances themselves (Engelhard, 2002). Conceptually, raters' observed scores are generated within the boundaries of a unique lens to each rater, equally affected by both raters' schema based upon unique cognitive and perceptual processes as well as the ecological environment (Brunswik, 1952).

This article seeks to investigate rater cognition in the context of music performance assessment by exploring rater types based upon differential severity and leniency associated with (a) raters' use of rating scale items, (b) rating scale category functioning, and (c) dimensions of music performance assessment (e.g. tone/intonation, balance, interpretation, rhythm, technical accuracy). The purpose of this study was to empirically identify a typology of rater types based upon systematic differential severity indices in the context of large ensemble music performance assessment. This study was guided by the following research questions:

1. Do individual raters maintain invariant levels of severity when rating high school concert band performances?
2. How does the structure of the rating scale designed to evaluate high school concert band performances vary across raters?
3. Does differential severity emerge for individual raters across items when evaluating high school concert band performances?
4. Does a meaningful typology exist based upon raters' differential severity indices when rating high school concert band performances?

Background

In the field of music, research concerning rater effects centers around methodologies that broadly represent a rater behavior-centered approach (Wolfe, 2004). These methodologies focus on the ecological content of human judgment and can be classified according to four distinct areas: (a) extra-musical effects related to the performer such as expressive variations (Repp, 1990, 1995), attractiveness and flair (Davidson & Coimbra, 2001; Wapnick, Darrow, Kovacs, & Dalrymple, 1997; Wapnick, Mazza, & Darrow, 1998), and body movement (Davidson, 1994, 2001; Davidson & Correia, 2002); (b) extra-musical effects related to the assessment context such as within-ensemble communication (Wesolowski, 2013; Williamon & Davidson, 2002), acoustics (Ando, 1988), social factors (Davidson, 1997), and audience support (Berliner, 1994; Monson, 1996); (c) rater-centered effects such as memory (Radocy, 1976),

first impressions (Stanley, Brooker, & Gilbert, 2002; Vasil, 1973), mood (Schubert, 1996), repertoire familiarity (Flores & Ginsburgh, 1996), and musical preference (Rentfrow, Goldberg, & Levitin, 2011; Rentfrow et al., 2012; Rentfrow & McDonald, 2009); and (d) non-musical effects such as stereotyping (Elliott, 1995; Morrison, 1998), performance order (Bergee, 2006, 2007; Flores & Ginsburgh, 1996), evaluation time (Thompson, Williamon, & Valentine, 2007), facets of musical expression, (Juslin, 2003), teaching-level and primary instrument (Hewitt & Smith, 2004). (See McPherson & Schubert, 2004; McPherson & Thompson, 1998 for process models implementing these areas.)

More recent investigations into rater behavior in the context of music performance assessment take a different approach, using empirically driven investigations into statistical indices that underscore the measurement process. The purpose of these approaches is to explore rater effects and quality of ratings from a psychometric perspective. According to Eckes (2012), rater variability under these conditions can stem from: (a) the degree to which raters comply with the measurement instrument; (b) the way raters interpret criteria in operational scoring sessions; (c) the degree of leniency and severity exhibited; (d) raters' understanding of the measurement instrument's rating scale categories; and (e) the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks. The application of Rasch measurement models has been shown to be a fruitful method to serve as a foundation for these empirically driven approaches (Engelhard, 2013). In the context of music performance assessment, Rasch methodology has been used successfully to empirically explore rater behavior with specific attention to (a) rater effects (Wesolowski, Wind, & Engelhard, 2016b); (b) rating scale structure and precision (Wesolowski, Wind, & Engelhard, 2016a); (c) differential rater functioning (Wesolowski, Wind, & Engelhard, 2015); and (d) time parameters (Wesolowski, Wind, & Engelhard, in press). In this study, the systematic investigation into rater types was explored using indices stemming from rater effects, rating scale structure, and differential rater functioning. First, rater effects (e.g. rater severity) were explored in order to quantify severity indices for each rater. Second, rating scale structure for each rater was explored in order to quantify precision indices. Third, differential rater functioning for each rater by item was explored in order to quantify raters' systematic differential severity indices in the use of each item. As a result, individual rater severity, indices of rating scale structure use, and indices of differential rater functioning were used as a foundation for identifying rater typologies. Full theoretical and technical explanations of the methodology and application for each rater index in the context of music performance assessment can be found in the above-cited papers.

Both rater behavior-centered approaches and empirically-driven approaches to investigating rater effects provide evidence of interesting cognitive phenomena affected by two judgmental predictors: (a) how well the cues within a measurement instrument function in relation to the raters' precise use of rating scale categories; and (b) the ecological validity of the measurement tool (Hammond, 1996). Building on the work of Freedman and Calfee (1983) in the field of writing assessment, Wolfe (1997) used both predictors to introduce a model of rater cognition consisting of two broad components driven by top-down cognitive processes: a *framework of scoring* and a *framework of writing*.

Wolfe's (1997) model is an information-processing model in the context of a psychometric setting. If this model is adapted to the context of music performance assessment, it can be surmised as having two components: *framework of scoring* and *framework for audition.* The *framework of scoring* can be described as raters' mental processes where "an [auditory] image is created, compared to the scoring criteria, and used as the basis for generating a scoring decision" (p. 89). This framework includes three hierarchical cognitive processes: interpretation of the auditory stimulus, evaluation of the auditory stimulus, and justification of the scoring

decision. First, interpretation of the auditory stimulus includes the cognitive processes where raters process a unique auditory image from the perceived auditory information. In this process, meaning is constructed through active listening. Second, evaluation of the auditory stimulus includes the cognitive process of mapping evaluative cues to the auditory information. The process of evaluation is where performance achievement levels are empirically defined based upon raters' varying levels of importance placed on particular items and dimensions of music performance (e.g. tone/intonation, balance, interpretation, rhythm, technical accuracy). The items underscoring each dimension are used as prompts to elicit raters' behavior stemming from cognitive processing. Third, justification of the scoring decision includes the self-monitoring and attention to raters' performances as evaluators, and how raters self-regulate and adjust their behavior based upon instructions, training, and other related input regarding the assessment process.

The *framework of audition* is the raters' mental schema of performance characteristics that represent varying levels of proficiency. A rater's cognitive schemata can play an important role (or even interfere) with how well he/she adopts the structure and framework of the measurement instrument due to rater behavior-centered effects and ecological context. According to Stanley et al. (2002):

> Initially, they [raters] adopt a "holistic" approach, relying on a "gut reaction", an "intuitive or emotional response which is basically one of enjoyment: Am I enjoying this playing?" This early process of global assessment frequently involves respondents arriving at a tentative grade. As one examiner noted: "I look at them and I say 'Distinction, high credit'. I have bands in my own mind and then the number is immaterial – to me the number is way more negotiable than the actual range." (p. 51)

According to Wolfe (1997), raters' mental representations within his framework for writing are all unique "because of differences in scoring experience, values, education, and familiarity with the scoring rubric" (pp. 89–90). Similarly in music, the uniqueness of listeners' mental representations of auditory stimuli have been demonstrated in terms of neurophysiological processes (Downar, Crawley, Mikulis, & Davis, 2001; Näätänen & Winkler, 1999), cognitive/perceptual processes (Gabrielsson & Lindstrom, 2001; Palmer, 1989; Persson, Pratt, & Robson, 1992; Serafine, Glassman, & Overbeeke, 1989; Sloboda, 1983), and psychometric processes (Wesolowski et al., 2015, 2016a, 2016b, in press). In both the context of writing assessment and music performance assessment, the raters' representation is not directly observable, and therefore can only be inferred from raters' behavior via the use of the measurement apparatus. As a result, careful attention to how raters engage specifically with each item in a measurement instrument and/or dimensions on a measurement instrument (e.g. tone/intonation, balance, interpretation, rhythm, technical accuracy) may provide valuable insight into raters' mental representations of the auditory stimuli being evaluated.

Both frameworks interact to result in unique rater variability influenced by individual rater schemata. Therefore, origins of the variability need to be carefully considered in making decisions and inferences based upon raters' observed data. Under hermeneutic scoring conditions, variability due to rater schemata can provide a desired variety of insightful diagnostic, formative, and summative information for the performers toward their improvement. On the other hand, under psychometric scoring conditions, it can provide conflicting perspectives of a working operational definition of performance achievement within the assessment system. Music research traditionally focuses on face value empirical interpretations of raters' consistency, consensus, accuracy, and precision to account for variability (Wesolowski et al., 2016a). The

empirical results, however, are driven by the interaction between raters' schemata and the measurement instrument. An understanding of the effects of rater schemata on scoring variability can provide better explanation for the empirical results and improve the overall evaluation process.

This study seeks to explore Wolfe's (1997) model in the context of music performance assessment. The application of the Many Facet Rasch Partial Credit (MFR-PC) measurement model was used to provide an empirical evidence of a *framework of scoring*. A combination of hierarchical and non-hierarchical clustering techniques was used to explore and define classes of rater schemata through a *framework of audition*.

Method

Apparatus

The measurement instrument used in this study was a 30-item rating scale to assess high school concert band performance (DCamp, 1980; see Figure 1). The items were developed and validated using a factor analytic approach to scale construction. The factor analysis yielded a five-factor solution: (a) tone/intonation ($n = 6$); (b) balance ($n = 6$); (c) interpretation ($n = 6$); (d) rhythm ($n = 6$); and (e) technical accuracy ($n = 6$). The original response alternatives in DCamp's scale included a 5-point Likert scale structure: *Strongly Agree, Agree, Not Applicable, Disagree, Strongly Disagree*. In this study, the response alternatives included a 4-point structure: *Strongly Agree, Agree, Disagree*, and *Strongly Disagree*. A 4-point scale was specifically chosen in order to eliminate a neutral category. The elimination of a neutral category provides a better measure of the intensity of participants' attitudes and opinions (Wright, 1977) and maintains a positive step ordering of rating scale categories (Linacre, 2002a).

Rater assessment structure

For this study, an a priori incomplete assessment structure was developed as suggested by Linacre and Wright (2004) and Wright and Stone (1979). The incomplete assessment network can be described specifically where each consecutive rater overlapped with the previous rater in evaluating the same two musical performances. As an example, Rater 1 evaluated performances 1, 2, 3, 4; Rater 2 evaluated performances 3, 4, 5, 6; Rater 3 evaluated performances 5, 6, 7, 8; etc.). The last rater (i.e. Rater 67) also evaluated performances 1 and 2, thereby linking to Rater 1 and creating a connection between all other raters in the model. This connectivity allowed for independent calibrations of all musical stimuli, rating scale items, and raters to be compared unambiguously. As raters accepted participation in the study, they were consecutively assigned to a rater number so as not to break the connectivity of the rater network.

Participants

A total of 67 ($n = 36$, male; $n = 31$, female) raters were solicited for the study on a voluntary basis. At the time of the study, each of the raters was an in-service secondary school-level ($n = 41$, high school, $n = 26$, middle school) concert band instructor with an average of 12.42 years ($SD = 4.65$) of school teaching experience. Each rater's specialized instruction area was instrumental music education. Each rater was asked to evaluate four full high school concert band musical performances. Anonymity was kept with all the participants in the study.

Dimension	Item					
1: tone/intonation	1.	Intonation is quite good.	SD	D	A	SA
1: tone/intonation	2.	Basic tuning is not good, band plays out of tune.	SD	D	A	SA
1: tone/intonation	3.	Tone is shallow.	SD	D	A	SA
1: tone/intonation	4.	The band plays with a well-defined pitch center.	SD	D	A	SA
1: tone/intonation	5.	The band has good control in high pitch registers.	SD	D	A	SA
1: tone/intonation	6.	Tutti chords are badly out of tune.	SD	D	A	SA
2: balance	7.	Excellent balance between parts in a section.	SD	D	A	SA
2: balance	8.	Excellent balance in all parts.	SD	D	A	SA
2: balance	9.	Lower parts balance the group well.	SD	D	A	SA
2: balance	10.	All sections are well balanced within.	SD	D	A	SA
2: balance	11.	Inner parts balance well.	SD	D	A	SA
2: balance	12.	Inner parts are too timid.	SD	D	A	SA
3: interpretation	13.	Performance shows careful attention to dynamics.	SD	D	A	SA
3: interpretation	14.	Performance has a good, wide variety in dynamics.	SD	D	A	SA
3: interpretation	15.	Needs to be more expressive.	SD	D	A	SA
3: interpretation	16.	Crescendo and diminuendo are properly graduated.	SD	D	A	SA
3: interpretation	17.	Performance lacks emotion.	SD	D	A	SA
3: interpretation	18.	Performance exhibits proper style.	SD	D	A	SA
4: rhythm	19.	Dotted rhythms played as triplets.	SD	D	A	SA
4: rhythm	20.	Dotted eighth-sixteenth pattern is inaccurate.	SD	D	A	SA
4: rhythm	21.	Syncopated patterns are correctly played.	SD	D	A	SA
4: rhythm	22.	Rhythmically accurate.	SD	D	A	SA
4: rhythm	23.	Performance is marked by unsteady rhythm.	SD	D	A	SA
4: rhythm	24.	Rhythmic figures are properly and distinctly executed.	SD	D	A	SA
5: technical accuracy	25.	Runs are played accurately and smoothly.	SD	D	A	SA
5: technical accuracy	26.	Notes in runs are inaccurate.	SD	D	A	SA
5: technical accuracy	27.	Fingering problems cause a lack in clarity.	SD	D	A	SA
5: technical accuracy	28.	Awkward and difficult passages are not prepared.	SD	D	A	SA
5: technical accuracy	29.	Entrances are not precise.	SD	D	A	SA
5: technical accuracy	30.	Technique is inadequate for performance of this music.	SD	D	A	SA

Figure 1. Thirty-item rating scale and associated factor dimensions for the evaluation of high school concert band performance (DCamp, 1980).

Stimuli

A total of 53 performances were rated, each randomly inserted into the assessment network using an electronic random generator. Audio recordings were gleaned from a pool of district and state large group music performance assessment performances in the United States.

Acceptability of audio stimuli quality was previously rated and verified by a cohort of music content experts using the audio component of the International Telecommunication Union's ITU-T Rating Scale (ITU, 2004).

Evaluation process

In order to provide a baseline of rater schema and to not influence interpretation, evaluation, or justification processes, raters were not trained nor were they provided anchor recordings. Raters were told that all musical examples were high school concert band performances, and asked to rate the performances to the best of their ability based upon their interpretation of the items, scoring criteria, and personal expectations of what constitutes a proficient high school concert band. Raters were asked to listen to the recordings using headphones at their desktop or laptop computer. They were allowed to listen to the recordings as many times as necessary in order to formulate a holistic auditory image of the performance being evaluated. The musical examples were distributed electronically and responses were collected via a web-based response collection service over a time span of two weeks. After responses were collected, responses stemming from negatively worded items were reverse coded.

Measurement model

The Many Facet Rasch Partial Credit (MFR-PC) measurement model was used to guide the measurement process of this study (Linacre, 1989; Masters, 1982). The benefit of using Rasch methodology is the five requirements of invariant measurement underscoring the analyses (Engelhard, 2013). When using observed data derived from rater-mediated assessment processes, the following five requirements are necessary for invariant measurement: (a) rater-invariant measurement of persons (i.e. the measurement of performances 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 achieving ensemble must always have a better chance of obtaining higher ratings from raters than a lower achieving ensemble); (c) person-invariant calibration of raters (i.e. the calibration of the raters must be independent of the particular ensembles used for calibration); (d) non-crossing rater response functions (i.e. any ensemble must have a better chance of obtaining a higher rating from lenient raters than from more severe raters); and (e) variable map (i.e. ensembles and raters must be simultaneously located on a single underlying latent variable). When adequate fit to the model is observed, rater-invariant measurement is achieved.

The MFR-PC model is specified as follows:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \lambda_i - \delta_j - \tau_{ik}, \tag{1}$$

where

$\ln[P_{nijk}/P_{nijk-1}]$ = the probability that Performance n rated by Rater i on Item j in level m receives a rating in category k rather than category k-1,

θ_n = the logit-scale location (e.g. achievement) of Performance i,

γ_i = the logit-scale location (e.g. severity) of Rater j,

δ_j = the logit-scale location (e.g. difficulty) of item m, and

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

Table 1. Summary statistics from the PC-MFR model.

	Facets		
	Performance (θ)	Rater (γ)	Item (δ)
Measure (Logits)			
Mean	−.05	.00	−.07
SD	.66	.47	.49
N	53	67	29
Infit *MSE*			
Mean	1.01	.99	.99
SD	.24	.17	.37
Std. Infit *MSE*			
Mean	−.10	−.10	−.70
SD	1.90	1.50	3.80
Outfit *MSE*			
Mean	1.04	1.04	1.04
SD	.28	.24	.47
Std. Outfit *MSE*			
Mean	.10	.20	−.50
SD	2.00	1.80	3.70
Separation statistics			
Reliability of separation	.98	.94	.98
Chi-square	1871.40*	1173.70*	1201.70*
Degrees of freedom	52	66	28

*$p < .01$.

The benefit of using the partial credit variation of the measurement model is to allow the rating scale categories to freely vary by each rater (Masters, 1982). This additional parameter allows for the investigation of rater differences in logit locations within the rating scale structure of each item. This provides a more precise estimate of ensemble true scores and better fit of the model to the data. The partial credit version of the measurement model was used because previous research in music performance assessment has verified that each rater demonstrates a unique rating scale structure for each item, resulting in an effect on overall ratings (Wesolowski et al., 2016a). Therefore, investigation into rating scale structure is warranted in this study.

Results

Summary statistics for the MFR-PC model

Table 1 provides the summary statistics for the MFR-PC model analysis of musical performances ($n = 53$), raters ($n = 67$), and items ($n = 29$). The analysis indicated an overall good model data fit with significant differences between performances ($\chi^2(52) = 1871.40$, $p < .01$), raters ($\chi^2(66) = 1173.70$, $p < .01$), and items ($\chi^2(28) = 1201.70$, $p < .01$). Reliability of separation statistics for all three facets are as follows: performances (*Rel* = .98), raters (*Rel* = .94), and items (*Rel* = .98). The separation statistic for performances can be interpreted similarly to Cronbach's alpha, indicating high reproducibility of relative measure locations. Separation statistics for raters and items can be interpreted as the separation verification of the hierarchy for

the elements within each facet (i.e. construct validity). Mean square (*MSE*) fit statistics demonstrate the overall randomness within the model (Linacre, 2002b). Perfect predictability is represented by the value of 1.0. Values less than 1.0 indicate too much predictability/redundancy in the data (i.e. muted data). Values above 1.0 indicate too little predictability in the data (i.e. sporadic data). In particular, infit mean square (*MSE*) fit statistics represent inlier-sensitive fit, where over- or under-fit for Guttman probabilistic patterns are detected. Outfit mean square (*MSE*) fit statistics represent outlier-sensitive fit, where over-fit for observations of model variance is detected. Overall, the separation statistics indicate that the sample confirms the hierarchy of item difficulty and rater severity. Indices of mean item fit for the performance facet (infit *MSE* = 1.01; outfit *MSE* = 1.04), rater facet (infit *MSE* = .99; outfit *MSE* = 1.04), and item facet (infit *MSE* = .99; outfit *MSE* = 1.04) center around 1.00 and are within the threshold of acceptability (0.5–1.5) for parameters. *MSE* statistics within the threshold indicate sufficient accuracy and predictability of model data fit, validity evidence for the variable map, and good productivity for measurement (Linacre, 2002b).

Variable map

The variable map is a visual representation of the hierarchical orderings of elements for each of the facets on the same log odds linear scale included in the measurement model (see Appendix A in Supplementary Materials online). Conceptually, the variable map is a visual representation of the operational definition of the latent construct. In this case, the latent construct can be described as high school concert band performance. The variable map includes facets for rater severity, item difficulty (i.e. rater endorsability of the items), and ensemble performance achievement. The variable map and following facet calibrations were estimated using *Facets* (Linacre, 2014).[1]

Research question 1

Calibration of rater facet. Table 2 provides rater calibration information. The rater facet was centered on the logit scale (mean of 0.00 logits) in order to provide a frame of reference for the interpretation of the item locations. As a result, the mean of the raters was 0.00 logits with a range of 1.36 for the most severe rater (Rater 52) to -1.87 for the most lenient rater (Rater 27). Expected values for facet-level mean square fit statistics are to center around 1.00. As indicated by Wright and Linacre's (1994) threshold for acceptable rater fit under the conditions of rating scale surveys (.60–1.40), none of the raters demonstrated any overly sporadic or muted behaviors.

Research question 2

The additional parameter from the partial credit formulation of the MFR model allows for the investigation of the rating scale structure for each rater's use according to each item. By reviewing these structures as well as frequency data, average observed and expected measures, and outfit *MSE* as provided in Table 3, one can evaluate specific rater behaviors and infer overall rater quality with valid, reliable, and linear empirical evidence. Table 3 provides the diagnosis of category structure by rater. The results do not support the hypothesis that the rating scale structure remains invariant across raters varying rating scale category thresholds. Specifically, Table 4 provides the Rasch-Andrich category threshold logit measures in relation to each rater's cluster assignment.

There are two important diagnostic points to discuss in evaluating the category diagnostics particular to this study (Linacre, 2002a). First, the single asterisk in Table 3 indicates violations

Table 2. Calibration of the rater facet.

Rater ID	Observed average	Measure	Standard error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
52	2.06	1.36	0.13	0.74	−2.51	0.76	−2.29
18	1.88	1.33	0.10	0.94	−0.73	0.97	−0.36
61	2.05	1.00	0.15	0.99	−0.05	1.08	0.56
14	2.49	0.84	0.18	0.98	−0.21	0.97	−0.25
17	2.00	0.58	0.11	1.04	0.35	1.07	0.65
40	1.88	0.57	0.10	0.80	−1.68	0.72	−2.02
41	1.79	0.57	0.10	1.04	0.31	1.08	0.54
7	2.04	0.49	0.09	0.96	−0.52	1.02	0.23
30	1.97	0.48	0.11	1.32	2.49	1.79	4.81
42	2.16	0.44	0.13	1.12	1.01	1.27	1.87
13	2.36	0.43	0.14	1.06	0.50	1.03	0.28
53	2.36	0.41	0.12	1.18	1.54	1.23	1.89
15	1.90	0.40	0.11	1.22	1.85	1.51	3.12
36	2.62	0.40	0.09	1.00	0.01	1.32	2.58
54	2.64	0.40	0.10	1.40	3.45	1.51	3.95
64	2.34	0.38	0.09	0.87	−1.29	0.84	−1.42
21	1.84	0.31	0.12	1.38	2.51	1.40	2.45
22	2.46	0.24	0.14	0.89	−0.85	0.88	−0.93
58	2.25	0.23	0.09	1.03	0.30	1.05	0.48
32	2.60	0.17	0.12	0.95	−0.41	0.95	−0.42
20	2.31	0.16	0.10	1.10	0.90	1.15	1.27
62	2.27	0.16	0.17	0.85	−1.22	0.84	−1.28
3	3.09	0.15	0.10	0.90	−0.86	0.84	−1.16
60	2.84	0.11	0.13	0.93	−0.49	0.92	−0.59
35	2.37	0.10	0.13	0.78	−1.99	0.77	−2.03
51	2.32	0.08	0.14	1.09	0.72	1.09	0.75
33	2.53	0.06	0.10	1.33	2.77	1.53	3.96
5	2.33	0.02	0.13	0.91	−0.81	0.89	−0.95
4	2.47	0.00	0.11	0.94	−0.54	0.92	−0.66
44	2.56	0.00	0.11	0.82	−1.66	0.80	−1.75
19	2.91	−0.04	0.10	0.85	−1.36	0.89	−0.85
23	2.45	−0.04	0.10	1.01	0.12	1.05	0.45
10	3.14	−0.05	0.10	0.95	−0.32	1.09	0.42
26	2.62	−0.06	0.10	1.05	0.51	1.10	0.93
57	2.71	−0.06	0.12	0.96	−0.30	1.00	0.03
29	2.41	−0.07	0.11	1.06	0.54	1.07	0.63
66	2.57	−0.07	0.17	1.00	0.07	1.00	0.07
46	2.27	−0.09	0.10	0.96	−0.35	0.94	−0.52
49	2.63	−0.09	0.12	1.07	0.60	1.12	1.03
16	2.38	−0.10	0.14	0.83	−1.45	0.81	−1.58
56	2.79	−0.10	0.11	1.29	2.41	1.33	2.59
9	2.50	−0.11	0.09	0.73	−2.72	0.71	−2.62
6	2.43	−0.13	0.12	0.88	−1.04	0.85	−1.30
59	2.53	−0.16	0.13	0.67	−3.08	0.67	−3.02
63	2.72	−0.16	0.13	1.29	2.25	1.39	2.95
1	2.59	−0.19	0.11	0.68	−3.18	0.69	−3.06

Table 2. (Continued)

Rater ID	Observed average	Measure	Standard error	Infit MSE	Std. Infit	Outfit MSE	Std. Outfit
8	2.61	−0.20	0.11	1.24	1.96	1.32	2.30
39	2.80	−0.22	0.11	0.92	−0.70	0.94	−0.53
27	2.79	−0.23	0.10	1.08	0.57	1.16	0.79
25	2.52	−0.24	0.11	0.78	−2.02	0.77	−2.17
48	2.61	−0.29	0.14	1.10	0.85	1.11	0.88
24	2.87	−0.30	0.10	0.84	−1.51	0.86	−1.09
28	2.84	−0.32	0.13	1.05	0.44	1.04	0.34
45	2.41	−0.33	0.10	1.28	2.47	1.32	2.66
67	2.80	−0.34	0.10	0.74	−2.48	0.72	−2.29
50	2.63	−0.36	0.10	1.04	0.37	0.98	−0.12
31	2.28	−0.37	0.10	1.04	0.39	1.24	1.49
2	2.44	−0.40	0.10	0.65	−3.79	0.65	−3.41
65	2.55	−0.46	0.09	0.85	−1.36	0.99	−0.02
38	2.85	−0.62	0.09	0.84	−1.55	1.00	0.03
11	3.00	−0.70	0.14	1.13	0.97	1.16	1.08
43	2.59	−0.70	0.11	0.99	−0.03	1.00	0.08
47	2.58	−0.76	0.12	0.91	−0.79	0.90	−0.82
37	3.03	−0.82	0.10	1.02	0.22	1.65	2.56
34	2.63	−0.83	0.12	1.03	0.26	1.02	0.20
55	2.45	−0.93	0.10	0.89	−1.00	0.89	−1.06
12	3.21	−0.95	0.13	1.18	1.39	1.30	2.19

of monotonicity. In these instances, person locations on the latent variable do not always increase with the rating scale categories. This illustrates a clear indication of raters' inaccurate use of specific rating scale categories on specific items. Second, the double asterisk indicates seven instances where Outfit *MSE* values are equal to or exceed 2.0 (e.g. Rater 12, category 1; Rater 15, category 2; Rater 33, category 1; Rater 36, category 1; Rater 37, category 1; Rater 54, category 1; and Rater 63, category 1). These cases indicate that there is substantive unpredictability in the ratings, ultimately producing an insufficient balance between redundancy and unpredictability based upon the expectations of the model.

Research question 3

Differential rater functioning: Bias interaction analysis. Bias interaction analysis is a secondary analysis that seeks to empirically define systematic leniency/severity in rater behavior. The benefit of using bias measures as a means to individually define systematic differences in rater behavior is that similar to each parameter in the model being statistically separable, so too are bias measures. Each bias measure resulting from rater by item interaction is freed from within group distribution resulting in estimates that are conditionally independent. This allows for independent and consistent estimates of differential rater functioning for each rater's response to an item. Substantively, bias measures provide empirical evidence for two important questions related to rater behavior: (a) are the ratings of each rater invariant over construct irrelevant components? and (b) are the raters invariant over construct irrelevant components for the overall assessment system? (Engelhard, 2013, p. 212).

Table 3. Category structure diagnostics: Rater behavior of category usage, average observed and expected logit measures, and outfit *MSE*.

Rater	Observed category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit *MSE*			
	1	2	3	4	1	2	3	4	1	2	3	4
1	15 (13)	37 (32)	44 (83)	20 (17)	−.54 (−.31)	−.22 (−.06)	.32 (.24)	.81 (.54)	.70	.70	.50	.70
2	32 (28)	27 (23)	31 (27)	26 (22)	−.76 (−.47)	−.04 (−.22)	.02 (.06)	.57 (.35)	.50	1.0	.60	.70
3	11 (9)	20 (17)	32 (28)	53 (46)	.32 (.30)	.41 (.43)	.38* (.61)	.98 (.84)	.90	.90	.50	.80
4	15 (13)	54 (47)	24 (21)	23 (20)	−.19 (−.14)	.03 (.03)	.20 (.25)	.58 (.49)	1.00	1.00	1.00	.80
5	19 (16)	45 (39)	47 (41)	5 (4)	−.86 (−.82)	−.70 (−.59)	−.18 (−.32)	−.27* (−.05)	.90	.60	.80	1.20
6	19 (16)	37 (32)	51 (44)	9 (8)	−.87 (−.72)	−.52 (−.43)	.07 (−.10)	−.09* (.23)	.80	.80	.70	1.20
7	76 (33)	79 (34)	68 (29)	9 (4)	−1.69 (−1.44)	−.63 (−.95)	−.45 (−.46)	−.78* (−.04)	.70	.60	1.10	2.20
8	25 (22)	27 (23)	32 (28)	32 (28)	−.29 (−.56)	−.23 (−.19)	.17 (.31)	.81 (.84)	1.70	1.00	1.60	1.00
9	36 (31)	18 (16)	30 (26)	32 (28)	−.51 (−.42)	−.44 (−.20)	.26 (.09)	.51 (.42)	.80	.20	.70	.70
10	23 (20)	8 (7)	15 (13)	70 (60)	−.13 (−.20)	−.02 (.18)	.33 (.57)	.98 (.93)	1.70	.50	.40	.70
11	6 (5)	32 (28)	34 (29)	44 (38)	.10 (−.27)	.03* (.09)	1.17 (1.05)	1.95 (2.06)	1.50	1.00	1.10	1.10
12	2 (2)	16 (14)	54 (47)	44 (38)	1.68 (.79)	1.12 (.93)	1.11 (1.11)	1.24 (1.34)	**2.10	1.30	1.20	1.10
13	14 (12)	49 (42)	50 (43)	3 (3)	−.82 (−.80)	−.60 (−.63)	−.39 (−.40)	−.83* (−.17)	1.00	1.10	1.00	1.20
14	–	61 (53)	53 (46)	2 (2)	–	−2.02 (−1.99)	−1.69 (−1.72)	−1.63 (−1.45)	–	1.00	1.00	1.00
15	58 (50)	18 (16)	34 (29)	6 (5)	−1.25 (−1.29)	−.85 (−.93)	−.46 (−.57)	−1.46* (−.27)	1.00	**2.00	.70	4.40
16	16 (14)	43 (37)	54 (47)	3 (3)	−1.72 (−1.49)	−.96 (−.86)	−.07 (−.25)	−.69* (.20)	.60	.90	.60	1.50
17	41 (35)	42 (36)	25 (22)	8 (7)	−.90 (−.85)	−.53 (−.66)	−.51 (−.43)	−.35 (−.21)	.90	1.10	1.20	1.20
18	78 (34)	106 (46)	47 (20)	1 (0)	−2.17 (−2.03)	−1.53 (−1.66)	−1.29 (−1.25)	−2.25* (−.86)	.80	.70	1.10	2.00
19	16 (14)	19 (16)	41 (35)	40 (34)	−.04 (−.06)	.02 (.16)	.33 (.44)	.93 (.75)	1.20	.70	.50	.80
20	36 (31)	29 (25)	30 (26)	21 (18)	−.57 (−.580	−.28 (−.33)	.01 (−.04)	.10 (.25)	1.30	.90	.70	1.30
21	47 (41)	49 (42)	12 (10)	8 (7)	−.85 (−1.04)	−.83 (−.75)	−.63 (−.44)	−.56 (−.15)	1.30	1.10	1.20	1.80
22	7 (6)	56 (48)	46 (40)	7 (6)	−.21 (−.29)	−.21 (−.12)	.14 (.11)	.72 (.34)	1.10	.80	.90	.70
23	27 (23)	32 (28)	35 (30)	22 (19)	−.65 (−.55)	−.11 (−.23)	.14 (.09)	.28 (.40)	.90	1.40	.80	1.20
24	15 (13)	26 (22)	34 (29)	41 (35)	−.16 (−.09)	.09 (.17)	.47 (.48)	.92 (.82)	1.10	.80	.60	.80
25	19 (16)	38 (33)	39 (34)	20 (17)	−.57 (−.51)	−.42 (−.22)	.35 (.19)	.82 (.69)	1.00	.40	.50	.90
26	24 (21)	28 (24)	32 (28)	32 (28)	−.14 (−.21)	−.04 (−.03)	.15 (.20)	.46 (.45)	1.30	.80	.90	1.10
27	31 (27)	16 (14)	15 (13)	54 (47)	−.45 (−.53)	−.17 (−.13)	.29 (.46)	1.02 (1.01)	1.20	.80	1.40	1.20
28	6 (5)	28 (24)	60 (52)	22 (19)	.09 (−.06)	.25 (.25)	.61 (.61)	.94 (.98)	1.10	1.00	1.20	1.00
29	22 (19)	47 (41)	24 (21)	23 (20)	−.45 (−.38)	−.01 (−.14)	−.05* (.15)	.49 (.48)	.90	1.70	1.20	.90
30	46 (40)	37 (32)	24 (21)	9 (8)	−1.13 (−1.14)	−.46 (−.70)	−.36 (−.35)	−1.11* (−.08)	1.00	.90	1.10	**4.10
31	44 (38)	22 (19)	24 (21)	26 (22)	−.92 (−.93)	−.48 (−.42)	.30 (.09)	.33 (.50)	.90	1.50	.30	**2.10
32	11 (9)	41 (35)	47 (41)	17 (15)	−.36 (−.42)	−.09 (−.04)	.32 (.35)	.87 (.71)	1.10	1.00	.90	.80
33	28 (24)	27 (23)	33 (28)	28 (24)	−.20 (−.44)	.02 (−.17)	−.17* (.16)	.47 (.49)	**2.10	1.50	1.70	1.00
34	19 (16)	26 (22)	50 (43)	21 (18)	−.92 (−.82)	−.24 (−.25)	.45 (.33)	.53 (.78)	.80	.90	1.10	1.30
35	17 (15)	47 (41)	44 (38)	8 (7)	−.89 (−.83)	−.66 (−.51)	.05 (−.05)	.90 (.47)	.90	.70	.80	.70
36	33 (28)	19 (16)	23 (20)	41 (35)	.03 (−.17)	−.28* (−.03)	−.15 (.16)	.52 (.39)	**2.20	.10	1.40	.60

Table 3. (Continued)

Rater	Observed category usage (%)				Average observed logit measure (Average expected logit measure)				Outfit *MSE*			
	1	2	3	4	1	2	3	4	1	2	3	4
37	23 (20)	15 (13)	14 (12)	64 (55)	.01 (−.17)	−.13* (.12)	.37 (.53)	.98 (.96)	**3.40	.30	.50	.70
38	25 (22)	18 (16)	22 (19)	51 (44)	−.23 (−.16)	.08 (.08)	.33 (.37)	.73 (.68)	1.00	1.50	1.10	.80
39	13 (11)	30 (26)	40 (34)	33 (28)	.00 (−.12)	.10 (.16)	.30 (.46)	.97 (.77)	1.30	.70	1.00	.70
40	57 (49)	30 (26)	15 (13)	14 (12)	−.76 (−.77)	−.75 (−.57)	−.26 (−.35)	.14 (−.14)	1.00	1.10	.60	.50
41	61 (53)	30 (26)	13 (11)	12 (10)	−.87 (−.85)	−.59 (−.65)	−.37 (−.430	−.36 (−.22)	1.00	.80	.90	1.30
42	34 (29)	33 (28)	45 (39)	4 (3)	−1.80 (−1.71)	−.68 (−.91)	−.42 (−.40)	−1.12* (−.07)	1.10	.50	1.10	**3.10
43	17 (15)	36 (31)	41 (35)	22 (19)	−.43 (−.44)	−.13 (−.12)	.29 (.26)	.62 (.65)	1.10	.80	1.20	.90
44	20 (17)	36 (31)	35 (30)	25 (22)	−.52 (−.52)	−.28 (−.15)	.23 (.280	.97 (.72)	1.00	.70	.60	.70
45	27 (23)	36 (31)	32 (28)	21 (18)	−.32 (−.48)	−.12 (−.24)	−.14* (.05)	.20 (.34)	1.30	1.60	1.30	1.20
46	35 (30)	32 (28)	32 (48)	17 (15)	−.73 (−.69)	−.43 (−.41)	.01 (−.09)	.14 (.21)	.90	.70	1.00	1.00
47	9 (8)	46 (40)	46 (40)	15 (13)	.01 (−.15)	−.09* (.04)	.31 (.27)	.70 (.52)	1.20	.70	.90	.80
48	7 (6)	42 (36)	56 (48)	11 (9)	−.13 (−.47)	.00 (−.05)	.25 (.35)	.78 (.69)	1.20	1.00	1.40	.90
49	16 (14)	36 (31)	39 (34)	25 (22)	−.69 (−.68)	−.12 (−.20)	.44 (.48)	.98 (1.04)	1.20	1.20	1.00	1.10
50	28 (24)	22 (19)	31 (27)	35 (30)	−.39 (−.42)	−.11 (−.11)	.19 (.25)	.64 (.61)	.00	.80	1.20	.90
51	15 (13)	54 (47)	42 (36)	5 (4)	−1.12 (−.90)	−.40 (−.58)	−.28 (−.18)	−.15 (.23)	.80	1.50	1.10	1.30
52	36 (31)	41 (35)	35 (30)	4 (3)	−1.64 (−1.42)	−.97 (−1.05)	−.35 (−.47)	.38 (.14)	.70	.60	.90	.80
53	25 (22)	38 (33)	39 (34)	14 (12)	−.65 (−.800	−.45 (−.41)	.09 (.03)	.08* (.42)	1.30	1.30	.70	1.50
54	23 (20)	28 (24)	33 (28)	32 (28)	.17 (−.23)	.10* (−.02)	−.12* (.24)	.51 (.52)	**2.00	1.20	1.90	1.00
55	25 (22)	34 (29)	37 (32)	20 (17)	−.47 (−.41)	−.24 (−.21)	.05 (.03)	.38 (.29)	.90	1.00	.90	.90
56	13 (11)	28 (24)	45 (39)	30 (26)	.22 (−.10)	.22 (.13)	.39 (.43)	.59 (.74)	1.60	1.10	.90	1.40
57	10 (9)	42 (36)	36 (31)	28 (24)	.26 (−.22)	−.10* (.07)	.46 (.54)	1.33 (1.15)	1.90	.50	.80	.80
58	39 (34)	30 (26)	26 (22)	21 (18)	−.34 (−.48)	−.47* (−.31)	−.25 (−.10)	.29 (.12)	1.20	.60	1.50	.80
59	10 (9)	45 (39)	51 (44)	10 (9)	−.67 (−.36)	−.32 (−.14)	.25 (.13)	.88 (.40)	.80	.50	.70	.70
60	7 (6)	24 (21)	66 (57)	19 (16)	−.16 (.08)	.27 (.23)	.41 (.430	.79 (.67)	.80	1.00	1.00	.90
61	18 (16)	77 (66)	18 (16)	3 (3)	−1.19 (−1.05)	−.72 (−.74)	−.14 (−.41)	−1.36* (−.10)	1.00	.80	.70	**2.00
62	11 (9)	64 (55)	40 (34)	1 (1)	−2.11 (−1.76)	−1.21 (1–.16)	−.23 (−.44)	−1.94* (.06)	.80	.70	.80	**2.00
63	8 (7)	39 (34)	47 (41)	22 (19)	.90 (−.46)	−.12* (.00)	.48 (.62)	1.32 (1.31)	**3.10	.60	1.20	1.00
64	4 (38)	18916)	25 (22)	29 (25)	−.46 (−.41)	−.21 (−.25)	−.11 (−.04)	.29 (.18)	.90	.80	1.10	.70
65	43 (37)	10 (9)	19 (16)	44 (38)	−.57 (−.47)	−.05 (−.17)	.23 (.16)	.49 (.48)	1.40	.90	.20	.80
66	2 (2)	49 (42)	62 (53)	3 (3)	−.55 (−.63)	−.08 (−.08)	.41 (.42)	.86 (.79)	1.10	1.00	1.00	1.00
67	17 (15)	30 (26)	28 (24)	41 (35)	−.19 (−.18)	−.12 (.08)	.38 (.45)	1.13 (.93)	1.10	.30	.50	.70

Note. Category 1 = "strongly disagree"; Category 2 = "disagree"; Category 3 = "agree"; Category 4 = "strongly agree".
* indicates violation of monotonicity; ** indicates Outfit *MSE* ⩾ 2.00.

Table 4. Rasch-Andrich category thresholds, cluster membership, and squared Euclidian distance measures by rater.

Rater	Category thresholds			Cluster	Distance
	1 to 2	2 to 3	3 to 4		
1	−1.09	−0.09	1.18	3	4.16
2	−0.17	−0.22	0.39	3	4.34
3	−0.25	0.04	0.21	3	3.33
4	−1.35	0.94	0.41	3	5.45
5	−1.57	−0.50	2.06	2	3.97
6	−1.23	−0.58	1.81	2	3.61
7	−1.24	−0.55	1.79	3	4.93
8	−0.47	−0.12	0.58	2	5.16
9	0.38	−0.57	0.19	3	4.64
10	1.04	−0.25	−0.79	2	3.94
11	−1.79	0.45	1.34	1	4.62
12	−1.23	−0.20	1.43	1	4.06
13	−1.98	−0.54	2.52	2	4.85
14	–	−1.71	1.71	2	5.46
15	0.06	−1.38	1.32	2	3.71
16	−2.15	−0.75	2.90	1	3.33
17	−0.79	−0.03	0.82	1	4.75
18	−2.16	−0.64	2.80	3	5.50
19	−0.13	−0.48	0.61	3	4.17
20	−0.24	−0.22	0.46	1	4.80
21	−0.94	0.82	0.12	2	4.68
22	−2.29	0.19	2.11	1	3.62
23	−0.56	−0.15	0.71	3	3.69
24	−0.51	0.05	0.46	2	3.61
25	−1.06	−0.05	1.11	3	4.08
26	−0.28	−0.05	0.33	2	3.96
27	0.31	0.22	−0.54	2	3.50
28	−1.45	−0.34	1.79	3	5.64
29	−1.03	0.67	0.36	3	4.47
30	−0.69	−0.08	0.78	2	5.05
31	0.01	−0.24	0.23	1	3.67
32	−1.56	0.01	1.55	1	4.07
33	−0.28	−0.21	0.49	1	4.38
34	−0.85	−0.60	1.45	3	5.02
35	−1.69	−0.22	1.92	2	4.47
36	0.44	−0.13	−0.31	3	4.18
37	0.39	0.39	−0.78	3	3.16
38	0.29	0.03	−0.31	1	4.00
39	−0.82	0.02	0.80	1	4.95
40	−0.04	0.22	−0.18	3	4.23
41	−0.05	0.29	−0.24	3	3.78
42	−1.26	−0.94	2.20	2	3.30
43	−1.03	−0.05	1.08	3	4.02
44	−0.93	0.09	0.84	3	4.48

Table 4. (Continued)

Rater	Category thresholds			Cluster	Distance
	1 to 2	2 to 3	3 to 4		
45	−0.64	0.03	0.62	3	5.78
46	−0.46	−0.25	0.70	2	5.05
47	−0.17	0.16	1.53	2	5.32
48	−2.04	−0.12	2.16	2	4.88
49	−1.27	0.05	1.22	3	5.49
50	−1.27	0.05	1.22	3	5.39
51	−2.03	−0.13	2.16	2	5.04
52	−1.39	−0.62	2.01	1	4.86
53	−1.03	−0.22	1.26	2	5.28
54	−0.34	−0.07	0.40	1	5.99
55	−0.61	−0.17	0.78	1	4.84
56	−0.77	−0.21	0.98	2	4.40
57	−1.53	0.44	1.09	3	4.37
58	−0.15	−0.07	0.22	2	5.78
59	−1.76	−0.14	1.89	3	4.25
60	−1.09	−0.69	1.79	1	4.25
61	−2.38	0.85	1.52	1	4.27
62	−3.23	−0.31	3.54	3	4.02
63	−1.83	0.11	1.72	1	5.62
64	0.56	−0.48	−0.08	3	4.25
65	1.15	−0.64	−0.51	3	3.61
66	−3.57	−0.06	3.63	3	5.79
67	−0.63	0.32	0.30	3	3.38

Note. Category 1 = "strongly disagree"; Category 2 = "disagree"; Category 3 = "agree"; Category 4 = "strongly agree". Cluster 1 = the syntactical rater, Cluster 2 = the expressive rater, Cluster 3 = the mental representation rater.

The interaction term between raters and items ($\lambda_i\gamma_j$) was added to the MFR-PC model as follows:

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

where $\lambda_i\gamma_j$ is the interaction between rater severity and item difficulty.

The interaction parameter tests the null hypothesis that the overall set of rater and item interactions is not significantly different from zero. Evidence of the significance of this omnibus test is a chi square statistic. Each interaction term is reflected as a *t*-statistic that represents the size of each pairwise comparison. The *t*-statistic can be conceptualized as a type of effect size, indicating usefulness of the data according to the model (Linacre, 2003). Bias measures above 0.00 indicate observed scores above what was expected by the model. Bias measures below 0.00 indicate observed scores below what was expected by the model. The *t*-statistic for each individual rater by item was used to define the rater clusters.

A differential rater functioning (rater by item) bias interaction analysis was conducted using *Facets* (Linacre, 2014). The analysis indicated an overall statistically significant differential measure ($\chi^2(1943) = 2322.20$, $p < .01$), explaining 24.83% of the variance within the measurable responses. A total of 1943 interactions were extrapolated.[2]

Research question 4

Cluster analyses. A hierarchical cluster analysis was conducted in order to meaningfully group raters by degree of association based upon differential leniency/severity measures. A hierarchical cluster analysis was used first as an exploratory tool to find the most meaningfully significant cluster solution possible. Appendix B in Supplementary Materials online presents the aggregate of raters represented by a hierarchical dendrogram. The dendrogram represents an agglomerative procedure, where each rater represents its own cluster. In order to partition the raters, Ward's linkage agglomerative clustering method was used. Ward's method was specifically used in order to determine the degree of acceptability in which clusters are linked together by maximizing intra-class similarity and minimizing inter-class similarity. Additionally, it is regarded as the most efficient linkage procedure through the minimization of within-cluster sums of squares over all partitions available (Ward, 1963). Squared Euclidian distances were selected as a method for computing the proximity between raters. The advantage of using this method is that other outlier objects do not affect distances between two objects. Additionally, it places progressively greater weights on raters further apart (Romesburg, 1984).

In order to target an appropriate number of clusters, Mardia, Kent, and Bobby's (1979) "rule of thumb" was considered where the number of clusters (k) is approximately equal to the square root of n divided by k. Additionally, Thorndike's (1953) elbow method was considered. Cluster solutions ranging from 2–8 were examined for frequency and the discernable and reasonable nature of substantive trends. Based upon the best meaningfully substantive interpretation, a three-cluster solution was elected.

A post hoc non-hierarchical K-means clustering was used to generate a three-cluster solution with the greatest possible case distinction. A K-means clustering technique was used in order to iteratively estimate cluster means based upon smallest distance to the cluster mean. Cluster centers (i.e. seeds) generated from the hierarchical clustering were used to pre-specify threshold distances. Table 5 provides centroid values for each cluster by item. Interpreting 0 as the mean/median for the values of each cluster dimensions provides an anchor for interpretation of the centroid values. Centroid values below 0 can be interpreted as having less value than other clusters. Conversely, centroid values above 0 can be interpreted as having more value than other clusters. As seen in Table 6, F values and significance levels for each item indicate how well the items discriminate the three clusters. The larger the observed significance, the less the variable contributes to the separation of the clusters.

Cluster labels. The results of the cluster analyses indicated a three-cluster solution. Cluster 1 (C1) can be labeled as *Syntactical rater-type* (N = 17, 25.40%). Based upon use of the rating scale, the syntactical rater seems to be more interested in the technical execution and accuracy of the notes, rhythms, balance, and sonority of the musical performance. It seems as though this rater reacts to when the syntax of the musical piece is executed improperly or not heard. In these instances, the limitations of the ensemble's execution may limit the rater from assembling the syntax into meaningful, logical structures. The salient dimension-level characteristics for C1 include: (a) tone/intonation of much importance; (b) balance is of importance (compared to Cluster 3); (c) interpretation is of importance (compared to Cluster 3); and (d) technical accuracy of importance. The salient item-level characteristics for C1 include: (a) excellent balance between parts in a section (Item 7) is of importance; (b) rhythmically accurate (item 22) is of importance; and (c) entrances are not precise (item 29) is of much importance.

Cluster 2 (C2) can be labeled as *Expressive rater-type* (N = 21, 31.30%). Based upon use of the rating scale, the expressive rater seems to be more interested in the items that reflect the eliciting

Table 5. Cluster centroids by item.

Item dimension	Item stem	Cluster		
		1	2	3
1	1. Intonation is quite good.	2.22	−1.55	.51
1	2. Basic tuning is not good, band plays out of tune.	1.35	−1.92	1.76
1	3. Tone is shallow.	.28	−.16	−.75
1	4. The band plays with a well-defined pitch center.	1.57	.03	.71
1	5. The band has good control in high pitch registers.	2.12	−.20	−.80
1	6. Tutti chords are badly out of tune.	1.21	−1.21	.27
2	7. Excellent balance between parts in a section.	1.46	.26	−.45
2	8. Excellent balance in all parts.	.41	−1.45	−1.45
2	9. Lower parts balance the group well.	.61	.92	−.33
2	10. All sections are well-balanced within.	.75	1.14	−.71
2	11. Inner parts balance well.	.86	−.40	−.05
2	12. Inner parts are too timid.	−.16	2.15	−1.45
3	13. Performance shows careful attention to dynamics.	.06	1.80	−1.36
3	14. Performance has a good, wide variety in dynamics.	.06	1.96	.32
3	15. Needs to be more expressive.	.30	2.09	−1.89
3	16. Crescendo and diminuendo are properly graduated.	−.63	1.27	−2.18
3	17. Performance lacks emotion.	.98	1.45	−1.96
4	20. Dotted eighth-sixteenth pattern is inaccurate.	−.20	−1.90	.38
4	21. Syncopated patterns are correctly played.	−.04	−1.24	−1.06
4	22. Rhythmically accurate.	1.17	−1.11	−.46
5	25. Runs are played accurately and smoothly.	1.41	−1.97	1.37
5	26. Notes in runs are inaccurate.	1.14	.44	2.06
5	27. Fingering problems cause a lack in clarity.	.25	.20	−.46
5	28. Awkward and difficult passages are not prepared.	1.49	.89	−.99
5	29. Entrances are not precise.	2.24	−.51	.73

Note. Construct dimensions separated via shaded areas. Dimension 1 = tone/intonation; Dimension 2 = balance; Dimension 3 = interpretation; Dimension 4 = rhythm; Dimension 5 = technical accuracy.

of modulations in the auditory cues related to expression. These include items pertaining to modulations in dynamics, intonation, and timbre. Additionally, this rater reacts to items that specifically use the terms "expression" and "emotion." The salient dimension-level characteristics for C2 include: (a) tone/intonation is of importance; (b) rhythm is of less importance; and (c) interpretation is of more importance. The salient item-level characteristics for C2 include: (a) inner parts are too timid (item 12) is of much importance; and (b) runs played accurately and smoothly (item 25) is of less importance.

Cluster 3 (C3) can be labeled as *Mental representation rater-type* ($N = 29$, 43.30%). Based upon use of the rating scale, the mental representation rater seems to rely on and engage within internal, musical imagery more so than the actual external sounds of the performance itself. This rater reacts less to items reflecting balance, intonation, interpretation, and sonority. The salient dimension-level characteristics for C3 include: (a) intonation is of more importance; (b) tone is of less importance; (c) balance of less importance; (d) interpretation of less importance; and (e) technical accuracy is of somewhat importance. The salient item-level characteristics for C3 include: (a) tone is shallow (item 3) is of less importance; and (b) the band has good control in high pitch registers (item 5) is of less importance.

Table 6. Analysis of Variance table.

	Cluster mean square ($df = 2$)	Error mean square ($df = 64$)	F	p
Item 1	6.06	.52	11.69	>.001
Item 2	10.61	.68	15.66	>.001
Item 3	1.67	.93	1.80	.17
Item 4	6.30	.77	8.15	>.001
Item 5	.92	.93	.99	.38
Item 6	2.44	.87	2.81	.07
Item 7	1.41	.53	2.66	.08
Item 8	4.62	1.11	4.18	.02
Item 9	.42	.72	.59	.56
Item 10	.60	.79	.75	.48
Item 11	9.86	1.30	7.58	>.001
Item 12	7.59	.54	14.08	.001
Item 13	5.19	.77	6.78	.002
Item 14	7.42	.91	8.12	.001
Item 15	7.51	.80	9.38	>.001
Item 16	7.40	.86	8.61	>.001
Item 17	1.80	.91	1.99	.15
Item 20	9.05	.91	9.99	>.001
Item 21	2.83	.86	3.28	.04
Item 22	7.21	1.01	7.14	.002
Item 25	9.05	.77	12.36	>.001
Item 26	15.72	1.13	13.91	>.001
Item 27	3.39	1.08	3.15	.05
Item 28	11.40	1.09	10.50	>.001
Item 29	7.02	.95	7.40	.001

Note. The *F* tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Cluster interpretations. A straightforward method to interpret the cluster behaviors is to consider at face value the various emphases outlined above as simple artifacts of raters' self-directing of the auditory spotlight: (a) raters in C2 consider interpretation of more importance than raters in C1 and C3; (b) raters in C3 place more of an emphasis on balance than C1; (c) raters in C1 and C3 value intonation more than C2, etc. However, a more broad speculation of raters' cognition and perception based upon musical structure and expectancy may shed some light on the various similarities and differences outlined above. Of course, the following is speculative and warrants further investigation through experimental manipulations of testing conditions, audio stimuli, and rater training/interventions.

As indicated in Namour's (1990) Implication-Realization (I-R) model, expectancy in music listening is driven by both bottom-up and top-down cognitive processes, where both processes interact, yet work separately. Bottom-up processes refer to the direct impact of sensory information input from the musical stimulus. Focus of these processes is on the innate, human cognitive and perceptual mechanisms. Top-down processes refer to perception driven by cognition, where the brain fills in the gaps of what is expected. Human perception of music is constrained by the processes of the auditory system, and is therefore influenced by how the system encodes and retains acoustic information (McDermott & Oxenham, 2008).

It can be hypothesized that C1 demonstrates cognitive constraint on the processing of pitch and timbre perception, which plays a significant role in the cognitive processing of tonal structure. C1 raters seem to be working from a top-down cognitive approach, where the auditory stimuli need to be clear and emphasized in the musical performances in order to perceive a tonal structure. Eliciting structure from multisensory cues in music is a contextually-driven, top-down sensory process requiring a system of complex auditory abilities to extract, organize, and interpret acoustic information (Baldwin, 2012). In particular, pitch and timbre (i.e. tone color) perception is based upon steady-state cues. Human perceptions of these steady-state cues are multidimensional, as they are affected by interactions between frequency regions and harmonic resonance (Ladefoged & Broadbent, 1957). Timbre discrimination becomes more difficult and confusing for the listener with the elimination of onset transients (Elliott, 1975; Saldanha & Corso, 1964). Interestingly, in the case of C1, raters placed a strong emphasis on item 29 (Entrances are not precise). Additionally, timbre consistency plays an important role in timbre perception. Iverson and Krumhansl (1993) found that timbres of different instruments (e.g. bassoon and French horn) are easily confused by performances of the same pitch in different registers. Notably, C1 demonstrates more emphasis on item 5 (The band plays with good pitch control in high registers). Additionally, related to pitch perception is the factor of contour, where patterns of relationships act as a facilitator of short musical memory (Dowling & Bartlett, 1981; Dowling & Fujitani, 1971). Evidence of this can be seen in C1's emphasis on the need to hear clear syntax of item 25 (Runs are played accurately and smoothly).

In contrast to C1, C3 places less overall emphasis on dimension 2: balance. It can be hypothesized that C3 is working more from a bottom-up cognitive approach that relies on auditory imagery and expectancy in order to "fill in the structural gaps" of what C1 was missing in the musical performances. This aligns with Meyer's (1956) view on embodied musical meaning, where the aesthetic power of the music comes from expectations of the listener. Similar to C1, the dimensions of intonation and technical accuracy are important in the performance. However, there is a drastic shift in the emphasis placed on balance dimension (i.e. items 7–12). C3 raters seem to be more reliant on the unique auditory images from perceived auditory information and distracted less than C1 on the dimensions of tone, intonation, balance, and interpretation.

C2 demonstrated drastically different characteristics than C1 and C3. Raters in C2 demonstrated a strong emphasis toward items under the interpretation dimension (e.g. items 13–17). Notably large indices were related to items reflecting musical expression and the modulations of expressive cues such as dynamics. It can be hypothesized that C2 raters, similar to C3 raters, have a strong reliance on auditory images, as little emphasis was placed on the dimensions of tone/intonation, balance, rhythm, and technical accuracy. In considering Frijda's (1988) Law of Comparative Feeling, the raters may be using their schemata as a frame of reference for the listening experience, comparing what they are hearing to past performances or similar qualities of past musical performances that moved them emotionally. The muted indices in the other dimensions may rule out the need for raters wanting increased modulations of auditory behavioral cues such as timing deviations, intensity, intonation, articulation, and timbre as a method to increase arousal.

A limitation of this study is the initial labeling of the clusters themselves. The use of qualitative labels to identify groups of raters in the context of this study is an initial step in identifying types of raters. As noted in the introduction to this article, the investigation falls under the category of an empirically driven investigation into statistical indices that underscore the measurement process. Clear empirical evidence exists of how statistical differences in raters exist based upon raters' differential severity of items, use of rating scale structure, and interpretation of dimensions in music performance; however, the true perceptual differences in mental

representations of the raters compared to the empirical results of the raters' rating scale behavior remains unclear. At this time, the labeling system should only be interpreted as an initial qualitative identifier to distinguish between clusters of raters demonstrating significant differences in rating processes.

Of course, psychological explanations for these occurrences are at best speculative and could certainly be limited by sample dependency. A retest of a different sample representing the same population may provide additional insight into the formulation of these clusters. Additionally, this methodology applied to samples representing different populations may provide additional insight into the listening and evaluation experiences.

Conclusion and implications

The purpose of this study was to identify a typology of rater types based upon systematic differential severity indices in the context of large ensemble music performance assessment. The first research question asked if individual raters maintain invariant levels of severity when rating high school concert band performances. Fit statistics from the MFR-PC model indicated that all raters demonstrated adequate fit to the model in the context of rating scale survey data. The second research question asked how rating scale structure varied across raters. Use of the partial credit formulation of the model allowed the rating scale structure to vary freely by rater. The results did not support the hypothesis that the rating scale structure remains invariant across raters varying rating scale category thresholds as evidenced by unique category thresholds for each rater. Additionally, some outfit MSE statistics brought to light violations of stochasticity and the comparison of observed versus expected logit measures indicated some violations of monotonicity. The third research question asked if differential severity emerges for individual raters according across items. The results indicated statistical indices unique to each rater by item that were found to be both above and below what was expected by the measurement model. The fourth research question asked if a meaningful typology existed based upon raters' differential severity. Results indicated a three-cluster solution demonstrating a distinct pattern of differential rater functioning based upon dimensions of music performance and related items. The rater types were labeled as (a) contextual rater type; (b) expressive rater type; and (c) mental representation rater type.

Underscoring the framework of scoring was empirical evidence gleaned from each raters' proficiency rating scale use. This includes indices measuring raters' overall leniency/severity, category usage, and differential leniency/severity by item (i.e. Rasch bias measures). These indices provide evidence of variability stemming from each raters' unique interpretation of the measurement instrument as a result of their schema. The framework of audition is a direct result of each individual rater's schemata based upon their experiential and environmental differences. In order to gain an empirical perspective of rater schemata, a cluster analysis based upon bias indices provided an empirical definition of distinct rater types.

Considering the rating process from a cognitive perspective can lead to the improvement of measure development, scoring processes, and training procedures. Ratings resulting from any rater-mediated assessment protocol are a representation of the rater, and not necessarily of the performance itself (Engelhard, 2002). If a clear methodology to managing rater mediated assessment data is not put into place, the threat of construct-irrelevant variability may obscure what is being measured, the use of the measurement apparatus itself, and the resulting raters' scores (Lane & Stone, 2006). Therefore, a clear empirical understanding of rater behavior is necessary for valid, reliable, and fair assessment practices in the evaluation of musical performances.

Measurement instruments that are carefully and strategically designed in conjunction with carefully trained and monitored raters can improve consistency and precision in the rating process. The incorporation of accuracy models and a more in-depth study of the effects of rater clusters on model fit may provide practical applications of the results of this research. In hermeneutic music performance scoring systems, selecting raters that represent a variety of rater types may provide a more broad and diverse (yet better controlled) perspective to the assessment process. Conversely, in psychometric music performance scoring systems, knowledge of rater types can inform and perhaps streamline the rater training process to provide more equitable and precise evaluations. The problem still exists, however, that little information is known about the relationship between rater cognition and rater proficiency. Investigation into rater training procedures related to rater types can provide insight into the connection of rater proficiency, environmental, and experiential implications to rater behavior. Additionally, investigation into the effects of rater type to the construction of a priori assessment network systems through evaluation of model fit may provide insightful and fruitful implications for developing large-scale performance music performance evaluation systems.

The approach underscoring cognitive modeling specific to rater behavior is to explain why raters respond to a particular item. In working with latent constructs such as music performance assessment, the most prominent validity concern lies in explaining the inferential gap between observed scores and the definition of the construct. Validity evidence can only truly be established when sources that affect response outcomes are explained. Validity, however, is not an omniscient knowledge of the construct of interest, as sources of error will always impede the measurement process. Therefore, the focus of validity in the context of psychometric measurement should not necessarily be on the validity itself, but on the *process of establishing validity.* The process of validity is a trifold relationship between inference, explanatory considerations, and evidence. As demonstrated in this article, raters' underlying cognitive processes are fundamental in explaining sources of variability. Further investigation into explanations of cognitive rating processes underscored by invariant psychometric measurement models with the aid of relevant statistical methods is essential for providing valid assessment in music performance and related psychological music research.

The ability to embed the Rasch model and related indices (e.g. rater severity, rating scale structure, differential rater severity) and analysis of cluster membership into the music performance assessment process would greatly improve its validity, reliability, and fairness. Specifically, the ability to identify raters' unique engagement with measurement instruments can help identify, diagnose, and correct potential weaknesses in their ability as a rater. Of course, there are instances when differences in rater opinion can provide a more holistic evaluation of a musical performance. However, in instances when quantitative data is used for purposes beyond that of improving teaching and learning, Rasch measurement has promise as a new and improved approach toward the advancement of music performance assessment protocols.

Funding

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

Notes

1. The calibration of the ensemble performances is not relevant to the research questions posed in this study and was therefore omitted in the presentation of results.
2. The pairwise interaction tables are available upon request.

References

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

Baldwin, C. L. (2012). *Auditory cognition and human performance: Research and applications*. Boca Raton, FL: CRC Press.

Bergee, M. J. (2006). Validation of a model of extramusical influences on solo and small-ensemble festival ratings. *Journal of Research in Music Education*, *54*(3), 244–256.

Bergee, M. J. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education*, *55*(4), 344–358.

Berliner, P. J. (1994). *Thinking in jazz: The infinite art of improvisation*. Chicago, IL: The University of Chicago Press.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago, IL: Chicago University Press.

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. (1997). The social in musical performance. In D. J. Hargraves & A. C. North (Eds.), *The social psychology of music* (pp. 209–228). Oxford, UK: Oxford University Press.

Davidson, J. W. (2001). The role of the body in the production and perception of solo vocal performance: A case study of Annie Lennox. *Musicae Scientiae*, *5*(2), 235–256.

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., & Correia, J. S. (2002). Body movement. In R. Parncutt & G. E. McPherson (Eds.), *The science and psychology of music performance* (pp. 237–249). Oxford, UK: Oxford University Press.

DCamp, C. B. (1980). *An application of the facet-factorial approach to scale construction in the development of a rating scale for high school band performance* (Unpublished doctoral dissertation). University of Iowa, Iowa City.

Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology*, *1*(1), 30–49.

Dowling, W. J., & Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies. *Journal of the Acoustical Society of America*, *49*, 524–531.

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2001). The effect of task relevance on the cortical response to changes in visual and auditory stimuli: An event-related fMRI study. *NeuroImage*, *14*(6), 1256–1267.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292.

Elliott, C. A. (1975). Attacks and releases as factors in instrument identification. *Journal of Research in Music Education*, *23*(1), 35–40.

Elliott, C. A. (1995). Race and gender as factors in judgments of musical performance. *Bulletin of the Council for Research in Music Education*, *127*, 50–56.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 261–287). Mahwah, NJ: Erlbaum.

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

Flores, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking. *The Statistician*, *45*(1), 97–104. doi:10.2307/2348415

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). New York, NY: Longman.

Frijda, N. H. (1988). The laws of emotion. *The American Psychologist*, *43*(5), 349–358.

Gabrielsson, A., & Lindstrom, E. (2001). The role of structure in the musical expression of emotions. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, and applications* (pp. 367–400). Oxford, UK: Oxford University Press.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York, NY: Oxford University Press.
Hewitt, M. P., & Smith, B. P. (2004). The influence of teaching-career level and primary performance instrument on the assessment of music performance. *Journal of Research in Music Education, 52*(4), 314–327.
International Telecommunication Union (ITU). (2004). *Objective perceptual assessment of video quality: Full reference television*. Geneva, Switzerland: ITU-T Telecommunication Standardization Bureau.
Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America, 94*(5), 2595–2603.
Juslin, P. N. (2003). Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music, 31*(3), 273–302. doi:10.1177/03057356030313003
Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*(1), 98–104.
Lane, S., & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education and Praeger.
Linacre, J. M. (1989). *Many facet Rasch measurement*. Chicago, IL: MESA Press.
Linacre, J. M. (2002a). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.
Linacre, J. M. (2002b). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transaction, 16*(2), 878.
Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions, 17*(1), 918.
Linacre, J. M. (2014). *Facets*. Chicago, IL: MESA Press.
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.
Mardia, K. V., Kent, J. T., & Bobby, J. M. (1979). *Multivariate analysis*. London, UK: Academic Press.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.
McDermott, J. H., & Oxenham, A. J. (2008). Music perception, pitch, and the auditory system. *Current Opinion in Neurobiology, 18*(4), 452–463. doi:10.1016/j.conb.2008.09.005
McPherson, G. E., & Schubert, E. (2004). Measuring performance enhancement in music. In A. Williamon (Ed.), *Musical excellence: Strategies and techniques to enhance performance* (pp. 61–82). Oxford, UK: Oxford University Press.
McPherson, G. E., & Thompson, W. F. (1998). Assessing music performance: Issues and influences. *Research Studies in Music Education, 10*(1), 12–24.
Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago, IL: The University of Chicago Press.
Monson, I. (1996). *Saying something: Jazz improvisation and interaction*. Chicago, IL: The University of Chicago Press.
Morrison, S. J. (1998). A comparison of reference responses of white and African-American students to musical versus musical/visual stimuli. *Journal of Research in Music Education, 46*, 208–222.
Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin, 125*(6), 826–859.
Namour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model*. Chicago, IL: The University of Chicago Press.
Palmer, C. (1989). Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception and Performance, 15*(12), 331–346.
Persson, R. S., Pratt, G., & Robson, C. (1992). Motivational and influential components of musical performance: A qualitative analysis. *European Journal of High Ability, 3*(2), 206–217. doi:10.1080/0937445920030209
Radocy, R. E. (1976). Effects of authority figure biases on changing judgments of musical events. *Journal of Research in Music Education, 24*, 119–128.
Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: A five-factor model. *Journal of Personality and Social Psychology, 100*(6), 1139–1157.

Rentfrow, P. J., Goldberg, L. R., Stillwell, D. J., Kosinski, M., Gosling, S. D., & Levitin, D. J. (2012). The song remains the same: A replication and extension of the MUSIC Model. *Music Perception, 30*(2), 161–185.

Rentfrow, P. J., & McDonald, J. A. (2009). Preference, personality, and emotion. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, application* (pp. 669–695). Oxford, UK: Oxford University Press.

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America, 88*(2), 622–641.

Repp, B. H. (1995). Expressive timing in Schumann's "Träumerei": An analysis of performances by graduate student pianists. *Journal of the Acoustical Society of America, 98*(5), 2413–2427.

Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning.

Saldanha, E. L., & Corso, J. F. (1964). Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America, 36*, 2021–2026.

Schubert, E. (1996). Enjoyment of negative emotions in music: An associative network explanation. *Psychology of Music, 24*, 18–28. doi:10.1177/0305735696241003

Serafine, M. L., Glassman, N., & Overbeeke, C. (1989). The cognitive reality of hierarchic structure in music. *Music Perception, 6*(4), 397–430.

Sloboda, J. A. (1983). The communication of musical metre in piano performance. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 35*(2), 377–396.

Stanley, M., Brooker, R., & Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Research Studies in Music Education, 18*(1), 46–56.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-dependent characteristics of performance evaluation. *Music Perception, 25*(1), 13–29.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika, 18*(4), 267–276.

Vasil, T. (1973). *The effects of systematically varying selected factors on music performing adjudication* (Doctoral dissertation). University of Connecticut, Storrs.

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*(3), 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*(4), 510–521. doi:10.2307/3345367

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244.

Wesolowski, B. C. (2013). Cognition and the assessment of interaction in jazz performance. *Psychomusicology: Music, Mind, and Brain, 23*(4), 236–242.

Wesolowski, B. C. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal, 101*, 77–85.

Wesolowski, B. C. (2015). Tracking student achievement in music performance: Developing student learning objectives for growth model assessments. *Music Educators Journal, 102*, 39–47.

Wesolowski, B. C., Wind, S. A., & Engelhard, G., Jr. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae, 19*(2), 147–170.

Wesolowski, B. C., Wind, S. A., & Engelhard, G., Jr. (2016a). Examining rater precision in music performance assessment: An analysis of rating scale structure using the Multifaceted Rasch Partial Credit Model. *Music Perception, 33*(5), 662–678.

Wesolowski, B. C., Wind, S. A., & Engelhard, G., Jr. (2016b). Rater analyses in music performance assessment: Application of the Many Facet Rasch Model. In T. S. Brophy, J. Marlatt, & G. K. Ritcher (Eds.), *Connecting practice, measurement, and evaluation: Selected papers from the 5th International Symposium on Assessment in Music Education*, (pp. 335–356). Chicago, IL: GIA.

Wesolowski, B. C., Wind, S. A., & Engelhard, G., Jr. (in press). Evaluating rater DRIFT in music performance assessment: Implications of a time parameter in monitoring rater severity and rating scale structure. *Bulletin of the Council for Research in Music Education.*

Williamon, A., & Davidson, J. W. (2002). Exploring co-performer communication. *Musicae Scientiae*, *5*(1), 53–72.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*(1), 83–106.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*(1), 35–51.

Wright, B. D. (1977). Solving measurment 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., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.