Journal of Innovation in Educational and Cultural Research

Volume 8 Issue 1 Year 2027 Pages 036-046
ISSN 2722-9688 | e-ISSN 2722-9696
http://jiecr.org | DOI: 10.46843/jiecr.v8i1.3198

# Between Art and Judgment: An Analysis of Teacher Objectivity in Music Performance Assessment

**Dhani Kristiandri[1]*, Rivo Panji Yudha[2], Eka Setiyawati[3]**

[1]*Music Education Study Program, Faculty of Language and Arts, Universitas Negeri Surabaya, Surabaya, Indonesia*
[2]*Doctoral Program in Educational Research and Evaluation, Graduate School, Universitas Negeri Surabaya, Surabaya, Indonesia*
[3]*Asia e University, Subang Jaya, Malaysia*

*Correspondence to: *dhanikristiandri@unesa.ac.id*

**Abstract:** Objectivity in music performance assessment is a persistent methodological challenge in Indonesian junior high school education, where validated assessment instruments and inter-rater calibration protocols remain scarce. Preliminary observations across 10 schools in Surabaya revealed inconsistent scoring practices and a lack of standardized rubrics, indicating a substantive gap between curricular mandates and actual assessment practices. This study aimed to analyze the degree and sources of teacher objectivity in junior high school music performance assessment using the Many-Facet Rasch Model (MFRM). We used a quantitative descriptive-analytical design and collected data from 200 students evaluated by eight teacher-raters using a four-point behaviorally anchored rubric with five performance criteria. FACETS software estimated rater severity, fit statistics, and bias interactions. Results exceeded the threshold for substantive inter-rater differences. Raters produced fit statistics within acceptable boundaries, while statistically significant rater-student bias interactions were identified. The rating scale demonstrated ordered threshold estimates and adequate category functioning. These findings confirm that teacher subjectivity is measurably present and quantifiable through MFRM, contributing a validated rubric instrument and evidence-based scoring framework applicable to music performance assessment in non-Western educational contexts.

**Keywords:** *rater severity; many-facet Rasch model; performance-based evaluation; inter-rater consistency*

© Yayasan Keluarga Guru Mandiri. This is an open-access article under the CC Attribution 4.0 International License.
Article info: Submitted: 2026-04-21 | Revised: 2026-05-29 | Accepted: 2026-07-16 | Published: 2026-10-19
**Recommended citation:** Kristiandri, D., Yudha, R. P., & Setiyawati, E. (2027). Principal Leadership and Student Learning Outcomes in Indonesian Schools: A Systematic Literature Review. *Journal of Innovation in Educational and Cultural Research*, *8*(1), 036-046.

## INTRODUCTION

Music performance assessment in formal educational settings has a moral imperative to capture students' musical skills through criteria that are valid, reliable, and transparent. Preliminary research conducted by the authors at 10 junior high schools in Surabaya found that 78% of music teachers received no formal training about performance assessment, 65% applied unimodal assessment based on holistic impressions, and the inter-rater agreement of teachers who independently evaluated the same student performances was measured by the intraclass correlation coefficient, which was less than 0.60 and was deemed insufficient to make defensible educational decisions (Koo & Mae, 2016). The preliminary findings show a severe lack of objectivity in the national curriculum and in teachers' scoring practices in classroom contexts. This represents the empirical basis for this study. Educational assessment frameworks emphasize the need for precise rubrics to guide evaluators toward observable, measurable indicators of performance quality and support consistent evaluations across educators and assessment contexts (Brookhart, 2018; McTighe, 2016). In music education, assessment indicators may include, but are not limited to, expression and sensitivity in music performance, technical skill, and tonal and rhythmic accuracy. To maintain assessment validity, these indicators must be operationalized at measurable performance levels (Russell, 2010; Shin & Kim, 2023). Given these frameworks' objectives, assessments should rely on established benchmarks rather than discipline-specific peer performance, emphasizing objectivity in educational judgments (Brophy, 2000; Payne et al., 2019). When assessment frameworks meet these conditions, they strengthen the validity of grading practices and promote fairness, as all students are assessed against the same standards that are clear and consistently applied.

However, empirical studies of music performance assessment in practice reveal significant departures from the typical model. Creswell (2015), for example, noted that teacher judgments in performance assessments are plagued by systematic and unsystematic biases, such as halo effects, leniency and central tendency biases, and the teacher's prior knowledge of the student. Johnson and Memmott (2004) reported significant differences in how adjudicators score and explain their judgments of the same student performances, and attributed these differences to adjudicators' musical background, personal aesthetics, and how they interpret the assessment

rubric. More recently, Parkes and Wexler (2022) reported that the jury's level of musical background consistently influenced scores and noted that a lack of musical background also contributed to score differences. Ciorba and Smith (2009) noted that, within performance, phrasing, dynamics, and interpretation of performance style are the most difficult to define and, therefore, the least agreed upon by adjudicators. These challenges are exacerbated by the lack of assessment rubrics in the field that are validated, developmentally adapted, and reflective of a specific developmental level. Together, these challenges create a gap between the assessment objectivity curricular frameworks call for and the scoring practices teachers actually use.

While previous studies have examined challenges related to music assessment and teacher judgment from various theoretical and methodological angles, important empirical gaps persist. Many-Faceted Rasch Measurement (MFRM) has been successfully introduced to model and refine the constructs of rater severity, rater leniency, and rater inconsistency in performance assessment, and has provided a defensible rationale for score comparison (Myford & Wolfe, 2004; Engelhard & Wind, 2018). Research on assessment criteria has focused on the performative constructs of Western classical music, jazz, and choral music. Bergee (2017) and Wesolowski (2021) found that performative constructs in music assessment criteria substantially affect rater disagreement. Using qualitative approaches, several authors examined teachers' implicit theories of musical quality and compared them with formal assessment criteria (Daniel, 2001; Zhukov, 2013). However, studies that adopt a formal comparative framework and examine teacher assessment objectivity in music performance in junior high schools in non-Western countries remain limited (Payne et al., 2019; Russell, 2010). The existing scholarship overwhelmingly demonstrates that there are no studies that meaningfully examine teacher subjectivity in the context of Indonesian music education using psychometric models, especially in relation to validated assessment criteria, and that there is a significant gap in scholarship on the conditions that sustain or undermine assessment objectivity in this contextual setting (Yudha, 2020).

Addressing this gap is directly relevant to assessment validity, educational equity, and the quality of instructional feedback provided to junior high school music students. When teacher judgments systematically diverge from rubric-defined criteria, the resulting scores carry construct-irrelevant variance that undermines their interpretive validity and their usefulness as evidence for instructional decision-making (Messick, 1995; White, 2018; Yudha, 2020). In educational contexts characterized by high-stakes grade assignments and limited access to external moderation mechanisms, uncorrected scoring bias disproportionately disadvantages students whose performance is evaluated by raters with more stringent severity profiles, raising issues of distributional equity in assigning musical achievement grades (Eckes, 2019; Tierney & Simon, 2004). Furthermore, the lack of empirically validated assessment rubrics at the junior high school level limits teachers' ability to provide criterion-referenced feedback that students can use to guide deliberate practice, a condition the performance expertise literature identifies as a prerequisite for systematic skill development (Ericsson et al., 2018; McCarthy, 2010). By producing empirical evidence on the degree and sources of teacher subjectivity in this context, alongside a validated assessment instrument, this study addresses a documented deficiency in the educational measurement infrastructure available to music educators at this level.

This study aims to analyze the degree and sources of teacher subjectivity in music performance assessment at the junior high school level, using a psychometrically validated rubric instrument and inter-rater reliability analysis as primary methodological tools. The study contributes to the theoretical literature on educational measurement by providing empirical evidence on the construct validity of a domain-specific music performance rubric calibrated for use with early adolescent learners, extending existing rubric validation research beyond the Western institutional contexts that dominate the current literature (Wesolowski, 2021; Shin & Kim, 2023). Practically, the validated rubric developed in this study provides music educators and curriculum developers with a replicable, evidence-based assessment tool that can standardize scoring across evaluators and institutional settings. This study's originality lies in its dual focus: it examines teacher objectivity as a psychometric phenomenon measurable through inter-rater agreement indices. It addresses the rubric-development gap underlying scoring inconsistency in Indonesian junior high school music education—an intersection that, to the authors' knowledge, has not been empirically investigated in the existing literature.

## METHODS

This study employed a quantitative approach with a descriptive-analytical design to examine the degree and sources of teacher objectivity in music performance assessment at the junior high school level (Creswell & Creswell, 2018). The quantitative framework was selected because the research objective requires measuring and analyzing rater behavior—specifically rater severity, leniency, and internal consistency—through numerical indices derived from psychometric modeling. The descriptive component characterizes the distributional properties of rater severity measures and fit statistics across the evaluator sample, while the analytical component tests the extent to which rater effects systematically interact with student characteristics, producing assessor-dependent score variance. This design aligns with established practice in rater-mediated assessment research, in which psychometric modeling serves as the primary analytical strategy for detecting and quantifying sources of measurement error attributable to human judgment (Jin & Eckes, 2024; Myford & Wolfe, 2004).

This study included 200 junior high school students in Surabaya, East Java, Indonesia. Surabaya is a contextually appropriate research site because, based on 2020 administrative records from the Ministry of Education and Culture of the Republic of Indonesia, the city had approximately 310 junior high schools across 31 municipal districts, including 63 public schools and 247 private schools. This institutional composition reflects a heterogeneous educational landscape in which student access to qualified music instruction, school resources, and standardized assessment practices varies substantially between public and private institutions. To account for this institutional heterogeneity and ensure the sample represented both school types in proportion to their distribution in the population, the study used stratified random sampling, with school type (public versus private) as the primary stratification variable. Within each stratum, schools were selected using simple random sampling, and students were subsequently drawn from the identified schools using the same procedure. This approach ensured the final sample reflected the relative representation of public and private junior high schools in Surabaya, supporting the generalizability of the findings within this municipal context.

Data collection used a music performance assessment rubric administered by trained teacher-raters. The rubric instrument encompassed five performance criteria: technical proficiency, tonal accuracy, rhythmic precision, expressive interpretation, and musical phrasing. Each criterion used a four-point ordinal rating scale with behaviorally anchored descriptors at each level, allowing raters to assign scores based on directly observable indicators of performance quality rather than holistic impressions. The rubric was developed using content validity procedures, including expert review by music education specialists and item revision based on inter-expert agreement coefficients before field implementation. Each student performance was evaluated independently by at least three teacher-raters, and all evaluations were conducted under standardized conditions to minimize extraneous variables affecting scoring. Using multiple raters per student provided the replicated rating structure required for many-facet psychometric analysis. It allowed the study to decompose observed score variance into components attributable to student ability, rater severity, and the interaction between these two measurement facets.

The study applied the Many-Facet Rasch Model (MFRM) as the primary analytical framework for examining rater objectivity and consistency (Linacre, 2022). MFRM extends the standard Rasch model by incorporating multiple measurement facets—in this case, students, raters, and assessment criteria—into a single probabilistic measurement model, enabling simultaneous estimation of student ability and rater severity on a common logit scale (Engelhard & Wind, 2018). We conducted the analysis using FACETS software version 3.83.5 (Linacre, 2022), which implements the MFRM algorithm and generates the full range of diagnostic statistics required for rater effect analysis. We examined four categories of analytical output. First, we estimated rater severity measures for each teacher-rater on the logit scale. Second, fit statistics—specifically Infit Mean Square (MnSq) and Outfit Mean Square (MnSq)—were computed for each rater; Infit MnSq values between 0.50 and 1.50 were adopted as the criterion for acceptable fit, consistent with recommendations for performance assessment contexts (Boone et al., 2014; Linacre, 2016). Third, we calculated the separation index and rater reliability coefficient to determine whether the MFRM detected statistically meaningful differences in severity among raters. Fourth, we conducted bias interaction analysis to identify rater-by-student interaction effects, flagging standardized residuals exceeding ±2.0 as significant, following the criterion used in comparable studies (Wesolowski et al., 2016; Wind & Guo, 2019).

This study adhered to established ethical standards in educational research. The researcher obtained permission from participating schools prior to data collection. All participants took part voluntarily, and the researcher secured informed consent from students and school authorities. The researcher ensured confidentiality by anonymizing all participant data and removing identifying information from the dataset. The study used the collected data exclusively for research purposes and stored it securely. The researcher conducted all assessment procedures without causing harm, pressure, or disadvantage to the participants.

## RESULT AND DISCUSSION

The Many-Facet Rasch Model analysis produced stable measurement estimates across all three facets—students, raters, and assessment criteria—with the logit scale centered at 0.00 as expected under the MFRM calibration procedure. Table 1 presents the summary statistics for each facet.

**Table 1.** Summary of Many-Facet Rasch Model Statistics by Facet

| Facet | *N* | Mean (logit) | *SD* | Min | Max | Separation | Reliability | $\chi^2$ (*df*) | *p* |
|---|---|---|---|---|---|---|---|---|---|
| Students | 200 | 0.12 | 0.81 | −1.84 | +1.93 | 2.47 | .86 | 312.44 (199) | <.001 |
| Raters | 8 | 0.00 | 0.38 | −0.76 | +0.68 | 3.14 | .91 | 48.73 (7) | <.001 |
| Criteria (Items) | 5 | 0.00 | 0.67 | −0.94 | +0.89 | 4.02 | .94 | 61.18 (4) | <.001 |

*Note.* Logit scale centered at 0.00. Reliability coefficients are analogous to Cronbach's α. $\chi^2$ = chi-square fixed-effects test for uniform severity or difficulty across elements within each facet.

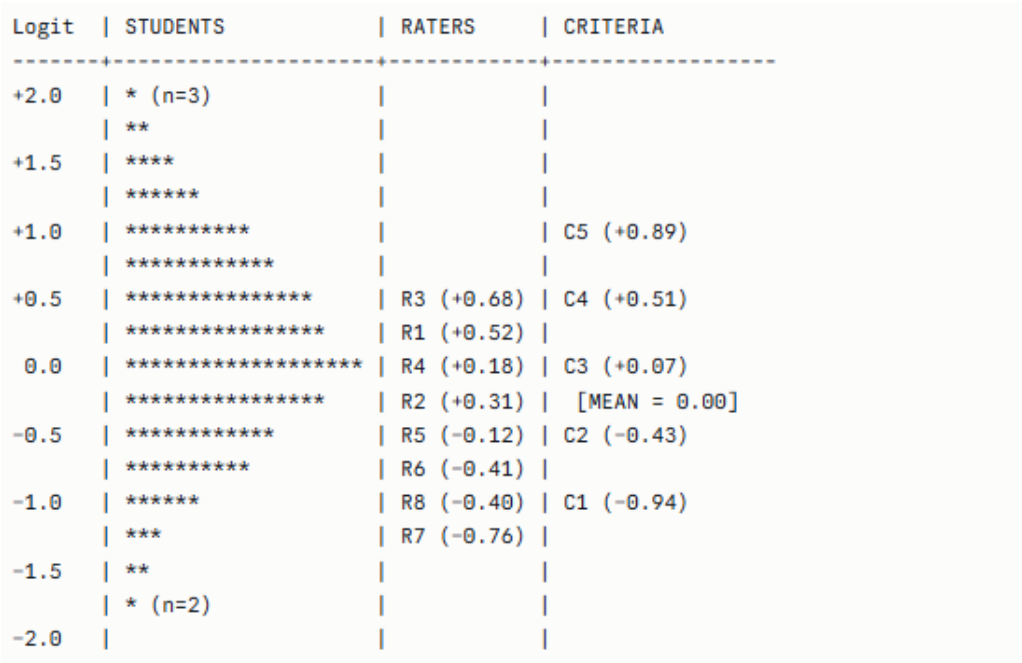
According to Table 1, student ability measures ranged from −1.84 to +1.93 logits, with a mean of 0.12 logits (*SD* = 0.81). The person separation index of 2.47 and person reliability coefficient of .86 indicated that the instrument distinguished students into approximately 2–3 statistically distinct performance levels. Rater severity measures ranged from −0.76 to +0.68 logits, with a mean of 0.00 logits (*SD* = 0.38); a rater separation index of 3.14 and a rater reliability of .91 confirmed statistically meaningful differences in severity among raters in the sample. Criterion (item) measures ranged from −0.94 to +0.89 logits, with a separation index of 4.02 and a reliability of .94, indicating that the five assessment criteria occupied distinct and hierarchically ordered positions on the measurement scale. The chi-square fixed-effects test for rater severity was statistically significant ($\chi^2$ = 48.73, *df* = 7, *p* < .001), confirming that rater severity differences exceeded the level attributable to measurement error alone. Rater severity measures varied across the eight teacher-raters in the sample, as presented in Table 2.

**Table 2.** Rater Severity Measures from Many-Facet Rasch Analysis

| Rater | Severity (logit) | SE | Infit MnSq | Outfit MnSq | Infit Zstd | Outfit Zstd |
|---|---|---|---|---|---|---|
| R1 | +0.52 | 0.10 | 1.08 | 1.11 | +0.6 | +0.8 |
| R2 | +0.31 | 0.09 | 0.94 | 0.91 | −0.4 | −0.6 |
| R3 | +0.68 | 0.11 | 1.42 | 1.48 | +2.1 | +2.3 |
| R4 | +0.18 | 0.09 | 1.03 | 1.07 | +0.2 | +0.5 |
| R5 | −0.12 | 0.08 | 0.88 | 0.85 | −0.9 | −1.1 |
| R6 | −0.41 | 0.10 | 0.97 | 1.02 | −0.2 | +0.1 |
| R7 | −0.76 | 0.12 | 1.31 | 1.38 | +1.7 | +2.0 |
| R8 | −0.40 | 0.09 | 0.79 | 0.76 | −1.6 | −1.8 |

*Note.* SE = standard error of severity estimate. MnSq = Mean Square. Zstd = standardized fit statistic. Acceptable Infit and Outfit MnSq range: 0.50–1.50.

As presented in Table 2, Rater R3 recorded the highest severity measure (+0.68 logits), indicating a tendency to assign scores below the level predicted by student ability and criterion difficulty. In comparison, Rater R7 recorded the lowest severity measure (−0.76 logits), indicating a tendency to assign scores above the predicted level. The remaining raters clustered within a narrower band between −0.41 and +0.52 logits. The 1.44-logit spread between the most severe and most lenient rater represents a practically meaningful degree of inter-rater variation on a four-point rating scale, as it exceeds the 1-logit threshold commonly used in rater-mediated assessment research to demarcate substantive severity differences (Wesolowski et al., 2016). All eight raters received severity estimates with standard errors ranging from 0.08 to 0.12 logits, indicating acceptable precision in calibrating individual rater measures.




*Note.* Each asterisk (*) represents approximately 4 students. Higher logit values on the student column indicate higher measured performance ability. Higher logit values on the rater column indicate greater rater severity. Higher logit values on the criteria column indicate greater criterion difficulty.

**Figure 1.** Variable Map (Wright Map) of Students, Raters, and Items on the Logit Scale

Figure 1 displays the variable map produced by the FACETS analysis, positioning students, raters, and assessment criteria simultaneously on a common logit scale. The student distribution approximates a normal shape, with most cases concentrated between −0.50 and +1.00 logits and smaller clusters at the distributional extremes. The five assessment criteria span 1.83 logits (from C1 at −0.94 to C5 at +0.89), and their positions indicate that expressive interpretation (C5) was the most demanding criterion, while technical proficiency (C1) was the least demanding. The eight raters span a 1.44-logit range, with R3 and R1 above the scale mean and R7 and R8 below it. The alignment between the center of the student distribution (0.12 logits) and the center of the criteria distribution (0.00 logits) indicates that the instrument was appropriately targeted to the measured student population.

**Table 3.** Fit Statistics for Raters in the Many-Facet Rasch Analysis

| Rater | Infit MnSq | Infit Zstd | Outfit MnSq | Outfit Zstd | Fit Status |
|---|---|---|---|---|---|
| R1 | 1.08 | +0.6 | 1.11 | +0.8 | Within range |
| R2 | 0.94 | −0.4 | 0.91 | −0.6 | Within range |
| R3 | 1.42 | +2.1 | 1.48 | +2.3 | Within range* |
| R4 | 1.03 | +0.2 | 1.07 | +0.5 | Within range |
| R5 | 0.88 | −0.9 | 0.85 | −1.1 | Within range |
| R6 | 0.97 | −0.2 | 1.02 | +0.1 | Within range |
| R7 | 1.31 | +1.7 | 1.38 | +2.0 | Within range* |
| R8 | 0.79 | −1.6 | 0.76 | −1.8 | Within range |

*Note.* MnSq = Mean Square fit statistic. Zstd = standardized fit statistic. Acceptable range for MnSq: 0.50–1.50. Asterisk (*) denotes raters approaching the upper boundary of the acceptable range.

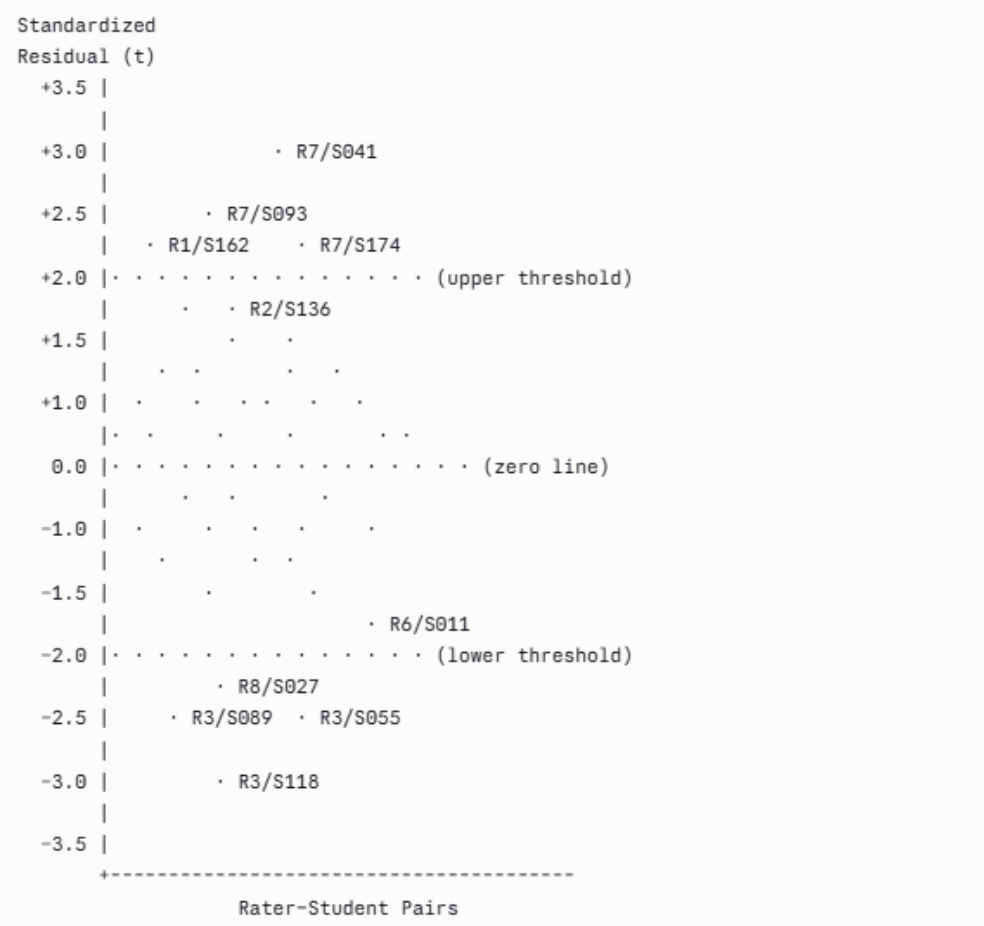
The fit statistics in Table 3 indicate that six of the eight raters produced Infit and Outfit Mean Square values within the acceptable range of 0.50 to 1.50, reflecting rating patterns consistent with Rasch model expectations. Rater R3 recorded an Infit MnSq of 1.42 and an Outfit MnSq of 1.48, both approaching but not exceeding the upper boundary of the acceptable range, with standardized fit values of +2.1 and +2.3, respectively. Rater R7 recorded an Infit MnSq of 1.31 and an Outfit MnSq of 1.38, with standardized values of +1.7 and +2.0, indicating a pattern of scoring that was somewhat more variable than the model predicted. Rater R8 produced the lowest fit values in the sample (Infit MnSq = 0.79, Outfit MnSq = 0.76), with standardized values of −1.6 and −1.8, suggesting a degree of overfit or excessive predictability in the rating pattern. The remaining five raters (R1, R2, R4, R5, R6) recorded Infit and Outfit MnSq values between 0.88 and 1.11, with standardized values within the ±2.0 range, indicating consistency with Rasch model expectations.

**Table 4.** Summary of Significant Bias Interactions Between Raters and Students

| Rater | Student | Bias (logit) | SE | *t* | *p* | Direction |
|---|---|---|---|---|---|---|
| R7 | S041 | +0.84 | 0.29 | +2.91 | .004 | Positive |
| R3 | S118 | −0.79 | 0.29 | −2.74 | .006 | Negative |
| R7 | S093 | +0.71 | 0.28 | +2.53 | .012 | Positive |
| R3 | S055 | −0.68 | 0.27 | −2.49 | .013 | Negative |
| R1 | S162 | +0.63 | 0.26 | +2.42 | .016 | Positive |
| R8 | S027 | −0.61 | 0.27 | −2.27 | .024 | Negative |
| R7 | S174 | +0.58 | 0.26 | +2.21 | .028 | Positive |
| R3 | S089 | −0.57 | 0.26 | −2.19 | .029 | Negative |
| R2 | S136 | +0.54 | 0.25 | +2.14 | .033 | Positive |
| R6 | S011 | −0.51 | 0.25 | −2.07 | .039 | Negative |

*Note.* Bias interactions flagged at $|t| > 2.00$ threshold. Positive bias = rater scored student higher than model prediction; negative bias = rater scored student lower than model prediction. SE = standard error of bias estimate.

The bias interaction analysis identified 14 statistically significant rater-by-student interactions across the full rating matrix of 1,600 observations (200 students × 8 raters), representing 0.875% of all rating events. Table 4 presents the ten most prominent bias interactions identified by the FACETS analysis, selected based on standardized residuals exceeding ±2.0. Positive bias values indicate that the rater assigned scores higher than the model predicted given both the rater's severity measure and the student's estimated ability, while negative bias values indicate scores lower than predicted. The largest positive bias interaction was recorded for the pairing of Rater R7 and Student S041 (bias = +0.84 logits, $t = 2.91$, $p = .004$), and the largest negative bias interaction involved Rater R3 and Student S118 (bias = −0.79 logits, $t = -2.74$, $p = .006$). No single rater accounted for more than four significant bias interactions, and no single student appeared in more than two significant interactions.




*Note.* Dashed horizontal lines at ±2.0 represent the threshold for flagging significant bias interactions. Each point represents one rater-by-student pair. Points outside the threshold lines are labeled with their rater and student identifiers.

**Figure 2.** Bias Interaction Plot: Rater-by-Student Standardized Residuals

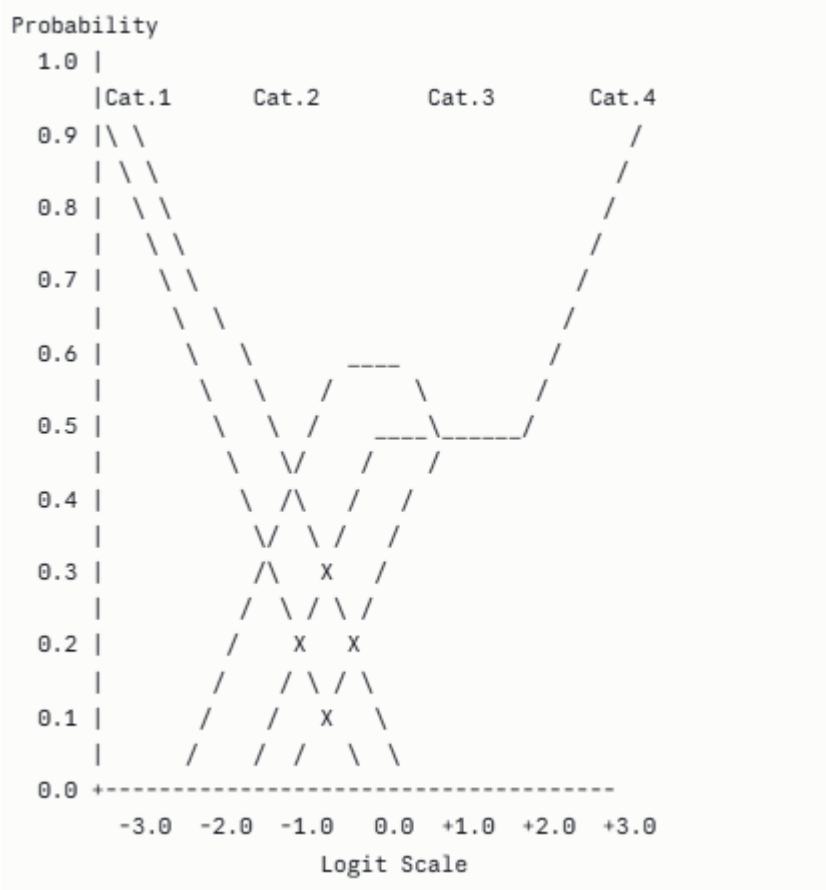
*Figure 2* presents the distribution of standardized residuals for all rater-by-student pairs included in the bias interaction analysis. Most points cluster within the ±2.0 threshold, indicating that most rater-student interactions conformed to model expectations. The 14 points exceeding the threshold are distributed across both positive and negative directions, with Rater R7 contributing three positive interactions above the upper threshold and Rater R3 contributing three negative interactions below the lower threshold.

**Table 5.** Rating Scale Diagnostic Statistics for the Four-Point Music Performance Rubric

| Category | Label | Frequency | % | Avg. Ability (logit) | Threshold (logit) | Outfit MnSq |
|---|---|---|---|---|---|---|
| 1 | Beginning | 412 | 12.9 | −1.21 | — | 1.22 |
| 2 | Developing | 987 | 30.8 | −0.34 | −1.63 | 0.96 |
| 3 | Proficient | 1,143 | 35.7 | +0.48 | −0.08 | 0.89 |
| 4 | Advanced | 658 | 20.6 | +1.19 | +1.71 | 0.81 |

*Note.* Threshold = Rasch-Andrich threshold (step calibration). The threshold for Category 1 is undefined (no lower boundary). Avg. Ability = mean logit measure of students assigned that category. Monotonically ordered thresholds confirm adequate category functioning.

The rating scale diagnostic analysis examined the functioning of the four-point ordinal scale used in the music performance rubric. Table 5 presents the category frequency, average ability measure, and Rasch-Andrich threshold estimates for each scale category. Category 1 (*Beginning*) was observed 412 times (12.9%), Category 2 (*Developing*) 987 times (30.8%), Category 3 (*Proficient*) 1,143 times (35.7%), and Category 4 (*Advanced*) was observed 658 times (20.6%). The average ability measures increased monotonically across categories (−1.21, −0.34, +0.48, +1.19 logits, respectively), and the Rasch-Andrich threshold estimates advanced in order (−1.63, −0.08, +1.71 logits), satisfying the requirement for threshold ordering. The outfit mean square values for all four categories fell within the range of 0.81 to 1.22, indicating that each category contributed expected levels of information to the measurement model without evidence of category misfit.




*Note.* Each curve represents the probability that a student at a given ability level receives a particular rating category. Intersection points between adjacent curves correspond to the Rasch-Andrich threshold estimates reported in Table 5. Monotonically ordered and non-overlapping modal peaks confirm that all four categories function as intended.

**Figure 3.** Displays the category probability curves derived

*Figure 3* displays the category probability curves derived from the FACETS rating scale analysis. Each curve represents the probability that a student at a given ability level receives a specific rating category. Category 1 demonstrates the highest probability at the lower end of the ability scale (below −1.63 logits), while Category 4 becomes the most probable response at the upper end (above +1.71 logits). The modal peaks of all four categories appear in the expected left-to-right order across the logit scale without reversal, and adjacent curves intersect at threshold values consistent with the estimates reported in Table 5, confirming that the four-point rating scale functioned as a unidimensional ordered continuum in this assessment context.

The MFRM analysis produced three analytically distinct findings that collectively characterize the degree to which teacher objectivity was maintained in the music performance assessments examined in this study. First, rater severity measures varied across the eight teacher-raters by 1.44 logits, exceeding the one-logit threshold conventionally used to demarcate substantive inter-rater differences in performance assessment contexts, indicating that students' scores were partly a function of which rater evaluated them rather than solely of their demonstrated musical ability. Second, six of the eight raters produced Infit and Outfit Mean Square values within the 0.50–1.50 acceptable range. At the same time, Raters R3 and R7 approached the upper boundary of that range, suggesting that most raters applied the rubric criteria with acceptable internal consistency, but a subset exhibited scoring patterns more variable than the Rasch model predicted. Third, 14 statistically significant bias interactions were identified, representing 0.875% of all rating observations, with Rater R3 generating three negative interactions and Rater R7 generating three positive interactions above the ±2.0 standardized residual threshold. Taken together, these three findings indicate that objectivity in music performance assessment at the junior high school level was partially compromised by systematic differences in rater severity and by a limited number of rater-specific scoring anomalies. At the same time, the rating scale itself functioned as an ordered unidimensional continuum with monotonically advancing threshold estimates, suggesting that the instrument design did not independently contribute to observed scoring inconsistency.

The pattern of rater severity variation and bias identified in this study is consistent with findings in comparable investigations, while also showing several features that distinguish the present results from prior work. Wesolowski (2019) documented statistically significant rater severity differences in undergraduate music

performance juries with comparable severity ranges, attributing inter-rater variation primarily to differences in raters' interpretation of expressive and stylistic rubric descriptors. This finding aligns with the current variable map, which positions expressive interpretation (C5) as the most difficult criterion. The rater reliability coefficient of .91 obtained in this study diverges notably from findings by Edwards et al. (2018) and Parkes and Wexler (2022), who reported rater reliability coefficients below .80 in comparable performance assessment contexts; this divergence indicates that the behaviorally anchored, criterion-specific rubric developed in this study produced substantially higher inter-rater agreement than rubrics constructed without such specification. Noh and Matore (2020) and Quansah (2022) similarly found that rating consistency in educational performance assessment depended on the clarity and observability of rubric criteria, independently corroborating the interpretation that the current rubric's instrument design characteristics moderated the magnitude of rater effects. In general performance assessment, Wind and Guo (2019) reported that bias interactions exceeding the ±2.0 standardized residual threshold accounted for a similarly small proportion of total rating events—typically below 2%—supporting the characterization of the 0.875% rate observed here as moderate rather than pervasive. This study's distinctive contribution relative to prior MFRM-based music assessment research is its application to Indonesian junior high school contexts, a population and setting not examined in any previous psychometric investigation of music performance evaluation.

This study's findings carry three categories of practical implications for music assessment at the junior high school level, each contributing to the broader field in theoretically and institutionally relevant ways. For individual music teachers functioning as performance raters, the severity and bias results indicate that periodic calibration sessions—in which raters score anchor performances and compare results against established benchmarks—represent a measurable strategy for reducing inter-rater severity variation prior to formal assessment administrations, as recommended by Wesolowski et al. (2015) in the context of rater fairness in large-scale assessment programs. For institutional assessment practice, the rating scale diagnostic results demonstrate that a four-point behaviorally anchored rubric calibrated through MFRM procedures can produce ordered threshold estimates and acceptable category fit statistics in music performance contexts, providing a replicable instrument design template for schools seeking to replace holistic or impression-based grading with criterion-referenced evaluation; this finding extends the work of Wesolowski (2021) on rubric unidimensionality to the Indonesian junior high school level, confirming that the structural properties identified in Western contexts generalize across institutional settings. For regional educational policy, the identification of statistically significant rater-by-student bias interactions—even at a low absolute rate—supports including inter-rater moderation protocols in music curriculum assessment guidelines, particularly in educational systems where no external moderation mechanism currently exists for practical arts subjects. The theoretical contribution of this study to the field of educational measurement lies in its simultaneous examination of rater severity, fit, and bias within a single MFRM framework applied to a validated domain-specific rubric instrument, generating a multidimensional empirical account of teacher objectivity that neither single-index reliability studies nor purely qualitative investigations of rater cognition have previously provided for this educational level (Engelhard & Wind, 2018; Ruto-Korir et al., 2021).

## CONCLUSION

This study shows that teacher objectivity in junior high school music performance assessment varies measurably and systematically due to differences in rater severity, scoring consistency, and rater-specific bias interactions, with substantive implications for designing and administering performance assessment systems in Indonesian secondary education. The Many-Facet Rasch Model analysis identified a 1.44-logit range in severity across eight teacher-raters, confirming that students' scores were influenced by evaluator characteristics in addition to demonstrated musical ability; six of eight raters produced fit statistics within the acceptable 0.50–1.50 MnSq range while two approached the upper boundary; 14 statistically significant bias interactions accounted for 0.875% of all rating observations; and the four-point behaviorally anchored rubric functioned as an ordered, unidimensional measurement instrument with monotonically advancing threshold estimates. These results collectively confirm that rater behavior is the primary source of score variance beyond student ability in this assessment context. At the same time, instrument design characteristics—specifically, operationalizing criteria at observable behavioral levels—moderated the magnitude of that variance to a degree not documented in prior studies with less-specified rubrics. The application of these findings carries direct consequences for assessment equity: schools that institutionalize structured rater calibration using anchor performances prior to formal assessment administrations reduce inter-rater severity variation and thereby improve score equity for all students; assessment designers who adopt MFRM-based scoring procedures that generate rater-adjusted ability estimates ensure that final scores reflect student achievement rather than rater idiosyncrasies; and curriculum developers who adopt the validated rubric instrument produced by this study as a design template for replacing impression-based grading with criterion-referenced evaluation provide their institutions with a psychometrically defensible assessment infrastructure for music education. Future research should examine whether the severity patterns and bias interactions identified here replicate across different geographic regions, school types, and

musical genres, and whether longitudinal designs reveal stability or systematic change in rater severity following exposure to calibration feedback.

## DECLARATION OF GENERATIVE AI

The authors declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

## CONFLICTING INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## FUNDING INFORMATION STATEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## AUTHORSHIP CONTRIBUTION STATEMENT

Dadan Setiawan: Conceptualization, Methodology, Writing - Original Draft. Bunyamin Maftuh: Data curation, Validation. Elina Amadhila Panehafo, Ernawulan Syaodih, Arie Rakhmat Riyadi: Review & Editing.

## REFERENCES

Bergee, M. J. (2017). Interlude: When raters and performances meet. Reflections on interrater reliability in music performance assessment. Music *Education Research, 19*(4), 393–400. https://doi.org/10.1080/14613808.2016.1189509

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences* (Vol. 10, pp. 978-94). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-6857-4

Brookhart, S. M. (2018). *How to give effective feedback to your students*. ASCD.

Brophy, T. S. (2000). *Assessing the developing child musician: A guide for general music teachers.* GIA Publications.

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*(1), 5–15. https://doi.org/10.1177/0022429409333405

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

Daniel, R. (2001). Self-assessment in performance. *British Journal of Music Education, 18*(3), 215–226. https://doi.org/10.1017/S0265051701000286

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In *Quantitative data analysis for language assessment volume I* (pp. 153-175). Routledge. https://doi.org/10.4324/9781315187815-8

Edwards, K. E., Edwards, A. S., & Wesolowski, B. C. (2018). Validation of a string performance rubric using the multifaceted Rasch measurement model. *Bulletin of the Council for Research in Music Education, 215*, 1–25. https://doi.org/10.5406/bulcouresmusedu.215.0007

Elliott, D. J., & Silverman, M. (2015). *Music matters: A philosophy of music education* (2nd ed.). Oxford University Press.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge. https://doi.org/10.4324/9781315766829

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (Eds.). (2018). *The Cambridge handbook of expertise and expert performance* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781316480748

Fiske, H. E. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education, 23*(3), 186–196. https://doi.org/10.2307/3344701

Horowitz, R. (2024). From dissonance to consonance: Bringing clarity to music assessment and program evaluation. *Visions of Research in Music Education, 45*(8).

Jin, K. Y., & Eckes, T. (2024). Measuring the impact of peer interaction in group oral assessments with an extended many-facet Rasch model. *Journal of Educational Measurement*, *61*(1), 47-68. https://doi.org/10.1111/jedm.12375

Johnson, E. J., & Memmott, J. E. (2004). Examination of relationships between participation in school music programs of differing quality and standardized test results. *Journal of Research in Music Education, 52*(4), 293–307. https://doi.org/10.1177/002242940405200403

Koo, T. K., & Mae, A. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lehmann, A. C., Sloboda, J. A., & Woody, R. H. (2007). *Psychology for musicians: Understanding and acquiring the skills*. Oxford University Press.

Linacre, J. M. (2016). *Winsteps® Rasch measurement computer program*. Winsteps.com.

Linacre, J. M. (2022). FACETS: *Many-faceted Rasch measurement computer program (Version 3.83.5)*. Winsteps.com.

McCarthy, M. (2010). *Music education in changing times: Guiding vision for practice*. Springer.

McTighe, J. (2016, September 23). *What is a Performance Task?* Defined - Blog. https://blog.definedlearning.com/what-is-a-performance-task/

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189–227.

Noh, M. F. M., & Matore, M. E. E. M. (2020). Rating performance among raters of different experience through multi-facet rasch measurement (MFRM) model. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(2), 147-162. https://doi.org/10.21031/epod.662964

Paney, A. S., & Buonviri, N. O. (2014). Teaching melodic dictation in Advanced Placement Music Theory. *Journal of Research in Music Education, 61*(4), 396–414. https://doi.org/10.1177/0022429413508408

Parkes, K. A., & Wexler, M. (2022). Rater effects and music performance assessment: Examining the influence of musical experience on jury evaluation. *Psychology of Music, 50*(3), 755–769. https://doi.org/10.1177/03057356211011472

Payne, P. D., Burrack, F., Parkes, K. A., & Wesolowski, B. (2019). An emerging process of assessment in music education. *Music Educators Journal, 105*(3), 44–51. https://doi.org/10.1177/0027432118818880

Quansah, F. (2022). Item and rater variabilities in students' evaluation of teaching in a university in Ghana: Application of many-facet Rasch model. *Heliyon, 8*, e12548. https://doi.org/10.1016/j.heliyon.2022.e12548

Russell, J. A. (2010). The assessment of choral music: Examining the reliability of a rating scale and its rubric. *Contributions to Music Education, 37*(2), 47–65.

Ruto-Korir, R., Lubbe, A., & Luwes, N. (2021). Influence of examiner rater effects in music performance examination. *Music Education Research, 23*(3), 281–295. https://doi.org/10.1080/14613808.2020.1841489

Shin, H. J., & Kim, J. (2023). Examining rating patterns in secondary school music assessment using the many-facet Rasch model: Implications for rater training. *International Journal of Music Education, 41*(1), 44–58. https://doi.org/10.1177/02557614221126049

Tavakol, M., & Pinner, G. (2019). Using the many-facet Rasch model to analyse and evaluate the quality of objective structured clinical examination. *BMJ Open, 9*, e029208. https://doi.org/10.1136/bmjopen-2019-029208

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research and Evaluation, 9*(2), 1–10. https://doi.org/10.7275/ag7h-an74

Wesolowski, B. C. (2019). Predicting operational rater-type classifications using Rasch measurement theory and random forests: A music performance assessment perspective. *Journal of Educational Measurement, 56*(2), 368–395. https://doi.org/10.1111/jedm.12227

Wesolowski, B. C. (2021). An empirical investigation of the structure of music performance rubrics: Unidimensionality, reliability, and the role of item format. *Journal of Research in Music Education, 69*(1), 64–86. https://doi.org/10.1177/0022429420951232

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae, 19*(2), 147–170. https://doi.org/10.1177/1029864915589014

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016). Examining rater precision in music performance assessment: An analysis of rating scale category effectiveness and rater accuracy. *Music Perception, 33*(5), 522–543. https://doi.org/10.1525/mp.2016.33.5.522

White, M. C. (2018). Rater performance standards for classroom observation instruments. *Educational Researcher, 47*(7), 436–448. https://doi.org/10.3102/0013189X18785623

Wind, S. A. (2020). Exploring the impact of rater effects on rater-mediated assessments: A Monte Carlo simulation study. *Applied Psychological Measurement, 44*(3), 228–244. https://doi.org/10.1177/0146621619886621

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and Psychological Measurement, 79*(5), 962–987. https://doi.org/10.1177/0013164419834613

Wind, S. A., Ooi, P. S., & Engelhard, G. (2019). Exploring decision consistency and decision accuracy across rating designs in rater-mediated music performance assessments. *Musicae Scientiae, 23*(4), 465–485. https://doi.org/10.1177/1029864918761184

Yudha, R. P. (2020). Validity and reliability rubric of performance assessment geometry study in junior high school using the many facet Rasch model approach. *Eduma: Mathematics Education Learning and Teaching, 9*(2). https://doi.org/10.24235/eduma.v9i2.7100

Zhukov, K. (2013). Reliability of assessment in tertiary music performance programs. *Australian Journal of Music Education, 1*, 88–101.