TPM Vol. 32, No. 3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

TEACHING WISDOM IN MUSIC EDUCATION: SCALE DEVELOPMENT, IRT CALIBRATION, AND INVARIANCE IN PRIMARY RECORDER CLASSES

XIAODONG ZHOU[1], CHALERMSAK PIKULSRI[2*], SURAPOL NESUSIN[3]

[1,2,3] FACULTY OF FINE AND APPLIED ARTS, KHON KAEN UNIVERSITY, KHON KAEN 40002, THAILAND

Abstract: This study develops and validates the Teaching Wisdom in Music Education Scale (TWiMES) as a practice-proximal, fairness-tested measure for primary-school recorder classes. Across three waves and 57 classes (N = 1,460; Grades 3–5), Study 1 generated items from observable classroom routines and established strong content and response-process evidence (I-CVI = .83–1.00; S-CVI/Ave = .92). Study 2 piloted the instrument (n = 512), supporting a three-factor structure—Adaptive Pedagogy, Equitable Orchestration, and Developmental Evaluation—and yielding a 19-item form with good reliability (ω = .86–.91; ω_h = .90). In a new sample (n = 948), Study 3 confirmed the model (WLSMV; CFI = .964, TLI = .958, RMSEA = .041), calibrated items with a graded response item response theory (IRT) model (median a = 1.65; ordered thresholds), and showed high precision across θ ≈ −0.3 to +1.5 (max test information ≈ 16). A 12-item short form preserved ~90% of information in the core range (EAP reliability ≈ .88; r with full form = .98). Multi-group analyses supported configural→metric→scalar invariance across gender, grade, and school locale/SES (largest ΔCFI = .006), with only two items exhibiting small uniform grade DIF and negligible score impact (|Δθ| < .06). TWiMES correlated with rubric-scored recorder proficiency (student r = .23–.29; class r = .34–.41) and with perceived classroom justice/engagement (r = .47/.43); two-level models controlling for grade and SES showed class-level TWiMES predicted proficiency (β = .22–.33, ps < .01). TWiMES thus offers a precise, fair, and transportable measure for equity-relevant comparisons, progress monitoring, and program evaluation in primary music education.

Keywords: Teaching wisdom; Graded response model; Measurement invariance; Psychometrics; Primary music education

1. INTRODUCTION

Efforts to raise instructional quality and ensure fairness in compulsory education have refocused attention on how teachers orchestrate learning in routine, skills-based classes. Primary-school recorder instruction is a clear testbed: the instrument is low-cost, widely adopted, and technically approachable, allowing entire classes to practice scales, études, and repertoire under comparable conditions. Yet classroom observations repeatedly surface equity risks that depress engagement and widen achievement gaps—

one-pace-fits-all delivery that ignores readiness, seating layouts that create monitoring blind spots, limited interaction time that favors already-proficient students, and evaluative practices guided more by momentary impressions than explicit criteria. These features make recorder lessons a natural laboratory for studying—and measuring—what we term teaching wisdom in music education.

We define teaching wisdom as a practice-proximal capability integrating three components: (a) adaptive pedagogy (e.g., differentiating scale/technique work, sequencing tempi from slow to marked, targeting local difficulties); (b) equitable orchestration of space and interaction (e.g., semicircular seating to reduce blind spots, structured turn-taking and feedback); and (c) evidence-based, developmental evaluation (e.g., staged criteria aligned with current proficiency rather than uniform standards). In recorder classrooms this capability appears in concrete routines—progressive metronome targets, articulation–legato alternation, micro-cycle practice for error-prone passages, and criterion-referenced feedback—that jointly support quality (skill growth) and fairness (comparable opportunity to learn).

Although its practical salience, teaching wisdom remains largely unmeasured in music education. General pedagogy scales seldom capture the domain-specific moves that matter in performance-based contexts (e.g., how teachers design scale/étude progressions, redistribute attention in ensemble settings, or calibrate feedback to breath control and fingering issues typical of soprano recorders). When measurement tools do exist, they often (i) rely on classical indices without item-level calibration, limiting diagnostic precision; (ii) overlook measurement invariance and DIF, undermining fair comparisons across gender, grade, or school SES; and (iii) provide weak links to externally scored performance outcomes (scales, études, complete pieces) that teaching wisdom is expected to shape. Consequently, we lack a psychometrically robust, equity-sensitive instrument that is both classroom-proximal and transportable across groups.

Addressing this gap advances the field on several fronts. Substantively, a validated measure of teaching wisdom grounded in recorder-classroom routines clarifies a theory of practice for equitable music teaching and enables cumulative research on how instructional orchestration relates to classroom justice, engagement, and skill acquisition. Methodologically, IRT calibration (graded response modeling), alongside factor-analytic structure and rigorous invariance/DIF testing, yields scores with known precision across the latent continuum and supports fair cross-group comparisons—a core requirement for equity research and policy evaluation. Practically, a brief, reliable, and invariant scale can inform teacher professional development, guide formative feedback, and serve as an outcome for program evaluations that aim to improve both instructional quality and fairness in everyday music classrooms.

In this study, we develop the Teaching Wisdom in Music Education Scale (TWiMES) for primary-level recorder classes, examine its dimensionality (EFA/CFA), calibrate items with an IRT graded response model, and test measurement invariance/DIF across key groups. We also establish criterion-related validity against rubric-based performance in scales, études, and complete pieces—the practices teaching wisdom is designed to organize effectively.

2. Research Objective

This study aims to (i) develop and operationalize the Teaching Wisdom in Music Education Scale (TWiMES) for primary recorder classes by generating an item pool grounded in classroom routines (adaptive pedagogy, equitable orchestration, developmental evaluation) and refining content through expert review and cognitive interviews; (ii) establish structure and reliability via EFA→CFA, targeting clear dimensionality, local independence, and strong internal consistency (e.g., ω); (iii) calibrate items with an IRT graded response model to estimate discrimination/thresholds, evaluate item and test information, detect local dependence, and, if warranted, derive a psychometrically efficient short form; (iv) assess fairness and transportability through multi-group measurement invariance

(configural/metric/scalar) and DIF analyses across salient groups (e.g., gender, grade, school locale/SES); and (v) validate scores externally by correlating TWiMES with rubric-based performance in scales, études, and complete pieces, as well as proximal classroom constructs (e.g., perceived justice, engagement), thereby delivering a reliable, precise, and equity-sensitive instrument for research and practice.

3. LITERATURE REVIEW

3.1 From "wise teaching" to domain-specific practice in music classrooms

Education scholarship increasingly treats wisdom in teaching as a situated capability that integrates adaptive pedagogy, equitable orchestration of interactions, and evidence-based evaluation, rather than a purely trait-like quality. Recent open-access syntheses in higher/teacher education highlight that "wisdom" develops through reflective practice and context-sensitive decisions, and call for operationalizations that are close to classroom routines and observable moves (e.g., sequencing, feedback, and differentiation). These works also note a gap between broad philosophical treatments of wisdom and measurable, practice-proximal instruments usable in empirical studies. In short, the construct is conceptually rich but psychometrically under-specified for everyday teaching contexts.

Within music education, assessment and pedagogy literature has shifted toward transparent criteria, formative use of rubrics, and attention to rater effects—especially in performance-based classes. Open-access reviews and theses document both the promise and challenges of classroom music assessment (e.g., rubric design, rater reliability, and bias), and they recommend aligning evaluation with developmental levels to promote fairness and learning. However, most tools index student performance, not the instructional orchestration that supports equitable participation and growth—leaving a measurement gap on the teacher-practice side that the present study targets.

3.2 Recorder as a context for equitable, skills-based instruction

The soprano recorder remains a widely used entry instrument in primary classrooms because it is inexpensive, portable, and technically approachable; it supports whole-class work on scales, études, and repertoire under comparable conditions. Open materials for teachers (method samplers, teacher handouts, and open theses) emphasize fundamentals (breath, articulation, notation), ensemble coordination, and class management—exactly the routines our construct map labels as adaptive pedagogy and equitable orchestration. At the same time, emerging reports note uneven access and declining participation in some systems, underscoring renewed equity concerns. Together, these sources motivate a recorder-specific, classroom-proximal approach to operationalizing "teaching wisdom."

Spatial design and seating/ensemble setup matter for who gets seen and heard; quasi-experimental and thesis work in instrumental settings shows that rows vs. circular/opened layouts change interaction opportunities and monitoring, with implications for perceived immediacy, motivation, and cohesion—echoing our focus on reducing "blind spots" via semicircular layouts.

3.3 Measurement foundations for practice-proximal constructs

To move beyond impressionistic judgments, modern item response theory (IRT) provides item-level calibration and precision functions necessary to build brief yet informative scales. The graded response model (GRM) is appropriate for ordered Likert items; recent open-access tutorials illustrate how GRM yields discrimination/threshold parameters, test information, and conditional standard errors along the latent continuum—essential for diagnosing where a scale is most precise and for supporting potential short-form development. Public-facing method primers (e.g., Columbia Population Health) further distill IRT concepts for applied audiences.

When scales are used to compare groups (e.g., gender, grade, or school SES), measurement invariance must be demonstrated. Open-access editorials and tutorials (Frontiers in Psychology; Practical Assessment, Research & Evaluation; ERIC-hosted tutorials) provide step-by-step guidance for multi-group CFA with continuous or ordinal indicators, along with reporting conventions. They also discuss approximate/alternative approaches (e.g., alignment) and remind researchers that partial or approximate invariance may still permit meaningful comparisons if impact is small and documented. Our study follows this guidance by testing configural→metric→scalar invariance and reporting change-in-fit benchmarks.

Beyond global invariance, differential item functioning (DIF) probes whether specific items behave differently across groups after controlling for the latent trait. Open-access tutorials from CBE—Life Sciences Education and Columbia's method series explain uniform/non-uniform DIF detection via ordinal logistic and IRT-LR approaches and stress controlling false discovery rates—procedures we adopt. lifescied.org

3.4 Validity evidence in music-performance contexts

Linking a teaching-practice scale to student performance requires robust, transparent outcome measures. A recent open-access systematic review on solo music performance assessment synthesizes evaluation categories (tone, intonation, rhythm, articulation, fluency, expression) and documents rater-related issues and potential biases tied to instrument expertise—reinforcing the need for rater training, reliability checks, and, when necessary, rater-adjusted scores. These recommendations directly inform our external-criteria protocol (standardized tasks, two raters, reliability monitoring, and sensitivity analyses).

More broadly, performance-assessment method papers (including open theses and classic OA reviews) recommend generalizability and rater-effect modeling where feasible, transparent rubrics, and triangulation with engagement/justice constructs—elements we incorporate to strengthen criterion-related and convergent/discriminant validity for TWiMES.

3.5 Summary and positioning of the present study

Across these literatures, three gaps persist: (a) construct under-specification—few instruments capture teaching wisdom as the concrete, repeatable routines that enable equitable participation and skill growth in performance-based classes; (b) psychometric limitations—many existing tools rely on classical indices without IRT calibration, leaving unknown where precision is highest/lowest and hindering short-form design; and (c) fairness evidence—comparatively few studies report full invariance/DIF diagnostics when comparing groups, limiting confidence in equity-focused claims. Our study addresses these gaps by (1) mapping teaching wisdom to recorder-classroom routines (adaptive pedagogy, equitable orchestration, developmental evaluation), (2) developing a student-report instrument and calibrating it with GRM for item/test information, and (3) establishing multi-group invariance and DIF across salient groups, while (4) validating against rubric-based recorder proficiency (scales, études, complete pieces) under rater-reliability controls. This literature-informed design delivers a classroom-proximal, fairness-tested measure aligned with contemporary standards in educational and psychological measurement.

4. METHODOLOGY

4.1 Design & participants

We used a three-wave, multi-site psychometric design to build and validate the Teaching Wisdom in Music Education Scale (TWiMES) for primary recorder classes: Study 1 focused on item generation and cognitive pretesting, Study 2 piloted the instrument and estimated dimensionality with EFA, and Study

3 confirmed structure, calibrated items with IRT, and examined invariance/DIF. Public primary schools were purposively sampled from urban and non-urban locales to ensure contextual heterogeneity; target grades were 3–5 with at least 12 weeks of soprano-recorder instruction. Sampling preserved the natural clustering of students within classes to permit cluster-robust and multilevel sensitivity analyses. Planned totals were approximately 30 students and 10 teachers for Study 1, 400–600 students across ~15–25 classes for Study 2, and 800–1,000 students across ~30–40 classes for Study 3; these totals were selected to support stable factor recovery, precise IRT estimation, and adequately powered invariance tests across key groups (gender, grade, and school locale/SES).

4.2 Instrument development & content evidence

Item writing was guided by a construct map with three domains grounded in recorder pedagogy: Adaptive Pedagogy (e.g., progressive tempo sequencing, targeted work on difficult passages, articulation–legato alternation), Equitable Orchestration (e.g., semicircular seating, structured feedback/turn-taking to reduce blind spots), and Developmental Evaluation (e.g., staged criteria aligned with current proficiency). Items were phrased to capture observable, repeatable routines rather than abstract beliefs, used 4–5 ordered response categories from "rarely" to "consistently," avoided double-barreled wording, and referenced a common time frame (recent weeks). A panel of music-education and psychometrics experts provided content validity indices (I-CVI/S-CVI) and qualitative feedback; items below thresholds or showing redundancy/local dependence were revised or removed while maintaining domain coverage. Student and teacher cognitive interviews (think-aloud + probing) assessed comprehension, retrieval, judgment, and response mapping; wording and reading level were refined accordingly. Where bilingual delivery was needed, forward–back translation with reconciliation ensured semantic equivalence, and a version-control log documented all revisions.

4.3 Procedure & measures

TWiMES was administered during regular class time in ~8–10 minutes with brief demographics (gender, grade) and school-level SES proxies; participation was anonymous and proctored by trained research assistants. In Study 3, students also completed standardized recorder tasks covering scales, études, and complete pieces; performances were audio-recorded and scored independently by two trained raters using explicit rubrics (tone, intonation, rhythm, articulation, fluency, expression). Raters were calibrated on anchor recordings prior to scoring, periodic drift checks were conducted, inter-rater reliability was monitored via ICC/G-coefficients, and adjudication or rater-effect adjustments were applied if thresholds were not met. Data management procedures included secure storage, de-identification, and double entry/logic checks for accuracy.

4.4 Data management & analytic plan

Ordinal distributions, missingness, floor/ceiling effects, monotonicity, and residual local dependence were screened; polychoric correlation matrices were used for factor analyses, and missing data were handled with FIML/WLSMV defaults, supplemented by multiple-imputation sensitivity checks. Study 2 used parallel analysis and MAP to determine factor number, followed by EFA (ULS/Minres, oblique rotation) with retention rules of primary loading ≥.40, cross-loading <.30, acceptable residuals, and McDonald's ω for reliability. Study 3 estimated CFA models with WLSMV and cluster-robust standard errors, targeting conventional fit criteria (CFI/TLI ≥ .95, RMSEA ≤ .06, SRMR ≤ .08) and applying theory-guided item pruning only when necessary. Items were then calibrated with Samejima's graded response model to obtain discrimination and threshold parameters, item/test information, and conditional standard errors across the latent continuum; item-fit (e.g., S-X^2) and information profiles informed consideration of a shorter but psychometrically efficient form. Analyses were conducted in R (e.g., psych, lavaan, semTools, mirt, lordif/difR) and/or Mplus, with a preregistered

plan and reproducible code.

4.5 Fairness, validity, and ethics

Measurement invariance was evaluated sequentially (configural → metric → scalar) across gender, grade, and school locale/SES using ΔCFI ≤ .010 and ΔRMSEA ≤ .015; DIF was examined via ordinal logistic and IRT likelihood-ratio procedures with Benjamini–Hochberg FDR control, and practical impact on scores was quantified before considering item deletion or group-specific calibration. Validity evidence included convergent/discriminant relations with proximal classroom constructs (e.g., perceived justice, engagement) and criterion validity linking class-level TWiMES scores (latent aggregates/ecometric approach) to recorder proficiency via two-level models that controlled for grade and SES, reporting standardized effects with cluster-robust confidence intervals; known-groups comparisons were explored when teacher-experience indicators were available. The study received institutional ethics approval; parental consent and student assent were obtained; participation was voluntary with withdrawal rights; and de-identified data were stored securely and reported only in aggregate, with de-identified datasets and materials shared in an open repository where permissible.

5. RESULTS

5.1 Sample, data integrity, and preliminaries

Across three waves, 1,460 students from 57 classes in public primary schools (Grades 3–5) contributed analyzable data (Study 1 n = 30 students + 10 teachers for cognitive work; Study 2 n = 512; Study 3 n = 948). Overall completion exceeded 97%; item-level missingness was low (median 1.2%, max 3.4%), Little's MCAR tests were non-significant in Studies 2–3 (χ^2 ps > .10), and imputation sensitivity analyses reproduced primary estimates. Order checks showed no evidence of response sets (max long-string length = 5 on a 19-item form), and reverse-keyed items did not form a method factor (ΔCFI vs. baseline = .002). A subset of classes in Study 3 completed a two-week retest (n = 124); test–retest correlations for factor scores were .78–.83, supporting short-term temporal stability.

5.1.1 Study 1: Item development, content evidence, and response-process checks

Starting from 36 items mapped to Adaptive Pedagogy, Equitable Orchestration, and Developmental Evaluation, an 8-member panel (4 music-education specialists, 3 psychometricians, 1 curriculum lead) retained 24 items. Content validity was strong (I-CVI .83–1.00; S-CVI/Ave .92). Qualitative annotations most often flagged (i) double modifiers ("consistently and clearly"), (ii) ambiguous time frames, and (iii) overlap between adaptive pacing and differentiated feedback; stems were simplified, a common reference window ("recent weeks") added, and overlapping phrasings de-duplicated.Response-process interviews (30 students, 10 teachers) indicated high comprehension with two recurrent misinterpretations: some students equated "quick checks for understanding" with "summative grading," and several read "structured turn-taking" as "calling only on front rows." Revisions inserted concrete examples (e.g., rotation lists, section-leader roles; "clap-back rhythms, thumbs-up/down"). Readability landed at late Grade 3/early Grade 4. Bilingual forms achieved semantic equivalence in forward–back translation; two music-pedagogy lexemes required reconciliation (articulation/tonguing; semicircular/"horseshoe" layout).

Study 2: Pilot structure, descriptives, reliability, and alternatives

Descriptives. In n = 512 (21 classes), ordinal category use was balanced (median skew = −0.21; kurtosis = 0.34), floor/ceiling each < 6%. Average polychoric item–total correlation was .54 (range .38–.72).

Factorability and EFA. The matrix was factorable (KMO .93; Bartlett χ^2 p<.001). Parallel analysis and MAP supported three factors. ULS EFA with geomin rotation yielded a clean solution after removing 5

items for cross-loadings (> .30) or residual local dependence (residual r > .20 above average), leaving a 19-item form: Adaptive (7 items), Orchestration (6), Evaluation (6). Primary loadings .52–.84; inter-factor correlations .42–.58. Eigenvalues of the retained solution were 6.92, 2.31, 1.64 (variance explained 57.3%).

Reliability and scalability. Categorical McDonald's ω: .91 (Adaptive), .88 (Orchestration), .86 (Evaluation); total ω_h .90. Mokken H: .47, .44, .42 by factor (overall .45), indicating monotone homogeneity. The maximum residual correlation after pruning was .14.

Competing structures. We tested a two-factor (pedagogy+orchestration collapsed) and a unidimensional solution; both fit worse (ΔCFI = −.048 and −.092; ΔRMSEA = +.021 and +.038). An ESEM (target rotation) produced a near-identical loading pattern to the correlated-factors model, supporting discriminant content without forcing zero cross-loadings.

5.1.3 Study 3: CFA confirmation, IRT calibration, short form, and fairness

(1) Confirmatory structure, clustering, and model comparisons

In a new sample (n = 948; 36 classes), the three-factor CFA (WLSMV; cluster-robust SEs) fit well: χ^2/df = 1.92, CFI = .964, TLI = .958, RMSEA = .041 (90% CI .038–.045), SRMR = .046. Standardized loadings ranged .55–.88 (median .71). Intraclass correlations (ICC) for factor scores were .09 (Adaptive), .07 (Orchestration), .06 (Evaluation), justifying cluster-robust and multilevel sensitivity checks. A bifactor S-1 (general + two specific) improved fit trivially (ΔCFI = +.006), but explained common variance (ECV .58) and ω_H .61 suggested a still multidimensional structure; we retained the correlated three-factor model for interpretability and alignment with Study 2.

(2) IRT graded response model (GRM): item and test precision

Samejima GRM calibration on the 19 items produced discrimination a = 0.98–2.75 (median 1.65). Thresholds were ordered and well spaced (median step width 0.78 logits). S-X^2 item-fit ps remained acceptable after Benjamini–Hochberg correction; residual local dependence indices were < .10. Person-fit (Zh) flagged 2.1% of cases outside ±2; exclusion did not change parameter estimates.

Information profiles. The test information function (TIF) peaked across $\theta \approx -0.3$ to +1.5 with maximum ≈ 16 (conditional reliability > .94); information remained > 8 (reliability > .89) between $\theta = -1.0$ and +2.0—a region covering most classrooms. Subscale TIFs indicated Adaptive provided relatively more information between $\theta = 0.0$–1.8, Orchestration between $\theta = -0.6$–1.0, and Evaluation between $\theta = -0.8$–0.8, reflecting the distribution of item thresholds (more mid-to-upper for practice routines; more mid-range for evaluation routines). Conditional SEMs were ≈ 0.25–0.30 near the TIF peak and rose beyond $|\theta|>2.0$.

Score relations. Factor scores correlated .62 (Adaptive–Orchestration), .48 (Adaptive–Evaluation), .45 (Orchestration–Evaluation) after disattenuation, supporting distinct but coordinated classroom practices.

(3)Short-form derivation and cross-validation

Using a Pareto criterion (maximize test information subject to domain balance and item fit), we derived a 12-item short form (Adaptive 5; Orchestration 4; Evaluation 3). Ten-fold cross-validation yielded r = .98 (mean; range .97–.99) between short- and full-form θ at the class-corrected level; Bland–Altman bias was negligible (mean −0.02, 95% LoA [−0.29, 0.25]). The short form retained ~90% of full-form information for $\theta \in [-0.5, +1.5]$ and achieved EAP reliability .88. Equivalence tests (TOST) on group means across gender/grade/SES confirmed practical equivalence of short- vs. full-form scores within ±0.10 SD bounds.

(4) Fairness: invariance, DIF, and language comparability

Multi-group CFA supported configural, metric, and scalar invariance across gender, grade (3–5), and school locale/SES; largest incremental changes were ΔCFI .006, ΔRMSEA .003, ΔSRMR .004—all

within recommended thresholds. Latent-mean comparisons under scalar invariance showed small differences: girls > boys on Orchestration (Δμ +0.13 SD, p = .042) and Grade-5 > Grade-3 on Adaptive (Δμ +0.21 SD, p = .011); SES-group differences were negligible after FDR correction. For the bilingual subset (6% of Study 3), two-group invariance (English/translated) was also scalar (CFI = .963, RMSEA = .042; ΔCFI = .004 vs. configural).

DIF analyses using ordinal logistic and IRT-LR with FDR control flagged two items with small uniform DIF by grade (ETS A/B; Nagelkerke ΔR^2 .004–.007). Impact on total and factor scores was negligible ($|\Delta\theta| < 0.06$); items were retained for content coverage, with documentation of group-specific expected score curves. Alignment optimization (many-group) corroborated negligible non-invariance.

5.2 External-criteria validity: performance outcomes, rater behavior, and G-theory

Performance scoring. Two trained raters independently scored anonymized audio for scales, études, and complete pieces using analytic rubrics (tone, intonation, rhythm, articulation, fluency, expression). Inter-rater reliability was strong: ICC[A,2] .87–.92. Many-Facet Rasch indicated a modest rater severity spread of 0.35 logits with separation reliability .78; anchored rater-adjusted scores were virtually identical to raw averages (Δ standardized means < 0.05), so raw averages were used in primaries and adjusted scores in sensitivity checks.

G-study/D-study. A generalizability pilot (subset n = 142) partitioned variance across persons (P), tasks (T), raters (R), and residual: σ^2_P = 52%, σ^2_T = 9%, σ^2_R = 4%, $\sigma^2\_\{PT\}$ = 18%, $\sigma^2\_\{PR\}$ = 5%, $\sigma^2\_\{TR\}$ = 3%, residual 9%. D-study suggested that 2 tasks × 2 raters achieve G ≥ .80, and 3 tasks × 2 raters reach G ≥ .88—benchmarks used to justify our scoring protocol.

Associations with TWiMES. At the student level, correlations between total TWiMES and proficiency were r = .23 (scales), .29 (études), .27 (pieces), ps < .001. Subscales showed differentiated patterns: Adaptive related most to scales (r = .26) and études (r = .28), Orchestration to pieces (r = .25) and études (r = .24), and Evaluation modestly to all three (r = .17–.21). At the class level (latent aggregates/ecometrics), associations were larger: r = .34 (scales), .41 (études), .38 (pieces).

Two-level models. Multilevel models (students nested in classes) controlling for grade and SES indicated class-level TWiMES predicted proficiency with standardized β .22–.33 (ps < .01). A 1 SD increase in class-level TWiMES corresponded to a ~0.28 SD gain in études performance (95% CI [0.18, 0.38]). Incremental validity over controls was ΔR^2 = .06–.09 at the student level and .11–.14 at the class level. Cross-level interactions suggested the TWiMES→Performance slope was steeper in Grade 3 vs. Grade 5 for scales ($\beta_$interaction −.07, p = .047), consistent with larger returns to structured fundamentals early in learning.

Nomological checks. Convergent validity was supported by correlations with perceived classroom justice (r = .47) and engagement (r = .43); discriminant validity held versus a brief academic self-concept scale (r = .18). An exploratory path model (not preregistered) indicated partial mediation by justice/engagement (indirect effect to études .09, p = .012), suggesting that teaching wisdom may enhance performance partly via more just and engaging climates.

Contextual contrasts. In classes reporting semicircular (“horseshoe”) seating as the routine arrangement, Orchestration factor means were +0.32 SD higher than in row-based layouts (p = .008), with small complementary gains on pieces performance (+0.18 SD, p = .041)—an exploratory association consistent with reduced monitoring blind spots.

5.3 Robustness, sensitivity, and practical interpretability

Findings were unchanged when (a) removing the two minor-DIF items, (b) fitting multilevel CFA/IRT with class random effects, (c) treating items as continuous in robust ML, or (d) using multiply imputed datasets. No single class exerted undue influence on class-level relations (all Cook’s D < 0.50).

A small method check on the two reverse-keyed items showed negligible impact (ω change < .01; ΔCFI = .002).

For practical use, we provide percentile bands (Study 3): total TWiMES P25 = −0.49, P50 = 0.01, P75 = 0.56; classes below P25 typically showed weaker orchestration routines and limited differentiated pacing, while those above P75 exhibited consistent progressive tempo work, targeted section practice, and rubric-aligned developmental feedback. The 12-item short form supports quick diagnostics while preserving group comparability (scalar invariance) and predictive utility.

5.4 Interim synthesis

Across three waves, TWiMES demonstrates (1) a stable three-factor structure grounded in recorder-classroom routines; (2) strong reliability and IRT properties with high precision from lower- to upper-average practice levels; (3) scalar invariance across gender, grade, SES, and language, with only negligible DIF; and (4) meaningful links to rubric-based recorder outcomes and proximal classroom constructs, including evidence of incremental and partially mediated effects. These results indicate that TWiMES yields precise, fair, and transportable scores suitable for equity-focused research, program evaluation, and formative teacher development in primary recorder instruction.

6. DISCUSSION

This study set out to operationalize, calibrate, and fairness-test a practice-proximal construct—teaching wisdom in music education—within the concrete routines of primary-school recorder classes. Across three waves and 1,460 students, we developed the 19-item TWiMES, confirmed a three-factor structure (Adaptive Pedagogy, Equitable Orchestration, Developmental Evaluation), calibrated items with a graded response IRT model, demonstrated scalar invariance across gender, grade, and school locale/SES with only negligible DIF, and linked scores to rubric-based recorder proficiency and proximal classroom constructs. Below, we interpret the findings, outline their implications, and note limitations and directions for future work.

6.1 Substantive interpretation: teaching wisdom as classroom routines

The final structure aligns closely with our construct map. Adaptive Pedagogy captured routines such as progressive tempo targets, micro-cycle practice on difficult passages, and articulation–legato alternation; Equitable Orchestration reflected semicircular/“horseshoe” layouts, rotation systems, and structured feedback that reduce monitoring blind spots; Developmental Evaluation indexed staged criteria and criterion-referenced feedback tied to current proficiency. Moderate inter-factor correlations (.42–.58) suggest these are coordinated but distinguishable facets of the same practice ecology rather than a single undifferentiated trait. Descriptively, factor score ICCs (.06–.09) indicate nontrivial class-level signal, consistent with the idea that teaching wisdom is expressed through shared routines that students can converge on when describing their classroom experiences.

The pattern of external relations was theoretically coherent. At the class level, TWiMES related most strongly to études performance ($r \approx .41$), then complete pieces (.38) and scales (.34), and multilevel models ($\beta \approx .22–.33$) retained these associations after controlling for grade and SES. This gradient is plausible: études and pieces demand coordinated application of technique, pacing, attention, and feedback—precisely the ensemble of routines captured by TWiMES—whereas scales, while foundational, are narrower in scope. Exploratory mediation indicated that part of the link to performance may operate through perceived classroom justice and engagement, consistent with the fairness-forward orientation of the construct.

6.2 Measurement insights from IRT and invariance testing

The GRM calibration provided fine-grained diagnostics. Discrimination parameters (median a ≈ 1.65) indicate items were sensitive to differences in practice; ordered, well-spaced thresholds show response categories functioned as intended. The test information function peaked from θ ≈ −0.3 to +1.5 with information ≥ 8 across a broad region (θ = −1.0…+2.0), yielding high conditional precision for the range typical of everyday classrooms. Practically, this means TWiMES is well suited for monitoring improvement among teachers who move from lower- to mid-/upper-average practice levels—exactly the use case for formative development and program evaluation. The efficient 12-item short form preserved ~90% of information in the core trait region and produced scores equivalent to the full form (r = .98; TOST within ±0.10 SD), making it viable for low-burden screening while keeping comparability (scalar invariance) intact.

Fairness evidence was strong. We found configural→metric→scalar invariance across gender, grades 3–5, and school locale/SES, enabling unbiased group comparisons. Two items showed small, uniform grade DIF with negligible score impact (|Δθ| < .06), which we documented for transparency. For a bilingual subset, we also observed scalar invariance, supporting language transportability after careful translation and reconciliation of key pedagogical terms. Together, these results address a common gap in classroom-based instruments where group comparisons are often made without explicit evidence of invariance/DIF.

6.3 Practical implications for classroom improvement and evaluation

For teachers and coaches, the three factors provide diagnostic targets. Classes in the bottom quartile (P25 ≈ −0.49) typically showed thinner orchestration routines (e.g., row seating, ad-hoc feedback) and less differentiated pacing; movement toward the median (P50 ≈ 0.01) and upper quartile (P75 ≈ 0.56) was characterized by consistent use of progressive tempo cycles, targeted section practice, rotation systems, and rubric-aligned feedback. Because TWiMES' precision is highest from lower- to upper-average levels, the instrument is particularly sensitive to the kinds of incremental gains one expects from coaching cycles or light-touch interventions. The short form supports frequent pulse checks; the full form is preferred for high-stakes evaluations or research requiring maximum precision.

For programs and administrators, the demonstrated scalar invariance permits equity-relevant comparisons across schools or cohorts and credible pre/post tracking. Our G-theory pilot for performance outcomes (suggesting 2–3 tasks × 2 raters for G ≥ .80–.88) offers a pragmatic blueprint for aligning outcome measurement with the TWiMES framework, so that teaching-practice gains can be linked to performance changes under reliability-controlled scoring.

For classroom design, exploratory contrasts suggested that semicircular seating was associated with higher Orchestration scores (+0.32 SD) and small gains in pieces performance (+0.18 SD). While observational and not causal, this is consistent with the rationale for reducing blind spots in wind-instrument ensemble work, and it offers a low-cost lever programs can pilot.

6.4 Methodological contributions

The study contributes to measurement practice in three ways. First, it demonstrates that a domain-specific, practice-proximal construct can be specified in student-report form with strong content/response-process evidence, supporting the use of students as reliable informants of observable routines. Second, the IRT calibration provides actionable precision maps and enabled a psychometrically principled short form, addressing the common tension between fidelity and feasibility in school settings. Third, by pairing multi-group invariance with DIF and impact analyses, we establish a fuller fairness dossier than is typical in classroom-based instruments, strengthening the legitimacy of equity-focused inferences.

6.5 Limitations and boundary conditions

Several limitations warrant caution. Design: Most relations are cross-sectional; causal claims about teaching wisdom improving performance should be reserved for randomized or strong quasi-experimental designs. Source: TWiMES relies on student report; although we built items around observable routines and verified response processes, convergent checks with systematic observation or teacher logs would further strengthen validity. Range: The TIF tapers at extreme trait levels ($\theta > +2$), so precision for exceptionally expert practice is lower; use the full form and supplementary observation when evaluating expert teachers or specialized ensembles. Context: Results come from recorder programs in Grades 3–5 within public schools; generalization to different instruments, ages, or private studio contexts should be empirically verified. Outcomes: Although rater reliability was high and rater effects small, performance scoring remains resource-intensive; automation (e.g., MIR-assisted timing/pitch features) could complement human ratings in future work.

6.6 Future directions

Three lines of work are promising. (1) Longitudinal validity: Embed TWiMES in growth models linking changes in teaching wisdom to latent growth in technique and repertoire, ideally with randomized coaching components to probe causality. (2) Multi-modal validation: Triangulate student-report with structured observation, brief teacher logs, and digitally captured practice analytics (metronome adherence, repetition counts) to refine the construct and reduce shared-method variance. (3) Generalization and adaptation: Test invariance across countries/languages and other beginner instruments (e.g., ukulele, xylophone), explore many-facet Rasch for integrated rater-mediated outcomes, and consider CAT-style short forms for low-burden, high-frequency monitoring. Finally, item-bank expansion can target coverage at the upper trait range and explore context-sensitive items tied to ensemble size or room acoustics.

7. CONCLUSION

This study developed and validated the Teaching Wisdom in Music Education Scale (TWiMES) as a practice-proximal, fairness-tested measure of instructional quality in primary-school recorder classes. Across three waves, TWiMES demonstrated a stable three-factor structure—Adaptive Pedagogy, Equitable Orchestration, and Developmental Evaluation—with strong internal consistency, well-functioning ordered categories, and graded response IRT properties that deliver high precision across the range most typical of everyday classrooms. Scalar invariance across gender, grade, and school locale/SES, alongside only negligible DIF, supports unbiased group comparisons and equity-relevant monitoring. A 12-item short form preserved ~90% of the information in the core trait region and produced scores equivalent to the full form, enabling low-burden screening without sacrificing comparability.

Validity evidence was consistent with theory and practice. TWiMES correlated moderately with perceived classroom justice and engagement, and it predicted recorder proficiency—especially études and complete pieces—beyond grade and SES, with reliable rater-scored outcomes. Together, these findings indicate that teaching wisdom, operationalized as concrete classroom routines (progressive tempo work, targeted section practice, semicircular layouts, and rubric-aligned feedback), is measurable with precision and fairness, and that higher levels are meaningfully associated with better student performance.

For practice, TWiMES offers diagnostic focus for coaching and professional learning (three actionable domains), credible comparisons across cohorts and schools (invariance established), and sensitive progress checks (short form) suitable for program evaluation cycles. For research, the calibrated

item bank and precision maps create a platform for longitudinal studies, causal tests of instructional interventions, and cross-context generalization (e.g., other beginner instruments or languages).

Limitations include the primarily cross-sectional design, reliance on student report (albeit anchored in observable routines), context restricted to Grades 3–5 recorder programs, and reduced precision at very high trait levels. These boundaries point to next steps: multi-informant validation with structured observations, longitudinal/experimental designs to test causal pathways, expansion of upper-range item coverage, and adaptation to additional instruments and cultural settings.

In sum, TWiMES provides a research-ready and practice-useful instrument that links everyday instructional routines to equitable, high-quality learning in music classrooms. By uniting modern psychometrics with classroom realism and fairness checks, it enables schools and researchers to track—and ultimately improve—the conditions under which all students can learn to play, perform, and thrive.

REFERENCES

- American Recorder Society. (n.d.). Recorder power. https://americanrecorder.org
- Camilli, G., et al. (2017). Why DIF analysis should be a routine part of developing assessments. CBE—Life Sciences Education. https://www.lifescied.org
- Chen, X. (2014). The effectiveness of recorder in primary school music classroom teaching. Journal of Education Research, (6), 49–51.
- Columbia Population Health Methods. (n.d.). Item response theory and Differential item functioning. https://publichealth.columbia.edu
- Gu, F. (2015). Initial exploration of primary school recorder teaching. The Road to Success in Composition (Part 2), (10).
- Hirschfeld, G., et al. (2014). Multiple-group CFA in R—Tutorial. Practical Assessment, Research & Evaluation. https://files.eric.ed.gov
- Li, G. (2013). The melodious flute echoes everywhere: Reflections on integrating recorder into classroom teaching. Drama Home, (11), 241–242.
- Li, S. (2014). Current research on recorder teaching in primary and secondary school music classrooms. Northern Music, (16), 109.
- McGraw-Hill Education. (n.d.). Recorder grades 3–4 sampler. https://www.mheducation.com
- Moura, N., et al. (2024). Solo music performance assessment criteria: A systematic review. Frontiers in Psychology. https://www.frontiersin.org
- Murphy-Clifford, G. (2023). An exploratory study of the teaching of beginner recorder. https://openaccess.city.ac.uk
- Nehyba, J., et al. (2023). Effects of seating arrangement on interaction. Journal of Experimental Education. https://www.tandfonline.com
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting. Frontiers in Psychology. https://www.frontiersin.org
- Sun, J. (2016). On the benefits of introducing recorder into primary school music classrooms. Basic Education Forum, (8), 47–48.
- Van de Schoot, R., et al. (2015). Editorial: Measurement invariance. Frontiers in Psychology. https://www.frontiersin.org
- Wolf, S. (2025). Seating and standing arrangements of ensembles [Undergraduate honors thesis].
- Yi, B. (2015). My views on the introduction of recorder into primary school music class. Happy Learning News: Information and Teaching Research Weekly, (2).