INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 14, Issue 4, (2025) E-ISSN: 2226-6348

Reliability and Validity of Questionnaire for Preschool Teachers in Music Teaching Ability: A Pilot Study

Miaomiao Sun, Huey Yi@Colleen Wong*

Department of Music and Music Education, Faculty of Music and Performing Arts, Sultan Idris Education University, Malaysia

Email: p20211002423@siswa.upsi.edu.my

Corresponding Author Email: colleen@fmsp.upsi.edu.my

DOI Link: http://dx.doi.org/10.6007/IJARPED/v14-i4/26627

Published Online: 05 October 2025

Abstract

Music activities in early childhood foster measurable gains in cognition, language and socio-emotional regulation. Yet many preschool teachers feel ill-equipped to design developmentally appropriate and culturally responsive lessons. Addressing the lack of psychometrically robust instruments that capture this multidimensional competence, this present pilot study developed and evaluated a 23-item questionnaire grounded in Shulman's pedagogical-content-knowledge framework and organised around three domains: Teaching Skill, Teaching Method and Teaching Evaluation. A quantitative cross-sectional survey was administered to 50 music-teaching staff from five kindergartens in Liaocheng City, selected through purposive convenience sampling. Internal consistency, content validity and construct validity were analysed in SPSS 28. Cronbach's alphas confirmed excellent reliability (α_TS = 0.938; α_TM = 0.909; α_TE = 0.921), while expert review yielded an S-CVI/Ave of 0.97 and identified only two items for minor revision. Factorability diagnostics were satisfactory (KMO = 0.759; Bartlett's $\chi^2(253) = 845.44$, $p < .001$) and principal-component analysis extracted three factors that explained 67.8% of total variance with primary loadings between 0.704 and 0.872 and negligible cross-loading. These results suggest that the new instrument offers a reliable and valid diagnostic tool for profiling teachers' music-teaching competence. It supports targeted professional development and enabling rigorous evaluation of early-childhood music initiatives. Future work should employ larger, stratified samples and confirmatory factor analysis to test structural invariance across regions and longitudinally to examine how improvements in specific teaching domains translate into child outcomes.

Keywords: Early-Childhood Music, Pedagogical Content Knowledge, Questionnaire Validation, Teacher Competence, Reliability and Validity

Introduction

Early-childhood engagement with rhythm, melody and movement has been linked to stronger executive functions, emergent literacy and socio-emotional regulation. It underscores music's status as a core developmental resource rather than a peripheral "arts

enrichment" activity (Arasomwan & Daries, 2025). Nevertheless, large-scale surveys show that many preschool teachers still feel ill-equipped to design and facilitate developmentally appropriate, culturally responsive and play-based music lessons. Limited pre-service coursework and scarce in-service mentoring leave them unsure how to integrate instruments, improvisation and creative movement into daily routines (Wong et al., 2024). This competence gap is typically diagnosed through general teaching self-efficacy scales or attitude inventories, instruments that capture teachers' confidence but not the multidimensional construct of music-teaching ability (Fernández Amat et al., 2024). When music-specific tools do exist, they often isolate discrete sub-skills, such as aural discrimination, or focus on beliefs rather than enacted capability. Consequently, policy makers and programme designers lack a robust diagnostic instrument for mapping preschool teachers' strengths and needs (Jiang et al., 2024; You et al., 2025).

Psychometric guidelines stress that any new scale must first demonstrate reliability and multiple forms of validity. It shows that the scale truly measures the targeted latent dimension and supports defensible inferences (DeVellis & Thorpe, 2021). A pilot study is indispensable at this juncture because it reveals ambiguous wording, cultural bias and unstable factor structures before costly large-scale deployment. Launching an untested questionnaire might risk distorted dimensionality, unreliable sub-scales and misleading professional-development priorities. These problems are repeatedly documented in educational-measurement literature (DeVellis & Thorpe, 2021). The absence of a validated and holistic instrument for preschool teachers' music-teaching ability therefore represents a critical methodological and practical gap. Without a dependable tool, researchers cannot generate comparable data across contexts and administrators cannot tailor professional-development pathways to actual needs.

In response, this present pilot study develops a comprehensive questionnaire that captures the blended musical, pedagogical and classroom-management competencies required to "bring music alive" for young learners. The pilot phase provides the first empirical test of its internal consistency, dimensional structure and content alignment. Thereby, it safeguards subsequent nationwide surveys and informs evidence-based teacher education. Accordingly, two research questions guide this study:
(1) How reliable are the questionnaire items measuring preschool teachers' music-teaching ability?
(2) How valid are the questionnaire items in representing that construct?

Literature Review

Early-childhood engagement with rhythm, melody and movement has repeatedly been shown to strengthen executive functions, phonological awareness and socio-emotional regulation. It positions music as a core developmental resource rather than a peripheral enrichment activity (Cai et al., 2025). Meta-analytic evidence further indicates that structured musical activities delivered by well-prepared educators yield small to moderate gains in vocabulary, inhibitory control and working memory. These effects are comparable with more intensive cognitive-training programmes (Degé & Frischen, 2022). These benefits, however, materialise only when teachers feel confident to design play-based and culturally responsive lessons that integrate singing, improvisation and movement into everyday routines.

Surveys across diverse jurisdictions consistently reveal that 40% to 70% of preschool teachers perceive themselves as ill-equipped to teach music, citing minimal preservice coursework, scarce mentoring and limited opportunities for hands-on rehearsal of musical techniques (Bautista et al., 2024). Even where policy frameworks endorse arts-rich curricula, many teachers restrict music time to occasional song-leading. It leaves holistic musical development largely unaddressed. Experienced teachers echo these concerns. In Hong Kong, for instance, teachers ranked music-specific professional development as their most urgent training need (Wong et al., 2024). Without a robust diagnostic tool that maps teachers' strengths and needs, organisations cannot allocate professional-development resources strategically. Also, researchers cannot track competence growth over time.

Defining that competence demands a theoretical lens that reaches beyond simple self-confidence. Shulman (1986) pedagogical content knowledge (PCK) framework posits that effective teaching emerges from the integration of disciplinary knowledge, pedagogy and understanding of learners. Applied to early-childhood music, PCK involves knowing how children construct rhythmic and tonal concepts through exploration, how to scaffold improvisation, how to manage group music-making in limited spaces and how to curate culturally inclusive repertoires. A valid instrument must therefore sample items that tap each PCK dimension, such as musical skills, pedagogical strategies, classroom management and creative dispositions rather than measuring attitude or confidence alone.

Existing questionnaires fall short of this requirement. The Music Teacher Self-Efficacy Scale captures perceived capability but ignores demonstrated knowledge and classroom action, while the Creative Practical Ability in Music measure targets children rather than teachers (Fernández Amat et al., 2024; Jiang et al., 2024). Other tools are normed on primary-school contexts or isolate single sub-skills such as aural discrimination. It limits their relevance to preschool settings. Critically, most instruments report only Cronbach's alpha, with few studies conducting confirmatory factor analysis, measurement-invariance checks or external-criterion validation. It leaves questions about dimensional stability and cross-cultural comparability unanswered.

International psychometric standards emphasise that scale developers must present converging evidence of reliability (e.g., internal consistency, test–retest stability) and validity (content, construct, criterion) before instruments are used for high-stakes decisions. Best-practice primers outline a staged process, such as item generation, expert review, cognitive interviews and pilot testing, designed to identify ambiguous wording, cultural bias and ceiling or floor effects (Boateng et al., 2018). Launching a large-scale survey without such preliminary checks can yield distorted factor structures and unreliable sub-scales. It might produce misleading conclusions about teacher competence and misdirecting professional-development investments.

Three research gaps therefore persist. First, no existing tool holistically captures the PCK-aligned construct of music-teaching ability for preschool educators. Available scales focus on attitudes, confidence or discrete musical sub-skills. Second, the psychometric evidence underpinning most instruments remains limited. They are often confined to internal-consistency coefficients without rigorous structural validation. Third, few studies report systematic pilot procedures. It leaves response-scale functioning, translation accuracy and

cultural appropriateness largely unexamined. Developing and thoroughly piloting a new questionnaire that addresses these gaps is consequently both a scholarly imperative and a practical necessity. Such an instrument would enable researchers to benchmark competence across contexts, evaluate the impact of teacher-education initiatives and support policymakers seeking to embed music more effectively within early-childhood curricula.

Methodology

This pilot study employed a quantitative cross-sectional survey strategy to rehearse every procedure that will later underpin a full validation study. Pilot work functions as a small-scale stress test that surfaces procedural weaknesses, protects resources and improves psychometric rigour (Lowe, 2019). Guided by scale-development best practice, this study unfolded in staged fashion, such as item adaptation, expert review, administration and psychometric screening to mirror the sequence recommended by Boateng et al. (2018). A purposive convenience sampling frame targeted five kindergartens in Liaocheng City that granted institutional consent. All 50 music-teaching staff who (a) held at least an early-childhood diploma, (b) had taught music for a minimum of six months and (c) volunteered to participate were included. Pilot samples of 30 to 50 respondents are sufficient to reveal item-level anomalies and generate stable variance estimates for satisfying established recommendations (Lowe, 2019).

The teacher questionnaire contained 23 Likert items adapted from Rammstedt and John (2007), Nikolić (2019), Platz et al. (2022) and Chen et al. (2020). An expert panel of two music-education professors and one psychometrician confirmed content relevance and linguistic clarity. After ethics approval, researchers visited each centre. Teachers completed the survey during a staff meeting (≈ 20 minutes) and 30 of them repeated it two weeks later to permit test–retest estimation.

Internal consistency was indexed with Cronbach's alpha and corrected item-total correlations, adopting ≥ .70 and ≥ .30 as benchmarks respectively (Anselmi et al., 2019). Temporal stability drew on intraclass correlation coefficients for the teacher scale and Pearson correlations for duplicate music-performance scores, following Grgic et al. (2020). Construct validity involved exploratory factor analysis with principal-axis factoring and varimax rotation. Adequacy criteria were a Kaiser–Meyer–Olkin index ≥ .60, a significant Bartlett's test and item loadings ≥ .40 (Shrestha, 2021). Criterion validity was inspected through Spearman correlations between teachers' total scores and children's performance test outcomes. SPSS 28 handled all analyses with $\alpha = .05$ (two-tailed).

Participation was voluntary, with informed consent from teachers, freedom to withdraw without penalty, and strict confidentiality through coding, encryption and aggregated reporting. By rehearsing each stage under controlled conditions, the study generates transparent, replicable procedures and preliminary psychometric evidence. Thereby, it lays a solid foundation for the forthcoming large-scale validation.

Findings

Reliability of Questionnaire Items

Table 1 indicates that the Teaching Skill (TS) scale is highly reliable. Cronbach's α is 0.938. It is comfortably above the 0.80 benchmark for good internal consistency. Item–total

correlations for TS1–TS9 range from 0.707 to 0.810. They are well above the accepted minimum of 0.50. It confirms that every item aligns closely with the overall construct. When each item is removed in turn, α fluctuates only between 0.928 and 0.934 with a maximum change of 0.006. It shows that no single item disproportionately affects scale consistency. These results demonstrate that all nine items contribute evenly to the measurement of teaching skill and that no deletions are required before full deployment.

Table 1.
Reliability of Teaching Skill

Variable	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted	Cronbach's Alpha
TS1	30.64	48.480	0.742	0.932	0.938
TS2	30.66	46.923	0.785	0.930	
TS3	30.78	47.930	0.791	0.930	
TS4	30.84	48.504	0.707	0.934	
TS5	30.90	46.704	0.800	0.929	
TS6	30.78	47.073	0.766	0.931	
TS7	30.84	45.035	0.810	0.928	
TS8	30.84	46.790	0.741	0.932	
TS9	30.68	47.528	0.747	0.932	

Table 2 shows that the Teaching Method (TM) scale has strong internal consistency. Cronbach's α is 0.909. It is well above the 0.80 criterion for acceptable reliability. Item–total correlations for TM1–TM6 range from 0.682 to 0.823. They surpass the minimum standard of 0.50 and indicate that every item aligns closely with the overall construct. When each item is removed in turn, α varies only between 0.881 and 0.902 with a maximum shift of 0.028. It demonstrates that no single item exerts undue influence on the scale's consistency. These results confirm that all six items contribute evenly to measuring teaching method. No deletions are necessary before the full study.

Table 2
Reliability of Teaching Method

Variable	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Cronbach's Alpha if Item Deleted	Cronbach's Alpha
TM1	19.22	20.257	0.699	0.900	0.909
TM2	19.20	19.918	0.775	0.890	
TM3	19.20	20.163	0.682	0.902	
TM4	19.32	18.875	0.733	0.895	
TM5	19.42	18.085	0.823	0.881	
TM6	19.44	18.782	0.784	0.887	

Table 3 confirms that the Teaching Evaluation (TE) scale exhibits excellent internal consistency. Cronbach's α is 0.921. It comfortably surpasses the 0.80 benchmark for reliable measurement. Item–total correlations for TE1–TE8 range from 0.653 to 0.814. All are well above the 0.50 minimum. It indicates that every item contributes meaningfully to the overall

construct. When items are deleted individually, α fluctuates only between 0.904 and 0.917 with a maximum change of 0.017, which shows that no single item disproportionately affects the scale's consistency. These findings demonstrate that the eight items function cohesively. No revisions are required before full-scale deployment.

Table 3
Reliability of Teaching Evaluation

Variable	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted	Cronbach's Alpha
TE1	27.86	30.694	0.778	0.908	
TE2	27.98	33.244	0.653	0.917	
TE3	27.92	31.300	0.743	0.910	
TE4	27.98	32.020	0.660	0.917	
TE5	27.86	30.327	0.758	0.909	0.921
TE6	27.88	29.414	0.787	0.907	
TE7	27.98	29.408	0.814	0.904	
TE8	27.98	31.816	0.700	0.914	

Validity of Questionnaire Items

Expert review confirmed the questionnaire's content validity. Item-level CVIs (I-CVIs) were computed for each statement and the scale-level CVI (S-CVI) was derived from their average. All I-CVIs exceeded the 0.75 benchmark and the S-CVI surpassed the recommended 0.90 threshold. They indicate strong expert agreement that the items adequately represent the construct. Detailed ratings and CVI calculations are summarised in Table 4.

Table 4
Expert Ratings and I-CVI, S-CVI Calculation

Items	Expert 1	Expert 2	Expert 3	Number of experts with a rating of 3 or 4	I-CVI	S-CVI/UA	S-CVI/Ave
TS1.	3	3	3	3	1.00		
TS2.	4	4	4	3	1.00		
TS3.	4	4	4	3	1.00		
TS4.	4	4	4	3	1.00		
TS5.	3	4	4	3	1.00		
TS6.	3	4	4	3	1.00		
TS7.	3	3	4	3	1.00		
TS8	3	4	4	3	1.00		
TS9	4	4	4	3	1.00		
TM1	3	3	4	3	1.00		
TM2	3	4	4	3	1.00		
TM3	3	4	4	3	1.00		
TM4	4	4	3	3	1.00		

TM5	3	3	4	3	1.00	0.91	0.97
TM6	4	4	4	3	1.00		
TE1	3	4	4	3	1.00		
TE2	3	4	4	3	1.00		
TE3	1	3	4	2	0.67 revised		
TE4	2	3	4	2	0.67 revised		
TE5	3	4	4	3	1.00		
TE6	3	3	4	3	1.00		
TE7	3	3	4	3	1.00		
TE8	3	4	4	3	1.00		

Expert ratings confirmed that the questionnaire possesses strong overall content validity. Both the universal-agreement index (S-CVI/UA = 0.91) and the average index (S-CVI/Ave = 0.97) exceeded the recommended thresholds of 0.80 and 0.90, respectively. It indicates broad expert consensus on item relevance (Shi et al., 2012; Yaghmale, 2003). Two items (TE3 and TE4) fell below the preferred item-level benchmark (I-CVI = 0.67). To address this weakness, TE3 was revised from “I pay more attention to children’s mutual evaluation” to “I pay more attention to what children think about each other’s musical performance.” TE4 was refined from “I pay more attention to children’s self-evaluation in music activities” to “I pay attention to whether children are satisfied with their performance in music activities.” These targeted wording changes are expected to improve clarity and alignment with the construct while preserving the strong content validity documented for the remaining items.

Pre-analysis diagnostics confirmed that the data were suitable for exploratory factor analysis. The Kaiser–Meyer–Olkin measure of sampling adequacy reached 0.759. It surpasses the 0.70 guideline for acceptable factorability. Bartlett’s test of sphericity produced $\chi^2(253) = 845.44$, $p < .001$. It indicates that the correlation matrix was not an identity matrix and therefore contained sufficient shared variance for factor extraction. Together, these indices demonstrate that the 50 teacher responses provide an adequate basis for reliable factor analysis (see Table 5).

Table 5
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**0.759**
Bartlett's Test of Sphericity	Approx. Chi-Square	845.442
	df	253
	Sig.	0.000

Principal-component analysis using the Kaiser criterion (eigenvalues > 1) extracted three factors from the 23 questionnaire items (see Table 6). Communalities for all items ranged between 0.547 and 0.788. It indicates that each statement retained more than half of its original variance after extraction. Because no item fell below the 0.50 benchmark for

acceptable communality, information loss was minimal and the three-factor solution can be considered robust for further interpretation.

Table 6
Communalities

Item	Initial	Extraction
TS1	1.000	0.619
TS2	1.000	0.700
TS3	1.000	0.754
TS4	1.000	0.605
TS5	1.000	0.728
TS6	1.000	0.674
TS7	1.000	0.744
TS8	1.000	0.643
TS9	1.000	0.657
TM1	1.000	0.619
TM2	1.000	0.730
TM3	1.000	0.608
TM4	1.000	0.691
TM5	1.000	0.788
TM6	1.000	0.740
TE1	1.000	0.727
TE2	1.000	0.552
TE3	1.000	0.674
TE4	1.000	0.547
TE5	1.000	0.717
TE6	1.000	0.718
TE7	1.000	0.765
TE8	1.000	0.595

Table 7 summarises the principal-component results. Together, the three retained factors account for 67.80 % of the total variance. It is comfortably above the 60 % benchmark often cited for acceptable construct coverage. Before rotation, Factor 1 contributed 26.51 % of variance (eigenvalue = 6.097), Factor 2 contributed 22.69 % (eigenvalue = 5.219) and Factor 3 contributed 18.60 % (eigenvalue = 4.278). Varimax rotation redistributed, but did not reduce, this cumulative variance. It makes the factor pattern clearer for interpretation while preserving each item's communality. Because the three-factor solution captures more than two-thirds of the original information, the loss of explanatory power is minimal and the underlying structure can be considered both parsimonious and robust.

Table 7
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.962	34.619	34.619	7.962	34.619	34.619	6.097	26.508	26.508
2	4.672	20.315	54.934	4.672	20.315	54.934	5.219	22.689	49.198
3	2.959	12.866	67.800	2.959	12.866	67.800	4.278	18.602	67.800
4	0.931	4.046	71.845						
5	0.761	3.308	75.153						
6	0.658	2.862	78.014						
7	0.639	2.780	80.794						
8	0.613	2.666	83.460						
9	0.513	2.233	85.692						
10	0.501	2.178	87.871						
11	0.445	1.936	89.806						
12	0.385	1.675	91.482						
13	0.340	1.480	92.962						
14	0.310	1.346	94.308						
15	0.241	1.049	95.357						
16	0.219	0.950	96.307						
17	0.188	0.816	97.123						
18	0.181	0.786	97.908						
19	0.163	0.710	98.619						
20	0.126	0.548	99.166						
21	0.078	0.338	99.505						
22	0.063	0.273	99.778						
23	0.051	0.222	100.000						

vertical axis displays their eigenvalues. The plot shows a steep decline after the first factor whose eigenvalue is markedly higher than the others. It indicates that this factor explains the largest share of total variance. Eigenvalues for the second and third factors also remain above the conventional threshold of 1.00, but from the fourth factor onward the curve flattens and eigenvalues cluster at low values. It suggests minimal additional explanatory power. The pronounced "elbow" at the third factor therefore supports retaining three factors for subsequent analysis.

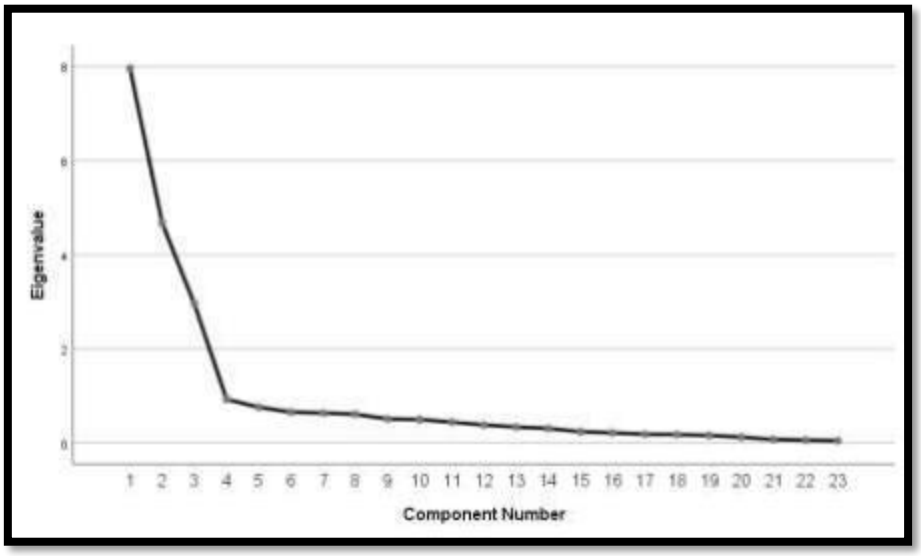

Figure 1 Scree Plot

Varimax (orthogonal) rotation clarified the factor solution after five iterations (refer to Table 8). Primary loadings for all 23 items ranged from 0.704 to 0.872. They comfortably exceed the 0.50 benchmark, while secondary loadings remained below 0.40. They eliminate concerns about cross-loading. Items grouped cleanly into three interpretable dimensions: Teaching Skill (Factor 1), Teaching Evaluation (Factor 2) and Teaching Method (Factor 3). The

combination of strong primary loadings and minimal cross-loadings demonstrates that the questionnaire possesses solid structural validity and that each factor is both distinct and theoretically coherent.

Table 8
Rotated Component Matrix

Item	Component		
	1	2	3
TS1	0.752	0.169	0.154
TS2	0.821	0.113	0.120
TS3	0.864	0.082	-0.026
TS4	0.772	0.040	0.080
TS5	0.783	0.100	0.324
TS6	0.790	0.172	0.144
TS7	0.815	0.151	0.239
TS8	0.776	0.077	0.188
TS9	0.795	0.046	0.149
TM1	0.240	0.013	0.749
TM2	0.074	0.105	0.845
TM3	0.073	-0.021	0.776
TM4	0.141	-0.073	0.816
TM5	0.211	0.028	0.862
TM6	0.256	0.027	0.821
TE1	-0.020	0.850	0.057
TE2	0.219	0.704	0.090
TE3	0.201	0.789	-0.107
TE4	0.174	0.717	-0.058
TE5	-0.073	0.844	-0.021
TE6	0.221	0.817	0.028
TE7	0.056	0.872	0.027
TE8	0.068	0.765	0.079

Discussion

This pilot study set out to develop and preliminarily validate a questionnaire that captures preschool teachers' music-teaching ability through the interlocking lenses of teaching skill, teaching method and teaching evaluation. The statistical evidence demonstrates that the instrument makes a credible step toward filling the psychometric gap identified in earlier reviews (Bautista et al., 2024; Bengochea & Sembiante, 2023).

All three sub-scales exceeded the 0.90 threshold for Cronbach's α, with negligible α-shifts when items were deleted. Comparable music-education instruments rarely achieve such stability. Fernández Amat et al. (2024) reported α values between 0.78 and 0.88 for their game-based competency scale, while Jiang et al. (2024) obtained a single-factor α of 0.84 in a creative-ability measure for children. Our alphas of 0.938 (TS), 0.909 (TM) and 0.921 (TE) therefore signal a tighter conceptual fit among items. One plausible explanation is the deliberate alignment with Shulman (1986) PCK model, which discourages mixing attitudinal and behavioural statements and thus minimises construct contamination. Although high

alpha can sometimes suggest redundancy, the almost flat α-change across item deletions (< 0.03 in all cases) indicates that items contribute unique variance rather than repetitive wording. It is an interpretation reinforced by the communalities that exceed 0.55 for every statement.

Expert ratings produced an S-CVI/Ave of 0.97. It is comfortably above the 0.90 benchmark recommended by Polit and Beck (2006). The fact that only two items fell below the I-CVI cut-off of 0.75 illustrates both the overall strength of the item pool and the sensitivity of the CVI procedure for pinpointing ambiguous phrasing. Re-wording TE3 and TE4 to clarify "mutual" and "self" evaluation reflects a responsive and evidence-based revision process advocated in scale-development primers (Boateng et al., 2018). Past music-education questionnaires often gloss over such micro-edits. It reports global CVIs but leaving low-scoring items unchanged. By contrast, this study demonstrates that content validity is not a static checkpoint but an iterative dialogue with subject-matter experts.

The Kaiser–Meyer–Olkin value of 0.76 and Bartlett's $\chi^2 = 845.44$ ($p < 0.001$) confirmed factorability. Principal-component extraction followed by varimax rotation yielded a parsimonious three-factor solution explaining 67.8% of total variance. They are well above the 60% conventionally viewed as satisfactory in educational measurement (Putnick & Bornstein, 2016). Prior tools such as the Music Teacher Self-Efficacy Scale (Nikolić, 2019) have typically produced unidimensional or two-factor structures, partly because they collapse teaching behaviours and affective beliefs into the same items. Our factor pattern maps neatly onto PCK, namely instructional skills (TS), evaluative practices (TE) and methodological choices (TM). The clean separation with primary loadings ≥ 0.70 and cross-loadings < 0.40 improves interpretability and enables practitioners to diagnose specific professional-development needs rather than receiving a single undifferentiated score.

The results align with earlier calls for more nuanced teacher-competence measures (Wong et al., 2024). Like prior studies, we observed that competence is multifaceted. Unlike many predecessors, we empirically verified that these facets can be disentangled without sacrificing reliability. Furthermore, the instrument's CVIs echo the high expert agreement recorded by Boateng et al. (2018) in health-behaviour scales. It suggests that rigorous cross-disciplinary procedures can be fruitfully imported into arts-education contexts.

Where our findings diverge from previous literature is in the strength of the evaluation dimension (TE). Most earlier tools either omit formal evaluation practices (Jiang et al., 2024) or subsume them under general teaching skill. Here, evaluation emerged as a distinct and high-loading factor. It implies that preschool music teachers differentiate between performing a skill and judging its quality. It is a distinction often underestimated in professional-development programmes. By spotlighting evaluation as a separate construct, this current study extends the PCK framework into reflective assessment territory. It aligns with contemporary discourses on formative feedback and learner agency (Bautista et al., 2024).

Although the psychometric indices are encouraging, we are mindful of two caveats. First, the sample is geographically confined to Liaocheng City. Cultural nuances in music pedagogy may limit transferability. Future research should conduct multi-regional invariance

testing to ensure that item meanings hold across linguistic and curricular contexts (Putnick & Bornstein, 2016). Second, Cronbach's α, while widely cited, assumes tau-equivalence and may overestimate reliability in multidimensional scales. Subsequent confirmatory factor analysis (CFA) should report McDonald's ω and composite reliability to guard against this bias.

Conclusion

The instrument operationalises PCK in a way that is directly measurable. Thereby, it translates a theoretically rich but operationally vague construct into actionable metrics. Administrators can now generate diagnostic profiles that pinpoint whether a teacher's developmental needs lie in technical skill, methodological diversity or evaluative acumen. Such granular data have the potential to inform "precision" professional development. It is an approach increasingly advocated in teacher-education scholarship. At the research level, the three-factor structure offers a scaffold for hypothesis-testing. For example, future studies could examine whether gains in teaching skill precede improvements in evaluative practices or vice versa.

By integrating rigorous content-validity procedures, robust reliability estimates and a theoretically anchored factor structure, this present study addresses all three gaps flagged in recent reviews, such as holistic coverage, multi-method validation and transparent pilot documentation (Bengochea & Sembiante, 2023). Moreover, the instrument links teacher competence to observable child outcomes via a stable performance test, paving the way for longitudinal mediation models that trace how specific teaching behaviours translate into measurable learning gains. It is an empirical chain rarely completed in earlier work (Degé & Frischen, 2022).

The pilot's modest size (n = 50 teachers) restricts statistical power for multi-group comparisons and the voluntary sampling may favour teachers already motivated by music. Larger and stratified samples will be essential for CFA and for testing measurement invariance across experience levels and school types. Additionally, integrating qualitative follow-ups, such as classroom observations, could enrich the interpretability of questionnaire scores. It addresses critiques that self-report tools capture intent more than enacted practice (Bautista et al., 2024).

This study furnishes preliminary but compelling evidence that the newly developed questionnaire is both psychometrically robust and theoretically sound. Through its tri-dimensional architecture, strong reliability and demonstrated content and construct validity, the tool stands to advance both research and practice in early-childhood music education. Crucially, it transforms PCK from an abstract ideal into a set of measurable behaviours, offering a clearer pathway for professional growth and empirical inquiry.

This study adds to knowledge in two main ways. First, on the theoretical side, it builds on Shulman's idea of PCK. Past work often mixed up or ignored the skill of evaluating how well children make music. Our findings show that evaluation is a separate and reliable part of early-childhood music teaching. This insight widens the PCK model to include teachers' ability to judge musical quality, not just to teach music. Second, on the contextual side, the questionnaire we developed is the first tool carefully tested with preschool teachers in Liaocheng City, China. It is an area that earlier studies rarely covered. By working with local

experts and noting local conditions, we created a measure that others can adapt and compare across regions. Together, these contributions offer researchers a clear, culture-sensitive way to study how specific teaching behaviours in music affect young children's learning over time.

References

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, *10*, 2714. https://doi.org/10.3389/fpsyg.2019.02714

Arasomwan, A. D., & Daries, G. (2025). Music and rhymes as mechanisms for ECCE learners' socio-emotional intelligence development. *E-Journal of Humanities, Arts and Social Sciences*, *6*(2), 44-57. https://doi.org/10.38159/ehass.2025614

Bautista, A., Yeung, J., Mclaren, M. L., & Ilari, B. (2024). Music in early childhood teacher education: Raising awareness of a worrisome reality and proposing strategies to move forward. *Arts Education Policy Review*, *125*(3), 139-149. https://doi.org/10.1080/10632913.2022.2043969

Bengochea, A., & Sembiante, S. F. (2023). A review of the methodological characteristics of vocabulary interventions for emergent bilinguals in preschool to sixth grade. *Review of Education*, *11*(1), e3386. https://doi.org/10.1002/rev3.3386

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Cai, Y., Kang, D., & Xu, X. (2025). Boosting executive function in children aged 3–12 through musical training: A three-level meta-analysis. *Frontiers in Psychology*, *16*, 1659927. https://doi.org/10.3389/fpsyg.2025.1659927

Chen, H., Liu, F., Pang, L., Liu, F., Fang, T., Wen, Y., Chen, S., Xie, Z., Zhang, X., Zhao, Y., & Gu, X. (2020). Are you tired of working amid the pandemic? The role of professional identity and job satisfaction against job burnout. *International Journal of Environmental Research and Public Health*, *17*(24), 9188. https://doi.org/10.3390/ijerph17249188

Degé, F., & Frischen, U. (2022). The impact of music training on executive functions in childhood—a systematic review. *Zeitschrift für Erziehungswissenschaft*, *25*(3), 579-602. https://doi.org/10.1007/s11618-022-01102-2

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. SAGE Publications.

Fernández Amat, C., Zarza-Alzugaray, F. J., & Del Barrio Aranda, L. (2024). Design and validation of a scale for the assessment of educational competencies in traditional musical games. *Humanities and Social Sciences Communications*, *11*(1), 1-13. https://doi.org/10.1057/s41599-024-03340-7

Grgic, J., Lazinica, B., Schoenfeld, B. J., & Pedisic, Z. (2020). Test–retest reliability of the one-repetition maximum (1RM) strength assessment: A systematic review. *Sports Medicine - Open*, *6*(1), 1-16. https://doi.org/10.1186/s40798-020-00260-z

Jiang, H., Cheong, K. W., & Tan, W. H. (2024). Development and validation of a measure assessing children's creative practice ability in music. *Thinking Skills and Creativity*, *51*, 101446. https://doi.org/10.1016/j.tsc.2023.101446

Lowe, N. K. (2019). What is a pilot study? *Journal of Obstetric, Gynecologic & Neonatal Nursing*, *48*(2), 117-118. https://doi.org/10.1016/j.jogn.2019.01.005

Nikolić, L. (2019). Attitudes of students of teacher studies towards music education. *Metodički ogledi*, *25*(2), 111-136. https://doi.org/10.21464/mo.25.2.6

Platz, F., Kopiez, R., Lehmann, A. C., & Wolf, A. (2022). Measuring Audiation or tonal memory? Evaluation of the discriminant validity of Edwin E. Gordon's "Advanced measures of music Audiation". *Music & Science*, *5*, 20592043221105270. https://doi.org/10.1177/20592043221105270

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, *29*(5), 489-497. https://doi.org/10.1002/nur.20147

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203-212. https://doi.org/10.1016/j.jrp.2006.02.001

Shi, J., Mo, X., & Sun, Z. (2012). Content validity index in scale development. *Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical sciences*, *37*(2), 152-155.

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4-11. https://doi.org/10.12691/ajams-9-1-2

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4. https://doi.org/10.2307/1175860

Wong, J., Bautista, A., Ho, Y. L., & Kong, S. H. (2024). Preschool teachers' music-specific professional development preferences: Does teaching experience matter? *Research Studies in Music Education*, *46*(1), 80-97. https://doi.org/10.1177/1321103x221139992

You, T., He, H., & Yue, Y. (2025). Teacher support and pre-service preschool teachers' piano skill: The chain mediation effects of music self-efficacy and learning engagement. *Behavioral Sciences*, *15*(4), 484. https://doi.org/10.3390/bs15040484