

융합형 수오나 성취도 검사의 심리측정학적 연구*:

중국 고등교육의 다문화 음악 교육 강화를 위한

CTT와 IRT의 통합적 적용

A Psychometric Study of a Fusion-Based Suona Achievement Test: Integrating CTT and IRT to Enhance Multicultural Music Pedagogy in Chinese Higher Education

장페이** · 사브리 샤피잔***

Zhang Fei  · Shafizan Sabri 

Abstract This study examines the psychometric validity and reliability of an achievement test designed for a fusion-based suona technique curriculum in Chinese universities. Grounded in constructivist theory, Vygotsky's zone of proximal development, and multiple intelligences theory, the assessment instrument was developed to measure diverse competencies ranging from theoretical knowledge to performance application. Utilizing a post-test-only design to collect validation data from an experimental cohort, the study employed a dual-framework psychometric analysis. Data were obtained from a 40-item multiple-choice test and performance-based tasks. Psychometric evaluation revealed a generally balanced distribution of item difficulty; however, Classical Test Theory (CTT) analysis indicated that while 25% of items demonstrated excellent discrimination, 50% required revision to improve distractor efficiency. Complementary Item Response Theory (IRT) analysis, using the Rasch rating scale model, demonstrated satisfactory internal consistency (Person Reliability = .776) and acceptable model-data fit as well as local independence (MADaQ3 $p = .182$). Wright Map analysis confirmed appropriate targeting for average-ability students but highlighted a need for additional items to adequately measure high-ability levels. These findings provide empirical evidence for the instrument's utility and offer a framework for refining assessment tools in multicultural and competency-based music pedagogy.

Key words: fusion-based music education, multicultural music pedagogy, psychometric evaluation, Classical Test Theory (CTT), Item Response Theory (IRT)

초록 본 연구는 중국 대학의 융합형 수오나(Suona) 교육과정을 위해 개발된 성취도 검사의 심리측정학적 타당도와 신뢰도를 분석하였다. 구성주의, 근접발달영역, 다중지능이론을 토대로 이론적 지식부터 실기 적용까지 다양한 역량을 측정할 수 있도록 평가 도구를 설계하였다. 융합 교육과정에 참여한 실험 집단으로부터 타당화 데이터를 수집하기 위해 사후 검사 설계를 적용하였으며, 40문항의 선다형 검사와 실기 평가 데이터를 바탕으로 이중 분석 프레임워크를 수행하였다. 분석 결과, 문항 난이도는 전반적으로 균형적이었으나 고전검사이론(CTT) 분석 결과 25%의 문항은 우수한 변별력을 보인 반면, 50%의 문항은 수정이 필요한 것으로 나타났다. 보완적으로 수행된 문항반응이론(IRT)의 라쉬(Rasch) 평정 척도 모형 분석에서는 피험자 신뢰도(.776)와 모형 적합도(MADaQ3 $p=.182$)가 양호하게 나타났다. 라이트맵(Wright map) 분석 결과 평균 능력의 학습자에 대한 적합도는 확인되었으나, 고능력 학습자를 변별하기 위한 문항의 보완 필요성이 드러났다. 결론적으로 본 연구는 개발된 평가 도구의 유용성에 대한 실증적 증거를 제공하며, 다문화 및 역량 중심 음악 교육에서 평가 도구를 개선하고 정교화하기 위한 기초 자료를 제시한다.

주제어: 융합 기반 음악 교육, 다문화 음악 교육, 심리측정학적 평가, 고전검사이론(CTT), 문항반응이론(IRT)

* This manuscript is derived from the first author's doctoral research.

** First author & Corresponding author, E-mail: zhangfei_feizhang@163.com

Ph.D. Student, Universiti Pendidikan Sultan Idris, Faculty of Music and Performing Arts, Tanjung Malim, Perak, Malaysia

*** Co-author, E-mail: shafizan@fmsp.ups.edu.my

Senior Lecturer, Universiti Pendidikan Sultan Idris, Faculty of Music and Performing Arts, Tanjung Malim, Perak, Malaysia

I . Introduction

The suona, a traditional Chinese double-reed woodwind instrument, has long held a central place in China's musical heritage, celebrated for its distinctive timbre, expressive range, and cultural symbolism (Cai, 2024; Chao & Nicolas, 2024). Originating in northern China and later spreading nationwide, the suona developed diverse regional styles and performance techniques, becoming indispensable in folk ensembles, ceremonial music, and local opera traditions (Xin & Taipanich, 2025). Historically, suona pedagogy relied on oral transmission and imitation, emphasizing the preservation of repertoire, technical mastery, and cultural authenticity (Cai, 2024; Xin & Taipanich, 2025).

In recent decades, higher music education in China has undergone a significant transformation, influenced by globalization, technological advancements, and increased intercultural exchange. This shift has prompted universities to explore fusion-based pedagogies that integrate Western harmonic theory, counterpoint, and modern performance techniques into traditional training (Bartleet et al., 2020; Bates et al., 2021; Campbell, 2018). Such integration expands students' technical and stylistic range while equipping them with the intercultural competencies required for participation in the global music profession (Cai, 2024). This approach aligns with contemporary views of traditional performing arts as essential conduits for multicultural music education, providing unique teaching methods that can bridge cultural divides and enhance global musical literacy (Kim, 2025). By contextualizing the suona within these broader pedagogical shifts, the development of a fusion-based curriculum moves beyond localized heritage toward a more inclusive, cross-cultural instructional model.

Research in multicultural and fusion music education consistently demonstrates benefits such as enhanced learner engagement, creativity, and cultural awareness (Bates et al., 2021). In the Chinese context, blending Western and Chinese traditions fosters innovation while strengthening cultural self-confidence (Xin & Taipanich, 2025). Nonetheless, implementing fusion pedagogy presents challenges, including the scarcity of specialized instructional resources and concerns about diluting traditional elements (Cai, 2024). Addressing these challenges requires structured curriculum design grounded in robust learning theories. The present study's curriculum is informed by constructivist theory, Vygotsky's zone of proximal development, and multiple intelligences theory, ensuring scaffolded, experiential, and multimodal learning experiences that both preserve traditional competencies and promote cross-cultural musicianship.

The successful implementation of such an innovative curriculum depends on the availability

of psychometrically sound assessment. In music education, valid and reliable measurement instruments are essential for accurately evaluating student achievement. Recent studies in the Korean context have emphasized the necessity of designing appropriate assessment plans to meet diverse learner needs (Won, 2022), a principle that parallels this study's aim to validate a fusion-based achievement test. Furthermore, the importance of ensuring that traditional music content is appropriately structured and aligned within the curriculum has been highlighted by Lee and Park (2024), who noted that materials must be carefully analyzed to meet specific instructional goals. This necessity of rigorous validation is further supported by Kwon et al. (2025), who highlighted the importance of examining validity and reliability when adapting assessment tools for specific cultural contexts. Similarly, the crucial role of performance assessment in predicting academic success in higher music education, as noted by Utomo et al. (2025), underscores the need for instruments that accurately reflect diverse learner competencies. To achieve such validation, specific psychometric frameworks are required. Classical test theory provides interpretable metrics such as item difficulty and discrimination, while item response theory yields sample-independent estimates, fit statistics, and detailed person-item targeting (Bond & Fox, 2015; Boone & Staver, 2020). These frameworks are crucial for detecting bias and establishing measurement invariance across diverse learner groups (Fischer & Karl, 2019; Kūkea Shultz & Englert, 2021). Without rigorous psychometric validation, assessment tools risk producing misleading comparisons or disadvantaging certain cultural groups (Alavi et al., 2025; Sosa Paredes & Andersson, 2025).

Although both classical test theory and item response theory have been applied in music education, they are rarely combined to assess traditional Chinese instruments. While Rasch models have been used for listening assessments (Chow et al., 2024) and performance evaluations (Ooi & Engelhard, 2019; Springer & Bradley, 2017), published studies that explicitly integrate classical test theory and item response theory within a single validation study are uncommon. To date, no peer-reviewed dual-framework validations exist for achievement tests targeting traditional instruments such as the suona. This omission hinders evidence-based curriculum design and limits the field's capacity to objectively evaluate instructional quality.

Therefore, the primary purpose of this study is to psychometrically validate a newly developed achievement test designed for the fusion-based suona curriculum. Specifically, the study aims to: (1) evaluate the item properties and reliability of the assessment using a dual-framework approach combining classical test theory and item response theory; (2) assess the model-data fit and person-item targeting to ensure the instrument accurately reflects learner ability; and (3)

identify items requiring revision to enhance measurement validity. By establishing the psychometric robustness of this instrument, this research seeks to provide empirical evidence that supports the refinement of multicultural music pedagogy in Chinese higher education.

II. Literature Review

1. Pedagogical and theoretical foundations for fusion-based suona instruction

A growing body of scholarship highlights the need for assessment practices in music education that are both psychometrically robust and culturally responsive, particularly in contexts where traditional instruments intersect with modern pedagogical innovations (Alavi et al., 2025; Chow et al., 2024; Fischer & Karl, 2019; Harrison et al., 2017; Kūkea et al., 2021; Ooi & Engelhard, 2019). In Chinese higher education, suona instruction historically emphasized the preservation of traditional techniques, repertoire, and stylistic authenticity through oral transmission and teacher-centered approaches (Xin & Taipanich, 2025). While these approaches remain valuable for cultural continuity, they provide limited opportunities for students to develop broader musicianship skills or engage in cross-cultural musical dialogue (Xin & Taipanich, 2025). Contemporary curriculum reform thus calls for an integrative model that combines the technical rigor of traditional suona training with Western theoretical constructs, fostering both cultural rootedness and global musical literacy (Bartleet et al., 2020; Bates et al., 2021; Cai, 2024).

The development of instructional materials and assessments for such a fusion-based curriculum requires a strong theoretical foundation. Constructivist theory provides a guiding framework by emphasizing that learners actively construct knowledge through meaningful interaction with content and context (Bruner, 1996). In music teacher education, Salvador et al., (2020) demonstrated how constructivist grounded theory can illuminate transformative learning processes when educators confront issues of cultural and social justice, underscoring the value of reflective practice and contextualized knowledge construction (Mygdanis, 2025). In suona pedagogy, this translates to experiential, hands-on practice in which students learn not merely by replication but through exploration, improvisation, and integration of new musical concepts.

Vygotsky's zone of proximal development further supports this approach by offering a mechanism for scaffolding enabling students to progress beyond their current skill level with guided support from instructors or peers (Chen, 2024; Vygotsky, 1978). This is particularly

relevant when introducing Western harmonic theory and counterpoint, as scaffolded instruction can help students gradually integrate these concepts into their performance practice.

Complementing these theories, Gardner's multiple intelligences theory recognizes that musical learning is multidimensional, encompassing not only technical skill but also auditory, kinesthetic, spatial, and interpersonal dimensions (Gardner, 2011, 2020; Shearer, 2020). Recent scholarship highlights both the potential and challenges of applying multiple intelligence theory within Chinese music education, noting that while multiple intelligence theory informed approaches can foster creativity and learner engagement, they require careful contextual adaptation to align with local teaching traditions (Xiong, 2025). Applying this perspective ensures that assessments capture the full breadth of musical understanding, from aural recognition and theoretical knowledge to expressive performance and creative application. Together, these three theoretical perspectives establish a comprehensive pedagogical and assessment framework that supports cognitive, technical, and cultural development within a fusion-based suona curriculum.

Building upon the pedagogical and theoretical foundations outlined above, the development of a fusion-based suona curriculum necessitates equally rigorous methods of evaluation. Theories such as constructivism, zone of proximal development, and multiple intelligences emphasize that assessment must not only measure technical proficiency but also capture growth, creativity, and holistic musicianship. To achieve this, psychometric frameworks capable of providing reliable, valid, and culturally sensitive measures are required. Classical test theory and item response theory, along with Rasch models, offer complementary analytic strategies that allow researchers to evaluate item performance, detect bias, and ensure alignment between curricular goals and learning outcomes. The following section reviews these approaches and highlights their applications in music education, establishing the methodological rationale for the dual-framework validation employed in the present study.

2. Psychometric approaches in music education: classical test theory, item response theory, and Rasch models

Contemporary assessment practice in education distinguishes two complementary psychometric traditions: Classical test theory and item response theory. Classical test theory remains widely used in music education because it is conceptually simple and provides straightforward item-level indices (e.g., difficulty, discrimination, distractor analysis) that are useful in early-stage test development and classroom assessment. However, classical test theory's sample dependence of

item statistics and limitations in modeling item functioning across ability levels constrain its utility for large-scale, high-stakes, or cross-cultural applications.

Item response theory and Rasch-family models address many of these limitations by locating items and persons on a common latent metric and by providing sample-independent item parameter estimates, information functions, and goodness-of-fit diagnostics. These properties make item response theory especially attractive for advanced measurement tasks in music education item banking, computerized adaptive testing, detection of differential item functioning, and fine-grained person-item targeting. Comprehensive reviews note the theoretical and practical advantages of IRT for educational measurement and recommend it for contexts requiring precision and comparability across samples (Boone & Staver, 2020).

In music-specific applications, item response theory and Rasch methods have yielded demonstrable benefits. P. Harrison et al. applied item response theory, computerized adaptive testing, and automatic item generation to melodic-discrimination testing, demonstrating improved measurement efficiency and the feasibility of adaptive testing for musical perception tasks (Harrison et al., 2017). More recently, Chow et al. used Rasch testlet models in large-scale music listening examinations to examine multidimensionality and local item dependency, showing how Rasch extensions yield richer diagnostic information for listening assessment design and score interpretation (Chow et al., 2024).

Performance assessment in music introduces additional measurement complexity because rater judgments (severity, centrality, halo effects) can significantly influence scores. Many-facets Rasch measurement is widely used to model and adjust for rater effects, thereby improving the fairness and interpretability of performance ratings. Empirical applications in higher-education performance contexts demonstrate many-facets Rasch measurement's utility for examining rater severity, scale functioning, and interactions among raters, tasks, and performer subgroups (Ooi & Engelhard, 2019; Springer & Bradley, 2017).

Despite the strengths of each approach, there is growing recognition that combined classical test theory/item response theory strategies can be psychometrically advantageous during instrument development: classical test theory provides accessible diagnostics (item difficulty, distractor functioning) while item response theory/Rasch offers sample-independent calibration and advanced modeling. Methodological literature and applied studies across education illustrate practical workflows that integrate both paradigms for robust test construction and validation (Bond & Fox, 2015; Boone & Staver, 2020).

However, within music education for traditional Chinese instruments, there remains a critical

gap. While sophisticated item response theory/Rasch applications exist for listening tests, melodic-discrimination tasks, and rater modeling, explicit peer-reviewed studies that combine classical test theory and item response theory within a single validation study for achievement tests of instruments such as the suona or erhu are lacking. Much existing research prioritizes performance analysis or computational/audio-recognition approaches rather than psychometric test development and validation underscoring the importance of the present study's dual-framework approach.

Given this context, the present study adopts a dual-framework psychometric approach to maximize the strengths of both traditions. Classical Test Theory was employed to provide accessible item-level diagnostics such as difficulty, discrimination, and distractor efficiency, allowing for targeted revision of underperforming items. In parallel, item response theory using a Rasch rating scale model was implemented to calibrate items and persons on a common latent continuum, evaluate model-data fit, and generate person-item maps that reveal the alignment of test difficulty with learner ability. This combined approach ensures a more robust, culturally sensitive validation of the fusion-based suona achievement test and provides actionable evidence to refine both the instrument and the instructional materials.

III. Methodology

1. Research design

This study adopted a pilot quasi-experimental design to evaluate the feasibility and psychometric soundness of fusion-based instructional materials integrating western and traditional Chinese music in suona education. The pilot framework emphasized early-stage validation of teaching strategies and assessment tools prior to large-scale implementation, enabling refinement of instructional components and measurement instruments within a controlled academic setting.

Focusing on a pilot study was methodologically significant for two reasons. First, it ensured that the fusion-based instructional materials, being newly developed, were tested and optimized in a limited yet authentic environment. Second, it allowed the achievement test to undergo rigorous item-level validation, enhancing its reliability and cultural relevance. Similar pilot studies in music education (Uludag & Satir, 2025; Wright et al., 2022) have demonstrated how small-scale trials can refine pedagogical design and validate assessment instruments effectively before broader deployment.

2. Participants

A total of 15 diploma-level students from the Shandong Arts Institute participated in the pilot. Participants were selected based on inclusion criteria related to prior musical experience, interest in traditional instruments, and availability during the study period. The pilot program consisted of eight instructional sections, each containing two lessons of two hours per session, held twice weekly on Tuesdays and Fridays from 09:00 to 11:00 a.m. Suona instruments were provided by the researcher to ensure accessibility and consistency across participants.

3. Instructional procedures

The instructional pilot was organized into eight modules, with each module encompassing two two-hour sessions. The program was delivered twice weekly on Tuesdays and Fridays (09:00-11:00) during the data collection period from January 21 to February 21, 2025. To maintain technical consistency and ensure accessibility, the researcher provided standardized suona instruments for all participants.

Instructional content utilized the fusion-based materials specifically developed for this study, which synthesize Western harmonic concepts, phrasing, and counterpoint with traditional Chinese suona pedagogy. The curriculum was designed to cultivate four primary competencies: technical proficiency, interpretative ability, musical creativity, and cultural understanding. The pedagogical implementation was grounded in a theoretical framework of constructivism, Vygotsky's zone of proximal development, and multiple intelligences theory. This approach facilitated a scaffolded, multidimensional learning environment, allowing for the systematic collection of performance data required to validate the suona achievement test.

4. Ethical considerations

Before initiating the study, the researcher obtained formal approval from the Human Research Ethics Committee of Sultan Idris Education University (reference number: 2024-0571-01). This approval authorized the research to proceed from 28 November 2024 to 27 November 2025.

Prior to participation, all students signed informed consent forms confirming their understanding of the study's objectives, voluntary participation, and assurance of confidentiality and anonymity. Approval was also obtained from course instructors to ensure that the study did not interfere

with regular academic responsibilities. All instructional sessions were conducted outside standard teaching hours to minimize disruption and uphold ethical integrity throughout the research process.

5. Research instruments

Data collection employed a triangulated set of instruments to capture both cognitive and practical learning outcomes within the fusion-based suona curriculum.

1) Achievement test

A 47-item test was designed to evaluate theoretical knowledge and stylistic understanding. The instrument comprised 40 multiple-choice items (Part A) and seven short-answer/performance-based items scored with partial credit (Parts B and C). Classical test theory (CTT) was applied to Part A for item difficulty, discrimination, and distractor analysis, while item response theory (IRT) using the rating scale model was employed for Parts B and C to examine person-item fit, reliability, and scale functioning.

2) Satisfaction questionnaire

A structured Likert-scale questionnaire was administered to gauge participants' perceptions of the instructional content quality, utility of the materials, and face validity of the assessment tasks. This self-report data provided complementary evidence of the curriculum's acceptability and perceived pedagogical value.

3) Individual and ensemble performance assessment

To evaluate practical musicianship, participants completed both solo and group performance tasks designed to reflect the dual emphases of technical precision and intercultural integration. Assessment criteria included tone production, technical accuracy, stylistic interpretation, ensemble cohesion, and responsiveness to both Western and traditional Chinese musical elements. Performance ratings were conducted by two independent expert assessors. Inter-rater reliability and potential severity/leniency effects were examined using the many-facets Rasch measurement model, an established method for improving fairness and interpretability in performance evaluation (Ooi

& Engelhard, 2019; Springer & Bradley, 2017). Including both individual and ensemble formats ensured that assessments captured personal technical achievement as well as collaborative, context-sensitive musicianship, skills highly valued in contemporary music education (Sim & Cheng, 2024; Uludag & Satir, 2025).

6. Data collection and analysis

At the conclusion of the pilot study, participants completed the achievement test and satisfaction questionnaire, while their individual and ensemble performance tasks were recorded and evaluated by two trained adjudicators using standardized rubrics.

The data analysis employed a dual-framework psychometric strategy. Classical test theory was utilized to evaluate the multiple-choice items (Part A) by calculating item difficulty, discrimination indices, and distractor efficiency. Complementing this, item response theory, specifically a Rasch rating scale model, was applied to the short-answer and performance tasks (Parts B and C). These analyses encompassed the estimation of person and item parameters, evaluation of model-data fit, and the construction of a Wright map to assess the alignment of item difficulty with learner ability.

The combined use of classical test theory and item response theory provided a comprehensive psychometric evaluation. Classical test theory yielded straightforward item-level diagnostics useful for pinpointing items with poor discrimination or ineffective distractors, while IRT contributed sample-independent parameter estimates and deeper insights into person-item interactions. Together, these complementary approaches established an evidence-based foundation for refining both the instructional materials and the assessment instrument, ensuring reliability, validity, and cultural responsiveness in measuring learning outcomes.

Focusing on a pilot study was methodologically significant for two reasons. First, it ensured that the fusion-based instructional materials being newly developed were tested and refined in a controlled setting before wider application. Second, it allowed the achievement test to undergo rigorous item-level validation, thereby strengthening its reliability and cultural relevance. Pilot-only designs are particularly valuable in educational intervention research where curriculum innovation intersects with psychometric evaluation, as they provide critical insights into feasibility, instructional alignment, and measurement robustness while minimizing risks to broader student populations. Evidence from prior music education research supports this approach: for example, pilot studies integrating mobile technologies into music theory instruction (Uludag & Satir, 2025),

teacher training interventions in music-based language programs (Sim & Cheng, 2024), and popular music media literacy (Wright et al., 2022) have demonstrated how small-scale trials allow researchers to assess feasibility, refine instructional strategies, and validate assessment tools before broader implementation. In this light, the present study's pilot design was essential to ensure that the fusion-based suona materials were both pedagogically appropriate and psychometrically sound.

IV. Results

1. Classical test theory analysis

The classical test theory analysis of the 40 multiple-choice items revealed variation in item discrimination and difficulty. Item discrimination indices showed that 25% of items fell into the excellent category ($> .40$), 25% were acceptable ($.20-.39$), and 50% were poor ($< .20$). This distribution highlights the need to revise or replace a substantial portion of items to strengthen their ability to differentiate between high- and low-performing students (<Table 1>).

<Table 1> Summary of item discrimination indices

Type	Range of D values	Number of items	Item numbers
Excellent	$D > .40$	10	A1, A2, A3, A5, A6, A8, A9, A15, A17, A27
Good	$.30 \leq D \leq .39$	0	-
Acceptable	$.20 \leq D \leq .29$	10	A4, A7, A10, A19, A21, A28, A32, A35, A37, A39
Poor	$D < .20$	20	A11, A12, A13, A14, A16, A18, A20, A22, A23, A24, A25, A26, A29, A30, A31, A33, A34, A36, A38, A40

As tabulated in <Table 2> majority of the items (52%) were classified as easy, while only 5% were difficult. The mean difficulty of .78 suggests that the test favored average to high-ability students but provided insufficient challenge for top performers. This reinforces the presence of a ceiling effect, indicating a need to balance item difficulty in future revisions to improve discrimination across the full ability spectrum.

<Table 2> Summary of item difficulty indices

Type	Range of difficulty (<i>p</i>) values	Number of items	Item numbers
Easy	$p \geq .80$	21	A1, A6, A11, A12, A13, A14, A16, A17, A18, A20, A22, A23, A25, A26, A30, A31, A33, A34, A36, A38, A40
Moderate	$.40 \leq p < .80$	17	A2, A3, A4, A5, A7, A8, A9, A19, A21, A24, A27, A28, A29, A32, A35, A37, A39
Difficult	$p < .40$	2	A10, A15

Analysis of distractor frequency patterns for top and low scorers revealed varying levels of effectiveness across items. Items were classified according to the number of functioning distractors options that were selected by at least 5% of low-performing students but avoided by top scorers.

<Table 3> Summary of distractor efficiency in the achievement test

Category	Criteria	Number of items	Item numbers	Interpretation
High distractor efficiency	≥ 3 functioning distractors	15	A1, A3, A4, A5, A7, A8, A10, A15, A24, A27, A29, A32, A35, A37, A39	All distractors attracted low scorers but few or no top scorers, indicating strong differentiation.
Moderate distractor efficiency	2 functioning distractors	14	A2, A6, A9, A11, A12, A13, A14, A16, A17, A19, A21, A28, A30, A31	Some distractors were effective, but one option was rarely chosen, suggesting partial redundancy.
Low distractor efficiency	≤ 1 functioning distractor	11	A18, A20, A22, A23, A25, A26, A33, A34, A36, A38, A40	Most items had non-functioning or implausible distractors, reducing diagnostic power and increasing ease.

Overall, 73% of items (high + moderate categories) contained at least two functioning distractors, demonstrating generally effective item construction. However, 27% of the items exhibited low distractor efficiency, often due to implausible or redundant options. This finding aligns with Chauhan et al. (2023), who established that the number of functioning distractors directly affects both item difficulty and discrimination. Consequently, revising or replacing weak distractors is essential to enhance the psychometric robustness and diagnostic accuracy of the achievement test.

2. Item response theory analysis: Rasch rating scale model

The item response theory analysis was conducted using the Rasch rating scale model to evaluate the seven short-answer and performance-based items (Parts B and C). The model provided sample-independent estimates and detailed insights into the functioning of the polytomous items, allowing for a deeper understanding of how students with varying abilities interacted with each task.

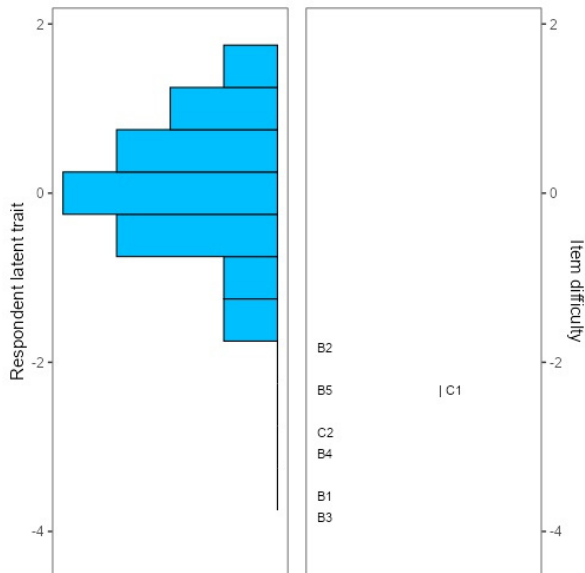
<Table 4> presents the summary of Rasch model fit statistics. The person reliability index (.776) indicates acceptable consistency across participants, while the item reliability index (.915) demonstrates a stable hierarchy of item difficulty. Both infit and outfit mean-square values fall within the acceptable range of 0.7-1.3, suggesting that response patterns conformed well to Rasch model expectations. The non-significant MADaQ3 value ($p = .182$) further confirms the absence of local item dependency.

<Table 4> Rasch model fit summary

Statistic	Value	Interpretation
Person reliability	.776	Acceptable reliability; consistent measurement across participants
Mean square infit	1.02	Within acceptable range (0.7-1.3), indicates good model fit
Mean square outfit	1.04	Acceptable; no major outlier effects detected
MADaQ3 p-value	.182	Non-significant, suggesting adequate local independence

The Rasch model demonstrated strong internal consistency and acceptable model-data fit. The person reliability coefficient (.776) reflected moderate stability across respondents, appropriate for a pilot-scale validation. Both infit and outfit mean square values (1.02 and 1.04, respectively) fell within the ideal range (0.7-1.3), confirming that item responses conformed to model expectations. The non-significant MADaQ3 statistic ($p = .182$) indicated no violations of local independence, reinforcing the construct validity of the assessment.

[Figure 1] presents the Wright map, illustrating the distribution of person abilities on the left and item difficulties on the right across a common logit scale. Each bar represents groups of participants, while labeled items indicate their relative difficulty levels. The alignment demonstrates that items were well-targeted for average learners but sparse at higher ability levels.



[Figure 1] Wright map of person-item distribution

A Wright map analysis further illustrated that the items were well-calibrated for average-ability students but offered limited challenge for advanced learners, aligning with the classic test theory findings. Together, the dual-framework approach of classical test theory and item response theory provided a comprehensive psychometric evaluation, confirming the test’s reliability while identifying targeted areas for refinement in future iterations.

Overall, the psychometric analyses provide compelling preliminary evidence that the fusion-based suona achievement test is both reliable and theoretically aligned with the instructional objectives of the pilot curriculum. The integration of classical test theory and Rasch-based item response theory allowed for a multidimensional validation, capturing both item-level performance and the broader construct of musical understanding. This dual-framework approach aligns with recent recommendations in music education and educational measurement emphasizing the need for culturally responsive, psychometrically sound assessment tools that balance traditional musicianship with contemporary pedagogical innovation (Boone & Staver, 2020; Fischer & Karl, 2019; Kūkea et al., 2021). While the results affirm the instrument’s overall measurement integrity, they also reveal the need for refining item difficulty and discrimination to better represent advanced learner proficiency, an observation consistent with prior validation studies in musical performance and listening assessment (Chow et al., 2024; Ooi & Engelhard, 2019). These findings

not only validate the feasibility of fusion-based assessment design but also establish a foundation for future large-scale implementations, informing pedagogical decisions and curriculum development in Chinese higher music education (Cai, 2024; Xin & Taipanich, 2025).

V. Discussion

The findings of this pilot study provide evidence that a fusion-based instructional design integrating Western and traditional Chinese musical elements can be both pedagogically and psychometrically viable in higher education suona instruction. The application of a dual-framework assessment approach combining classical test theory and Rasch-based item response theory enabled a robust evaluation of the instrument's validity and reliability while also illuminating areas requiring refinement. The moderate person reliability (.776) and strong item fit indices confirmed the internal coherence of the achievement test, supporting its suitability for measuring foundational competencies in fusion-based suona performance. The results of this study confirm that the suona achievement test is a generally reliable instrument. This alignment with rigorous psychometric standards resonates with recent calls in the Korean context for designing appropriate assessment plans to meet diverse learner needs (Won, 2022). By validating this fusion-based test, we address the critical need for empirically verified assessment tools in traditional music education.

From a theoretical perspective, the results align closely with Constructivist principles emphasizing active, context-driven learning (Bruner, 1996; Salvador et al., 2020). The iterative process of developing, piloting, and refining the instructional materials reflects a constructivist cycle of knowledge construction where learners engage in meaning-making through the interplay of new and prior musical experiences. By exposing students to both Western harmonic structures and traditional suona phrasing, the fusion-based curriculum encouraged learners to construct deeper conceptual and technical understanding, consistent with experiential and discovery-based learning paradigms.

Vygotsky's (1978) zone of proximal development provides further insight into how scaffolding supports student progression in a hybrid musical learning environment. The pilot results suggest that while the test effectively captured competencies among average-ability learners, additional high-difficulty items are needed to assess advanced performance within students' upper zone of proximal development boundaries. This finding highlights the importance of guided practice and

differentiated instruction in supporting learners as they internalize increasingly complex musical relationships between Chinese and Western systems (Chen, 2024). Through instructor mediation, students are able to advance from imitation-based traditional learning toward more autonomous creative interpretation, which represents an essential outcome of fusion-based pedagogy.

Multiple Intelligences Theory (Gardner, 2011; Shearer, 2020; Xiong, 2025) further reinforces the multidimensional nature of musical learning observed in this study. The suona performance context demands not only logical-mathematical and auditory intelligence but also kinesthetic coordination, spatial awareness, and interpersonal communication within ensemble settings. The fusion curriculum's emphasis on creative improvisation and interpretive performance provided opportunities for these diverse intelligences to manifest, thereby supporting a more holistic model of musical competence. However, the Rasch analysis indicated limited item coverage for high-ability learners, suggesting that future test revisions should include tasks assessing advanced integrative skills such as harmonic improvisation and cross-cultural stylistic adaptation.

Taken together, these results advance current understanding of assessment design in multicultural music education. Our findings align with recent methodological shifts in the field, moving beyond simple score reporting to in-depth instrument validation. Kwon et al. (2025) demonstrated the utility of preliminary validity studies in ensuring that assessment tools function correctly across cultural contexts, a principle mirrored in our dual-framework analysis of the fusion-based suona curriculum. Additionally, the positive correlation between performance skills and academic achievement noted by Utomo et al. (2025) supports our inclusion of both theoretical and performance-based components to provide a holistic measure of student musicianship. The integration of classical test theory and item response theory facilitated an evidence-based evaluation that transcends the limitations of single-framework analyses, echoing calls within psychometrics for culturally responsive and theoretically grounded testing models (Boone & Staver, 2020; Fischer & Karl, 2019; Kūkea et al., 2021). The findings underscore the pedagogical potential of fusion-based instruction not only to enhance technical proficiency but also to cultivate broader intercultural competence and creative musicianship.

Beyond the psychometric outcomes, this study demonstrates how theoretically informed design principles can bridge traditional and contemporary educational paradigms in Chinese music education. The suona, once a symbol of localized musical heritage, is recontextualized as a vehicle for cross-cultural learning, aligning with national efforts to modernize curriculum while preserving cultural identity (Cai, 2024; Xin & Taipanich, 2025). Future studies should extend this work by implementing the refined test in larger cohorts across multiple institutions and incorporating

longitudinal analyses to assess sustained learning outcomes and transferability of fusion-based competencies.

VI. Conclusion

This study provides empirical support for the psychometric soundness and pedagogical value of a fusion-based suona instructional approach in Chinese higher education. The integration of classical test theory and Rasch-based item response theory enabled a comprehensive validation of the achievement test, confirming its internal consistency and cultural relevance while identifying areas for item refinement. The findings affirm that fusion-oriented curriculum design, grounded in constructivist learning theory, Vygotsky's zone of proximal development, and multiple intelligences theory, can effectively promote cross-cultural musical literacy and learner engagement.

By combining traditional suona pedagogy with Western theoretical constructs, this research bridges cultural traditions and modern educational paradigms, contributing to the growing discourse on multicultural assessment in music education. The pilot study demonstrates the feasibility of such integrative approaches and offers a validated framework for future curriculum development and large-scale assessment studies.

However, several limitations should be acknowledged. The pilot's modest sample size and single-institution scope limit the generalizability of the findings. Additionally, the test items primarily targeted average-ability learners, resulting in a mild ceiling effect that constrained measurement precision among advanced performers. Future research should expand the participant pool across diverse institutions, introduce more challenging items to capture higher proficiency levels, and incorporate longitudinal designs to examine learning retention and curriculum transferability.

Overall, this study establishes a foundational model for psychometric and pedagogical innovation in traditional Chinese instrument instruction. It contributes to the broader goal of advancing culturally responsive and evidence-based music education practices that honor heritage while embracing global musical literacy.

References

- Alavi, M., Le Lagadec, D., & Cleary, M. (2025). Challenges of cross-cultural validation of clinical assessment measures: A practical introduction. *Journal of Advanced Nursing*, 82, 941-949. <https://doi.org/10.1111/jan.16906>
- Bartleet, B. L., Grant, C., Mani, C., & Tomlinson, V. (2020). Global mobility in music higher education: Reflections on how intercultural music-making can enhance students' musical practices and identities. *International Journal of Music Education*, 38(2), 161-176. <https://doi.org/10.1177/0255761419890943>
- Bates, V. C., Shevock, D. J., & Prest, A. (2021). Cultural diversity, ecodiversity, and music education. In *The politics of diversity in music education* (pp. 163-174). Springer. https://doi.org/10.1007/978-3-030-65617-1_12
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.
- Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Springer.
- Bruner, J. (1996). *The culture of education*. Harvard University Press.
- Cai, G. (2024). The art of Chinese suona performance from an international perspective: An exploration of cross-cultural communication and educational value. *The Educational Review, USA*, 8(6), 847-851. <https://doi.org/10.26855/er.2024.06.011>
- Campbell, P. S. (2018). *Teaching music globally* (2nd ed.). Oxford University Press.
- Chao, C., & Nicolas, A. (2024). Evolution and literacy transmission of the suona ensemble in Southwest Shandong Province, China. *International Journal of Education and Literacy Studies*, 12(2), 87-93. <https://doi.org/10.7575/aiac.ijels.v.12n.2p.87>
- Chauhan, G. R., Chauhan, B. R., & Vaza, J. V. (2023). Relations of the number of functioning distractors with the item difficulty index and the item discrimination power in multiple-choice questions. *Cureus*, 15(7). <https://doi.org/10.7759/cureus.42492>
- Chen, Y. (2024). Effective strategies and methods to enhance music practical teaching in higher vocational preschool education. *Pacific International Journal*, 7(1), 162-167. <https://doi.org/10.55014/pij.v7i1.545>
- Chow, J. K. F., Jin, K.-Y., & Siu, W. L. (2024). What do music listening tests assess? Applying Rasch testlet models to the HKDSE data. *International Journal of Music Education*. <https://doi.org/10.1177/02557614241297277>

- Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology, 10*, 1507. <https://doi.org/10.3389/fpsyg.2019.01507>
- Gardner, H. (2011). *Frames of mind: The theory of multiple intelligences* (3rd ed.). Basic Books.
- Gardner, H. (2020). *A synthesizing mind: A memoir from the creator of multiple intelligences theory*. MIT Press.
- Harrison, P. M. C., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerized adaptive testing, and automatic item generation. *Scientific Reports, 7*, 3618. <https://doi.org/10.1038/s41598-017-03586-z>
- Kim, S. (2025). The value of Korean traditional music and its teaching methods as multicultural music education outside Korea: Centered around Royal Court Dance, Cheoyoungmu. *Korean Journal of Research in Music Education, 54*(1), 73-95. <https://doi.org/10.30775/KMES.54.1.73>
- Kūkea Shultz, P., & Englert, K. (2021). Cultural validity as foundational to assessment development: An indigenous example. *Frontiers in Education, 6*, 701973. <https://doi.org/10.3389/educ.2021.701973>
- Kwon, S., Yang, H., & Choi, J. (2025). The preliminary validity study for the development of the Goldsmiths Musical Sophistication Index in Korean (Gold-MSI-K). *Korean Journal of Research in Music Education, 54*(1), 1-30. <https://doi.org/10.30775/KMES.54.1.01>
- Lee, D., & Park, J. (2024). Analysis of Korean traditional music content in ‘Pleasant Life’, the 2022 revised integrated textbook for 1st and 2nd grade elementary school students. *Korean Journal of Research in Music Education, 53*(4), 133-163. <https://doi.org/10.30775/KMES.53.4.133>
- Mygdanis, Y. (2025). Design-based research in music education: Theoretical foundations, methodological perspectives, and practice implications. *Futurity Education, 5*(2), 90-114. <https://doi.org/10.57125/FED.2025.06.25.05>
- Ooi, P. S., & Engelhard, G., Jr. (2019). Examining rater judgements in music performance assessment using the Many-Facets Rasch rating scale measurement model. *Journal of Applied Measurement, 20*(1), 79-99.
- Salvador, K., Paetz, A. M., & Tippetts, M. M. (2020). “We all have a little more homework to do.” A constructivist grounded theory of transformative learning processes for

- practicing music teachers encountering social justice. *Journal of Research in Music Education*, 68(2), 193-215. <https://doi.org/10.1177/0022429420920630>
- Shearer, C. B. (2020). Multiple intelligences in gifted and talented education: Lessons learned from neuroscience after 35 years. *Roeper Review*, 42(1), 49-63. <https://doi.org/10.1080/02783193.2019.1690079>
- Sim, T., & Cheng, J. L. A. (2024). Music-based language programme teacher training on preschool teachers' music teaching self-efficacy and delivery performance: A case study. *Southeast Asia Early Childhood*, 13(1), 133-151. <https://doi.org/10.37134/saecj.vol13.1.9.2024>
- Sosa Paredes, Y. K., & Andersson, B. (2025). Validity of test score interpretations and cross-cultural comparisons in the First and Second International Science Studies. *Educational Assessment, Evaluation and Accountability*, 37, 231-254. <https://doi.org/10.1007/s11092-024-09444-7>
- Springer, D. G., & Bradley, K. D. (2017). Investigating adjudicator bias in concert band evaluations: An application of the Many-Facets Rasch Model. *Musicae Scientiae*, 22(3), 377-393. <https://doi.org/10.1177/1029864917697782>
- Uludag, A. K., & Satir, U. K. (2025). Seeking alternatives in music education: The effects of mobile technologies on students' achievement in basic music theory. *International Journal of Music Education*, 43(2), 172-188. <https://doi.org/10.1177/02557614231196972>
- Utomo, U., Rachman, A., & Suharto (2025). A study on the relationship between music college entrance performance test and academic achievement of prospective music educators in Indonesia. *Korean Journal of Research in Music Education*, 54(3), 139-166. <https://doi.org/10.30775/KMES.54.3.139>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Wright, C. L., Branch, R., Ey, L. A., Hopper, K. M., & Warburton, W. (2022). Popular music media literacy: A pilot study. *Journal of Media Literacy Education*, 14(3), 29-38. <https://doi.org/10.23860/JMLE-2022-14-3-3>
- Xin, L., & Taipanich, C. (2025). Suona teaching and research in music university. *Asia Pacific Journal of Religions and Cultures*, 9(1), 275-284.
- Xiong, Z. (2025). Exploring the potential and challenges of applying multiple intelligences theory in Chinese music education. *Teaching and Teacher Education*, 160, 105031. <https://doi.org/10.1016/j.tate.2025.105031>