Multimodal Music Teaching Based on Artificial Intelligence

Kuang Sheng[1], Liqiong Duan[2]
[1]College of Music, Hengyang Normal University, Hengyang 421001, Hunan, China
[2]Admissions Office, Zhangjiakou Open University, Zhangjiakou 075000, Hebei, China

Corresponding author: Liqiong Duan (e-mail: momo_918@163.com)

ABSTRACT With the rapid development of artificial intelligence and multimodal information processing technologies, personalized learning has become an important research direction in intelligent educational systems. This study proposes an AI-enhanced framework for personalized multimodal music teaching based on data-driven learner profiling. Multimodal learning interaction data, including visual, auditory, textual, interactive, social, and temporal behavioral features, are analyzed using the K-means++ clustering algorithm to identify heterogeneous learning patterns. Three learner profiles, namely Audiovisual-Interactive, Text-Structure Dependent, and Auditory-Rhythm Sensitive, are extracted and used to guide the design of differentiated instructional strategies. To evaluate the effectiveness of the proposed framework, simulation experiments are conducted using virtual learner agents with distinct behavioral characteristics. Results demonstrate that personalized multimodal instruction significantly improves learning efficiency and engagement compared with conventional non-personalized approaches. The proposed framework provides a data-driven methodology for adaptive multimodal learning and highlights the potential of artificial intelligence in multimodal information perception, behavioral pattern analysis, and intelligent human–machine interaction. These findings offer valuable insights for the development of intelligent educational environments and multimodal signal-driven learning systems.

INDEX TERMS multimodal teaching, artificial intelligence, K-means++ clustering, intelligent learning systems

I. INTRODUCTION

SINCE the 1990s, the advancement of computer technology has led to increasingly widespread software applications. Scholars have recognized that meaning is conveyed not only through language but also through a variety of non-verbal semiotic resources. This understanding has given rise to multimodal discourse analysis, which integrates verbal and non-verbal signs, expanding the methods of discourse analysis. The application of these principles in education has yielded positive outcomes, leading to the development of Multimodal Teaching (MMT). From a multimodal perspective, the physical properties of teaching tools, are increasingly recognized as important non-verbal carriers of information. In the context of music education, MMT can stimulate students' multiple senses, enhance their interest, facilitate knowledge acquisition, and create a dynamic and coherent classroom atmosphere conducive to talent development. The integration of Artificial Intelligence (AI) with education offers transformative possibilities for addressing pedagogical challenges. In music education, AI applications range from intelligent tutoring systems to user acceptance modeling. Prior research has highlighted the importance of high-quality teacher professional development [1], explored demographic profiles of music students [2], and examined shifts to online learning [3]. However, there is a notable gap in research that systematically leverages AI to enable personalized multimodal music instruction—i.e., adapting MMT strategies to the heterogeneous learning preferences and behaviors of individual students [4,5]. Personalization is crucial because students interact with multimodal resources (e.g., video, audio, text, interactive tools) in diverse ways. Some may thrive on audiovisual and interactive content, while others prefer structured textual information or are particularly sensitive to auditory and rhythmic patterns. A one-size-fits-all MMT approach may not optimize learning for all. Therefore, the key challenge is to accurately identify these distinct learner profiles and tailor instructional strategies accordingly. To tackle this challenge, this study employs a data-driven approach centered on AI clustering techniques. The research focuses on applying MMT to secondary school music instruction and addresses the personalization problem through the following methodology: First, multimodal learning data (encompassing visual, auditory, textual, interactive, assessment, social, temporal, and modal-transition features) are extracted and preprocessed from a public educational dataset (EdNet KT1). Second, a K-means++ clustering algorithm is applied to this data to identify distinct, data-informed learner profiles. Third, based on the characteristics of each cluster, personalized multimodal teaching strategies are designed. Finally, to validate the effectiveness of these strategies within a controlled and scalable environment, a simulation experiment is conducted. This experiment compares the learning performance and engagement of virtual agent groups receiving personalized instruction with those receiving a standard, non-personalized curriculum. The results indicate that the AI-clustering framework can effectively segment learners into meaningful groups and that the corresponding personalized strategies significantly improve simulated learning efficiency and engagement. This study contributes a robust theoretical and methodological foundation for implementing AI-enhanced, personalized multimodal music instruction, offering practical insights for future educational innovation.

II. MULTI MODAL TEACHING METHOD BASED ON AI

A. ARTIFICIAL INTELLIGENCE

With the development of computer technology, a variety of new intelligent teaching systems came into being. Intelligent monitoring system, teaching robot, voice recognition and other new devices have been applied to the field of school education, effectively improving the working efficiency of teachers. It creates a more fair and harmonious learning environment for students (Figure 1). However, with the advent of the information age, many traditional teaching methods have been unable to meet the needs of modern teaching. For example, the traditional blackboard+chalk+textbooks and other models can no longer meet the needs of the times. Computer network courses and video courses have also become the main way for students to learn. With the development trend of integration of AI and education industry becoming increasingly obvious, intelligent teaching system has become an effective way to solve the problem of traditional teaching mode. For data analysis, an AI clustering algorithm was utilized. Then, according to the feedback information of teachers on students' classroom performance, this text automatically judged students' mastery of knowledge and created a set of exclusive student performance evaluation system for them. This text analyzed the application status and effect of AI clustering algorithm in teaching.

B. CLUSTERING ALGORITHM

To achieve objective clustering of student groups, this study employs the K-means++ clustering algorithm to analyze the preprocessed student feature vectors. This algorithm improves upon the traditional K-means algorithm, which is sensitive to the initial clustering centers. By keeping the centers away from each other during the initialization phase, it enhances the stability of the algorithm and the convergence speed of the results. In this study's application, the algorithm measures the similarity between student feature vectors using Euclidean distance, which reflects the overall differences among students in eight multimodal learning behavior dimensions. To determine the optimal number of clusters, the study comprehensively applied both the elbow method and the silhouette coefficient method. The elbow method makes a preliminary judgment by observing the inflection points of the total squared deviation curve under different K values, while the silhouette coefficient quantitatively evaluates the closeness of each sample point to its own cluster and other clusters. The validation results of both methods point to K=3. When K=3, the clustering results exhibit the highest intra-cluster sample compactness and the greatest separation between different clusters, which confirms from a data perspective the existence of three subpopulations of learners with significantly different behavior patterns, providing a solid data foundation for subsequent personalized teaching strategies.

C. MULTIMODAL TEACHING

Multimodality, in the context of learning, emphasizes the integration of multiple sensory and semiotic resources for meaning-making. In the 1990s, Western scholars proposed the specific application of MMT in social semiotics, systemic functional grammar and traditional discourse analysis. The results of the study show that in other countries' language teaching, a variety of models can be used. In painting, music, dance and other arts, it can use the form of nonverbal symbols to produce the meaning people need. Multimodal teaching can enable students to obtain more symbolic stimuli, such as language, body movements, expressions and modern science and technology. Various symbol models can effectively stimulate students' feelings, stimulate their initiative, stimulate their imagination and association, and thus improve their knowledge reserves and technical level. The effect of this is better than forcing students to remember all the knowledge. The main feature of MMT is to use multimedia technology as an auxiliary means, through the size, color, animation, body language, etc. of voice, picture, font and other multiple modes. Through animation, body language, etc., it makes language concrete and visual, so that students can better understand and remember what they have learned, create a relaxed and happy classroom atmosphere, and stimulate students' enthusiasm for learning [6,7]. The MMT method has many advantages, so many teachers are willing to adopt MMT. However, when using MMT, the selection of modes is a key issue [8,9]. The diagram of MMT topology system is shown in Figure 2. There are many modes, which can be selected according to their functions. When people want to describe a group of events, they can use text mode. When people want to show the relationship between things, they can choose the picture mode [10,11]. Different modes also differ in the accuracy of meaning expression. People just want to use some gestures to convey a general meaning, such as asking the direction of others. If people want to express their meaning accurately, they should use language patterns. In addition to these, people can also choose the mode according to the goals they want to achieve [12-14].

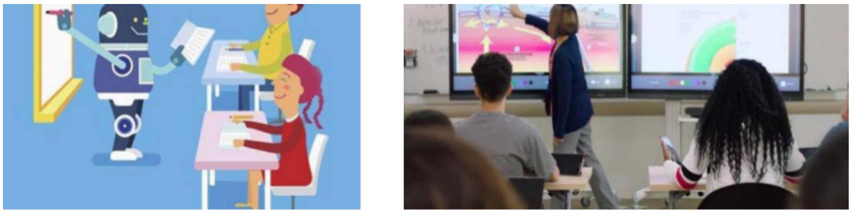

FIGURE 1. Application of artificial intelligence in teaching

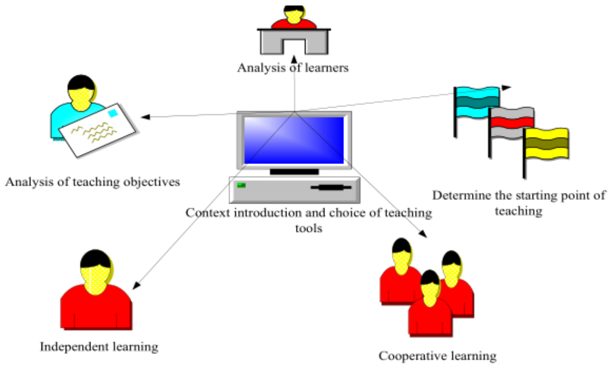

FIGURE 2. Topology of MMT

III. SIMULATION VERIFICATION OF MULTIMODAL MUSIC TEACHING MODE BASED ON PUBLIC DATA

A. DATA SOURCES AND PREPROCESSING

The study utilizes the EdNet KT1 public education dataset, which captures large-scale, anonymized student interaction logs from an online learning platform. From this dataset, all records associated with music theory and appreciation courses were extracted, resulting in 28,541 valid interaction logs from 1,824 unique learners [15]. The preprocessing pipeline, illustrated in Figure 3, consists of four sequential stages designed to transform raw interaction data into a structured, analyzable format suitable for multimodal clustering.

1) Stage 1: Record Filtering

Relevant learning logs were isolated from the broader EdNet KT1 corpus based on course metadata. This initial filtering yielded a focused subset of 28,541 records corresponding to music-related learning activities.

2) Stage 2: Feature Engineering

An eight-dimensional multimodal feature vector was created for each student to measure different types of behavior such as how they use each sense and how they interact with others:

Visual activity (video completion rate)

Auditory activity (the amount of time spent listening to audio materials)

Textual activity (the average time spent on written resources)

Interactive activity (number of interaction with the instructor in class)

Assessment activity (the average percentage of correct answers on quizzes)

Social activity (the number of times participated in the discussion forum)

Temporal Pattern (Frequency of Learning Session Sequence)

Modal Transition (How often types of resources are switched)

3) Stage 3: Data Cleansing

To reduce noise, invalid sessions that lasted less than 30 seconds were removed from the analysis. The continuous features were standardised by applying Z-score normalisation; the categorical variables, such as resource types, were encoded using a one-hot encoding process so that they could be used with distance-based clustering methods.

4) Stage 4: Missing Value Imputation

The missing data points were filled using the average of the corresponding attribute based on the cohort for the learner's course. This cohort level approach maintained the distributional characteristics of the variables while also maintaining data integrity. This pipeline resulted in a clean and standardized dataset, which was ready for the cluster analysis process. The structure of the dataset allowed for distinct profiles of the learners, identified through their use of multimodal interacting with the learning resources over time.

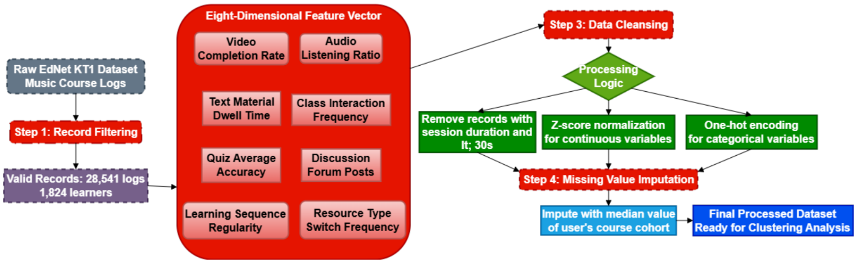

FIGURE 3. Schematic diagram of feature engineering process

B. CLUSTERING ANALYSIS AND PERSONALIZED LEARNING MODES

In order to develop an accurate grouping of the learner profiles and create an individualised teaching strategy for each profile, cluster analysis was conducted on the pre-processed eight-dimensional feature vectors using the K-means++ algorithm to provide the best possible starting point in terms of initial centroids, avoiding the issue of local optima. The optimum number of clusters (K=3) was found using a combination of the elbow method and the silhouette coefficient method to ensure that the clusters exhibited a good amount of coherence with each other, while still remaining separate from other clusters. The algorithm calculated distance between individual data points, where the distance between data points was determined by using Euclidean distance, and converged to an error of 1e-4 within 300 iterations. To further reveal the specific differences among the three types of learners in various behavioral dimensions, Figure 4 presents radar charts of the three cluster centers based on eight standardized features.

These distinct behavioral profiles, as quantified in Figure 4, are visually separated in the reduced dimensional space shown in Figure 5.

The radar chart visually demonstrates that learners of Type C1 exhibit peaks in the two dimensions of "Visual Activity" and "Interactive Activity"; learners of Type C2 demonstrate outstanding performance in "Textual Activity" and "Temporal Pattern"; learners of Type C3 are significantly higher than the other two types in "Auditory Activity" and "Assessment Activity". This characteristic profile provides a direct quantitative basis for subsequent cluster naming and personalized strategy design. The resulting three clusters of learners contain differing behavioral patterns that are shown in Figure 4. All three clusters were characterized by different profiles based on their defining features, which are as follows:

Cluster C1: Audiovisual-Interactive Dominant Type (Red); Learners in this cluster had a higher than average completion rate of videos and frequently engaged in classroom activities. Learners within this cluster were especially engaged with video and preferred a more dynamic and interactive method of learning through video materials.

Cluster C2: Text-Structure Dependent Type (Teal); Learners within this cluster displayed an extremely long average duration of staying within the text learning materials, consistently completing learning sequences as well as consistently using the same learning resources. The preference of these learners for a structured and linear approach to learning was demonstrated through the preference for well-organized and easily read text.

Cluster C3: Auditory-Rhythm Sensitive Type (Blue); Learners within this cluster preferred audio learning materials over other types and excelled on tests related to rhythm. Learners comprising this cluster demonstrated a keen awareness of their auditory surroundings and responded positively to the rhythmic and musical aspects of their environment. In the 3D scatter plot of data clusters based off of the first three Principal Components (PC1, PC2, PC3 - 64.7%, 16.6%, 13.3% respectively) that explain 84.6% of the variation in the dataset (as shown in Figure 4) shows clear and separate locations of the three different colour point clouds which provides a clear visual representation that the clustering has been done correctly. Using a yellow star to indicate the centre of each of the clustered point clouds, you will find that these Yellow Stars are indicative of the prototypical learner for each of the three groups and provide an anchor point for creating personalised types of instructional approaches. Based on the feature numerical profiles of the three clustering centers (Figure 5), this study has designed targeted core teaching strategies for each category of learners, aiming to strengthen their dominant channels or compensate for their relative weaknesses. For C1 learners, who exhibit high values in visual and interactive activities, the core strategy is 'interaction enhancement'. This strategy directly aligns with their high 'interactive activity' characteristic values, emphasizing the deep integration of interactive elements into video content. For instance, incorporating pop-up quizzes at video nodes or adopting a branching narrative structure. Additionally, providing real-time feedback tools such as virtual musical instruments maintains a high level of engagement that matches their behavioral characteristics. In response to the characteristics of high numerical values in textual activities and regular temporal patterns among C2 learners, the core strategy is 'structural explicitation'. This strategy directly serves their need for structured and sequential information. Specific measures include providing knowledge maps presented in text and graphics, clearly mapping the steps of music symbol interpretation to the hierarchical levels of musical notation writing, and presenting the relationships between musical

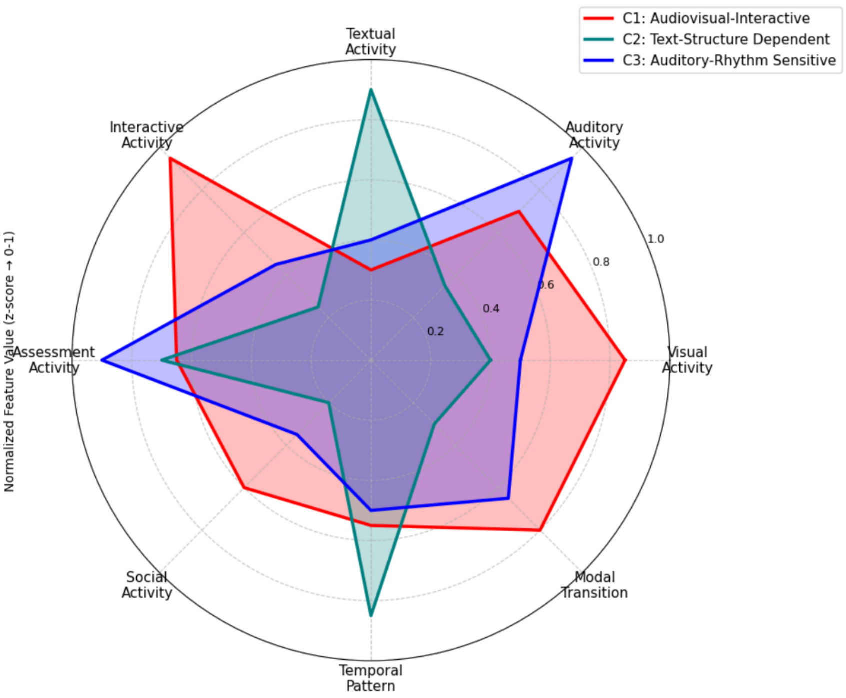

FIGURE 4. Radar chart of cluster centers across eight multimodal features

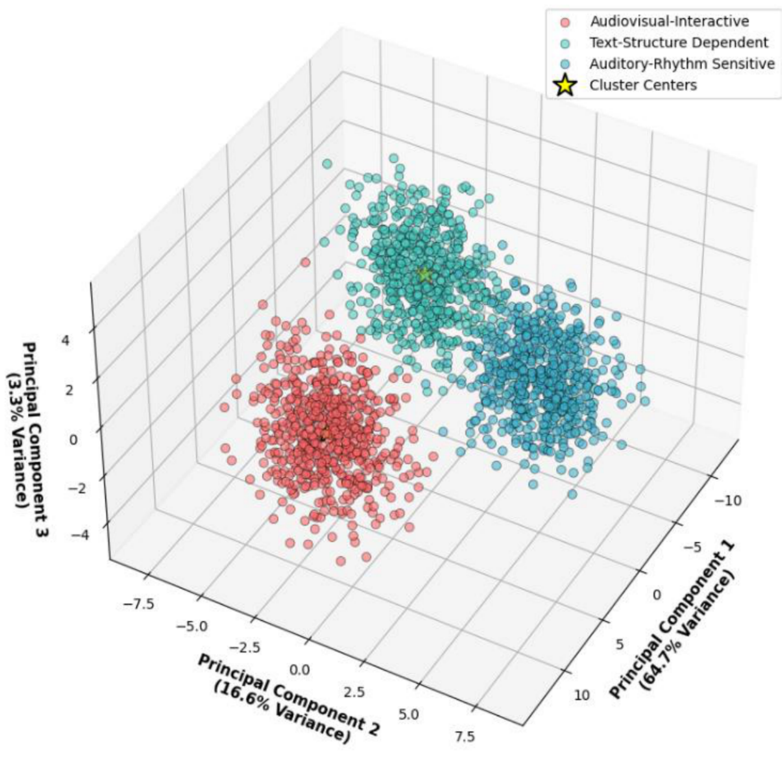

FIGURE 5. 3D feature space clustering scatter plot

elements using comparative tables. Given the high scores in auditory activities and assessment tasks among C3 learners, the core strategy is 'auditory reinforcement and visualization'. This strategy aims to amplify their auditory strengths and aid comprehension. The resources provided include audio tools that allow for playback speed adjustment, sound editing modules for creating multi-track mixes, software that converts sound into visual waveforms, and audio games that focus on rhythm recognition, all of which serve to reinforce the learning effects achieved through the auditory channel.

C. ANALYSIS OF SIMULATION EXPERIMENT RESULTS

To assess the effectiveness of our personalized multimodal teaching strategy, we conducted a simulation experiment using intelligent agents. We compared a control group, which received a standard, non-personalized set of resources, with an experimental group that received content tailored to their specific learning profiles. We created a simulation environment with 100 virtual agents for each of three learner clusters (C1, C2, and C3). The decision-making rules of each type of agent are parameterized based on the feature vectors of their corresponding cluster centers. For example, when selecting learning resources, type C1 agents are assigned a higher probability to trigger interactive video content; type C2 agents are set to prefer completing text materials in a linear order; and when encountering rhythm-related assessment tasks, type C3 agents have a higher simulated accuracy benchmark. This parameterized setting enables the behavior patterns of agents to approximately reflect the typical characteristics of their corresponding learner personas. These agents were designed with behavioral attributes based on their respective cluster characteristics. They then participated in a simulated music course unit, which included video, audio, text, and quiz resources. We used two main metrics to evaluate performance:

Knowledge Mastery Efficiency: This measured how quickly an agent reached a predefined mastery level (test accuracy of 85% or higher).

Learning Engagement: This quantified the average number of interactive actions an agent took per time step (e.g., replaying content, taking notes, doing extra exercises). The results of the 20-run simulation, shown in Table 1, clearly demonstrate that the personalized approach significantly outperformed the standard approach across all metrics.

The experimental group consistently outperformed the control group in terms of knowledge acquisition speed, as detailed in Table 1. The personalized strategy accelerated mastery, reducing the average time by 17.5%, from 82.7 steps to 68.2 steps. This improvement was observed across all learner clusters, with the Auditory-Rhythm Sensitive (C3) cluster showing the most significant reduction, at 24.1%. Furthermore, the personalized approach significantly boosted learning engagement. The experimental group averaged 4.3 actions per step, a 38.7% increase compared to the control group's 3.1 actions per step. This heightened engagement was particularly noticeable among the Audiovisual-Interactive (C1) learners, who averaged 4.8 interactions per step, indicating that the tailored interactive elements effectively catered to their preferred learning style. These findings provide compelling quantitative evidence that aligning instructional methods with learners' individual multimodal preferences, as identified through cluster analysis, leads to more efficient knowledge acquisition and increased active participation. The results support the central hypothesis of this study: that AI-driven personalization can optimize the effectiveness of multimodal music education.

TABLE 1. Comparison of simulation experiment results (mean $\pm$ standard deviation)

Cluster Type	Control Group Mastery Time (Steps)	Experimental Group Mastery Time (Steps)	Efficiency Improvement	Control Group Engagement (Actions/Step)	Experimental Group Engagement (Actions/Step)
C1: Audiovisual-Interactive	78.3 ± 5.2	65.1 ± 4.8	16.9%	3.4 ± 0.3	$4.8 \pm 0.5\,2$
C2: Text-Structure Dependent	85.1 ± 6.1	70.5 ± 5.3	17.2%	$.8 \pm 0.4$	$3.9 \pm 0.4\,3.$
C3: Auditory-Rhythm Sensitive	84.7 ± 5.8	64.3 ± 4.5	24.1%	0 ± 0.3	$4.2 \pm 0.5\,3.1$
Overall	82.7 ± 5.7	68.2 ± 4.9	17.5%	± 0.3	4.3 ± 0.5

IV. CONCLUSION

The emergence of multimodal discourse analysis has led to the emergence of MMT models in music classes. In order to improve the quality of teaching, teachers use a variety of teaching models in the classroom. In response to the low level of personalization in current multimodal teaching, this article combines artificial intelligence and uses clustering algorithms to personalize student groups, and based on this, develops differentiated teaching strategies. The experimental results showed that compared with traditional teaching classes, the experimental group had a significant improvement in music learning effectiveness. Overall, the application of artificial intelligence technology in music education helps to achieve personalized delivery of teaching content and intelligent management of the teaching process, providing new ideas for the innovative development of music education in the future. However, the application of multimodal teaching methods in classroom instruction requires careful selection of modalities. Multimodal teaching integrates various symbolic forms and media resources to aid student comprehension and memory. Multimodal teaching is to combine various forms of symbols with various media resources, and then carry out teaching to help students understand and remember what they have learned. However, due to the diversity of modes and the complexity of media, as well as teachers' lack of multimodal theoretical knowledge, teachers are often at a loss when making choices, or blindly do what they know and see, resulting in the complexity and incongruity of the content. Therefore, teachers need to learn deeply about MMT knowledge and multimedia. From the experimental results, it can know that the rational use of MMT can effectively improve the quality of teaching, providing a theoretical foundation for the integrated application of AI, multimodal pedagogy in the future.

AUTHOR CONTRIBUTIONS

Kuang Sheng designed,collected and analyzed the data, and drafted the manuscript.Kuang Sheng and Liqionq Duan conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Liqiong Duan participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This work was supported by MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No.18YJC760035).

ACKNOWLEDGEMENTS

No other supporting information.

REFERENCES

[1] A. Bautista, X. Yau, and J. Wong, “High-quality music teacher professional development: A review of the literature,” Music Education Research, vol. 19, no. 4, pp. 455–469, 2017, doi: 10.1080/14613808.2016.1249357.

[2] K. Elpus and C. R. Abril, “Who enrolls in high school music? A national profile of US students, 2009-2013,” Journal of Research in Music Education, vol. 67, no. 3, pp. 323–338, 2019, doi: 10.1177/0022429419862837.

[3] A. Daubney and M. Fautley, “Editorial Research: Music education in a time of pandemic,” British Journal of Music Education, vol. 37, no. 2, pp. 107–114, 2020, doi: 10.1017/S0265051720000133.

[4] A. Moreno, “Elements of music based on artificial intelligence,” Acta Informatica Malaysia, vol. 4, no. 2, pp. 30–32, 2020, doi: 10.26480/aim.02.2020.30.32.

[5] W. Zha, “A Study on the User Acceptance Model of Artificial Intelligence Music Based on UTAUT,” Journal of the Korea Society of Computer and Information, vol. 25, no. 6, pp. 25–33, 2020.

[6] J. Palkki and W. Sauerland, “Considering gender complexity in music teacher education,” Journal of Music Teacher Education, vol. 28, no. 3, pp. 72–84, 2019, doi: 10.1177/1057083718814582.

[7] K. Navbakhor, “The role of music lessons in the formation of national and intercultural competence in students,” Mental Enlightenment Scientific-Methodological Journal, pp. 130–139, 2020.

[8] A. P. Montgomery, A. Mousavi, M. Carbonaro, D. V. Hayward, and W. Dunn, “Using learning analytics to explore self-regulated learning in flipped blended learning music teacher education,” British Journal of Educational Technology, vol. 50, no. 1, pp. 114–127, 2019, doi: 10.1111/bjet.12590.

[9] F. E. Majidovich, “The Three Keys to a Musical Treasure,” Mental Enlightenment Scientific-Methodological Journal, pp. 173–182, 2021, doi: 10.51348/tziuj2021219.

[10] D. B. C. Kilic, “An Investigation of Music Teacher Candidates' Performance Anxiety Levels in Piano Examinations,” Journal of Education and Learning, vol. 7, no. 1, pp. 299–308, 2018, doi: 10.5539/jel.v7n1p299.

[11] K. Sutela, J. Ojala, and M. Kielinen, “Developing agency through music and movement,” Research Studies in Music Education, vol. 43, no. 2, pp. 195–211, 2021, doi: 10.1177/1321103X20934084.

[12] S. M. Demorest, J. Kelley, and P. Q. Pfordresher, “Singing ability, musical self-concept, and future music participation,” Journal of Research in Music Education, vol. 64, no. 4, pp. 405–420, 2017, doi: 10.1177/0022429416680096.

[13] Y. P. Wong and W. M. Lau, “Perceived attributes of music teaching effectiveness among kindergarten teachers: Role of personality,” Australian Journal of Teacher Education (Online), vol. 43, no. 9, pp. 28–38, 2018, doi: 10.14221/ajte.2018v43n9.2.

[14] K. Salvador, “Sustaining the flame: (Re) igniting joy in teaching music,” Music Educators Journal, vol. 106, no. 2, pp. 28–36, 2019, doi: 10.1177/0027432119873701.

[15] T. Raschdorf, B. N. May, and A. Searcy, “Integrating social-emotional learning into our "new normal" teaching elementary general music,” General Music Today, vol. 34, no. 2, pp. 42–48, 2021, doi: 10.1177/1048371320961372.