JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 12(2), 83–104. https://doi.org/10.18608/jla.2025.8835

Exploring Automated Assessment of Primary Students' Creativity in a Flow-Based Music Programming Environment

Zifeng Liu[1], Wanli Xing[2*], Chenglu Li[3], Fan Zhang[4], Hai Li[5] and Victor Minces[6*]

Abstract

Creativity is a vital skill in science, technology, engineering, and mathematics (STEM)–related education, fostering innovation and problem-solving. Traditionally, creativity assessments relied on human evaluations, such as the consensual assessment technique (CAT), which are resource-intensive, time-consuming, and often subjective. Recent advances in computational methods, particularly large language models (LLMs), have enabled automated creativity assessments. In this study, we extend research on automated creativity scoring to a flow-based music programming environment, a context that integrates computational and creative thinking. We collected 383 programming artifacts from 194 primary school students (2022–2024) and employed two automated approaches: an evidence-centred design (ECD) framework–based approach and an LLM-based approach using ChatGPT-4 with few-shot learning. The ECD-based approach integrates divergent thinking, complexity, efficiency, and emotional expressiveness, while the LLM-based approach uses CAT ratings and ECD examples to learn creativity scoring. Results revealed moderate to strong correlations with human evaluations (ECD-based: $r = 0.48$; LLM-based: $r = 0.68$), with the LLM-based approach demonstrating greater consistency across varying learning examples ($r = 0.82$). These findings highlight the potential of automated tools for scalable, objective, and efficient creativity assessment, paving the way for their application in creativity-focused learning environments.

Notes for Practice

- Evidence-centred design (ECD)–based frameworks and large language models (LLMs) can provide scalable and efficient solutions for evaluating creativity, but it is recommended to involve humans in the loop to reduce reliance on resource-intensive and subjective human assessments.
- The LLM-based approach demonstrated consistency across different learning examples, likely because its assessments reflect inherent patterns of judgment or are limited by a focus on structural and text-based aspects of creativity.
- Applying automated assessment methods to contexts like flow-based music programming can help educators better evaluate and support creativity in computational learning environments. These approaches enable real-time feedback and personalized support to foster student creativity.

Keywords

Creativity, automated assessment, generative AI, music programming, flow-based programming, K–12

Submitted: 26/12/2024 — **Accepted:** 02/07/2025 — **Published:** 29/08/2025

[1] *Email: liuzifeng@ufl.edu Address: College of Education, University of Florida, Gainesville, USA. ORCID iD: https://orcid.org/0009-0005-5833-2141*

[2*] *Corresponding author email: wanli.xing@coe.ufl.edu Address: College of Education, University of Florida, Gainesville, Florida, USA. ORCID iD: https://orcid.org/0000-0002-1446-889X*

[3] *Email: chenglu.li@utah.edu Address: Department of Educational Psychology, University of Utah, Salt Lake City, Utah, USA. ORCID iD: https://orcid.org/0000-0002-1782-0457*

[4] *Email: f.zhang1@ufl.edu Address: College of Education, University of Florida, Gainesville, Florida, USA. ORCID iD: https://orcid.org/0000-0003-0221-040X*

[5] *Email: li.ha@ufl.edu Address: College of Education, University of Florida, Gainesville, Florida, USA. ORCID iD: https://orcid.org/0009-0004-7299-2042*

[6*] *Corresponding author email: vminces@ucsd.edu Address: Department of Cognitive Science, University of California, San Diego, La Jolla, California, USA. ORCID iD: https://orcid.org/0000-0002-7733-0601*

1. Introduction

Creativity is widely recognized as the cognitive ability to generate innovative solutions and produce original, valuable outcomes, serving as a cornerstone of human innovation across disciplines and influencing many facets of life (Amabile, 1983; Loveless, 2002). While this definition emphasizes novelty and usefulness, it is important to acknowledge that creativity is a multifaceted construct with no single, universally agreed-upon definition (Runco & Acar, 2012; Sternberg, 1999; Kaufman & Beghetto, 2009). For example, Amabile's componential theory highlights domain-relevant skills, creativity-relevant cognitive processes, and intrinsic motivation as essential components of creative performance (Amabile, 1996). Sternberg (1999) defines creativity as the ability to produce work that is both novel and appropriate. Kaufman and Beghetto (2009) further propose the four C model, recognizing creativity at different levels, from everyday problem-solving to historically eminent contributions. Although creativity manifests differently depending on the discipline and context, it shares universal attributes that transcend fields (L. D. Newton & Newton, 2014). On a personal level, creativity empowers individuals, fostering self-fulfillment and enabling them to lead more meaningful lives (Kind & Kind, 2007; L. D. Newton & Newton, 2014). On a broader scale, creativity is recognized as a critical 21st-century skill that is essential for addressing complex contemporary challenges (National Advisory Committee on Creative and Cultural Education (NACCCE), 1999; Celik et al., 2024) and has been a driving force behind the advancement of human civilizations (Henriksen et al., 2018). Despite its significance, measuring creativity has long been a widely debated topic, particularly in education, as it serves as a crucial first step toward fostering creativity effectively.

1.1 Creativity Assessment: From Traditional Methods to Automated and AI-Supported Approaches

Several well-established methods have been used to assess creativity (shown in Table 1), each with distinct strengths and limitations. Divergent thinking (DT) tests, such as the Torrance tests of creative thinking (TTCT), are widely used to evaluate fluency, originality, flexibility, and elaboration (Plucker & Makel, 2010; Runco & Acar, 2012). However, these dimensions are often intercorrelated, and scoring can be time-consuming.

Table 1. *Existing creativity assessment methods.*

Literature	Method	Advantages	Disadvantages
Plucker and Makel (2010); Runco (2011); Runco and Acar (2012)	DT tests	Easy to administer; widely used; measures fluency, originality, flexibility, elaboration	High intercorrelation among dimensions; labour-intensive scoring; subjectivity in evaluation
S. Mednick (1962); S. A. Mednick (1968)	RAT	Emphasizes remote associative ability; helps distinguish truly creative ideas	Measures only one aspect of creativity; language-dependent
Amabile (1982); Kaufman et al. (2008); Turkman (2016)	CAT	Assesses actual creative products; theory-independent; considered the "gold standard"	Requires expert judges; resource-intensive; lacks predictive validity
Dumas et al. (2020); Beaty and Johnson (2021); Acar et al. (2024); Rahimi et al. (2024)	Automated scoring approaches	Scalable; reduces subjectivity; efficient across domains	Depends on model quality; potential bias; limited transparency
Kenett and Faust (2019); Heinen and Johnson (2018)	Semantic distance analysis (e.g., LSA)	Based on associative theory; objective measure of originality	Sensitive to vocabulary and dataset; overlooks structural creativity
Barbot (2018)	MTCI	Tracks creativity over time; captures development dynamics	Complex to implement; may not generalize across contexts
Doshi and Hauser (2024); Henriksen et al. (2018)	LLMs for co-creation	Real-time support for ideation; fosters human-AI collaboration	May reduce diversity of ideas; reflects training data bias
Beaty and Johnson (2021)	*SemDis* platform	Calculates semantic distance; automates feedback	May miss structural elements of creativity
Acar et al. (2024)	MOTES	Game-like format; strong correlation with human ratings; education-friendly	High technical dependency; low explainability; potential algorithmic bias
DiStefano et al. (2024)	LLMs predicting human ratings	High accuracy; outperforms traditional metrics	Computationally expensive; concerns over fairness, bias, and ethics

To better identify truly creative responses, Mednick's remote associates test (RAT) emphasizes the ability to form connections between distantly related concepts (S. Mednick, 1962). Another influential approach, the consensual assessment technique (CAT), relies on expert judges to evaluate creative products and is often considered the "gold standard" (Amabile, 1982; Turkman, 2016). While CAT focuses on real-world creative output and avoids theory dependence, it faces challenges such as the need for qualified raters and limited predictive validity (Kaufman et al., 2008).

To address subjectivity and scalability issues, researchers have developed automated scoring methods using computational

tools (Dumas et al., 2020; Beaty & Johnson, 2021). Techniques like latent semantic analysis (LSA) help measure semantic distance as a proxy for originality (Kenett & Faust, 2019; Heinen & Johnson, 2018), while the multi-trial creative ideation (MTCI) framework tracks creativity development over time (Barbot, 2018).

More recently, large language models (LLMs) such as ChatGPT have reshaped creativity assessment and support. LLMs act as co-creators, aiding idea generation and providing real-time support in educational settings (Doshi & Hauser, 2024; Henriksen et al., 2018). However, their reliance on training data may reduce diversity in outputs. LLMs are also used in automated creativity assessment platforms. For example, the *SemDis* platform applies natural language processing (NLP) to evaluate semantic distance and provide objective feedback on creativity (Beaty & Johnson, 2021). The MOTES tool, designed for elementary students, uses fine-tuned LLMs to assess creativity in a game-like format, showing strong alignment with human ratings (Acar et al., 2024). Furthermore, LLMs like RoBERTa and GPT-2 have achieved high accuracy in predicting human judgments in metaphor generation tasks (DiStefano et al., 2024).

Despite their promise, LLM-based methods raise concerns about computational costs, algorithmic bias, limited transparency, and ethical implications in high-stakes educational contexts (Beaty & Johnson, 2021; Barbot, 2018).

1.2 Creativity for Teaching and Learning

From an educational perspective, fostering creativity has been consistently emphasized by researchers (National Advisory Committee on Creative and Cultural Education (NACCCE), 1999; L. D. Newton & Newton, 2014). Boden (2004) underscores the pervasive nature of creativity across all disciplines, while the National Advisory Committee on Creative and Cultural Education (NACCCE) (1999) report and Robinson, chair of the UK government's National Advisory Committee on Creative and Cultural Education (as cited in Azzam (2009)), advocate for a holistic integration of science, technology, arts, and humanities within the curriculum to nurture creativity comprehensively. Since the early 2000s, research on creativity in education has expanded significantly, encompassing diverse areas such as teaching creativity (Bowkett, 2007), its relationship with learning variables (Runco, 2014; Colton & Wiggins, 2012), factors influencing creativity (D. P. Newton, 2013), and the design of creativity-supporting learning environments (Rahimi, 2023; Rahimi et al., 2024). However, existing research varies widely in how creativity is defined, what units of analysis are used, and which methods are applied for data collection and analysis. This diversity reflects the complexity of studying creativity in educational contexts (L. D. Newton & Newton, 2014; Turkman, 2016; Rahimi et al., 2024).

Researchers widely acknowledge that creativity is an essential skill in science, technology, engineering, and mathematics (STEM) fields (Henriksen et al., 2018; Lou et al., 2017). However, some scholars argue that insufficient attention has been given to the supportive environments necessary for nurturing creativity, which may hinder efforts to foster it effectively (Lin, 2011). Increasingly, empirical research is needed to assess and enhance creativity within STEM learning environments, particularly those designed to prepare individuals for success in STEM-related careers (Rahimi et al., 2024). For example, in computer science education (CSE), creativity is defined as the ability to recognize or produce something "different, new, or innovative." It involves generating adaptive, useful solutions that are novel within the relevant context (Kaufman & Sternberg, 2010; Sharmin, 2021). The role of creativity in CSE is an emerging area with significant potential for exploration, and educators across various subjects generally acknowledge its importance (Noh & Lee, 2020; Sharmin, 2021).

1.3 The Present Research

While a few studies have demonstrated novel approaches to measuring creativity in CSE, such as assessing student originality in programming (Chou et al., 2024; Noh & Lee, 2020; Sharmin, 2021), significant gaps remain in understanding and addressing creativity in this field (Rahimi et al., 2024). The complexity of creativity as a construct, coupled with the relatively nascent nature of CSE research compared to other fields, presents unique challenges in understanding and evaluating creativity effectively (Sharmin, 2021). Prior research highlights the need to assess and foster creativity in STEM learning environments. It also underscores the importance of objective, automated tools for creativity assessment. Building on this foundation, this study seeks to expand existing work by exploring innovative methods to evaluate students' creativity within a flow-based music programming environment. Flow-based music programming is a novel approach for engaging young students in early programming learning and has been found to improve their learning attitudes (Song et al., 2023; Liu et al., 2025). Creativity is particularly important in CSE because it fosters problem-solving, innovation, and the ability to approach challenges from multiple perspectives—skills that are essential for success in the rapidly evolving technological landscape (Henriksen et al., 2021; Hershkovitz et al., 2019; Israel-Fishelson & Hershkovitz, 2022). Moreover, creativity in programming supports students in generating novel and effective solutions that are critical not only for academic achievement but also for real-world applications (Li et al., 2022; Venckutė et al., 2020; Rubenstein et al., 2022).

The central research question driving this investigation is: *How can students' creativity be automatically assessed within a flow-based music programming environment?* To explore this, we utilized an open, web-based flow-based programming environment called M-Flow (V. H. Minces et al., 2023). From 2022 to 2024, we collected programming artifacts from 194 primary school students to analyze their creativity within these programs. To address the research question, we propose and

develop two automated approaches: (1) an evidence-centred design (ECD) framework–based approach, drawing on prior work (R. G. Almond et al., 2015; Rahimi et al., 2024), and (2) an LLM-based approach utilizing few-shot learning (Parnami & Lee, 2022). To validate and benchmark these methods, CAT was also employed. The ECD approach integrates four key dimensions—divergent thinking, complexity, efficiency, and emotional expressiveness—to provide a structured, multifaceted assessment model. Meanwhile, the LLM approach uses ChatGPT-4o, incorporating ECD assessment examples and CAT results as learning examples for creativity scoring. This dual-method strategy offers both a theoretically grounded framework and an AI-driven solution to advance creativity evaluation in programming education. An overview of the assessment methods used in this study is shown in Figure 1. Details can be found in Section 2.

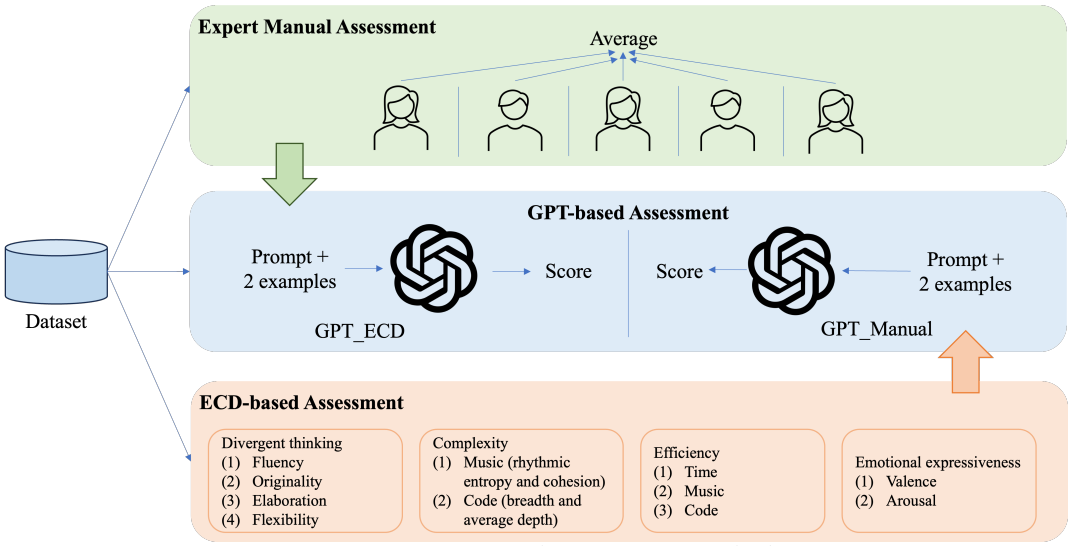


Figure 1. *Overview of the assessment methods.*

2. Method

2.1 M-Flow: A Flow-Based Music Programming Platform

M-Flow is an open, web-based, flow-based programming tool designed and developed to support the creation of music and sound compositions while providing early, accessible, and authentic computer science (CS) learning experiences for students from low-income and underrepresented backgrounds. Most programming languages, including Scratch (Zhang & Nouri, 2019) and Python-based EarSketch (Engelman et al., 2017), use an imperative programming paradigm, where commands are executed sequentially along the script. In contrast to imperative programming, flow-based programming is a programming paradigm in that the code is a direct visualization of the program's structure, with boxes representing processes connected by arrows representing the program's flow. This approach is particularly beneficial for learners with minimal CS experience, enabling them to quickly develop functional applications and focus more on creative exploration (Szydlo et al., 2017; V. H. Minces et al., 2023). Because of its intuitiveness, a music-oriented flow-based programming platform is well suited to be adopted in primary school general education classrooms, where students who might never have been interested in CS or exposed to it can engage in an authentic CS learning experience. Music makes programming activities more engaging, enhancing students' motivation and participation (V. Minces et al., 2021; Siva et al., 2018), and programming can help students better understand musical compositions (Repenning et al., 2020). Furthermore, making music is inherently a creative activity that can inspire creative thinking during programming (Bănuț et al., 2022; V. H. Minces & Akshay, 2023). Figure 2 shows M-Flow's interface; detailed descriptions of the platform's functions and example artifacts created by students are provided in Appendix A of the supplementary material. More information can be found on the M-Flow website[1] and other papers describing the tool (V. H. Minces et al., 2023; Song et al., 2023).

2.2 Participants and Data Description

To investigate the automated assessment of primary students' creativity within a flow-based music programming environment, data were gathered from the M-Flow platform. Participants included 208 upper-primary students in fourth grade, attending a school that predominantly serves Latinx students (88%). This school also supports a large proportion of underrepresented and

[1] https://mflow.sciencemusic.org/

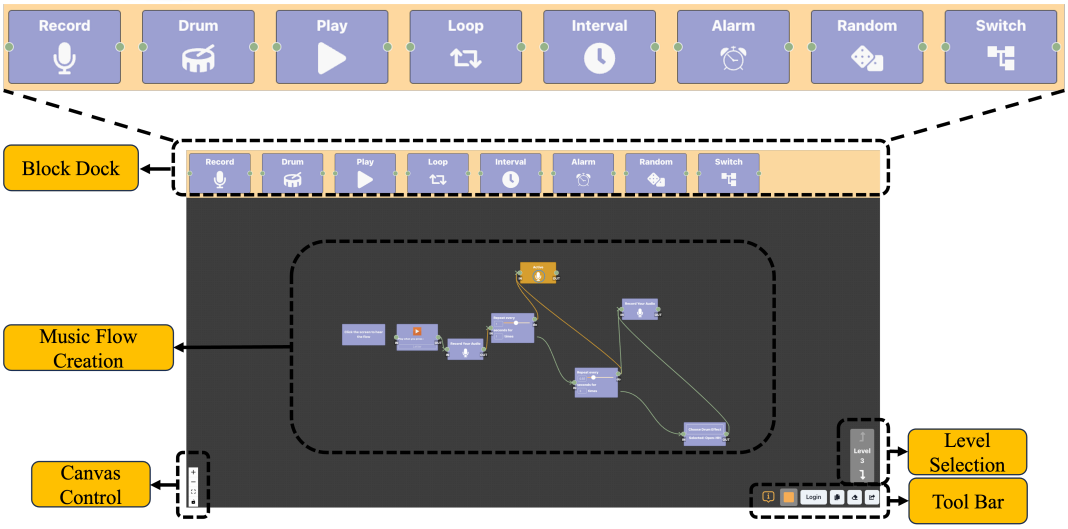


Figure 2. *The M-Flow interface.*

low-income students, with 48% identified as English language learners (ELLs) and 74% qualifying for free or reduced-price meals. The study was approved by the Institutional Review Board (IRB) of the University of Florida.

The classroom implementation was integrated into a science class, where teachers introduced the course background and instructed students on using the M-Flow platform to create music compositions. In these classroom settings, students engaged with the M-Flow platform and were tasked with creating sound compositions designed to convey specific emotions. Data collection spanned three years, from 2022 to 2024, with one or two rounds of classroom implementation conducted annually. This process resulted in programs created by 194 unique students (excluding 14 students who did not create any programs). The dataset includes a total of 383 programs (flows) and 6,002 blocks used within these compositions. The distribution of programs across the three years is as follows: 66 in 2022, 158 in 2023, and 159 in 2024.

All programs were stored as JSON files that show the nodes and connections that form a program. The JSON includes all the student IDs, the flows with IDs created by the students, and the elements (blocks and arrows) in each program (see supplementary material, Appendix B). Appendix Figure 1 showcases examples of student-created artifacts that can be retrieved through unique identifiers (UIDs) and flow IDs recorded in the M-Flow platform for manual inspection and further analysis. Through the JSON files and retrieved visual programs on M-Flow, this study seeks to explore automated methods for analyzing students' creativity within a flow-based music programming environment. A simplified JSON data sample is included in the supplementary material, Appendix B.

2.3 Theoretical Foundation for ECD-Based Assessment

To measure students' creativity in a flow-based music programming environment, we used the ECD framework (R. G. Almond et al., 2015). The ECD framework is composed of three interrelated models. The competency model outlines the "unobservables," or competencies that are being assessed, as well as their sub-facets and interrelationships. The evidence model (EM) defines observable indicators that demonstrate competency (R. Almond et al., 2020). It also incorporates evidence criteria, such as automated grading rubrics, and statistical models for data aggregation. The task model (TM) identifies the tasks or learning contexts that are intended to elicit the evidence specified by the EM. By incorporating these models into a digital learning environment, users are constantly producing evidence that reflects their knowledge and skills (Shute et al., 2020; Rahimi et al., 2023). The ECD framework can facilitate the development of psychometrically sound assessments while assuring validity, reliability, and fairness (R. G. Almond et al., 2015; Rahimi et al., 2024).

Based on the ECD framework, we developed the creativity competency and EMs within the M-Flow platform (shown in Figure 3). Our approach operationalizes the creativity competency model, drawing inspiration from the construct of divergent thinking (Guilford, 1956), which serves as a proxy for creativity or creative potential through individuals' responses (Runco, 2014). This model evaluates creativity along the dimensions of fluency, flexibility, originality, and elaboration. Additionally, we integrated complexity, supported in the literature as a critical component of creative products (Amabile, 1982; Rahimi, 2023). Furthermore, we incorporated an analysis of efficiency—encompassing time, code, and music efficiency—alongside students' emotional expressiveness, particularly in the context of tasks requiring music creation to convey emotion. Creativity and efficiency are inherently intertwined, as creative solutions often involve the discovery of more efficient methods for problem-solving or achieving desired outcomes (Williamon et al., 2006; Runco, 2014; Colton & Wiggins, 2012). Emotion, as an influential factor in creativity (Baas et al., 2008; D. P. Newton, 2013), is also included in the model as students' given

task is using the M-Flow tool to create compositions to express emotion. To quantify students' emotion expression and their relationship with creative expression during programming tasks, we employed the dimensional emotion model (Russell, 1980), which offers a continuous framework for quantifying emotions based on valence and arousal. Valence refers to the intrinsic positivity or negativity of an emotion, ranging from unpleasant (e.g., sadness, anger) to pleasant (e.g., happiness, excitement). Arousal represents the level of physiological and psychological activation associated with an emotion, ranging from low arousal (e.g., calmness, relaxation) to high arousal (e.g., excitement, agitation). The emotional expressiveness aspect helps capture the nuanced interplay between emotion and creative outputs (Amabile, 1983).

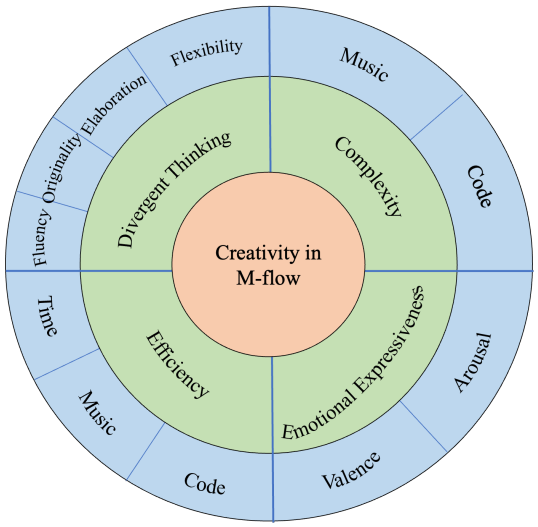


Figure 3. *The creativity competency and EMs in M-Flow.*

Note: This creativity competency model and its associated EMs consist of four aspects and 11 variables. Divergent thinking is represented by four variables, complexity by two variables, efficiency by three variables, and emotional expressiveness by two variables. The TM is based on the flow-based learning environment M-Flow, where students use the platform to create compositions aimed at expressing emotions.

2.4 ECD-Based Automated Assessments

This section introduced how we implement the automated creativity assessment based on the creativity and EMs in M-Flow (shown in Figure 3). Based on the theoretical foundation above, we identified observable indicators for each construct mentioned. Figure 3 illustrates the four constructs with 11 indicators used to analyze creativity in a flow-based programming environment.

2.4.1 Divergent Thinking

Inspired by the work of Rahimi and colleagues (2024), divergent thinking was operationalized through computational measures of fluency, flexibility, originality, and elaboration in this study. Grounded in Guilford (1956)'s conception of fluency as idea generation, fluency was quantified by counting the number of sound blocks (i.e., *audioNode*, *drumNode*) and tracks used in each program. Each distinct block or track is treated as a discrete "musical idea," providing an automatable proxy for idea generation within a flow-based programming context. For instance, the programs in Appendix Figure 1 contain 4, 5, 9, and 5 sound blocks, and 2, 1, 6, and 1 tracks, respectively.

Originality was assessed by comparing each program to others using a directed graph representation. Each flow is represented as a directed graph, with nodes characterized by their types (e.g., audioNode, drumNode). The algorithm first calculates the distribution of node types within each flow and measures their similarity to other flows using cosine similarity, capturing component-level resemblances. To assess structural similarities, it identifies all simple paths in the directed graph for each flow and determines the longest path shared between any two flows, normalized by the number of nodes in the compared graphs. An overall similarity score is then calculated for each pair of flows by averaging the node type distribution similarity and the normalized longest common path. The originality of a flow is inversely proportional to its average similarity with all other flows, computed as $(1 - \text{Average Similarity})$, where lower similarity to other flows implies higher originality.

Elaboration was evaluated by calculating the length of each program, including audio length and program length. The audio length considers the durations of audioNode and drumNode audio elements and applies loop multipliers from loopNode and intervalNode metadata to account for structural repetitions. For audio nodes (audioNode), the duration is computed by subtracting the `startTime` from the `stopTime` (both provided in milliseconds), shown in the script in Section 2.5, and converting the result into seconds. For drum nodes (drumNode), a default duration of 1 second is assigned because the drum

audio clips last approximately 1 second. For the program length, the number of blocks in the program was calculated; for example, in Appendix Figure 1, the program lengths are 7, 7, 11, and 7, respectively. Similarly, in the work of Rahimi and colleagues (2024), the elaboration was also quantified by the length of a project in seconds and measures.

Finally, the final flexibility score is computed as the total count of unique block types in the flow, providing a measure of the diversity in the student's programming artifacts. Each flow is represented in a JSON structure (see Appendix B in the supplementary material), where "elements" describe the individual nodes and their attributes. To determine flexibility, the algorithm processes the elements of the flow, extracting the "type" attribute from each node. This attribute specifies the block's functionality (e.g., audioNode, drumNode, loopNode, etc.). Only unique block types are considered. Connections between nodes, which are characterized by "source" and "target" attributes, are excluded from the computation as they do not constitute distinct block types. A higher score indicates greater flexibility, showcasing the student's ability to employ a broader range of elements within the programming environment.

2.4.2 Complexity

For complexity, we included both musical complexity and code complexity. Musical complexity consists of rhythmic and beat complexity, and code complexity is measured by the project's breadth and average depth. To calculate musical complexity, we generated audio files corresponding to each program by constructing and exporting composite audio based on the logical structure and attributes of elements within a flow. This process involved representing the flow data as a directed graph, where nodes correspond to elements such as audioNode, drumNode, loopNode, and intervalNode, while edges represent the connections between these elements. Audio content associated with recordNode and drumNode was sourced either from external URLs or from local audio files. The flow traversal began at root nodes, defined as nodes without incoming edges. In cases where no explicit root node was identified, a default node was randomly selected. The traversal method recursively constructed composite audio by summing and concatenating audio segments from downstream nodes. For nodes involving looping operations, a multiplier was applied to extend the duration of audio segments proportionally. If a node lacked valid audio content, a silent audio segment of equivalent duration was substituted to maintain the logical flow. The resulting composite audio for each flow was then exported as an MP3 file, enabling the subsequent analysis of rhythmic and beat complexity.

The rhythmic complexity is calculated using entropy. Thul (2008) noted entropy as a potential measure of a musical sample's rhythmic complexity, and Rahimi and colleagues (2024) also implemented entropy to measure rhythmic complexity in a music programming environment. We followed the example set by Thul (2008) and Rahimi and colleagues (2024) and used entropy as a measure of rhythmic complexity. Entropy measures the uncertainty in a given distribution and has been found to correlate with human perception of rhythm (de Fleurian et al., 2017; Pearce, 2018). Nagaraj and Balasubramanian (2017) also demonstrated that entropy captures a key dimension of complexity across various sequences and correlates with distinctions between ordered and chaotic patterns. We computed entropy for each clip within a measure (a unit of musical time divided into beats) and then averaged the results to obtain an entropy score per measure. A higher score indicates greater uncertainty in that distribution.

Cohesion in this context is calculated as a measure of the rhythmic regularity within an audio composition, specifically through the variability of intervals between detected beats. The process begins by loading the audio file and converting it into a numerical waveform representation using `librosa`, a Python library for analyzing and processing audio signals. The algorithm then detects beats through `librosa.beat.beat_track`, which identifies the temporal positions of beats based on the tempo and specified tracking parameters. These beat positions, initially represented as frame indices, are converted into time values to calculate the intervals between consecutive beats. Cohesion is then determined as the standard deviation of these intervals, with lower values indicating greater rhythmic consistency and higher values reflecting irregularity in the beat structure.

The code complexity is calculated using breadth and average depth. The calculation of breadth and average depth aims to quantify the structural complexity of a flow within the given programs. The process begins by extracting elements from the flow and constructing a directed graph. Nodes represent distinct components of the flow, while edges represent connections between these components. To calculate breadth, the algorithm identifies the starting nodes and performs a breadth-first search (BFS) traversal. The breadth at each level of the graph is defined as the number of nodes being processed at that level. The maximum breadth observed during the traversal is recorded as the breadth of the flow. Average depth measures the typical path length from the starting nodes to the terminal nodes (nodes without successors). During the BFS traversal, the depth of each terminal node is accumulated, and the number of terminal nodes is counted. The average depth is computed by dividing the total depth by the number of terminal nodes.

2.4.3 Efficiency

Efficiency was assessed by calculating students' time efficiency, as well as the music and code efficiency within their programs. Time efficiency evaluates the rate at which users construct flows within the programming environment, serving as a measure of productivity and engagement. This metric is derived by calculating the number of blocks created per class during the flow's

development. Specifically, the total time spent on constructing each flow is calculated as the difference between its creation timestamp (`createdAt`) and its last modification timestamp (`updatedAt`), shown in the JSON example script in Section 2.5. If this difference exceeds 45 minutes, 45 minutes is used as the total time spent on the program. This approach accounts for the possibility that students may modify their programs at any point after their initial creation, leading to time gaps that could span months or even years. Standardizing the time to a single class period effectively reduces such discrepancies. The number of blocks is determined by counting the entries in the `elements` field of the flow. Efficiency is then calculated as the ratio of the number of blocks to the total time, yielding a value that reflects the pace of block creation. It is important to note that students can freely add, delete, and modify blocks throughout the programming process. Due to data limitations, we do not have access to the full edit history or detailed logs of these intermediate steps. Therefore, our efficiency analysis should be interpreted as an outcome-based proxy.

Music efficiency integrates multiple dimensions of musical and temporal characteristics, including cohesion, entropy, and audio duration, to assess the efficiency of music flows. Cohesion and entropy data were calculated previously when computing the music complexity. For each music flow, the audio file corresponding to its identifier was loaded and its duration was extracted using the `librosa` library in Python. Then the music efficiency metric was defined as

$$\text{music efficiency} = \frac{\text{cohesion} + \frac{1}{\text{entropy}}}{\text{audio duration}}, \tag{1}$$

where inverse entropy ($1/\text{entropy}$) captures rhythmic variation and cohesion reflects the consistency of beat intervals. The resulting metric was normalized and aggregated across flows to provide a scalable and comprehensive measure of how cohesively and efficiently a flow utilizes musical and temporal resources. The proposed formula for music efficiency aims to capture the balance between structural unity (cohesion), creative complexity (inverse entropy), and temporal efficiency (audio duration) to provide a holistic metric to evaluate the effectiveness of music compositions in flow-based programming environments.

Code efficiency is determined through four key components. First, isolated blocks are identified via graph traversal, with penalties applied for any nodes disconnected from the overall structure. Second, the position and connectivity of playNodes are evaluated to ensure that they appear early in the flow and maintain appropriate downstream connections to audioNodes or drumNodes. Third, link complexity is calculated as the ratio of the total number of edges to the total number of blocks, providing a measure of structural compactness. Lastly, specific node properties, such as the presence of valid blobURLs for audioNodes or loop settings for loopNodes, are validated to ensure functional integrity. Any violation of these criteria results in a one-point deduction from the overall efficiency score.

2.4.4 Emotional Expressiveness

The emotional expressiveness of music flows was assessed using an open-source deep learning–based speech emotion recognition model built on the Wav2Vec2 framework (Wagner et al., 2023). Each audio file generated from the music flows was processed to extract the key emotional dimensions: *arousal* and *valence*. The model expects a raw audio signal as input and outputs predictions for arousal and valence in a range of approximately 0 to 1.

The predictions for *arousal* (reflecting energy or intensity) and *valence* (positivity or negativity) were recorded for each flow. The two dimensions collectively represent the emotional expressiveness of the music, capturing subtle nuances in how emotions are conveyed through sound. Emotional expressiveness is quantified using the arousal and valence dimensions from the dimensional emotion model (Russell, 1980). The calculation combines these two dimensions into a single scalar value by computing the Euclidean distance from the origin in a two-dimensional emotional space, as follows:

$$\text{emotional expressiveness} = \sqrt{\text{arousal}^2 + \text{valence}^2}. \tag{2}$$

In conclusion, this study employed Python to extract and compute variables (observables) from the JSON data, which were subsequently used to facilitate our analyses. The aggregation of ECD-based overall creativity estimates for each program followed a multi-step process. Initially, low-level indicators, illustrated as the blue area in Figure 3, were standardized. These standardized indicators were then averaged to calculate the sub-facet variables, including fluency, flexibility, originality, elaboration, music complexity, code complexity, time, musical efficiency, code efficiency, valence, and arousal. Next, divergent thinking and complexity scores were computed by averaging their respective sub-facet variables. For instance, fluency, flexibility, originality, and elaboration were averaged to derive each student's divergent thinking score, while music and code complexity indicators were combined to calculate the complexity score. Subsequently, the overall ECD-based automated creativity score for each program was obtained by averaging the divergent thinking score, complexity score, efficiency metrics, and emotional expressiveness. Table 2 provides detailed definitions for all the variables used to compute the creativity sub-facets. Both the

sub-facet variables (e.g., fluency, flexibility, complexity) and the overall automated creativity scores were employed to address the research questions posed in this study.

Table 2. *The variables calculated for analysis.*

Category	Variable name	Description
Fluency	Number of sounds	Number of sound blocks per program
	Number of tracks	Number of tracks per program
Originality	Avg. similarity	Average similarity between a program and other programs
Elaboration	Length (audio)	Audio duration in a program
	Length (block)	Total number of blocks in a program
Flexibility	Unique block	Number of unique blocks per program
Music complexity	Entropy	The uncertainty in a beat track
	Cohesion	Similarity between two given click tracks
Code complexity	Breadth	Max number of nodes at each level of the program
	Avg. depth	Average path length from start nodes to end nodes
Time efficiency	Time score	Number of blocks created per unit time
Music efficiency	Music score	Defined by cohesion, entropy, and audio duration
Code efficiency	Code score	Calculated through graph structure analysis
Valence	Valence	Measures the pleasantness or positive/negative emotion expressed
Arousal	Arousal	Indicates the intensity or energy level of the emotional expression

2.5 LLM-Based Automated Assessment

The LLM-based automated assessment uses an LLM to evaluate students' music-programming flows based on predefined creativity dimensions and illustrative examples. In this study, ChatGPT-4o was selected as the LLM because it is the most advanced GPT model and has demonstrated exceptional performance in analyzing textual data and programming tasks (OpenAI et al., 2024). The assessment process begins by extracting and simplifying each student's programming flow into a standardized JSON format. This simplification isolates nodes and their connections while excluding irrelevant metadata (e.g., timestamps) to address the token limit of the ChatGPT API (application programming interface)[2]. Each flow is represented only by its key components—such as node types (`audioNode`, `playNode`) and their interconnections—ensuring that the essential structural and functional details are retained.

The ChatGPT-based assessment evaluates creativity along four predefined dimensions: *divergent thinking*, *complexity*, *efficiency*, and *expressiveness*. A structured prompt guides the evaluation by providing operational definitions for each dimension together with a 1 to 6 rating scale. Additionally, the prompt incorporates two selected few-shot examples, one representing high creativity and one representing low creativity (chosen so that their scores lie within one standard deviation above or below the dataset mean; see Table 3). The GPT model outputs both a creativity score and a detailed rationale. The exact prompt used for generating the creativity score is given below (placeholders indicate where the target JSON and example flows are inserted):

```
``I want to evaluate the creativity of music programming flows. Each flow consists of multiple nodes
and their connections (arrows), representing a student's creation.
The evaluation is based on four dimensions:
1. Divergent Thinking: Measures the diversity of the creation.
2. Complexity: Evaluates the structural sophistication.
3. Efficiency: Assesses how efficiently the flow achieves its purpose.
4. Expressiveness: Examines the degree of personal style or innovation.
Each dimension is scored on a scale of 1 to 6.
Provide an overall score (1 to 6) and justification for the score.
Here are some example evaluations:
{examples_text}

Evaluate the following flow:
Flow ID: `{flow_id}`
Simplified JSON:
{json.dumps(simplified_flow_data, indent=4)} ".
```

To address concerns about the robustness of LLM scoring, two ablation analyses were conducted (see details in supplementary material Appendix D):

1. Alternative-example sets: We replaced the original few-shot pair with five high/low combinations drawn from the same creativity strata (high = 3.6–5.0; low = 1.2–1.8) and re-scored 10 target programs in the GPT_MANUAL condition. Cronbach's α was 0.92, indicating stable scores across different anchor pairs.

[2] https://platform.openai.com/docs/models/gpt-4o

2. Test–retest with fixed examples: We repeated the evaluation five times with the original anchors, yielding Cronbach's $\alpha = 0.98$; this indicates very high within-run stability.

We implemented two ChatGPT-based assessment conditions: GPT_ECD, which used anchors derived from ECD assessments (scores 4.2 and 2.3), and GPT_MANUAL, which used anchors based on human ratings (scores 3.8 and 1.8). Appendix C in the supplementary material shows an example of the response obtained from the LLM in the GPT_ECD condition.

2.6 Manual Assessment

To validate and compare automatic assessment approaches with human evaluations, this study employed CAT (Amabile, 1982) to manually assess students' flow-based music programming artifacts ($n = 383$). CAT is a widely recognized method for assessing creativity, and its central idea is that creativity, while subjective, can be reliably assessed by a group of experts within a given domain (Amabile, 1982; Baer & McKool, 2009). Five educational experts, all of whom possessed extensive familiarity with the flow-based music environment and had been working on related projects, were tasked with rating each program on a scale from 1 to 6 (1 = very uncreative, 6 = very creative). According to Amabile (1996), a panel of five to 10 experts is considered sufficient for reliably scoring creativity.

During the review process, the experts were provided with links to these programs, allowing them to click and view the program content on the M-Flow platform (as illustrated in Appendix Figure 1) and play the audio within each program. Drawing upon prior research (Rahimi et al., 2024), we provided the raters with specific instructions (see Appendix E).

Assessing creativity or originality is often regarded as a normative task rather than a purely objective one (Organisciak et al., 2023). This perspective suggests that, given a sufficient number of raters, a consensus on originality ratings typically emerges, even though individual judgments may vary, particularly when distinguishing between highly and moderately original responses. While creativity assessment inherently involves subjectivity, the CAT enhances both the reliability and the validity of the evaluation process. In this study, inter-rater reliability was measured using Cronbach's alpha ($\alpha = 0.81$), indicating a high level of internal consistency among raters. The creativity score for each program was derived by averaging the five individual ratings, with the resulting value rounded to the nearest 0.1.

2.7 Comparison of the Three Assessment Approaches

To validate the two automated creativity assessment approaches and compare them with the CAT-based assessment, we first analyzed the mean, standard deviation, and score distributions generated by the ECD-based, LLM-based (GPT_ECD and GPT_MANUAL), and CAT-based methods. These distributions were visualized using density plots and histograms to provide a clear comparative perspective. Pearson correlation coefficients were then calculated to examine the relationships between the scores generated by the ECD-based and LLM-based methods and those assigned by individual raters on their averaged scores. This analysis aimed to assess the validity of the automated methods in comparison to the CAT-based approach. Additionally, a one-way analysis of variance (ANOVA) was conducted to determine whether significant differences existed among the scores produced by the three methods.

3. Results

3.1 Description of the Variables

Table 3 shows the key descriptive results of calculated variables. For the four types of creativity measurements (i.e., ECD-based, GPT_ECD, GPT_MANUAL, and Manual), the ECD-based creativity score, derived from the above variables, had mean = 3.20, SD = 0.89. GPT_ECD creativity, scored using two ECD-based examples for learning, had mean = 2.68, SD = 0.80. GPT_MANUAL creativity, scored using two manually rated examples for learning, had mean = 3.06, SD = 0.87. Manual creativity, rated using CAT, had mean = 2.68, SD = 0.88.

3.1.1 ECD-Based Automated Assessments Results

Figure 4 shows the structural relationships between various variables contributing to ECD-based creativity. It can be observed that most variables represented by blue ovals show moderate to high correlation coefficients ($r = 0.24$ to 0.93) with the following four variables: divergent thinking, efficiency, complexity, and emotional expressiveness. The correlation coefficients between these four variables and the final ECD-based creativity score are 0.31, 0.74, 0.39, and 0.68, respectively. To verify whether the four variables—divergent thinking, efficiency, complexity, and emotional expressiveness—used to measure ECD-based creativity are consistent with manual-based creativity, we calculated their correlation coefficients with manual creativity scores. The results are 0.36, 0.29, 0.53, and 0.07, respectively. Furthermore, the correlation between manual-based creativity and ECD-based creativity is $r = 0.48$, $p < 0.001$, indicating a moderate relationship between the two.

Table 3. *Descriptive statistics.*

Variable	Min	Max	Mean	SD
Number of sounds	1.00	20.00	5.66	3.34
Number of tracks	1.00	35.00	2.57	3.02
Originality score	0.57	0.94	0.63	0.06
Length (audio)	0.71	300.00	39.34	75.35
Length (block)	1.00	22.00	7.56	3.82
Entropy	0.56	2.42	1.38	0.38
Cohesion	0.14	0.70	0.24	0.13
Breadth	1.00	17.00	1.98	1.71
Avg. depth	1.00	19.00	4.79	3.02
Time efficiency	0.02	1.02	0.32	0.18
Music efficiency	0.01	4.95	0.58	1.00
Code efficiency	3.00	4.61	3.48	0.29
Valence	0.12	0.88	0.55	0.15
Arousal	0.27	0.97	0.63	0.18
Divergent thinking	0.00	1.00	0.45	0.17
Complexity	0.00	1.00	0.47	0.16
Efficiency	0.00	1.00	0.58	0.16
Emotional expressiveness	0.00	1.00	0.47	0.23
ECD-based creativity	1.00	6.00	3.20	0.89
GPT_ECD creativity	1.00	4.25	2.68	0.80
GPT_MANUAL creativity	1.00	5.40	3.06	0.87
Manual creativity	1.00	5.40	2.68	0.88

Note: The variables were normalized before being aggregated into higher-level variables. SD represents standard deviation. Avg. means average.

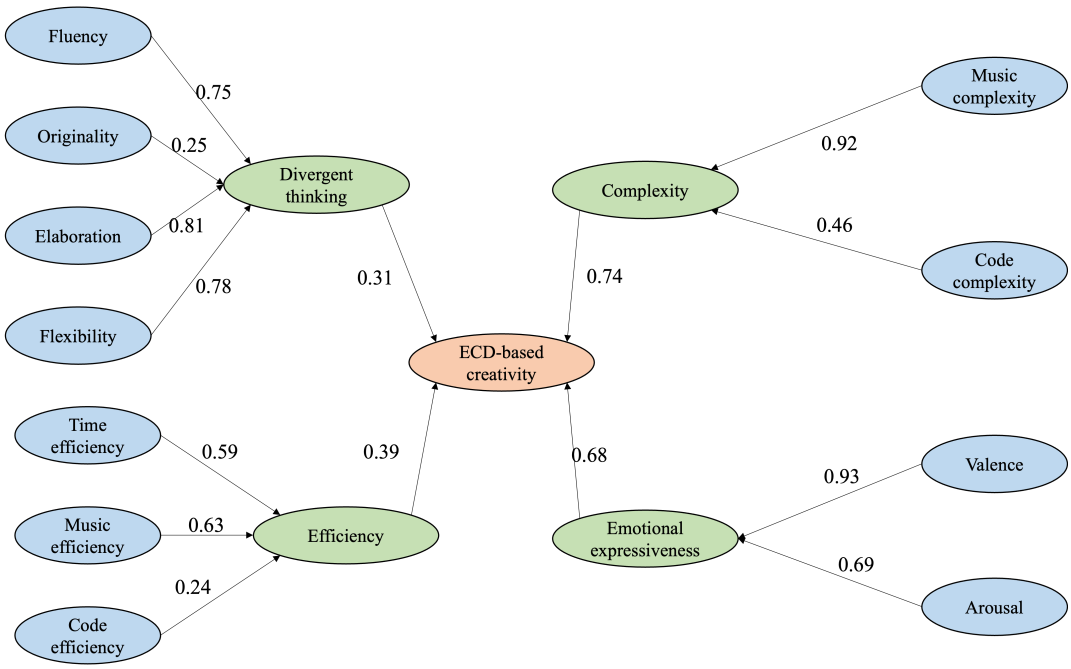


Figure 4. *ECD-based creativity dimensions and variables correlations.*

Note: The figure illustrates the correlations among the variables in the ECD-based creativity assessment. Each variable demonstrates a moderate or strong correlation with its corresponding higher-level variable.

3.1.2 LLM-Based Automated Assessment Results

In this study, the GPT-4o API was employed to perform few-shot learning for evaluating the creativity of all 383 student programs. Two distinct GPT-based assessments were conducted: the GPT_ECD assessment used examples derived from the ECD evaluations (scores of 1.9 and 3.5), while the GPT_MANUAL assessment used examples from human raters (scores of 3.6 and 4.4). Notably, the same pair of examples was employed across both assessments to ensure consistency in the learning context.

To evaluate the similarities and differences between the two ChatGPT-based assessment methods, we first calculated the Pearson correlation coefficient, which yielded $r = 0.81$, $p < 0.001$, indicating a strong positive correlation between the GPT_ECD and GPT_MANUAL scores. To further compare the two methods, the creativity scores were categorized into three tertile-based groups: low ($0 <$ score < 0.33), medium ($0.33 <$ score < 0.66), and high ($0.66 <$ score < 1) creativity. The GPT_ECD assessment classified 152 programs as low, 105 as medium, and 126 as high. Similarly, the GPT_MANUAL assessment grouped 151 programs as low, 115 as medium, and 117 as high, demonstrating a comparable distribution with minor variations in group sizes. Figure 5 presents the cross-tabulation of group classifications, offering a detailed view of the consistency between the two methods. Specifically, 31.19% of the programs were consistently classified as low-low, 12.37% as medium-medium, and 19.07% as high-high. In total, approximately 62.63% of the programs received consistent classifications across both methods, underscoring a strong level of agreement between the GPT_ECD and GPT_MANUAL assessments.

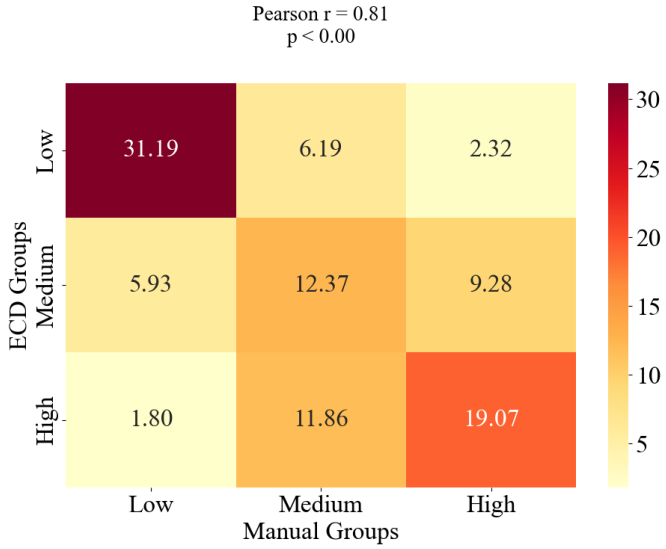


Figure 5. *Distribution of group combinations (%).*

Note: The diagonal represents the overlap between the two GPT-based models in classifying programs into the three categories: low, medium, and high score.

3.1.3 Manual Assessment

Table 4 presents the descriptive statistics for each expert, including the minimum, maximum, mean, and standard deviation of their scores.

Table 4. *Descriptive statistics of scores by each evaluator.*

Evaluator	Min	Max	Mean	SD
Evaluator 1	1	6	2.45	0.98
Evaluator 2	1	6	3.34	1.66
Evaluator 3	1	5	2.60	1.06
Evaluator 4	1	6	3.10	1.09
Evaluator 5	1	6	2.82	1.18

Note: Min represents the minimum score, Max the maximum score, and SD the standard deviation of the scores assigned by each evaluator.

The final manual creativity assessment score represents the average of ratings provided by five expert evaluators. As shown in Figure 6, the distribution of overall creativity scores ranges from 1 to 5.4, with a mean of 2.70 and a standard deviation (SD)

of 0.88. The overall ratings indicate a central tendency toward moderate levels of creativity.

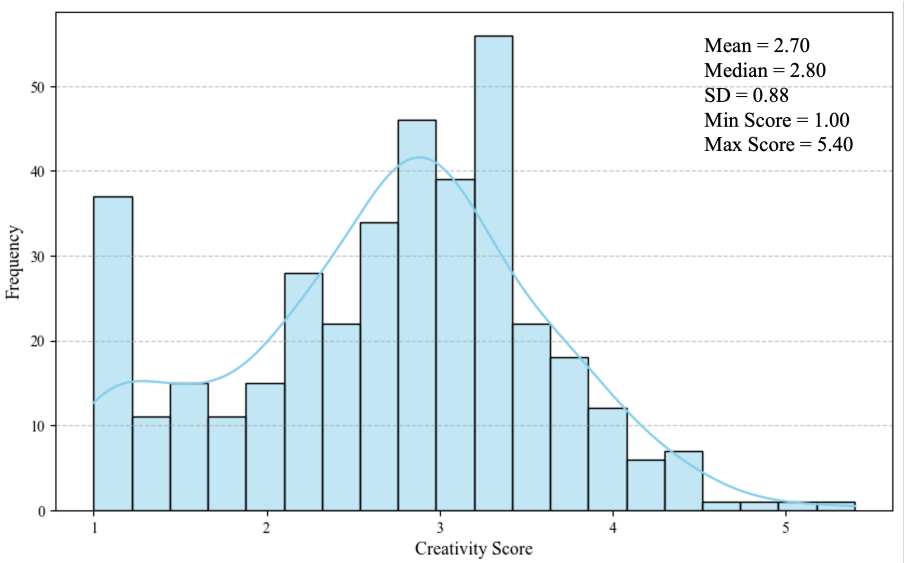


Figure 6. *Distribution of manual creativity score results.*

3.2 Comparison of Assessment Results

3.2.1 Mean Scores

Table 5 summarizes the results of Tukey's HSD post hoc tests. ECD yielded significantly higher scores than both GPT_ECD and Manual, while GPT_MANUAL was not significantly different from ECD. The largest divergence appeared between GPT_MANUAL and GPT_ECD.

Table 5. *Pairwise comparisons of mean creativity scores (Tukey's HSD).*

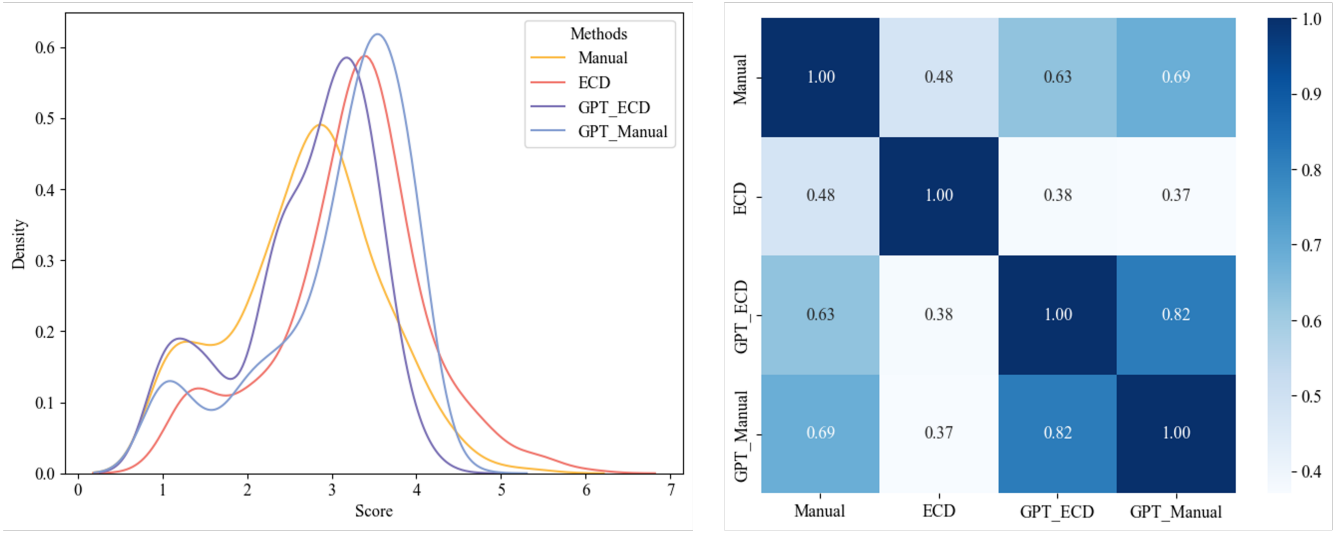
Comparison	Mean Difference (ΔM)	p-value	Significance
ECD vs. GPT_ECD	0.53	< 0.001	Significant
ECD vs. Manual	0.52	< 0.001	Significant
ECD vs. GPT_MANUAL	0.14	0.12	Not significant
GPT_ECD vs. GPT_MANUAL	0.38	< 0.001	Significant
GPT_ECD vs. Manual	0.0036	1.000	Not significant
GPT_MANUAL vs. Manual	0.38	< 0.001	Significant

3.2.2 Density Distributions

We further visualized the score density of the four assessment methods to explore differences and similarities in their distributions. For comparability, the ECD scores were scaled to a range of 1 to 6. The left panel of Figure 7 presents the density distributions of creativity scores across the four methods: Manual, ECD, GPT_ECD, and GPT_MANUAL. The ECD method exhibits a right-skewed distribution, with the majority of scores concentrated between 3 and 4. Similarly, GPT_MANUAL demonstrates the highest density around a score of 3.5. The GPT_ECD distribution closely mirrors the ECD distribution but shows a slightly lower majority of scores (around 3). Lastly, the Manual method displays a peak at a lower score than the other three methods and also exhibits a reduced density in the mid-score range, relatively. We calculated the Pearson correlation coefficients among the four assessment methods. The right panel of Figure 7 presents a heatmap illustrating these correlations.

3.2.3 Score Difference

Figure 8 presents the absolute score difference (ASD) between the ECD and Manual assessment methods (upper panel, represented by blue points) and between the two GPT-based assessments (lower panel, represented by red points). The x-axis represents the 383 programs, while the y-axis indicates the ASD values. The mean score difference between ECD and Manual is 0.84 (SD = 0.61), with a maximum difference of 4.26 and a minimum of 0.01. Similarly, the mean score difference between GPT_ECD and GPT_MANUAL is 0.48 (SD = 0.42), with a maximum difference of 2.58 and a minimum of 0.



(a) Score Density Comparison

(b) Score Correlations

Figure 7. *Distribution and correlation of different creativity assessment scores.*

We further calculated the number of programs with ASD exceeding thresholds of 1, 2, and 3 for these methods. Table 6 shows that, when comparing Manual and ECD, there are 128 programs (33.33%) with ASD $>$ 1, 22 programs (5.69%) with ASD $>$ 2, and only 1 program (0.27%) with ASD $>$ 3. For the comparison between GPT_MANUAL and GPT_ECD, the differences are smaller, with only 27 programs (6.96%) having ASD $>$ 1, three programs (0.77%) having ASD $>$ 2, and no programs showing ASD $>$ 3. These results are also reflected in Figure 8.

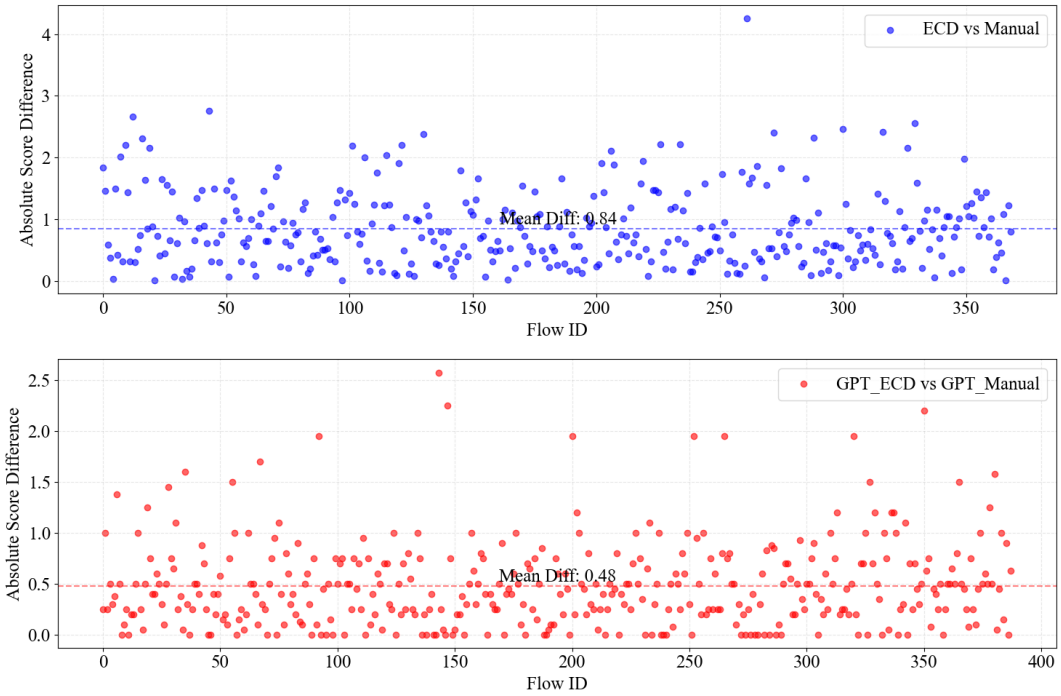


Figure 8. *ASD between different assessments.*

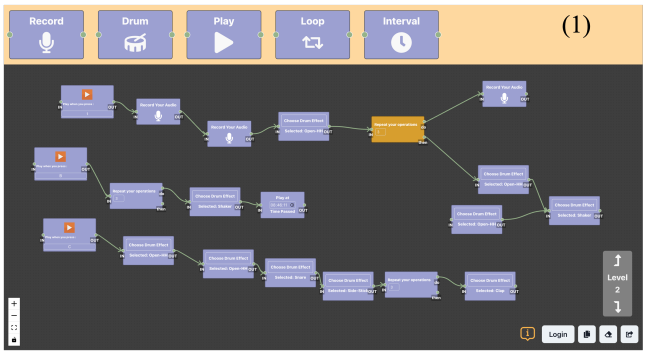
Note: The figure shows the absolute score difference (ASD) between the ECD and manual assessment methods (upper) and between the two GPT-based assessments (lower).

Table 6. *Summary of score differences for ECD vs. Manual and GPT_ECD vs.* GPT_MANUAL

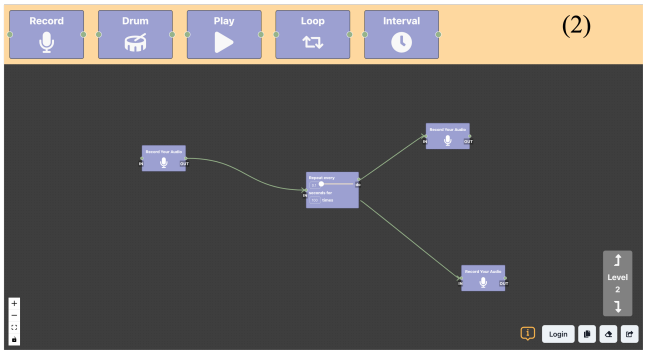
Comparison	Difference	Sample (Proportion)
ECD vs. Manual	ASD >1	128 (33.33%)
	ASD >2	22 (5.69%)
	ASD >3	1 (0.27%)
GPT_ECD vs. GPT_MANUAL	ASD >1	27 (6.96%)
	ASD >2	3 (0.77%)
	ASD >3	0 (0%)

3.2.4 Example Comparison

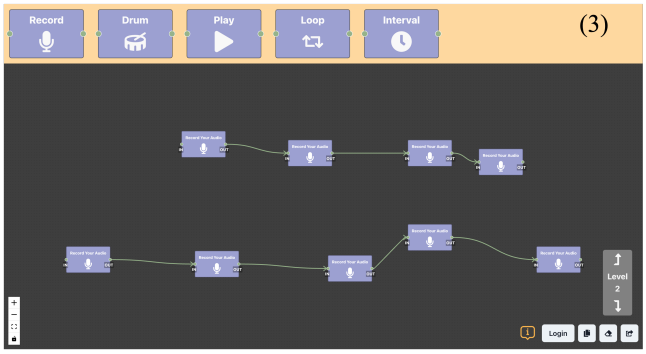
To gain deeper insights into the creativity measurement methods, we selected four programs as examples: two with very similar scores and two with very different scores from the four assessment methods. In Figure 9 (1), the program received the highest creativity score from the ECD method (score = 3.60) and the lowest score from GPT_ECD (score = 3.40), with a difference of 0.2. According to the ECD method, the program was rated as follows: divergent thinking (0.38), complexity (0.30), efficiency (0.53), and emotional expressiveness (0.87). In contrast, GPT_ECD rated the program as divergent thinking (0.50), complexity (0.36), efficiency (0.58), and emotional expressiveness (0.50). The difference stems primarily from the emotional expressiveness dimension. In Figure 9 (2), the program achieved the highest creativity score from the ECD method (score = 3.18), with ratings of divergent thinking (0.16), complexity (0.54), efficiency (0.30), and emotional expressiveness (0.74). Meanwhile, the lowest score was given by the manual method (score = 2.80), with individual evaluators assigning scores of 2, 3, 3, 3, and 3, respectively. In Figure 9 (3), the highest creativity score came from the ECD method (score = 5.18), while the lowest score was assigned by GPT_ECD (score = 1.55), resulting in a difference of 3.63. The ECD method rated the program as divergent thinking (0.58), complexity (0.95), efficiency (0.87), and emotional expressiveness (0.94). In comparison, GPT_ECD rated the program as divergent thinking (1.0/6), complexity (1.0/6), efficiency (3.0/6), and expressiveness (1.0/6), with a total score of 1.5/6. The differences were substantial across all four dimensions, particularly in complexity and emotional expressiveness. Finally, in Figure 9 (4), the highest score was provided by the manual method (score = 5.4, with individual ratings of 6, 6, 5, 4, and 6), while the lowest score was assigned by the ECD method (score = 2.27), with ratings of divergent thinking (0.33), complexity (0.30), efficiency (0.26), and emotional expressiveness (0.13).



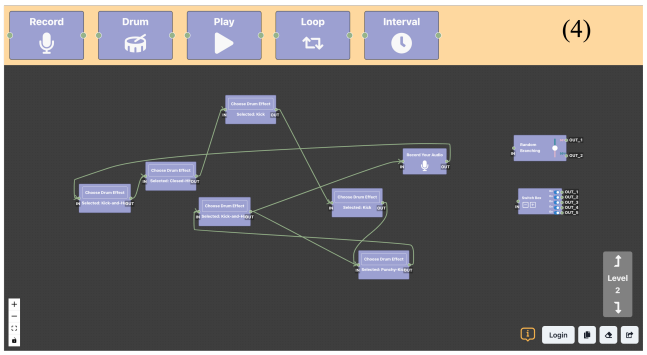
Highest Score: ECD (3.60), Lowest Score: GPT_ECD (3.40), Difference: 0.20



Highest Score: ECD (3.18), Lowest Score: Manual (2.80), Difference: 0.38



Highest Score: ECD (5.18), Lowest Score: GPT_ECD (1.55), Difference: 3.63



Highest Score: Manual (5.40), Lowest Score: ECD (2.27), Difference: 3.13

Figure 9. *Distribution and correlation of different creativity assessment scores.*

4. Discussion

This study investigates automated methods for assessing students' creativity in a flow-based music programming environment. Specifically, it addresses the following question: *How can students' creativity be automatically assessed within a flow-based music programming environment?* To answer this, we collected students' artifacts from a flow-based music programming platform, encompassing 383 projects created by 194 students. Two automated assessment approaches were proposed and employed: one grounded in an ECD framework and the other using ChatGPT. Both methods were compared against human manual creativity assessments to evaluate their validity and reliability. The findings demonstrate the feasibility of using automated tools to assess creativity in educational programming environments, significantly reducing the time and resources required for traditional assessment methods. The main contributions of this study include (1) proposed and implemented automated methods for assessing students' creativity in a flow-based music programming environment, bridging the gap between traditional assessment approaches and advanced computational techniques, and (2) evaluation of the validity and reliability of these methods by comparing them against human manual assessments, demonstrating their potential to provide accurate and consistent results. In the following sections, we discuss the the results and the implications of automated creativity assessment.

4.1 Effectiveness of the ECD-Based Creativity Assessment

The ECD framework evaluates student creativity along four programmatic dimensions (divergent thinking, complexity, efficiency, and emotional expressiveness). Divergent thinking and complexity have long been recognized as cornerstones of creative performance in programming contexts (Rahimi et al., 2024; Guilford, 1956). Building on this foundation, we introduced efficiency and emotional expressiveness to yield a more holistic rubric (Figure 3). Efficiency couples tightly with creativity because novel solutions often entail more economical ways of achieving a goal (Williamon et al., 2006; Runco, 2014; Colton & Wiggins, 2012), whereas emotional expressiveness taps the affective impact of musical choices (Amabile, 1983).

The ECD-based assessment aggregate score correlates moderately with manual ratings ($r = 0.48$, $p < 0.001$), mirroring prior work that relied only on divergent thinking and complexity ($r = 0.47$; Rahimi et al., 2024). Dimension-level analyses on divergent thinking, complexity, efficiency, and emotional expressiveness show moderate-to-high correlations with the global ECD score ($r = 0.31$–0.74), which means that all of these constructs contribute to the final creativity score. Moreover, ECD-based divergent thinking, complexity, and efficiency align reasonably with human judgment (correlation $r = 0.29$–0.53), whereas emotional expressiveness shows a weaker link—likely because human raters place less weight on affective nuances than on structural novelty or technical sophistication (Baas et al., 2008). Human raters may place greater emphasis on visual or structural features, such as the arrangement of code blocks, the use of sprites, or the overall functionality of the program, rather than auditory expressiveness. In contrast, our ECD-based creativity score assigns equal weight to each dimension (e.g., originality, structural complexity, efficiency, and emotional expressiveness), regardless of how raters might subjectively prioritize them. Prior research supports this interpretation. Kovalkov and colleagues (2021) found that, among experts scoring Scratch program creativity, 80% placed roughly twice as much emphasis on visual aspects as on audio elements. Additionally, recent work in music performance evaluation has consistently demonstrated a sight-over-sound effect, where evaluators rely more on visual expressiveness than auditory cues (Samma et al., 2025; Tsay, 2013). These findings validate the ECD framework in assessing creativity in a flow-based music programming environment while exposing aspects of creativity (e.g., emotional expressiveness) that remain difficult to quantify objectively.

4.2 Capabilities and Limitations of GPT-Based Creativity Assessment

We further explored few-shot GPT-4 prompts to emulate creativity judgments. We supplied two distinct exemplar sets—one derived from ECD scores and the other from manual ratings—to create *GPT_ECD* and GPT_MANUAL conditions. The results show that there is a similar score density distribution between the GPT-based scores and human expert scores, which indicates that LLMs are helpful for human-in-the-loop creativity assessment.

Moreover, the two GPT-based scoring approaches show (1) a closer Pearson correlation ($r = 0.80$) than their respective training reference sources. Specifically, the correlation between GPT_MANUAL and manual ratings is $r = 0.69$, whereas the correlation between *GPT_ECD* and ECD is lower at $r = 0.38$. The results also show that (2) GPT's correlation with ECD is weaker than with manual ratings. There are two plausible contributing factors. First, while GPT can learn to evaluate creativity from examples, its assessments still reflect its own inherent patterns of judgment. Although from previous studies there is evidence that LLM scores can shift with the choice and order of reference examples (DiStefano et al., 2024; Yoshida, 2024), prior research has also shown that bias in LLMs persists even when few-shot examples are varied or when state-of-the-art calibration techniques are applied (Reif & Schwartz, 2024), suggesting that such biases are inherent and not easily mitigated through prompting strategies alone. Another study by Gupta and colleagues (2023) has found that some LLMs, such as GPT-4, exhibit strong robustness (over 90% accuracy) to majority label bias in in-context learning settings, even when the few-shot examples are heavily skewed toward a dominant label. In our study, GPT models may exhibit a potential bias by placing

greater emphasis on the structural and functional aspects of programs when evaluating creativity, while underrepresenting other dimensions such as emotional expressiveness or aesthetic novelty. Future work could enhance LLM-based creativity assessments by incorporating multimodal inputs (such as audio analysis or paired visual-audio prompts) to better capture the full spectrum of creative expression. Second, one possible explanation for the stronger correlation between GPT_MANUAL scores and human ratings is that LLMs and human raters may both implicitly emphasize visual and structural features of student work over more subjective or less visible attributes like emotional expressiveness. For LLMs, this may reflect limitations in their ability to "perceive" auditory or affective elements from textual data. For humans, as mentioned above, prior studies show a similar bias: experts assessing Scratch program creativity weighted visual features nearly twice as heavily as auditory ones (Kovalkov et al., 2021). Related work in music performance has also identified a persistent "sight-over-sound" effect, where raters prioritize visual expressiveness over sound (Samma et al., 2025; Tsay, 2013). In contrast, our ECD-based computational score applies equal weight to all rubric dimensions, including visual and audio aspects. This mismatch in weighting likely contributes to the weaker alignment between GPT-based scores and ECD-based scores.

4.3 Comparisons and Alignment across Assessments

The comparison among ECD-based, manual, and GPT-based assessment methods reveals both convergence and divergence across metrics, distributions, and example-level outcomes. First, one-way ANOVA followed by Tukey's HSD test confirmed significant differences in mean creativity scores across the four methods ($F = 35.13$, $p < 0.001$). The ECD method produced significantly higher scores than both GPT_ECD and Manual assessments, while GPT_MANUAL exhibited the highest alignment with ECD, showing no significant difference in mean score ($\Delta M = 0.14$, $p = 0.12$). Further analysis of the score distributions (Figure 7) supports these findings. While all methods display similar distributional ranges (scaled to 1 to 6), their density curves differ in shape and central tendency. ECD and GPT_MANUAL tend to concentrate scores around 3.5 to 4, while Manual and GPT_ECD scores skew slightly lower. One possible explanation is that the ECD method applies consistent rubric-based scoring with algorithmically derived features, which may lead to higher overall scores—particularly in dimensions like efficiency or emotional expressiveness—compared to human raters, who tend to score more conservatively. In contrast, GPT_MANUAL, trained on human-labelled examples, captures human-like reasoning but redistributes scores in a manner that aligns more closely with ECD's overall scale, despite inheriting its structure from Manual assessments.

Each of these assessment methods has strengths and limitations. Automated systems typically assess only the final product, overlooking the creative process (e.g., iteration and experimentation), which can offer important insights into student creativity (Colton, 2008). However, in programming education, automated methods may prioritize syntactic correctness and output functionality, while human raters may value code readability, design elegance, and algorithmic originality (Granåsen, 2018; Brown et al., 2023). In music programming, subjective elements like sound combinations, timing, and expressive flow influence human perception of creativity but remain difficult for computational models to quantify. The absence of standardized creativity criteria across domains further complicates the alignment between human and machine evaluations (Jordanous, 2012; Hershkovitz et al., 2019). Discrepancies are more pronounced in tasks requiring flexible reasoning, contextual interpretation, or affective sensitivity (Hone et al., 1999).

4.4 Implications

4.4.1 Educational Implications

Automated creativity assessment methods, such as ECD-based and GPT-based evaluations, offer scalable solutions for evaluating students' creativity in interactive programming learning environments. These methods enable real-time, detailed feedback on students' creative outputs, allowing educators to tailor instructional strategies to individual needs. For instance, students struggling in specific dimensions like complexity or effectiveness can receive targeted interventions aimed at enhancing these aspects of creativity. By providing a systematic and equitable approach to creativity assessment, automated methods address the limitations of traditional manual evaluations that are time-intensive and prone to evaluator bias (Kenett & Faust, 2019; Heinen & Johnson, 2018). GPT-based methods, particularly GPT_MANUAL, demonstrate strong alignment with human evaluators, making them a reliable alternative in educational settings. For example, in a 45-minute class, students re-imagined a classic fairy tale in Scratch. GPT_MANUAL delivered feedback on each project's originality and narrative coherence within seconds, allowing the teacher to regroup the class: students with high originality but weak coherence peer-reviewed one another's scripts, whereas those with strong coherence but low originality rewrote their storylines. Furthermore, these tools democratize creativity assessment by ensuring consistent evaluations, even in large-scale classrooms or online learning platforms. These assessment models can be used to generate feedback and can serve as a prompt for reflection and revision. For example, a student who receives a low "efficiency" score may be encouraged to revisit their code for redundant logic or overly complex structures. Similarly, a high "emotional expressiveness" score can validate a student's use of storytelling or musical elements, reinforcing affective aspects of creativity. Educators can use these dimensional scores to tailor targeted mini-lessons, peer-critique sessions, or guided revisions based on specific areas of need. Last but not least, in educational settings, ethical use of automated creativity assessments requires transparent disclosure, periodic human review, and opportunities for students to contest or

appeal algorithm-generated scores. To help students interpret scores, scaffolding is essential. Examples include (a) displaying anonymized projects that span the score range and jointly analyzing what makes them distinctive, and (b) guiding students to set concrete improvement goals such as, "Add two new elements that have not appeared in any classmate's project."

4.4.2 Creativity Assessment Implications

The application of AI-driven assessments, particularly through GPT-based methods, demonstrates the potential of LLMs to mirror human judgment in creativity evaluation. These findings suggest that AI tools could play a central role in automating creativity assessments, reducing the time and effort required for manual evaluations while maintaining a high degree of accuracy. Moreover, GPT-based methods offer an equitable and scalable solution for assessing creativity in larger student populations, addressing the limitations of resource-intensive manual assessments.

Despite these advancements, the differences observed among ECD-based, GPT-based, and manual evaluations highlight persistent challenges in achieving fully consistent creativity scoring. For instance, variations in emotional expressiveness scores between ECD- and GPT-based methods, as well as discrepancies between GPT-based and manual assessments, indicate potential misalignments in how each approach interprets this dimension. Addressing these inconsistencies necessitates the development of robust validation mechanisms to harmonize the assessment methodologies. Combining the precision and objectivity of ECD-based models, which emphasize measurable features, with the nuanced interpretative capabilities of GPT-based methods gives us an opportunity to create a more holistic and balanced creativity assessment framework. Such an integrated approach could capitalize on the strengths of both methodologies, enhancing creativity evaluation and fostering a deeper understanding of students' creative potential (Rahimi, 2023; Shute et al., 2016).

5. Conclusion, Limitations, and Future Work

This study analyzed 383 programming artifacts collected from 194 primary school students over three years (2022–2024) using two automated assessment methods: an ECD-based approach and an LLM-based approach using ChatGPT-4 with few-shot learning. The ECD-based method evaluated creativity across four key dimensions: divergent thinking, complexity, efficiency, and emotional expressiveness. The LLM-based method employed examples from both CAT ratings and ECD assessments for learning. The findings showed that automated assessments have moderate to strong correlations with human evaluations, with the LLM-based approach exhibiting greater consistency across diverse learning examples. Besides, GPT-based methods, trained on human-assessed examples, demonstrate strong alignment with human evaluators, illustrating their capacity to complement and support human judgment in creativity assessment. The findings highlight the potential of automated methods, including ECD-based and GPT-based approaches, for assessing creativity within a flow-based music programming environment. By employing a multidimensional framework encompassing divergent thinking, complexity, efficiency, and emotional expressiveness, the ECD-based method offers a systematic and structured approach to evaluating creativity. This research contributes to the growing body of literature on creativity assessment by showcasing the feasibility and validity of automated tools in educational settings. Besides, it provides practical insights into how ECD-based and LLM-based approaches can effectively support creativity assessment, enabling educators to deliver timely, objective, and consistent feedback.

Despite these contributions, several limitations exist. First, while this study used students' program artifacts to measure creative activity, we were unable to collect self-reported creativity data from students or include external measures, such as traditional creativity tests, to complement the findings. Future research should incorporate these additional data sources and try other diversity-oriented metrics to investigate further and contextualize students' creativity. Additionally, we employed CAT to evaluate students' creativity in the flow-based music programming environment, with ratings provided by five expert evaluators. While previous studies suggest that five to 10 experts are sufficient for reliable CAT scoring (Turkman, 2016), increasing the number of evaluators can improve the accuracy and robustness of assessments. Future research should strive to include a larger pool of experts to enhance the reliability of CAT-based evaluations and conduct inter-rater reliability studies that benchmark automated metrics against human judgments. Moreover, comparing the LLM-based scoring with alternative models (e.g., LSA in Table 1) can provide deeper insights into the strengths and limitations of different approaches. Finally, the aggregation process used in this study was simplified to align with prior research (Rahimi et al., 2024). Future work could explore alternative weighting strategies, including expert-defined or data-driven approaches, to examine whether some dimensions contribute more strongly to perceived creativity. Future work should also focus on refining the alignment between automated and manual assessments, exploring broader applications across diverse educational contexts, and addressing ethical considerations surrounding the implementation of AI-driven creativity assessments.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The publication of this article received financial support from the National Science Foundation's ITEST program (22-585).

References

Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2024). Measuring original thinking in elementary school. *Journal of Educational Psychology*, *116*(6), 953–981. https://doi.org/10.1037/edu0000844

Almond, R., Shute, V., Tingir, S., & Rahimi, S. (2020). Identifying observable outcomes in game-based assessments. In H. Jiao & R. W. Lissitz (Eds.), *Innovative psychometric modeling and methods* (pp. 163–192). Information Age Publishing. https://myweb.fsu.edu/vshute/pdf/marcs2020.pdf

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer. https://doi.org/10.1007/978-1-4939-2125-6

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*(5), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology*, *45*(2), 357–376. https://doi.org/10.1037/0022-3514.45.2.357

Amabile, T. M. (1996). *Creativity in context: Update to 'The social psychology of creativity.'* Westview Press. https://doi.org/10.4324/9780429501234

Azzam, A. M. (2009). Why creativity now? A conversation with Sir Ken Robinson. *Educational Leadership*, *67*(1), 22–26.

Baas, M., De Dreu, C. K. W., & Nijstad, B. A. (2008). A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus? *Psychological Bulletin*, *134*(6), 779–806. https://doi.org/10.1037/a0012815

Baer, J., & McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In C. S. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65–77). IGI Global. https://doi.org/10.4018/978-1-60566-667-9.ch004

Bănuț, M., Albulescu, I., & Simion, A. (2022). Creativity pedagogy: Students' expression through music and programming. In *Education, reflection, development—ERD 2022, vol 6. European Proceedings of Educational Sciences* (pp. 306–321). European Publisher. https://doi.org/10.15405/epes.23056.28

Barbot, B. (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02529

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with *SemDis*: An open platform for computing semantic distance. *Behavior Research Methods*, *53*(2), 757–780. https://doi.org/10.3758/s13428-020-01453-w

Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Routledge. https://doi.org/10.4324/9780203508527

Bowkett, S. (2007). *100 ideas for teaching creativity*. Continuum. https://books.google.com/books?id=GJdLAAAAYAAJ

Brown, N., Messer, M., Kölling, M., & Shi, M. (2023). Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education*, *24*(1). https://doi.org/10.1145/3636515

Celik, I., Gedrimiene, E., Siklander, S., & Muukkonen, H. (2024). The affordances of artificial intelligence-based tools for supporting 21st-century skills: A systematic review of empirical research in higher education. *Australasian Journal of Educational Technology*, *40*(3), 19–38. https://doi.org/10.14742/ajet.9069

Chou, E., Fossati, D., & Hershkovitz, A. (2024). A code distance approach to measure originality in computer programming. In O. Poquet, A. Ortega-Arranz, O. Viberg, I.-A. Chounta, B. McLaren, & J. Jovanovic (Eds.), *Proceedings of the 16th International Conference on Computer Supported Education* (CSEDU 2024), 2–4 May 2024, Angers, France (pp. 541–548, Vol. 2). SciTePress—Science and Technology Publications. https://www.scitepress.org/Papers/2024/126321/126321.pdf

Colton, S. (2008). Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Intelligent Systems* (AAAI 2008), 26–28 March 2008, Palo Alto, California, USA (pp. 14–20). AAAI. https://cdn.aaai.org/Symposia/Spring/2008/SS-08-03/SS08-03-003.pdf

Colton, S., & Wiggins, G. A. (2012). Computational creativity: The final frontier? In L. D. Raedt, C. Bessiere, D. Dubois, P. Doherty, & P. Frasconi (Eds.), *Proceedings of the 20th European Conference on Artificial Intelligence* (ECAI 2012), 27–31 August 2012, Montpellier, France (pp. 21–26). ACM. https://dl.acm.org/doi/10.5555/3007337.3007345

de Fleurian, R., Blackwell, T., Ben-Tal, O., & Müllensiefen, D. (2017). Information-theoretic measures predict the human judgment of rhythm complexity. *Cognitive Science*, *41*(3), 800–813. https://doi.org/10.1111/cogs.12347

DiStefano, P. V., Patterson, J. D., & Beaty, R. E. (2024). Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*. https://doi.org/10.1080/10400419.2024.2326343

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, *10*(28). https://doi.org/10.1126/sciadv.adn5290

Dumas, D., Organisciak, P., & Doherty, P. (2020). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, *16*(4), 665–678. https://doi.org/10.1037/aca0000355

Engelman, S., Magerko, B., McKlin, T., Miller, M., Edwards, D., & Freeman, J. (2017). Creativity in authentic STEAM education with EarSketch. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (SIGCSE 2017), 8–11 March 2017, Seattle, Washington, USA (pp. 183–188). ACM. https://doi.org/10.1145/3017680.3017763

Granåsen, D. (2018). Towards automated assessment of team performance by mimicking expert observers' ratings. *Cognition, Technology & Work*, *21*, 253–274. https://doi.org/10.1007/s10111-018-0499-6

Guilford, J. P. (1956). *Fundamental statistics in psychology and education* (3rd ed.). McGraw-Hill.

Gupta, K., Roychowdhury, S., Kasa, S., Kasa, S., Bhanushali, A., Pattisapu, N., & Murthy, P. (2023). How robust are LLMs to in-context majority label bias? *arXiv preprint arXiv:2312.16549*. https://doi.org/10.48550/arXiv.2312.16549

Heinen, D. J. P., & Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(2), 144–156. https://doi.org/10.1037/aca0000125

Henriksen, D., Creely, E., Henderson, M., & Mishra, P. (2021). Creativity and technology in teaching and learning: A literature review of the uneasy space of implementation. *Educational Technology Research and Development*, *69*(4), 2091–2108. https://doi.org/10.1007/s11423-020-09912-z

Henriksen, D., Henderson, M., Creely, E., Ceretkova, S., Černochová, M., Sendova, E., Sointu, E. T., & Tienken, C. H. (2018). Creativity and technology in education: An international perspective. *Technology, Knowledge and Learning*, *23*(3), 409–424. https://doi.org/10.1007/s10758-018-9380-1

Hershkovitz, A., Sitman, R., Israel-Fishelson, R., Eguíluz, A., Garaizar, P., & Guenaga, M. (2019). Creativity in the acquisition of computational thinking. *Interactive Learning Environments*, *27*(5–6), 628–644. https://doi.org/10.1080/10494820.2019.1610451

Hone, A., Williamson, D., & Bejar, I. (1999). "Mental model" comparison of automated and human scoring. *Journal of Educational Measurement*, *36*(2), 158–184. https://doi.org/10.1111/j.1745-3984.1999.tb00552.x

Israel-Fishelson, R., & Hershkovitz, A. (2022). Studying interrelations of computational thinking and creativity: A scoping review (2011–2020). *Computers & Education*, *176*, 104353. https://doi.org/10.1016/j.compedu.2021.104353

Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, *4*, 246–279. https://doi.org/10.1007/s12559-012-9156-1

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, *20*(2), 171–178. https://doi.org/10.1080/10400410802059929

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four C model of creativity. *Review of General Psychology*, *13*(1), 1–12. https://doi.org/10.1037/a0013688

Kaufman, J. C., & Sternberg, R. J. (Eds.). (2010). *The Cambridge handbook of creativity*. Cambridge University Press. https://doi.org/10.1017/9781316979839

Kenett, Y. N., & Faust, M. (2019). A semantic network cartography of the creative mind. *Trends in Cognitive Sciences*, *23*(4), 271–274. https://doi.org/10.1016/j.tics.2019.01.007

Kind, P. M., & Kind, V. (2007). Creativity in science education: Perspectives and challenges for developing school science. *Studies in Science Education*, *43*, 1–37. https://doi.org/10.1080/03057260708560225

Kovalkov, A., Paaßen, B., Segal, A., Pinkwart, N., & Gal, K. (2021). Automatic creativity measurement in Scratch programs across modalities. *IEEE Transactions on Learning Technologies*, *14*(6), 740–753. https://doi.org/10.1109/TLT.2022.3144442

Li, Y., Kim, M., & Palkar, J. (2022). Using emerging technologies to promote creativity in education: A systematic review. *International Journal of Educational Research Open*, *3*, 100177. https://doi.org/10.1016/j.ijedro.2022.100177

Lin, Y.-S. (2011). Fostering creativity through education: A conceptual framework of creative pedagogy. *Creative Education*, *2*(3), 149–155. https://doi.org/10.4236/ce.2011.23021

Liu, Z., Zhang, S., Israel, M., Smith, R., Xing, W., & Minces, V. (2025). Engaging K–12 students with flow-based music programming: An experience report on its impact on teaching and learning. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education* (SIGCSE TS 2025), 26 February–1 March 2025, Pittsburgh, Pennsylvania, USA (pp. 708–714, Vol. 1). ACM. https://doi.org/10.1145/3641554.3701902

Lou, S.-J., Chou, Y.-C., Shih, R.-C., & Chung, C.-C. (2017). A study of creativity in CaC_2 steamship-derived STEM project-based learning. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(6), 2387–2404. https://doi.org/10.12973/eurasia.2017.01231a

Loveless, A. (2002). *Literature review in creativity, new technologies and learning* (tech. rep.) (A NESTA Futurelab Research report—report 4). Futurelab. https://telearn.hal.science/hal-00190439

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. https://doi.org/10.1037/h0048850

Mednick, S. A. (1968). The remote associates test. *The Journal of Creative Behavior*, *2*(3), 213–214. https://doi.org/10.1002/j.2162-6057.1968.tb00104.x

Minces, V., Booker, A., & Khalil, A. (2021). Listening to waves: Engaging underrepresented students through the science of sound and music. *Connected Science Learning*, *3*(4), 12318697. https://doi.org/10.1080/24758779.2021.12318697

Minces, V. H., & Akshay, N. (2023). STEAM for all: A vision for STEM and arts integration. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International encyclopedia of education* (4th ed., pp. 10–18). Elsevier. https://doi.org/10.1016/b978-0-12-818630-5.13053-2

Minces, V. H., Xing, W., & Li, C. (2023). Work in progress: Mflow, a flow-based music programming platform for young children. In C. da Rocha Brito & M. M. Ciampi (Eds.), *2023 IEEE World Engineering Education Conference* (EDUNINE 2023), 12–15 March 2023, Bogota, Columbia (pp. 1–4). IEEE. https://doi.org/10.1109/EDUNINE57531.2023.10102852

Nagaraj, N., & Balasubramanian, K. (2017). Three perspectives on complexity: Entropy, compression, subsymmetry. *The European Physical Journal Special Topics*, *226*(15–16), 3251–3272. https://doi.org/10.1140/epjst/e2016-60347-2

National Advisory Committee on Creative and Cultural Education (NACCCE). (1999). *All our futures: Creativity, culture and education* (tech. rep.). Department for Education and Employment. London, UK. https://eric.ed.gov/?id=ED440037

Newton, D. P. (2013). Moods, emotions and creative thinking. *Thinking Skills and Creativity*, *8*, 34–44. https://doi.org/10.1016/j.tsc.2012.05.006

Newton, L. D., & Newton, D. P. (2014). Creativity in 21st-century education. *Prospects*, *44*(4), 575–589. https://doi.org/10.1007/s11125-014-9322-1

Noh, J., & Lee, J. (2020). Effects of robotics programming on the computational thinking and creativity of elementary school students. *Educational Technology Research and Development*, *68*(1), 463–484. https://doi.org/10.1007/s11423-019-09708-w

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 technical report* (tech. rep.). https://arxiv.org/abs/2303.08774

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, *49*, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Parnami, A., & Lee, M. (2022). Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv.2203.04291*. https://doi.org/10.48550/arXiv.2203.04291

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, *1423*(1), 378–395. https://doi.org/10.1111/nyas.13654

Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (pp. 48–73). Cambridge University Press. https://doi.org/10.1017/9781316979839.005

Rahimi, S., Almond, R. G., & Shute, V. J. (2023). Getting the first and second decimals right: Psychometrics of stealth assessment. In M. P. McCreery & S. K. Krach (Eds.), *Games as stealth assessments* (pp. 125–153). IGI Global. https://doi.org/10.4018/979-8-3693-0568-3.ch006

Rahimi, S. (2023). Going beyond the brick: Assessing and supporting creativity using AI-powered digital games. *Creativity Research Journal*, *37*(2), 275–283. https://doi.org/10.1080/10400419.2023.2241779

Rahimi, S., Smith, J. B., Truesdell, E. J. K., Vinay, A., Boyer, K. E., Magerko, B., Freeman, J., & Mcklin, T. (2024). An automated, unobtrusive, formative assessment of creativity in a computer science and music remixing learning environment [Advance online publication]. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000683

Reif, Y., & Schwartz, R. (2024). Beyond performance: Quantifying and mitigating label bias in LLMs. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 16–21 June 2024, Mexico City, Mexico (pp. 6784–6798, Vol. 1, Long Papers). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.378

Repenning, A., Zurmühle, J., Lamprou, A., & Hug, D. (2020). Computational music thinking patterns: Connecting music education with computer science education through the design of interactive notations. In H. C. Lane, S. Zvacek, & J. Uhomoibhi (Eds.), *Proceedings of the 12th International Conference on Computer Supported Education* (CSEDU 2020), 2–4 May 2020, online (pp. 641–652, Vol. 1). SciTePress. https://doi.org/10.5220/0009817506410652

Rubenstein, L. D. V., Thomas, J., Finch, W. H., & Ridgley, L. M. (2022). Exploring creativity's complex relationship with learning in early elementary students. *Thinking Skills and Creativity*, *44*, 101030. https://doi.org/10.1016/J.TSC.2022.101030

Runco, M. A. (2011). Divergent thinking. In M. A. Runco & S. R. Pritzker (Eds.), *Encyclopedia of creativity* (2nd ed., pp. 400–403, Vol. 1). Academic Press.

Runco, M. A. (2014). *Creativity: Theories and themes: Research, development, and practice*. Academic Press. https://doi.org/10.1016/C2012-0-06920-7

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal*, *24*(1), 66–75. https://doi.org/10.1080/10400419.2012.652929

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161. https://doi.org/10.1037/h0077714

Samma, T., Honda, K., & Fujii, S. (2025). Sight-over-sound effect depends on interaction between evaluators' musical experience and auditory-visual integration: An examination using Japanese brass band competition recordings. *PLoS ONE*, *20*(4), e0321442. https://doi.org/10.1371/journal.pone.0321442

Sharmin, S. (2021). Creativity in CS1: A literature review. *ACM Transactions on Computing Education (TOCE)*, *22*(1), 1–26. https://doi.org/10.1145/3459995

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. *Educational Assessment*, *21*(1), 34–59. https://doi.org/10.1080/10627197.2015.1127752

Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., Kuba, R., Liu, Z., Yang, X., & Sun, C.-L. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity, and learning supports in educational games. *Journal of Computer Assisted Learning*, *37*(1), 127–141. https://doi.org/10.1111/jcal.12473

Siva, S., Im, T., McKlin, T., Freeman, J., & Magerko, B. (2018). Using music to engage students in an introductory undergraduate programming course for non-majors. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (SIGCSE 2018), 21–24 February 2018, Baltimore, Maryland, USA (pp. 975–980). ACM. https://doi.org/10.1145/3159450.3159468

Song, Y., Xing, W., Barron, A., Oh, H., Li, C., & Minces, V. (2023). M-flow: A flow-based music creation platform improves underrepresented children's attitudes toward computer programming. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (IDC 2023), 19–23 June 2023, Chicago, Illinois, USA (pp. 233–238). ACM. https://doi.org/10.1145/3585088.3589383

Sternberg, R. J. (1999). Handbook of creativity. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511807916

Szydlo, T., Brzoza-Woch, R., Sendorek, J., Windak, M., & Gniady, C. (2017). Flow-based programming for IoT leveraging fog computing. In S. M. Reddy, W. Cellary, & M. Fugini (Eds.), *2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises* (WETICE 2017), 21–23 June 2017, Poznan, Poland (pp. 74–79). IEEE. https://doi.org/10.1109/WETICE.2017.17

Thul, E. (2008). *Measuring the complexity of musical rhythm* [Master's thesis, McGill University]. https://www-cgrl.cs.mcgill.ca/~godfried/teaching/mir-reading-assignments/Eric-Thul-Thesis.pdf

Tsay, C.-J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(36), 14580–14585. https://doi.org/10.1073/pnas.1221454110

Turkman, B. (2016). *Subjective and objective measurement in creativity: Comparison studies* [Doctoral dissertation, University of Georgia]. https://openscholar.uga.edu/record/14629?v=pdf

Venckutė, M., Berg Mulvik, I., Lucas, B., Bacigalupo, M., Cachia, R., & Kampylis, P. (2020). *Creativity, a transversal skill for lifelong learning—An overview of existing concepts and practices—Final report* (tech. rep.). Publications Office of the European Union. https://publications.jrc.ec.europa.eu/repository/handle/JRC122016https://publications.jrc.ec.europa.eu/repository/handle/JRC122016

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10745–10759. https://doi.org/10.1109/tpami.2023.3263585

Williamon, A., Thompson, S., Lisboa, T., & Wiffen, C. (2006). Creativity, originality, and value in music performance. In I. Deliège & G. A. Wiggins (Eds.), *Musical creativity* (pp. 177–196). Psychology Press. https://doi.org/10.4324/9780203088111-22

Yoshida, L. (2024). The impact of example selection in few-shot prompting on automated essay scoring using GPT models. In A. Olney, I. Chounta, Z. Liu, O. Santos, & I. Bittencourt (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky. AIED 2024. Communications in computer and information science* (pp. 61–73, Vol. 2150). Springer. https://doi.org/10.1007/978-3-031-64315-6_5

Zhang, L., & Nouri, J. (2019). A systematic review of learning computational thinking through Scratch in K–9. *Computers & Education*, *141*, 103607. https://doi.org/10.1016/j.compedu.2019.103607