

An Automated, Unobtrusive, Formative Assessment of Creativity in a Computer Science and Music Remixing Learning Environment

Seyedahmad Rahimi¹, Jason Brent Smith², Erin J. K. Truesdell³, Ashvala Vinay², Kristy Elizabeth Boyer⁴, Brian Magerko³, Jason Freeman², and Tom Mcklin⁵

¹ School of Teaching and Learning, College of Education, University of Florida

² Georgia Tech Center for Music Technology, Georgia Institute of Technology, Atlanta, Georgia, United States

³ Expressive Machinery Laboratory, Georgia Institute of Technology

⁴ Engineering Education, Computer and Information Science and Engineering, University of Florida

⁵ The Findings Group, Decatur, Georgia, United States

Creativity is one of the most crucial skills for success in the 21st-century workforce. Specifically, creativity is an important skill to have in science, technology, engineering, and mathematics (STEM)-related fields, and more empirical studies are needed to assess and improve creativity in STEM-related learning environments. In this study, we designed and validated an automated, unobtrusive, formative assessment of creativity in EarSketch, a computational music remixing platform where students learn to write Python or JavaScript code to create pieces of music. Using an existing data set of EarSketch projects ($n = 53$), we addressed two research questions: (Research Question 1) To what extent is the automated assessment of creativity that we designed in EarSketch psychometrically sound (focusing on validity and reliability), and (Research Question 2) what variables (i.e., divergent thinking, complexity, and self-report variables) predict students' creativity in EarSketch? Our main findings show that (a) the automated assessment of creativity has reasonable convergent validity ($r = .47$) and discriminant validity; (b) the automated assessment of creativity has a reliability estimate of .70; and (c) divergent thinking and the students' confidence in learning how to code significantly predicted students' creativity scores in an external, consensual assessment of creativity by EarSketch experts. Providing learning environments that can assess and support essential skills such as creativity alongside other STEM-related skills such as programming and computational thinking holds great promise for developing the next generation of the workforce who is not merely aware of STEM concepts and principles, but is creative and innovative in pursuing STEM solutions.

Keywords: creativity, automated assessment, computer programming, EarSketch, music

Creativity is generally defined as the ability to transcend conventional ideas and create products (e.g., ideas, solutions, rules, patterns, and music) that are novel/original, appropriate/useful, and valuable (Amabile, 1983; Kaufman & Beghetto, 2009; Piffer, 2012; Plucker

et al., 2004). Throughout history and especially in the 21st century, creativity has been one of the crucial drivers that move human civilizations forward (Craft, 2010; Glăveanu et al., 2020; Hennessey & Amabile, 2010; Henriksen et al., 2018; Runco & Jaeger, 2012).

Emphasizing the importance of creativity, many organizations started to list creativity as one of the most important skills in the 21st century. For instance, the Partnership P21 (2019) included creativity as one of the essential skills for success in modern life. Also, the World Economic Forum (Gray, 2016) listed creativity as the third most important skill people need to succeed in the fourth industrial revolution. Moreover, developing an educated and creative workforce is necessary for any society to meet economic needs, rise to new challenges, and continue to be a leader in the global community (Florida, 2002; Hennessey & Amabile, 2010; Shaheen, 2010).

The importance of creativity from an individual's point of view (i.e., helping one increase their quality of life by solving their daily problems and having more job opportunities) and a socioeconomic perspective (i.e., enhancing the quality of life in societies through scientific and life-changing discoveries and inventions) has attracted the attention of researchers from multiple disciplines to study, assess, and attempt to improve people's creativity (e.g., Guilford, 1950; Hennessey & Amabile, 2010; Kaufman & Beghetto, 2009; Rahimi & Shute, 2021; Shute & Rahimi, 2021; Simonton, 1997; Sternberg, 1988). However, there are still many underinvestigated areas and

This article was published Online First May 16, 2024.

Seyedahmad Rahimi  <https://orcid.org/0000-0001-9266-758X>

Jason Brent Smith  <https://orcid.org/0000-0002-7075-6132>

Erin J. K. Truesdell  <https://orcid.org/0000-0001-8692-4172>

Ashvala Vinay  <https://orcid.org/0000-0002-2487-2052>

Kristy Elizabeth Boyer  <https://orcid.org/0000-0003-3434-3450>

Brian Magerko  <https://orcid.org/0000-0003-1900-4020>

Jason Freeman  <https://orcid.org/0000-0003-3827-1060>

Tom Mcklin  <https://orcid.org/0000-0001-6120-7222>

This material is based upon the work supported by the National Science Foundation (Awards DRL-1814083 and DRL-1813740). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors have no conflicts of interest to declare. EarSketch is available online at <https://ears sketch.gatech.edu>.

Correspondence concerning this article should be addressed to Seyedahmad Rahimi, School of Teaching and Learning, College of Education, University of Florida, 2821 Norman Hall, P.O. Box 117048, Gainesville, FL 32611, United States. Email: srahimi@ufl.edu

open questions related to creativity. For instance, creativity is an important skill to have in science, technology, engineering, and mathematics (STEM)-related fields (Henriksen et al., 2018; Lou et al., 2017; Mayasari et al., 2016) and more empirical studies are needed to assess and improve creativity in STEM-related learning environments that prepare individuals to perform well in STEM-related fields. To foster creativity in such learning environments, one must first accurately assess it. Also, when STEM-related learning environments embrace creativity (i.e., allow learners to show their creativity), assessing creativity in a formative manner makes the most sense rather than assessing it using traditional tests of creativity or independent of the learning environment.

As future jobs become more automated and replaced by artificial intelligence (AI)-powered machines (e.g., replacement of a customer service division with systems equipped with GPT-4, i.e., generative pre-trained transformer), the future STEM workforce critically needs competencies that include creativity and complex problem-solving skills (or a mix of the two, i.e., creative problem-solving) above manual skills or memorized content (Education & Human Resources, 2020). In other words, we need a workforce that is not merely aware of STEM concepts and principles but is creative and innovative in pursuing STEM solutions. Some scholars have introduced a type of co-creativity of humans with AI systems (Vinchon et al., 2023; Wingstrom et al., 2023) that can enhance human creativity. Those scholars also talked about some uses of AI that may not lead to enhanced human creativity and can be considered unethical or even plagiaristic (Vinchon et al., 2023). The type of use of AI that can enhance human creativity is referred to as intelligence augmented (Dede et al., 2021). Such positive uses of AI need to be investigated further. Moreover, interdisciplinary efforts (e.g., between the fields of psychological science and computer science) can enrich AI-powered digital environments leading to unexpected, yet promising human-AI co-creativity (Rafner, Beaty, et al., 2023). However, we first need to create environments that can accurately assess humans' creativity. Then, we can be more successful in augmenting humans' creativity using advanced, AI-based systems. The current study aims to design and validate an automated, formative assessment of creativity using an automated assessment method derived from the method of stealth assessment (Shute, 2011) in a web-based, programming-learning environment for music remixing called EarSketch (Magerko et al., 2016).

Background

Creativity in Music

Since the context of our study is related to music, we first start by defining creativity in music. Hickey and Webster (2001) define musical composition as the result of multiple stages of creative decision-making and identify a variety of evaluation schemes as appropriate for evaluating musical creativity. Hickey (2001) also tested the reliability of Amabile's Consensual Assessment Technique using a student's musical compositions as the evaluated creative product. Employing music teachers, composers, theorists, seventh-grade children, and second-grade children as judges, Hickey reported inter-judge reliabilities for creativity ratings on 11 compositions: composers, .04; all music teachers, .64; music theorists, .73; seventh-grade children, .61; and second-grade children, .50. These results revealed evidence for the reliability of this assessment technique across judge groups, with notably weak or negative correlations between composers' scores and those of other groups,

suggesting potential challenges in achieving consensus on assessments of musical creativity.

When examining how people understand creativity in music, Truax (1992) describes a shift from linear, sound-based models toward a multidimensional model of structural and compositional elements. According to Truax, this combination leads to an understanding in which the context or intention behind a musical work is linked to the musical elements within it. Structural aspects of music that can be used to evaluate complexity include melody, harmony, and rhythm (Streich, 2006). Thul (2008) notes entropy as a potential measure of a musical sample's rhythmic complexity.

In brief, we may consider a musical product creative if it contains structural and compositional complexity, is expressive within a given context, and may contain elements of unpredictability. Also, we acknowledge that musical creativity is more complex than any list of characteristics but that for the sake of assessment, we need to operationalize it using the affordances in the given learning environment at hand which we discuss later in this article. Next, we briefly discuss creativity assessment.

Creativity Assessment

Creativity research studies fall on a spectrum from domain-general to domain-specific (Baer & Kaufman, 2005; Kaufman & Baer, 2005; Sternberg & Lubart, 1996). Domain-general studies tend to focus on psychometric measurements of individual differences in creativity, as exemplified by the work of Torrance (1972). Conversely, domain-specific studies involve a deeper examination of creativity within a specific social and environmental domain. The domain-general model of creativity that Guilford (1956) introduced, for example, operationalized creativity as divergent thinking, with four subfacets: fluency (the ability to produce a large number of ideas, solutions, or products), flexibility (the ability to come up with ideas and other products from different themes, genres, or categories), originality (the ability to produce products or come up with statistically novel ideas), and elaboration (the ability to implement an idea in detail and high quality). Multifaceted models like this can assist researchers in developing precise creativity assessments and identifying particular subfacets of creativity to enhance (e.g., Rahimi & Shute, 2021; Rahimi et al., 2023; Shute & Rahimi, 2021).

Assessments of creativity based on a product include aspects such as aesthetics and being of high quality (Sternberg & Lubart, 1996), though Cropley and Cropley (2011) and Runco (2003) also indicated that creativity aims to go beyond aesthetic properties. For instance, Richard Hewitt demonstrated creativity by ingeniously combining a bicycle and a large container to invent the SpinCycle washing machine (Hewitt, 2019) in a remote African village where electricity and financial resources were scarce. The SpinCycle invention is innovative and also practical and valuable for the community, that is, creativity is not just about artistic work. Other aspects of a creative product in the literature include complexity. According to Guilford (1950), complexity is characterized as the degree to which a unified entity is formed by incorporating diverse and separate components. For example, when constructing a building with materials like glass, metal, and wood, it leads to the formation of a complex structure.

Other than these traditional assessments of creativity that are usually about assessing divergent thinking, for example, the Torrance Test of Creative Thinking (Torrance, 1974); the Alternative Uses Test (Wallach & Kogan, 1965), one of the frequently used

assessment techniques is called the Consensual Assessment Technique (Amabile, 1996). In this technique, multiple experts (at least two; see the Method section for a definition of the expert within the context of this study) rate products of interest on their creativity (with no training on how to rate creativity). If, collectively, those raters agree that a product is creative, we must accept that product as creative. This assessment technique is one of the most accurate methods of creativity assessment; however, it is very resource intensive (i.e., many raters who are experts in the domain at hand are needed, and rating the products takes a lot of time and effort). New assessment techniques that can objectively automate the assessment of a creative product are needed. In this study, inspired by the traditional definitions of creativity in terms of creative thinking and the aspects of a creative product, we operationalized creativity (i.e., developed the competency model [CM] of creativity in EarSketch—discussed later in this article), designed an automated, formative assessment of creativity based on that operationalization, and validated it using the Consensual Assessment Technique. Next, we discuss some recent efforts in the automation of traditional assessments of creativity.

Automated Scoring of Creativity Assessments

Advances in technology, computer sciences, and creativity research have led to success in automatic scoring approaches in domain-general creativity assessments. These automated scoring approaches of creativity assessments are related to verbal creative thinking (Acar et al., 2024; Beaty & Johnson, 2021; Beaty et al., 2022, 2021; Patterson, Merseal, et al., 2023), figural creative thinking (Barbot, 2018; Patterson, Barbot, et al., 2023; Rafner, Wang, et al., 2023), and some in domain-specific creativity assessments (Marrone et al., 2023). While our examination of automated creativity assessment is not exhaustive, we delve into some of the endeavors within this research area in this section.

As an example of verbal creative thinking assessment, Beaty et al. (2022) used semantic distance analysis to automatically assess the originality of responses for an Alternative Uses test. The core idea of semantic distance analysis is how best to quantify the creativity quality of responses to a creativity test. Specifically, semantic distance measures the originality or novelty of human responses to a creative thinking test by assessing conceptual dissimilarity, indicating the degree to which concepts exhibit a considerable separation from each other (Beaty et al., 2022). For instance, the words “fork” and “eat” have a low semantic distance value compared to the words “fork” and “laptop.” Beaty et al. (2021) created the SemDis platform (<https://semdis.wlu.psu.edu>), which includes five semantic models trained on a large corpus of English text data, and can automatically compute the semantic distance value of words within a response and calculate the overall originality score of that response. Similarly, a multi-lab, large study ($n = 6,552$; Patterson, Merseal, et al., 2023) used an automatic semantic distance analysis approach to create and validate an automatic, general, verbal assessment of creativity in 12 languages: Arabic, Chinese, Dutch, English, Farsi, French, German, Hebrew, Italian, Polish, Russian, and Spanish. The authors gathered 107,672 responses from the participants and used two deep-learning, semantic-based models to compute the semantic distance values for the responses. Results showed positive evidence for the validity of the automated scores of verbal creativity.

With the advent of large language models, researchers started to enhance the automated scoring of traditional verbal creativity tests

(e.g., Alternative Uses Test). A recent study by Acar et al. (2024) designed and validated an automated assessment of originality called a measure of original thinking for elementary school (MOTES). MOTES comprises 24 game-like items divided into three types: uses (i.e., come up with alternative uses for an ordinary object), examples (i.e., come up with surprising examples), and sentences (i.e., complete incomplete sentences). This automated assessment uses a large language models to score the originality of students’ responses. MOTES was described to students as a game (i.e., “The Game of Surprises”) alluding to what the students had to do: generating surprising responses. Results of a validation study showed a high level of agreement with independent subjective human ratings across all three item types ($r_s = .79, .91$, and $.85$ for uses, examples, and sentences, respectively).

Automatic assessment of figural or drawing tasks (e.g., incomplete shape tasks) can be challenging. However, with advances in AI systems, we will see more progress in the automated assessment of figural creativity in the near future. Despite the challenges, some creative efforts on this front have been made. For example, Barbot (2018) constructed a structured figural assessment of creative thinking which allowed for quantification of the drawing completed by participants. Then, dealing with numbers, the structured figural creativity assessment could automate the scoring process as well as capture data related to the creative process (i.e., exploration, production, and verification). Similarly, using machine learning approaches, Patterson, Barbot, et al. (2023) created an automated drawing assessment platform called AuDrA which automatically scores drawing productions on their creativity. AuDrA was trained using a large collection of line drawings and human creativity ratings. Results of a validation study (Patterson, Barbot, et al., 2023) with 60 raters and over 13,000 drawings (e.g., from incomplete shape tasks) showed medium to strong correlations between human ratings and the automated scores by AuDrA ($r = .65-.81$).

Apart from the domain-general automated assessment of creativity discussed above, researchers have advanced in assessing creativity in specific domains in automated ways. For instance, in an exploratory study, Marrone et al. (2023) examined an automated assessment of mathematical creativity using natural language processing (NLP). The authors used the creative production mathematics task (Tubb et al., 2020) given to 441 Australian secondary school students. For example, the students were asked “What is the most creative way you can write some math questions or equations where the answer is 8 (e.g., ‘ 2×4 ,’ or ‘ $18 - 10$ ’ are not very creative answers).” Students’ responses were scored for creativity and later compared to an NLP model predicting responses’ creativity. Results showed that the NLP model had an accuracy of close to 0.7 and a fair level of agreement (Cohen’s $\kappa = 0.41$). These findings indicate that NLP applied to the assessment of mathematical creativity achieved an accuracy level comparable to that of human judges.

The research on automated scoring of creativity assessments is promising. However, what all of the discussed automated assessments have in common is that they all look like traditional tests and the participants are aware of the fact that they are being assessed (i.e., there is a question to respond to). The current study aims to assess participants’ creativity unobtrusively, embedded within a computer programming learning environment. Also, what the current study aims for is to create a formative assessment of creativity that can later be used to enhance the creativity in students’ products (i.e., pieces of music they create using EarSketch). Such automated, unobtrusive, formative assessments could entail a higher degree of ecological validity.

Automated, Unobtrusive, Formative Assessment

The automated, unobtrusive assessment we discuss in this article is based on a formative assessment technique (i.e., assessment for learning; see [Shute & Rahimi, 2017](#)) called stealth assessment, introduced by [Shute \(2011\)](#). Stealth assessment uses digital learning environments (historically, digital games) as vehicles for assessing and supporting various competencies such as basic knowledge and skills including physics ([Rahimi & Shute, 2023](#); [Shute & Rahimi, 2021](#)); and other hard-to-measure skills such as creativity ([Rafner, 2021](#); [Shute & Rahimi, 2021](#)), persistence ([DiCerbo, 2014](#); [Rahimi et al., 2021, 2022](#)), and problem-solving ([Akram et al., 2018](#)). This assessment method was originally called stealth because it is woven into the learning process and is invisible to the learners. Not all automated assessments use stealth assessment principles. However, an automated assessment can be similar to a stealth assessment when the participants do not feel they are being tested (e.g., they are deeply engaged in gameplay or computer programming). Also, such stealth or automated, unobtrusive assessments are formative and less high-stakes. If a learning environment can move learners away from focusing on the fact that they are being tested, lower the stakes of testing, and help learners focus on performance and learning, the results of a formative assessment in such a learning environment may be more accurate (without test anxiety), showing learners weaknesses and strengths. Then, such ongoing assessment results can be used to support learning in a personalized manner. Stealth assessment (or as we call it automated, unobtrusive, formative assessment) is different than other automated assessments that look like traditional assessments (i.e., there is an item to respond to such as the Alternative Uses Test) but the scoring is automated without human intervention as we discussed earlier.

Since there is a negative connotation with the word stealth, we refer to this type of assessment as automated assessment in this article (not to be confused with other automated assessments as we discussed above). An automated assessment system, when implemented, collects various user interaction data in the log files, through an evidence identification process, identifies and scores those interactions, and finally through an evidence accumulation process creates learner models that show the level of competency (i.e., low, medium, or high) on a particular skill at hand in a granular way. Automated assessment estimates, then, can diagnostically be used to support learners in enhancing their skills, or to adapt the learning environment challenges to the level of learners' ability. This adaptation could lead to maximizing enjoyment, engagement, and eventually learning ([Shute et al., 2020](#)). This research primarily concerns the design and validation of the creativity assessment in EarSketch. The use of the validated assessment to provide support to EarSketch users can be addressed in future work (see the Limitations and Future Research section).

Automated assessment uses an evidence-centered design framework of assessment (ECD; [Almond et al., 2015](#); [Mislevy et al., 2004](#)). ECD is designed to help assessment designers generate assessments that are psychometrically sound (i.e., valid, reliable, and fair). Validity refers to the extent of accuracy of the estimated scores from an assessment. Reliability refers to the level of consistency of those scores. Fairness refers to the extent to which an assessment is fair for different groups of learners (e.g., groups by gender or ethnicity). According to [Messick \(1994\)](#), "...validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles,

they are social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made" (p. 13). The quality of an assessment is important since assessment results lead to some following consequences (e.g., selection or promotion of people in high-stakes tests; adapting the challenges, and/or providing learning support in low-stakes, formative assessments). Such consequences would be unjust and inappropriate if based on low-quality assessments.

ECD uses three main models: (a) the CM defines the "unobservables," or what is being assessed (i.e., what are the subfacets and their relationships defining a competency); (b) the evidence model (EM) defines what indicators (i.e., observables) can provide evidence for the competency of interest ([Almond et al., 2020](#)). EM also includes rules of evidence (i.e., scoring rubrics that can be programmed and automatized) and statistical model (i.e., an aggregation or accumulation method; as sophisticated as Bayesian Networks or as simple as tallies of numbers). Finally, (c) the task model (TM) defines what tasks (or learning environments) can elicit the evidence for the EM. Once these three models are defined and embedded in a digital learning environment, the more the users generate data (i.e., evidence) for what they know and can do, the more accurate the estimates of an automated assessment will become. To read more about the steps to create an automated assessment please see [Shute et al. \(2020\)](#) and [Rahimi, Almond, and Shute \(2023\)](#).

Generally, there are two ways to design and develop an automated assessment. First, one can design a digital learning environment from scratch equipped with automated assessment machinery. This way, the researchers and assessment designers have full control over what can be done. For instance, they can plan for logging specific data that can help with the accuracy (validity) of the assessment. Second, one can use an existing digital learning environment and identify the best indicators for the competency of interest. The second approach introduces some limitations: for example, sometimes the researchers cannot find appropriate indicators for some aspects of their CM and they need to adjust their CM accordingly (i.e., only assess the parts of the CM that they can collect log data for). However, existing digital environments that have undergone multiple tests could function properly with minimal glitches. Such well-functioning, robust existing learning environments can improve the accuracy of automated assessments in the end as they allow learners to show their best performance without interruptions because of technical issues. Therefore, there is always a trade-off between the two paths mentioned above. In this study, we used an existing learning environment called EarSketch to design and validate an automated assessment of creativity.

EarSketch Platform

EarSketch is an expressive computing learning environment engaging high school students in computer science through music sample remixing ([Magerko et al., 2016](#)). This computer programming learning environment enables creative expression, targeting various learning objectives, and emphasizes technology's role in fostering creativity and communication ([Dede et al., 2021](#); [Wardrip-Fruin, 2009](#)). Designed for beginners, EarSketch allows users to code in Python or JavaScript, manipulating a library of 4,000 professionally produced sounds on a multitrack timeline. This web-based learning environment also includes a curriculum covering programming concepts in line with state and national

standards (e.g., functions, loops, and debugging). Studies show positive effects on students' attitudes and persistence in computing, especially among underrepresented populations (Engelman et al., 2017; Magerko et al., 2016; McKlin et al., 2018).

The version of EarSketch used in this study included a cocreative AI learning assistant called CAI. CAI contains an analysis module, which represents the student's current project based on its abstract syntax tree as well as a Sound Profile object (explained in more depth later) which represents a hierarchical musical structure. Based on that analysis module, CAI generates music and code-related suggestions through a dialogue tree that users can navigate through dialogue prompts (see Figure 1). Students initiate conversation with CAI, and the frequency of their interaction with CAI varies per student. CAI also outputs suggestions when students run their code, and these suggestions vary per project in part because all projects are slightly different and students run their projects with different frequencies. Therefore, CAI is not deterministic and its suggestions are not fixed; the creativity of students and how their projects unfold determine how CAI interacts with them.

Here are two example music suggestions: "Let's put in some effects with *setEffect*, like a *filter* or *volume mixing*." "Our project looks like 'AB' form. How about making it 'ABA' by adding a *section*?" And, here are two example code suggestions: "adding a *function* can help us reuse our code multiple times"; "we could let the user control something about our song by using *console input*." These suggestions include links (shown as *italic*) to curriculum chapters about those concepts.

The Current Study

This study investigates the validity of an automated assessment of creativity using a sample of projects created in EarSketch. We also aim to investigate what variables can predict students' creativity in EarSketch. Specifically, we will address the following research questions:

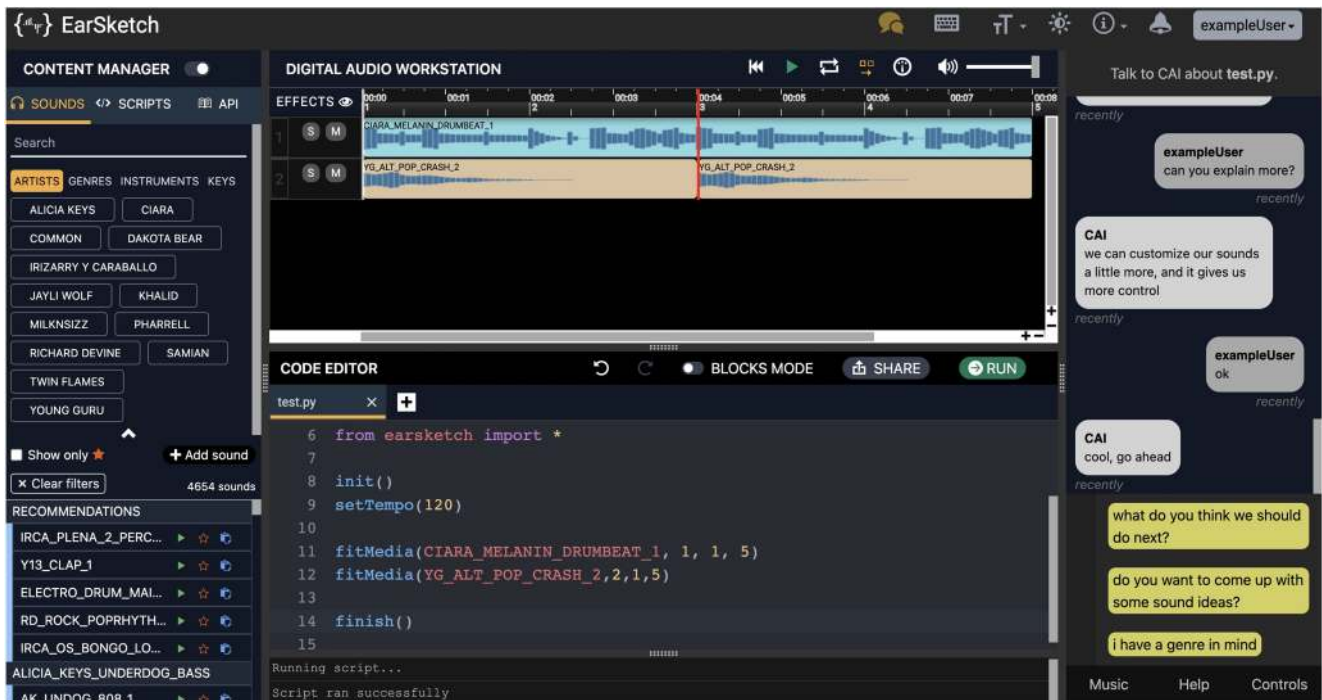
Research Question 1 (RQ1): To what extent is the automated assessment of creativity that we designed in EarSketch psychometrically sound (focusing on reliability and validity)?

Research Question 2 (RQ2): What variables (i.e., divergent thinking, complexity, and self-report variables) predict students' creativity in EarSketch?

We used the Consensual Assessment Technique (Amabile, 1996), discussed in detail in the Creativity Assessment section, where we asked a group of experts to rate the final product of the sample projects (i.e., the pieces of music only; not the code) for creativity. Then, we used those ratings for validation purposes. The findings of this study can provide evidence for a new automated assessment of creativity. This assessment, if psychometrically sound, can be used formatively and locally in the EarSketch environment, or it can be used for assessing creativity in general. Moreover, this study shows an example of how we can focus on two important areas in education, that is, STEM knowledge and creativity, at the same time. Usually, the focus of similar learning environments is on one or the other skill. However, in EarSketch, students learn to write programming code

Figure 1

The EarSketch Application's Sound Library (Left), Code Editor (Bottom), DAW (Top), and Experimental CAI Interface (Right)



Note. DAW = digital audio workstation; CAI = co-creative AI learning assistant; AI = artificial intelligence. See the online article for the color version of this figure.

while also expressing their creativity through the pieces of music they create in this learning environment.

Method

Participants and Data Set

The data were drawn from high school students' projects created in sessions held in five classrooms (each class had a different teacher) in the southeastern United States from our earlier studies investigating student use of CAI, a co-creative AI built for the EarSketch platform (Truesdell et al., 2021). EarSketch was designed for high school students to accommodate the level of knowledge needed in their computer programming classes. Therefore, the inclusion of high-school students in this study is justified. All projects in our data set were created on an experimental version of EarSketch that contains the CAI interface (see Figure 1) and allowed students to communicate with either the CAI agent or a researcher acting as CAI in a Wizard of Oz experiment (Dahlback et al., 1993). We used the data from 53 students (females = 11; males = 39; not listed = 3). Students identified as Asian ($n = 8$), Black/African American ($n = 11$), Hispanic/Latino ($n = 9$), and White ($n = 21$). More than half of the students ($n = 27$) had not taken any other courses related to programming; while some students had taken one other programming course ($n = 20$), and few students had taken two or more other programming courses ($n = 6$). Moreover, our sample included students from various grades (i.e., ninth = 29, 10th = 10, 11th = 8, and 12th = 6).

Procedure

Students were instructed to create a project (i.e., write code to produce a piece of music) in EarSketch, using the insertMedia, makeBeat, or fitMedia EarSketch application programming interface functions to add audio samples and setEffect for audio effects such as Reverb. They were instructed to use loops to repeat the audio samples. The assignments also included items encouraging students to use more complex coding structures such as loops if they have additional time. Further instructions directed students to vary their projects based on measure number or user input and create a percussion line for songs of at least eight measures or 15 s. Moreover, the students were instructed to use custom functions to define musical sections and use those functions for an A-B-A musical form (one with a section that repeats after a second, different section). Students were not instructed to be creative, as the goal of the assignment was more related to writing computer code. However, the expressive, open-ended nature of EarSketch allowed for creativity.

The study took place in five different existing computer science classrooms, and students were given the full duration of a class period (between 45 and 90 min) to complete their assignments. Participation in this study acted as participation in class for that day for consenting students, and no additional incentives were given to the students. After an introduction to the assignment by their teacher, they then used the remainder of the class period to work on their projects individually.

During these sessions, students completed their projects with CAI's assistance. All students whose projects were included in the data set had access to support in the form of the CAI agent or a Wizard of Oz communicating through the CAI interface. The CAI agent and wizards acting as the CAI agent are not able to edit user code; all project code included in the data set was written by the students themselves.

The current version of CAI does not provide suggestions that are intended to improve students' creativity. Additionally, since students made an average of 2.4 sound sample requests per project and 1.1 code/musical structure requests per project, the effect of CAI on the creativity of their projects was probably minimal. Therefore, since this work is intended to evaluate the creativity evident in the product (i.e., pieces of music) and it is not important in this validation study how that product was created (i.e., with or without the help of CAI), the inclusion of the CAI agent in student interfaces poses no risk to the validity of the results of the study. Assignments varied by classroom and were selected based on students' experience, teachers' curricular needs and the CAI features being tested during each session. This variation in assignments does not introduce a threat to the validity of our assessment because the external assessment of creativity (i.e., ratings of the experts) was solely based on the final piece of music, not the code written by the students.

Students' EarSketch code was saved after every time they ran the script, logged out of the page, or manually saved via keyboard commands. To create the music data set with 53 pieces of music (used for both the automated and external assessment of creativity in this study), we used the output of the student's last saved script version with a length longer than 0 measures/0 s. In some cases, multiple named scripts were saved for each student. All of these cases were from students who had renamed a script while working on it, and all saved names pertained to the same project; the most recently named version of the project was kept in the data set and earlier versions were excluded. Specifically, we excluded 10 projects because they were earlier versions of a project that we included in this study.

Research Design

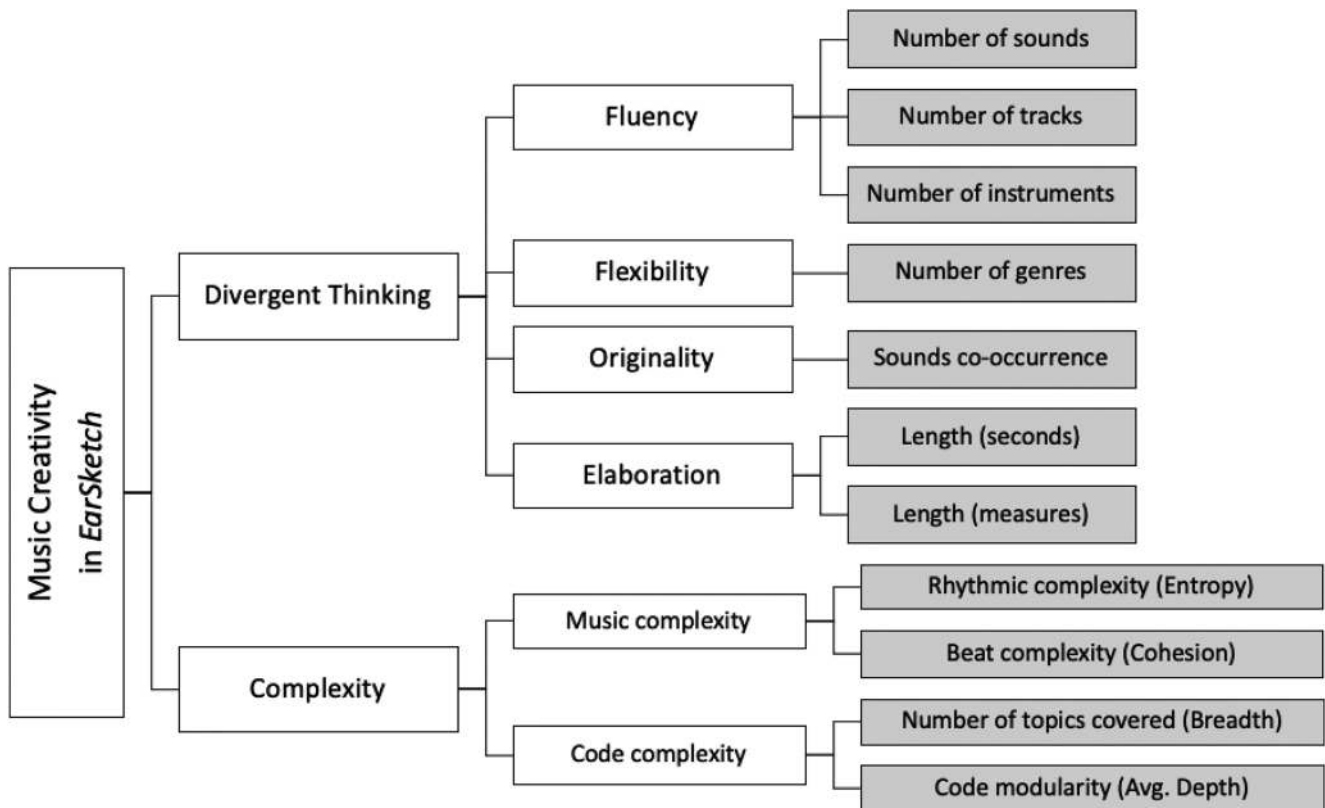
We used a one-group, within-subject, correlational research design where we used the data from 53 students and investigated the relationships among multiple variables (indicators/observables) extracted from log data in EarSketch, the self-report measures, and the external assessment of creativity to address our research questions.

The Competency, Evidence, and TMs

As we discussed earlier, in this study we operationalize the creativity CM inspired by the divergent thinking construct (Guilford, 1950) which is a proxy for creativity or creativity potential through individuals' responses (Runco, 2014): fluency, flexibility, originality, and elaboration. Moreover, we included complexity which is supported by the literature as an aspect of a creative product (Amabile, 2018; Guilford, 1950; Rahimi et al., 2023). After identifying the CM for creativity (unobservables), we brainstormed a list of potential indicators (observables; shown in gray boxes in Figure 2) for each CM subfacet—that is, the EM.

Figure 2 shows the CM and EM of creativity in EarSketch. Specifically, for the divergent thinking aspect of our CM, and for fluency, we included the number of sounds, tracks, and instruments. For flexibility, we only included the number of different genres in each project. For originality, since we could not compare the students' projects for statistical rarity, we created a variable called sound co-occurrence (i.e., the rareness of the occurrence of the sounds in a project—discussed later). For elaboration, we included the length of a project in seconds and measures. For complexity, we included both music complexity which consisted of rhythmic

Figure 2
The Creativity Competency and EMs in EarSketch



Note. EM = evidence model; Ave. = average.

and beat complexity (discussed later), and code complexity (breadth and average depth) of a project. Also, this assessment can be seen as a general creativity assessment. Therefore, the ability to show complexity (both from a music and code perspective) can be seen as evidence for creative thinking. We explain the rules of evidence for each of these variables in the next section.

Since we used an existing learning environment rather than building one from scratch, we did not design or modify the TM (i.e., the EarSketch environment). When one creates an automated assessment using ECD in an existing learning environment, the TM is already in place and the assessment designer may not have an opportunity to modify the TM; rather, they will look for the existing tasks and their features for collecting evidence (i.e., data) needed. EarSketch already provides numerous opportunities for students to show evidence of the creativity CM. Next, we explain the rules of evidence (i.e., scoring rules—a part of the EM) related to each indicator we used from EarSketch log data.

Rules of Evidence for Scoring the Observables

Rules of evidence in ECD refer to a set of rules (i.e., a rubric) that clearly define the observables and how they should be scored. Some of the observables can simply be computed by counting the number of times they occurred. However, some of the observables should be computed using a complex algorithm or sometimes using complex AI models (e.g., classification models). In the following, we define

and explain the rules of evidence for the automated assessment of creativity in EarSketch.

Code Complexity Assessment

An updated version of EarSketch's code complexity assessment tool was used to analyze the complexity of each project in the data set. The complexity analysis tool analyzes an abstract syntax tree generated from a user's EarSketch code in order to identify the complexity level at which the student utilizes programming concepts such as functions, loops, console input, and conditionals. Topics covered in the complexity rubric are drawn directly from the EarSketch curriculum, which aligns with the learning objectives for the advanced placement Computer Science Principles course (Magerko et al., 2016). The complexity analysis tool analyzes a user's project to estimate the user's knowledge of 17 concepts from the EarSketch curriculum.

In previous work on computer science assessment, general taxonomies of learning (such as the Bloom and Structure of Observed Learning Outcomes, i.e., SOLO, taxonomies) have been adapted to evaluate the level of complexity at which novice programmers understand computing topics (Fuller et al., 2007; Ginat & Menashe, 2015; Sheard et al., 2008; Starr et al., 2008; Thompson et al., 2008). Similarly, the complexity assessment tool assigns complexity scores between 0 and a value of 1–3 for each concept. Zero indicates that the concept does not appear in the user's code. Higher values indicate increasing levels of complexity of use. For example,

a conditional statement that only includes an if statement will receive a score of 1; a conditional that adds an else component will receive a score of 2; and a conditional that includes else if components to follow three or more code paths receive a score of 3. Some components have a maximum score of 1 or 2, such as the EarSketch function `makeBeat()`, which includes simple and advanced overrides (scored as 1 and 2, respectively). A sample of score levels for the “Functions—Repeat Execution” code concept appears in Table 1, and scripts demonstrating these score levels are shown in Figure 3.

The complexity analysis tool walks recursively through the nodes of the abstract syntax tree generated by student code, which enables analysis to run in the background of the EarSketch webpage without interrupting the user. This tool has been used to support a number of code analysis activities in EarSketch, including providing information about students’ code to a co-creative AI (J. B. Smith et al., 2020; Truesdell et al., 2021) and for bulk offline analysis of student scripts.

The complexity analysis system was originally designed to represent students’ use of code topics to provide information to an AI agent that allows it to target opportunities for code-related suggestions. We use summary metrics based on this analysis to provide a holistic view of students’ code sophistication as represented by the number of topics covered (breadth) and code modularity (average depth).

The complexity calculator includes measures of code breadth and depth, which can be understood as proxy values to cover the number of concepts used in a project and the modularity of the code in the project, respectively. The breadth score represents the number of concepts evaluated that are given a score of at least 1 (i.e., that are present in the project at all). The breadth score can thus be viewed as a representation of how many code concepts the student has utilized in their project.

The average depth value acts as a representation of code modularity. This value is calculated by assigning each nonblank, noncommented line of code a depth score. A depth score can be likened to the number of indentations present in a script line. A score of 1 indicates that the student is not using any modular constructs. This is most frequently seen when students make their music using a large number of calls to built-in EarSketch functions but do not engage with loops, conditionals, or custom functions. A depth score of 2 indicates that the student has used a modular construct (such as a loop or function) in their code, and the current line is part of that construct. A score of 3 or more indicates that the student has nested at least one modular construct inside another (e.g., a custom function that contains a for loop). The average depth is computed using the depth scores for an entire script. An increase in average depth indicates that the student has utilized a large number of modular structures, such as loops and functions, and/or that the student has minimized the number of top-level lines of code, both of which indicate an increase in code modularity. An example of code with the same musical output at different levels of modularity is shown in Figure 4.

Table 1
Knowledge Levels for the “Functions—Repeat Execution” Concept

Score	Description
0	Does not use a custom function
1	Uses a custom function to execute code once
2	Uses a custom function to execute code multiple times
3	Uses a custom function to execute code multiple times with parameters to alter output

A high level of breadth and average depth signifies code optimization (i.e., refining the code to enhance its efficiency and quality) which is valued in computer programming. The creation of an optimized and valuable product, such as code, inherently involves creativity. Code complexity, when contributing to optimization and efficiency, can be viewed positively. However, it is crucial to balance complexity with readability and manageability, ensuring that introduced intricacies align with optimization goals without compromising clarity. Code that is overly complex and difficult to understand can lead to maintenance challenges, debugging issues, and reduced collaboration among team members. In our study, code complexity serves as a measure reflecting the level of code optimization and efficiency, providing evidence for creativity.

Sound Profile Generation

In addition to the complexity calculator, EarSketch contains tools to model the musical contents of a project as a series of measure-by-measure musical and transitional elements that determine a song’s structure (J. B. Smith et al., 2020). By adjusting a distance threshold, a project can be divided into varying amounts of musical sections and subsections. These are used to create a Sound Profile (Truesdell et al., 2021), a nested dictionary comprising section-based representations of musical content. The Sound Profile data structure is designed to allow CAI to refer to a project’s musical structure when discussing it with a student. Each musical section is represented as an object with a list of sounds, effects, the measures and lines of code with which they appear, and the subsections contained within that section.

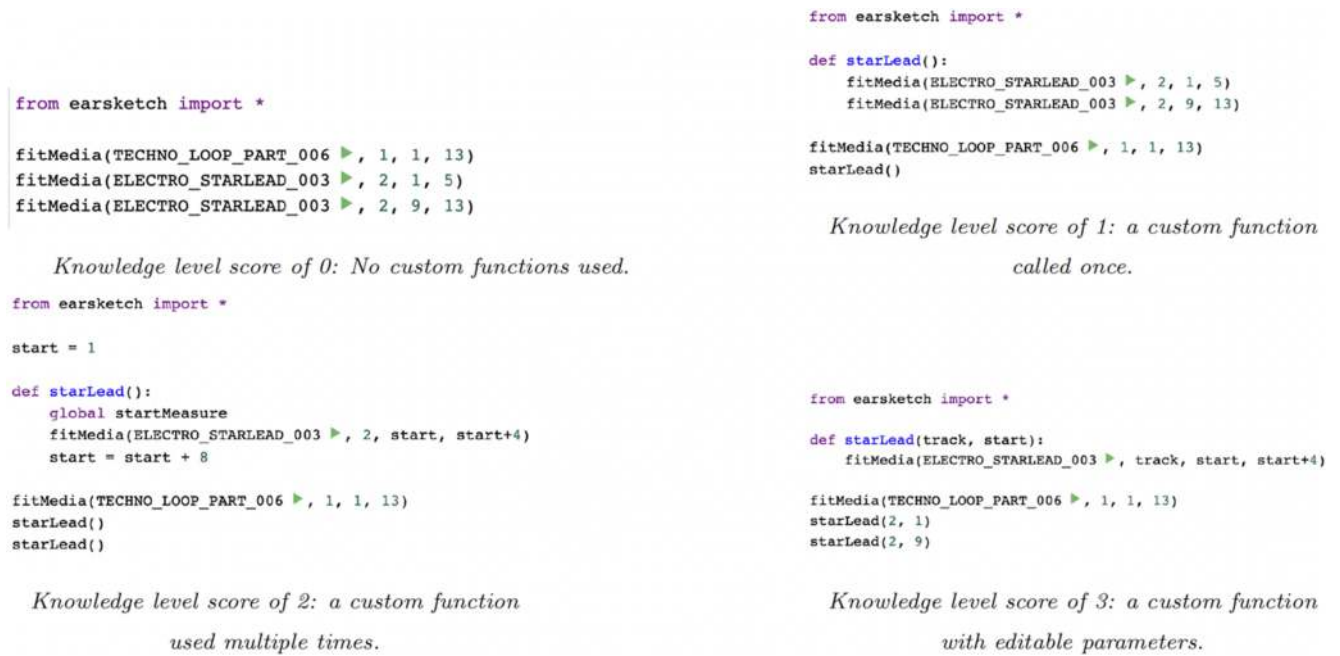
The complexity calculator and Sound Profile analysis tools were used to generate a series of values reflecting the code and music properties of the student projects in the data set. The values generated for each project included scores for each creativity subfacet (see Figure 2), breadth and average depth scores, number of unique sound samples used in the project (e.g., DUBSTEP BASS WOBBLE 001, a dubstep bass sample, and Y24 DRUMPAD 1, a short series of cymbal hits), the number of the unique genre (e.g., pop, rock, and hip-hop) and instrument (e.g., drums, bass, and guitar) tags present among the sounds in the project, the maximum number of sounds overlapping by being used simultaneously in separate tracks (each track could have multiple sounds and genres included, but a track cannot contain temporally overlapping sounds), the length of a project’s output song in both seconds and measures, and average co-occurrence scores of sound samples in the project. Some of these variables could be computed simply by counting the number of times they occurred (e.g., the number of sounds, tracks, genres, and instruments), other variables needed more computation (e.g., length in seconds and measures), and some other variables needed more complex computations such as the variables described in the Sound Co-Occurrence to Time on Task (Effort) sections.

Sound Co-Occurrence

We investigate unique sound selections as a proxy for originality. Each pair of sounds is given a co-occurrence score, based on the number of times the two sounds have been used at the same time in a large collection of previous EarSketch projects by various users (J. B. Smith et al., 2019). Using these scores, we analyze the data set of student projects to determine if they contain a lower

Figure 3

Four EarSketch Scripts of Increasing “Functions—Repeat Execution” Ratings With Identical Musical Output



Note. See the online article for the color version of this figure.

average co-occurrence score, which means that they contain statistically rarer combinations of sounds.

Rhythmic Complexity and Beat Cohesion

We followed the example set by Thul (2008) and used entropy as a measure of rhythmic complexity. Entropy measures the uncertainty in a given distribution, which in our case was a beat track, and is found to correlate with human perception of rhythm (de Fleurian et al., 2017). We computed entropy for each clip in a measure (a unit of musical time, divided into beats) and then averaged it to get an entropy score per measure. The higher the score, the higher the uncertainty in that distribution. This entropy score is listed as “rhythmic complexity” in Table 3.

For every sound clip in EarSketch, we computed a “click track” representation, that is, a binary vector where 1 indicates the presence of a beat at the timestamp and 0 indicates the absence of such a beat. Details on how this is computed are in J. B. Smith et al. (2023).

We define cohesion as the similarity between two given click tracks. To compute beat cohesion, we computed the hamming distance between every possible pair of successive sounds in a measure, as suggested by Toussaint (2004). The higher the hamming distance, the lower the similarity. Therefore, the lower the distance, the better the two tracks might be rhythmically cohesive. This rhythmic cohesion score is listed as “beat complexity” in Table 3.

Time on Task (Effort)

As an indicator of mental effort, the time-on-task percentage has been used in a variety of psychological user evaluations in human-computer interaction research and evaluations of students (Karweit, 1984). It can be calculated as the ratio of time intervals (in an even

division from the start to the end of a task) that contain actions or attention by a user related to the task at hand. In EarSketch, users demonstrate on-task behavior by editing a project or searching for content in the browsers or curriculum. Time-on-task percentage has been measured alongside the assessment of creativity (Benedek et al., 2013). We did not include effort (i.e., time on task) as an indicator of creativity in EarSketch. However, the effort can be seen as an indirect indicator of creativity and an indicator of motivation which research shows can affect creativity (Amabile et al., 2005).

The version of EarSketch used for studies also logs keystroke-level interactions between the user, the agent interface (either CAI or a human agent in the back), and the EarSketch interface. Between the initial start of a project and its submission, actions for each project are grouped into 10-s windows. If any actions, such as changes in the Code Editor, playing the project, or sending a message to CAI, are found between any two timestamps, then that corresponding point in the vector is assigned a “true” value. Each project is given a time-on-task percentage, determined by the ratio of points that contain actions by the user and the points that do not contain any actions logged. As time on task was calculated using these server logs, other actions taken by subjects outside of the EarSketch application, such as brainstorming with pen and article or conversations with their teacher, were not factored into time-on-task calculation.

Among the 53 projects in our data set, the average time-on-task percentage was 47.3%, with a *SD* of 18.5.

External Assessment of Creativity in Music Created in EarSketch

We used the Consensual Assessment Technique (Amabile, 1982) to externally assess the pieces of music ($n = 53$) and use the results from this assessment for validation purposes. Amabile (1982) asserts that if

Figure 4

Three EarSketch Scripts of Increasing Average Depth With Identical Musical Output

```
from earsketch import *

#Section A
fitMedia(KHALID_NORM_KEYS_STACK, 1, 1, 5)
fitMedia(Y20_PERCUSSION_5, 2, 1, 5)

#Section B
fitMedia(COMMON_LOVE_THEME_JUNO_1, 1, 5, 9)
fitMedia(YG_POP_TAMBOURINE_1, 2, 5, 9)
fitMedia(TECHNO_MAINLOOP_005, 3, 5, 9)

#Section A'
fitMedia(KHALID_NORM_KEYS_STACK, 1, 9, 13)
fitMedia(Y20_PERCUSSION_5, 2, 9, 13)
makeBeat(OS_KICK03, 3, 9, "0-000-000---0+--")
makeBeat(OS_KICK03, 3, 11, "0-000-000---0+--")

#Section B
fitMedia(CIARA_SET_THEME_PIANO, 1, 13, 17)
fitMedia(YG_POP_TAMBOURINE_1, 2, 13, 17)
fitMedia(TECHNO_MAINLOOP_005, 3, 13, 17)
```

Average Depth: 1.00

```
from earsketch import *

def sectionB(start, end):
    fitMedia(COMMON_LOVE_THEME_JUNO_1, 1, start, end)
    fitMedia(YG_POP_TAMBOURINE_1, 2, start, end)
    fitMedia(TECHNO_MAINLOOP_005, 3, start, end)

#Section A
fitMedia(KHALID_NORM_KEYS_STACK, 1, 1, 5)
fitMedia(Y20_PERCUSSION_5, 2, 1, 5)

#Section B
sectionB(5,9)

#Section A'
fitMedia(KHALID_NORM_KEYS_STACK, 1, 9, 13)
fitMedia(Y20_PERCUSSION_5, 2, 9, 13)
for measure in range(9, 12, 2):
    makeBeat(OS_KICK03, 3, measure, "0-000-000---0+--")

#Section B
sectionB(13,17)
```

Average Depth: 1.31

```
from earsketch import *

def sectionA(start, end, addDrums)
    fitMedia(KHALID_NORM_KEYS_STACK, 1, 9, 13)
    fitMedia(Y20_PERCUSSION_5, 2, 9, 13)
    if(addDrums):
        for measure in range(9, 12, 2):
            makeBeat(OS_KICK03, 3, measure, "0-000-000---0+--")

def sectionB(start, end):
    fitMedia(COMMON_LOVE_THEME_JUNO_1, 1, start, end)
    fitMedia(YG_POP_TAMBOURINE_1, 2, start, end)
    fitMedia(TECHNO_MAINLOOP_005, 3, start, end)

sectionA(1, 5, False) #Section A

sectionB(5,9) #Section B

sectionA(9, 13, True) #Section A'

sectionB(13,17) #Section B
```

Average Depth: 1.73

Note. See the online article for the color version of this figure.

a group of experts in the domain at hand believes a product (e.g., a piece of music) is creative, we should accept that product as creative. We asked 10 EarSketch experts who were very familiar with the music created using the platform to independently rate the 53 pieces of music on a 1–6 scale (1 = *very uncreative*, 2 = *uncreative*, 3 = *somewhat uncreative*, 4 = *somewhat creative*, 5 = *creative*, 6 = *very creative*) in random order. These experts have been judging the complexity of projects created in EarSketch in the past several years. Also,

they have been involved in research projects conducted using EarSketch. Thus, they were very well qualified as experts on what is possible with high school students creating music using EarSketch. We provided the raters with specific instructions:

Thank you for accepting to be one of our music-creativity raters. There is only one criterion in rating these pieces of music: creativity. We realize that creativity does not exist in a vacuum, and to some extent, creativity

probably overlaps other criteria one might apply—aesthetic appeal, organization, the richness of the music, the sophistication of expression, novelty of choice, appropriateness of notes and genres, for example—but we ask you to rate the pieces of music solely based on your thoughtful-but-subjective opinions of their creativity. The point is, you are the expert, and you do not need to defend your choices or articulate a definition of creativity. What creativity means to you can remain a mystery—what we want you to do is use that mysterious expert sense to rate the music for creativity. Note that you should judge these pieces of music based on what students can create using the *EarSketch* platform and not a radio-level, professional creativity. In the following pages, you will see 53 pieces of music with a Likert-scale rating question.

For our analyses, we averaged the 10 ratings for each piece of music and used that single score (range = 1–6) as our external measure of creativity. The reliability of the ratings was reasonable (Cronbach's $\alpha = .82$) indicating that the raters were consistent in their judgment for the 53 pieces of music (i.e., each rater rated all the 53 pieces of music for creativity) as an external assessment of creativity in *EarSketch*.

Demographic and Background Questionnaire

After working with *EarSketch*, the students responded to a demographic and background questionnaire which included questions about students' gender, ethnicity, grade level, and age. Moreover, the background questions asked students about the number of courses they took which included computing (possible choices: “no other course,” “one other course,” “two other courses,” and “more than two other courses”), the number of hours students usually listen to music (possible choices: “almost never,” “less than 1 hr per day,” “1–2 hr per day,” “3–4 hr per day,” and “5 hr per day”), and four Likert-scale items (from 1 = *strongly disagree* to

5 = *strongly agree*) about music enjoyment (Item: “I enjoy listening to music”), coding enjoyment (Item: “I enjoy coding”), making-music confidence (Item: “I have a lot of confidence when it comes to making music”), and learning-to-code confidence (Item: “I have a lot of confidence when it comes to learning to code”). We used the data from these variables in our analyses—discussed next.

Data Preparation

We used Python to identify and score the variables (observables) from the log data and prepare the final data set (i.e., merging the data set with the observables and the data set with the demographic and background data) for analysis. We then used SPSS software (IBM, 2017) to conduct our analyses.

To aggregate the automated assessment overall creativity estimates per student (i.e., shown as creativity [automated] in Tables 3 and 4), we first standardized the low-level indicators—shown in gray boxes in Figure 2—(computing the *z*-scores for each indicator). Then, we averaged those standardized estimates to compute the subfacet variables: fluency, flexibility, originality, elaboration, and music and code complexities. Next, we computed the divergent thinking and complexity scores by averaging the scores of their relevant subfacets (e.g., averaged fluency, flexibility, originality, and elaboration for computing each student's divergent thinking skill; and averaged music and code complexity indicators to compute each student's complexity score). Finally, we averaged the divergent thinking and complexity scores to obtain the overall automated creativity score per student. Table 2 defines all the variables we used to compute the subfacets of creativity. We used the accumulated subfacet variables (e.g., fluency, flexibility, complexity, etc.) and the overall automated assessment of creativity scores to address our research questions.

Table 2
The Main Variables We Used for Analysis

Variable name	Description
Creativity (external)	Average ratings from the human raters
Number of sounds	Number of unique sound samples per project
Number of tracks	Number of sample tracks per project
Number of instruments	Number of unique instrument tags on sound samples per project
Number of genres	Number of unique genre tags on sound samples per project
Avg. sounds co-occurrence	Average co-occurrence score between all sample pairs per project
Length (s)	Music length in seconds
Length (measures)	Music length in measures
Rhythmic complexity	The uncertainty in a beat track (i.e., entropy)
Beat complexity	Similarity between two given click tracks (i.e., cohesion)
Breadth	Number of topics covered. The complexity of the topics per project
Avg. depth	Average depth of a node in a project (i.e., code Modularity)
Time on task (effort)	The ratio of time intervals that contain actions by a user
Creativity (automated)	Overall creativity in <i>EarSketch</i>
Age	Age of student
Music making enjoyment	Students' self-report on “I enjoy making music” ^a
Coding enjoyment	Students' self-report on “I enjoy coding” ^a
Hours listening to music	Students' self-report on the hours per day they listen to music on average (i.e., almost never, <1, 1–2, 3–4, and ≥5 hr)
Music making confidence	Students' self-report on “I have a lot of self-confidence when it comes to making music” ^a
Coding learning confidence	Students' self-report on “I have a lot of self-confidence when it comes to learning to code” ^a

Note. Avg. = average.

^a 1–5 scale with 1 = *strongly disagree* to 5 = *strongly agree*.

Table 3
Descriptive Statistics (n = 53)

Variable name	Min	Max	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
Creativity (external)	2.50	5.30	3.86	0.70	−0.30	−0.88
From EarSketch log data						
Number of sounds	1.00	14.00	5.79	2.71	0.33	0.52
Number of tracks	1.00	101.00	6.72	13.47	6.84	48.59
Number of instruments	1.00	7.00	2.98	1.34	0.54	0.39
Number of genres	1.00	7.00	2.60	1.56	1.08	0.85
Avg. sound co-occurrence	0.00	0.10	0.01	0.02	3.17	9.33
Length (s)	2.00	118.00	32.48	21.04	1.69	4.25
Length (measures)	1.00	59.00	16.10	11.18	1.94	4.74
Breadth	0.00	8.00	1.62	1.99	1.50	1.86
Avg. depth	0.00	2.03	1.19	0.35	0.10	1.78
Rhythmic complexity	0.00	2.63	1.78	0.70	−1.24	0.85
Beat complexity	0.00	0.03	0.02	0.01	−0.60	−0.09
Effort (percentage time on task)	12.96	78.90	48.32	15.21	−0.37	−0.25
Divergent thinking	−0.72	1.49	0.05	0.56	0.69	−0.14
Complexity	−1.06	1.64	0.01	0.60	0.41	−0.01
Creativity (automated)	0.00	1.71	0.89	0.45	−0.08	−0.97
From self-report data						
Age	14.00	18.00	15.40	1.28	0.76	−0.49
Music making enjoyment	1.00	5.00	3.68	1.19	−0.41	−0.84
Coding enjoyment	1.00	5.00	3.81	0.96	−0.95	1.25
Hours listening to music	1.00	5.00	3.68	1.03	−0.50	−0.41
Music making confidence	1.00	5.00	3.08	1.17	0.07	−0.64
Coding learning confidence	1.00	5.00	3.53	0.87	−0.73	1.33

Note. Min = minimum; Max = maximum; Avg. = average.

Results

Preliminaries

The descriptive statistics related to the main variables we used for our analyses are shown in Table 3. Generally, students created pieces of music that were considered somewhat creative on average ($M = 3.86$, $SD = 0.70$). The time on task (effort) was about 48% on average ($SD = 15.21$) across the 53 students (see the Time on Task [Effort] section). That is, on average, the students spent 48% of their time producing code, testing their code by listening to their projects or reading the

EarSketch curriculum. While some of the low-level variables (i.e., shown in gray boxes in the CM; Figure 2) are somewhat skewed, the computed, main variables (i.e., divergent thinking, complexity, creativity [external], and creativity [automated]) that we used for our main analyses are normally distributed with Skewness and Kurtosis between -2 and 2 .

To examine the reliability of the automated assessment of creativity in EarSketch we computed the Cronbach's α coefficient using nine variables: number of sounds, number of tracks, number of instruments, number of genres, sound co-occurrence, length

Table 4
Correlation Matrix as Evidence for Convergent and Discriminant Validity of the Automated Assessment of Creativity in EarSketch (n = 53)

Variable name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Creativity (external)	—															
2. Fluency	.37	—														
3. Flexibility	.30	.48	—													
4. Originality	.07	.08	.15	—												
5. Elaboration	.21	.43	−.08	−.20	—											
6. Music complexity	.23	.23	.18	.05	.23	—										
7. Code complexity	.26	.13	−.06	−.17	.19	.08	—									
8. Divergent thinking	.39	.79	.67	.50	.42	.28	.01	—								
9. Complexity	.33	.24	.07	−.09	.28	.68	.79	.18	—							
10. Creativity (automated)	.47	.66	.47	.25	.45	.63	.53	.75	.78	—						
11. Effort	.25	.19	−.04	−.10	.32	−.04	.03	.13	−.01	.08	—					
12. Age	.03	−.06	.03	−.06	−.13	.15	.28	−.09	.30	.14	.03	—				
13. Music making enjoyment	−.01	.07	−.12	−.16	.18	−.17	.11	−.04	−.02	−.04	.09	.06	—			
14. Coding enjoyment	.08	−.23	−.10	−.03	−.28	.07	.32	−.26	.28	.03	.01	.09	.20	—		
15. Hours listening to music	.05	.12	−.05	−.03	.04	−.05	.07	.03	.02	.03	.21	−.02	.24	−.04	—	
16. Music making confidence	.06	.00	−.05	.09	−.18	−.01	.09	−.05	.06	.01	−.15	.21	.61	.25	−.01	—
17. Coding learning confidence	.30	−.19	.00	.01	−.13	.18	.31	−.11	.34	.16	.10	.31	.17	.67	−.07	.22

Note. The correlations related to convergent validity are boldfaced; those related to discriminant validity are boldfaced and italicized.

(second), length (measures), music complexity, and code complexity. Results showed a Cronbach's α of .70. It is worth noting that the nature of this assessment is formative and needs more longitudinal data per student for higher reliability (DiCerbo et al., 2016).

We investigated the construct validity of our automated assessment of creativity by looking into its convergent validity and discriminant validity (Cronbach & Meehl, 1955; G. T. Smith, 2005). To investigate the convergent validity of the automated assessment of creativity in EarSketch, we correlated the observables computed from the log data in EarSketch with the external assessment of creativity (the expert ratings).

Results (see Table 4) show that all but the originality subfacet of creativity positively correlated with the external assessment of creativity (the average human ratings). Specifically, fluency ($r = .37$), flexibility ($r = .30$), elaboration ($r = .21$), music complexity ($r = .23$), and code complexity ($r = .26$) have reasonable, small to medium correlations with the external assessment of creativity. On a higher level, the divergent thinking score ($r = .39$) and the complexity score ($r = .33$) showed positive, medium correlations with the external creativity assessment. Moreover, effort ($r = .25$) also showed a small to medium correlation with the external creativity assessment indicating the effect of students' engagement and the creativity of their project. Finally, the overall automated assessment of creativity (the average score of the divergent thinking skills and complexity) correlated positively ($r = .47$) with the external assessment of creativity (the average of expert ratings). All these positive correlations provide good evidence for the convergent validity of our automated assessment of creativity in EarSketch. We then turned our attention to examining the discriminant validity evaluation of our automated assessment of creativity in EarSketch (see Table 4). We correlated the external assessment of creativity, the automated assessment of creativity, and its subfacets (i.e., fluency, flexibility, originality, elaboration, complexity, and effort) with students' age and the other self-report measures (i.e., music-making enjoyment, coding enjoyment, hours listening to music, music-making confidence, and coding confidence). Results show that there is no relationship between the self-report measures and the automated assessment of creativity EarSketch and its subfacets. This finding provides evidence that the automated assessment of creativity in EarSketch may be independent of those variables. For instance, students' confidence in making music did not correlate with the automated assessment of creativity. In short, these results indicate that the automated assessment of creativity has an acceptable level of discriminant validity (i.e., constructs that theoretically should not be correlated are not, actually, related).

Since we used an existing data set and could not compute a power analysis beforehand, we conducted a post hoc power analysis using G*Power 3 (Faul et al., 2007). After selecting the correlation, point

biserial model from the statistical test drop-down, and selecting the "Post hoc: Compute achieved power given α , sample size, and effect size" in G*Power, we calculated the effect size for the smallest ($r = .21$) and the largest ($r = .47$), significant correlations we found and computed the achieved power for both separately. We selected α to be .05 and our sample size to be 53. The computed effect sizes for our small and large correlations were .46 and .69, respectively. Results showed that we achieved acceptable powers of .96 and .99, respectively, which indicate our sample size was sufficient to detect the effect we were looking for.

Finally, we ran a step-wise multiple regression analysis with the external assessment of creativity (i.e., experts' rating) as the dependent variable and the self-report measures (i.e., music-making enjoyment, coding enjoyment, hours listening to music, music-making confidence, coding learning confidence, and age) along with the two subfacets of our automated assessment of creativity (divergent thinking and complexity sub facets) as predictors. As shown in Table 4, the predictors are not highly correlated. Moreover, the tolerance and variance inflation factor shown in Table 5, raise no concerns related to multicollinearity of the predictors.

After removing the nonsignificant predictors (i.e., complexity, music-making enjoyment, coding enjoyment, hours listening to music, music-making confidence, and age), the best model (see Table 5) with an adjusted R^2 of .24 (i.e., 24% of the variability in the external assessment of creativity is explained by the model), $F(2, 50) = 9.23, p < .001$, showed that the divergent thinking subfacet ($B = 0.43, p < .001$) and learning coding confidence ($B = 0.35, p = .006$) significantly predicted students' creativity. That is, with one unit change in students' divergent thinking, their score on the external assessment of creativity changes by 0.43 SDs controlling for learning coding confidence. Also, with one unit change in students' learning coding confidence, their score on the external assessment of creativity changes by 0.35 SDs controlling for their divergent thinking. Therefore, divergent thinking has a higher predictive power predicting students' creativity in EarSketch than students' confidence in learning coding. However, their confidence in learning coding (a proxy for their coding skills) is essential for their creativity in EarSketch. Next, we discuss these findings in detail.

Discussion

In this study, we used an existing digital learning environment called EarSketch (an expressive computing platform designed to engage high school students in computer science through the computational remixing of music samples from popular artists and genres) to design and validate an automated assessment of creativity. The data from 53 students were used to address our two research questions regarding the psychometric qualities of the automated assessment of creativity

Table 5
Regression Model Predicting Students' Creativity Rated by the Experts ($n = 53$)

Predictors	B	SE	B	t	p	95% CI for B		Collinearity statistics	
						LB	UB	Tolerance	VIF
Constant	2.84	0.36		7.90	<.001	2.12	3.56		
DT	0.53	0.15	0.43	3.50	<.001	0.23	0.84	0.99	1.01
CLC	0.28	0.10	0.35	2.86	.006	0.08	0.48	0.99	1.01

Note. CI = confidence interval; LB = lower bound; UB = upper bound; VIF = variance inflation factor; DT = divergent thinking; CLC = coding learning confidence.

(RQ1), and the predictive power of the assessment and other self-report variables we included in this study (RQ2). This section discusses our findings, the limitations, and promising future directions.

RQ1: Psychometric Qualities of the Automated Assessment of Creativity

Regarding RQ1, the reliability and validity of our automated assessment of creativity should be examined from various standpoints. First, since the nature of this automated assessment of creativity is meant to be formative, more data across several student projects are needed to establish a robust reliability estimate of the assessment. Second, this type of assessment of creativity has more ecological validity (i.e., the extent to which the estimates of an assessment are generalizable to the real world, such as situations or settings typical of everyday life (Runco, 2014) compared to traditional assessments of creativity (e.g., list as many alternative uses for a pen). In this study, we used an actual, real-world situation or environment (e.g., EarSketch) to embed a performance-based assessment of creativity, and thought about what behaviors or products could be seen as creative in this situation. Therefore, the automated assessment of creativity has good ecological validity. Second, using ECD, consulting the literature on creativity, and consulting with music experts in EarSketch, we were able to establish the face validity and content validity of our CM for creativity in EarSketch. That is, we included relevant subfacets of creativity in our CM, and we included sufficient indicators per subfacet from the EarSketch environment per subfacet (i.e., the coverage of our assessment was reasonable).

Moreover, using empirical investigation, we found evidence for the construct validity of our automated assessment of creativity (i.e., convergent and discriminant validity; Cronbach & Meehl, 1955; G. T. Smith, 2005). In terms of convergent validity, our results showed that the automated assessment of creativity in EarSketch correlated very well with our external assessment of creativity both on an overall and subfacet level (ranging from $r = .21$ to $r = .47$). These results are promising as we only had one project per student, and this was our first pass at designing and examining this assessment. If we could include multiple projects' data per student, the accuracy of our automated assessment could improve. In terms of discriminant validity, our results suggest that there are no relationships between the automated assessment estimates and the self-report data (e.g., students did not produce creative pieces of music because they enjoyed listening to music or coding). Overall, we can conclude that our automated assessment of creativity in EarSketch has a reasonable, promising level of validity based on the current data.

In this study, we provided evidence for our automated assessment accuracy (i.e., convergent validity and discriminant) or what Kane (2001) refers to as criterion validity. We also provided evidence for the modern view on validity which is evidence-based for each step of the assessment—from every step of designing the assessment to checking the accuracy of the assessment after the fact (Kane, 2001, 2013). It is worth noting that this automated assessment of creativity is meant for formative rather than summative purposes. Moreover, the nature of the complex learning environment EarSketch may have provided evidence that includes multiple factors with differentially correlated errors. Such evidence can negatively influence the computation of a reliability coefficient. Therefore, we need to approach the reliability estimation using more longitudinal data which can help minimize error (DiCerbo et al., 2016). For instance,

if we could include multiple projects per student for several attempts, the reliability estimates could be improved. Additionally, collecting more data from students' projects can enhance both the validity and reliability of the automated assessment of creativity in the future.

RQ2: Predictors of Students' Creativity

We addressed our second research question (What variables can predict students' creativity?) by conducting a regression analysis. Results showed that our automated assessment of creativity together with students' confidence in learning coding could significantly predict students' creativity in EarSketch. This finding aligns well with the amusing park theoretical model of creativity (Baer & Kaufman, 2005). Amusing park theoretical includes general factors (i.e., initial requirements) that affect creativity in all domains and several levels of domain-specific factors that affect creative performance in specific activities (i.e., general thematic areas, domains, and micro-domains). In other words, to be creative in a domain one must know the knowledge and skills required to a reasonable level. We did not have an external measure of students' programming knowledge and skills. However, the code complexity analyses along with the coding confidence self-report measure may serve as suitable proxies for students' programming skills. Thus, our findings provide evidence for creativity being both domain-general and domain-specific (Baer & Kaufman, 2005, 2019; Hong & Milgram, 2010; Kaufman & Baer, 2004; Meusbarger et al., 2009; Plucker et al., 2004).

Despite years of research on the topic of creativity domains and contexts, there is still an ongoing debate about whether creativity is truly domain-general, domain-specific, or a mixture, and how creativity manifests itself in various contexts (Kaufman & Baer, 2004, 2005; Plucker et al., 2004; Qian et al., 2019). For many years, some studies identified creativity as more domain-general than specific (Plucker et al., 2004; Qian et al., 2019), while other studies showed that creativity is domain-specific (Baer, 1994, 1998; Julmi & Scherm, 2015; Palmiero et al., 2015). For instance, Palmiero et al. (2015) found that visual creativity is domain- and even task-specific. However, most creativity scholars think that creativity is neither domain-specific nor domain-general, but rather has elements of both (Baer & Kaufman, 2005, 2019; Hong & Milgram, 2010; Kaufman & Baer, 2004; Meusbarger et al., 2009; Plucker et al., 2004). Since our CM of creativity includes general creative thinking aspects (i.e., divergent thinking facets) and domain-specific, product-based aspects, and the regression analysis showed both divergent thinking (i.e., general creative thinking skills) and confidence in learning how to code (i.e., a proxy for domain-specific knowledge and interest) as good predictors of creative performance, our study's findings support the point of view of creativity being both domain-general and domain-specific. Next, we discuss the possible implications of this study.

Implications

Overall, this study can be used as a model for other researchers who want to design and develop STEM learning environments that synergically assess and improve learners' creativity and STEM knowledge and skills. Moreover, STEM learning environment designers who focus on music and programming can take various pieces of the automated creativity assessment we designed (e.g., the CM) and embed them into their learning environments. In the following, we elaborate

on the implications of our study from educational and creativity assessment perspectives.

Educational Implication

Creativity holds a paramount position in the revised Bloom's taxonomy of learning objectives (Krathwohl, 2002), indicating its significance as the highest level of learning in education. Despite this importance, creativity is often overlooked or even discouraged in formal education (Kaufman & Sternberg, 2007). The development of expressive learning environments, such as EarSketch, which seamlessly integrates a challenging subject such as computer programming with creativity, stands as a successful example of blending creativity and learning in education. This integration brings creativity to the forefront of the learning process, prompting creativity researchers, learning scientists, and educational technologists to consider how to assess and nurture creativity. Moreover, creativity acts as a catalyst for enhancing students' intrinsic motivation, leading to increased engagement and improved learning outcomes, particularly in challenging subjects. This study serves as an exemplary illustration of synergistically including creativity in formal education. The incorporation of creativity within EarSketch has the potential to reshape educators' perceptions, encouraging a paradigm shift that views creativity as a valuable skill complementing technical expertise rather than a disruptive element—especially in STEM-related learning environments. Fostering creativity in STEM education is essential for developing a workforce capable of addressing complex and novel problems and driving innovation in STEM fields.

Creativity Assessment Implication

As discussed earlier, unlike traditional assessments of creativity, the automated, formative creativity assessment in EarSketch entails higher ecological validity. The process we employed, particularly the use of ECD, can be adapted for similar learning environments, paving the way for the creation of comparable assessments of creativity. For example, the globally renowned block-based programming platform, Scratch (Resnick, 2018; Resnick et al., 2005), offers numerous affordances for students worldwide to express their creativity. It serves as an excellent candidate for designing and implementing an automated, formative assessment of creativity. These ecologically valid assessments of creativity can inspire other researchers in the field to develop new tools, pushing the boundaries of creativity assessment (Rahimi, 2023). We contend that capturing students' creativity across multiple learning environments and using ecologically valid assessments longitudinal, provides a more comprehensive representation of their creativity—both domain-specific and domain-general—compared to isolated, lab-based, one-time assessments. Advances in technologies and the intersection of psychometrics, creativity, and learning sciences enable us to create, evaluate, and implement such assessments of creativity at scale (Shute et al., 2016). Next, we discuss the limitations and future research directions.

Limitations and Future Research

We faced several limitations during this study. First, since we used an existing learning environment for creating the automated assessment of creativity, we were bound by the log data that existed in EarSketch. Second, since the data were collected before in another study for other purposes, we did not have the option to include other

external measures of creativity or creativity in general. Therefore, we used a Consensual Assessment Technique for assessing students' creativity in producing music in EarSketch. Still, the Consensual Assessment Technique was useful and reasonable since the raters rated the pieces of music based on their definition of creativity and not based on a rubric that included similar aspects of the music generated by the students (e.g., the number of sounds, the number of genres). Future studies could benefit from the inclusion of external measures such as traditional tests of creativity for validation purposes. Another future direction could be the use of statistical models such as Bayesian Networks to improve the accuracy of the accumulation process. In those models, we could use the correlations we found in this study as the weights for each subfacet of our CM rather than considering equal weights for all the subfacets. These considerations can help with the accuracy of our assessment. Finally, finding 10 EarSketch experts for the consensual assessment was challenging. Ideally, we were looking for 15 raters, however, our efforts to increase the number of experts, to more than 10 raters, were not successful. Future studies, depending on the domain at hand, could overcome this challenge.

We also were not able to conduct appropriate analyses related to fairness (e.g., Differential Item Functioning analyses such as the Mantel-Haenszel test) because of our small sample size for each categorical variable (e.g., gender, race, and ethnicity). Future studies can help us improve the validity and reliability and further investigate the other psychometric parameters such as the fairness of our automated assessment using a larger sample size.

Moreover, in the future, we can enhance the accuracy of our creativity automated assessment by including other indicators in our EM. For instance, we can incorporate creativity indicators within students' project code. This might entail assessing student-generated code for signs of originality or departure from the instructed norms. Although we made progress in automatically evaluating students' code using a complexity measure, there is potential to identify additional indicators of divergent thinking within their code. Additionally, we can improve the accuracy of our originality indicator (i.e., sound co-occurrence) as more data gets collected by EarSketch and more sounds get used in the projects. Moreover, future studies can examine the effect of CAI on students' learning and the creativity of their products.

One other important promising use of automated assessment is to leverage the automated assessment estimates in real time to inform personalized support. For example, in real time the automated assessment can guide the suggestions made by CAI (e.g., suggesting to the students to increase the variety of their sounds); and holds the potential to inform some form of adaptation to maximize learning (e.g., Shute et al., 2020). One natural next step could be putting this automated assessment of creativity to use in EarSketch. Moreover, this automated assessment could be replicated in other similar learning environments. When the automated assessment functions in real time, we will have a complete cycle of automated assessment in place as this assessment method is meant to be mostly used for formative rather than summative purposes (Rahimi & Shute, 2023).

Conclusion

In conclusion, the automated, formative assessment of creativity presents a promising approach to measuring creativity objectively and accurately in a variety of settings. By improving our ability to assess creativity, we can better understand how it can be developed and enhanced in individuals across different domains and

environments. The results of this research have the potential to inform the design of innovative STEM-rated educational programs and interventions that foster learners' creativity and innovation. Moreover, the method we used in this study can be used by other researchers to design and validate automated assessments of other competencies in other digital learning environments. As we continue to explore new ways of assessing creativity, we can pave the way for a more creative and innovative society, where individuals are equipped with the skills they need to tackle the challenges of the future.

References

- Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2024). Measuring original thinking in elementary school. *Journal of Educational Psychology*, 1–29. Advance online publication. <https://doi.org/10.1037/edu0000844>
- Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2018). *Improving stealth assessment in game-based learning with LSTM-based analytics*. In International Conference on Educational Data Mining, Buffalo, NY USA.
- Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian Networks in educational assessment*. Springer.
- Almond, R. G., Shute, V. J., Tingir, S., & Rahimi, S. (2020). Identifying observable outcomes in gamebased assessments. In H. Jiao & R. W. Lissitz (Eds.), *Innovative psychometric modeling and methods* (pp. 163–191). IAP.
- 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*. Westview Press Harper Collins Publishers.
- Amabile, T. M. (2018). Creativity and the labor of love. In R. J. Sternberg, J. C. Kaufman (Eds.), *The nature of human creativity* (pp. 1–15). Cambridge University Press.
- Amabile, T. M., Barsade, S. G., Mueller, J. S., & Staw, B. M. (2005). Affect and creativity at work. *Administrative Science Quarterly*, 50(3), 367–403. <https://doi.org/10.2189/asqu.2005.50.3.367>
- Baer, J. (1994). Divergent thinking is not a general trait: A multidomain training experiment. *Creativity Research Journal*, 7(1), 35–46. <https://doi.org/10.1080/10400419409534507>
- Baer, J. (1998). The case for domain specificity of creativity. *Creativity Research Journal*, 11(2), 173–177. https://doi.org/10.1207/s15326934crj1102_7
- Baer, J., & Kaufman, J. C. (2005). Bridging generality and specificity: The amusement park theoretical (APT) model of creativity. *Roeper Review*, 27(3), 158–163. <https://doi.org/10.1080/02783190509554310>
- Baer, J., & Kaufman, J. C. (2019). Assessing creativity with the consensual assessment technique. In I. Lebeda & V. P. Glăveanu (Eds.), *The Palgrave Handbook of social creativity research, Palgrave studies in creativity and culture* (pp. 27–37). Springer.
- Barbot, B. (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in Psychology*, 9, 1–8. <https://doi.org/10.3389/fpsyg.2018.02529>
- Beatty, 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>
- Beatty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal*, 34(3), 245–260. <https://doi.org/10.1080/10400419.2022.2025720>
- Beatty, R. E., Zeitlen, D. C., Baker, B. S., & Kenett, Y. N. (2021). Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity*, 41, Article 100859. <https://doi.org/10.1016/j.tsc.2021.100859>
- Benedek, M., Mühlmann, C., Jauk, E., & Neubauer, A. C. (2013). Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top ideas and time-on-task on reliability and validity. *Psychology of Aesthetics, Creativity, and the Arts*, 7(4), 341–349. <https://doi.org/10.1037/a0033644>
- Craft, A. (2010). *Creativity and education futures: Learning in a digital age*. ERIC.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>
- Cropley, A., & Cropley, D. (2011). Creativity and lawbreaking. *Creativity Research Journal*, 23(4), 313–320. <https://doi.org/10.1080/10400419.2011.621817>
- Dahlback, N., Jonsson, A., & Ahrenberg, L. (1993). *Wizard of oz studies: Why and how*. Proceedings of the 1st International Conference on Intelligent User Interfaces (pp. 193–200). Orlando, Florida, USA.
- Dede, C., Etemadi, A., & Forshaw, T. (2021). *Intelligence augmentation: Upskilling humans to complement AI*. The Next Level Lab at the Harvard Graduate School of Education. President and Fellows of Harvard College.
- 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>
- DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology & Society*, 17(1), 17–28. <https://www.jstor.org/stable/jeductechsoci.17.1.17>
- DiCerbo, K. E., Shute, V., & Kim, Y. J. (2016). The future of assessment in technology-rich environments: Psychometric considerations. In M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology* (pp. 1–21). Springer. https://doi.org/10.1007/978-3-319-17727-4_66-1
- Education & Human Resources. (2020). *STEM education for the future: A visioning report* [Technical report]. National Science Foundation.
- 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* (pp. 183–188). Association for Computing Machinery.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Florida, R. L. (2002). *The rise of the creative class: And how it's transforming work, leisure, community and everyday life*. Basic Books.
- Fuller, U., Johnson, C. G., Ahoniemi, T., Cukierman, D., Hernan-Losada, I., Jackova, J., Lahtinen, E., Lewis, T. L., Thompson, D. M., Riedesel, C., & Hernán-Losada, I. (2007). Developing a computer science-specific learning taxonomy. *ACM SIGCSE Bulletin*, 39(4), 152–170. <https://doi.org/10.1145/1345375.1345438>
- Ginat, D., & Menashe, E. (2015). SOLO taxonomy for assessing novices' algorithmic design. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education* (pp. 452–457). ACM.
- Glăveanu, V. P., Hanchett Hanson, M., Baer, J., Barbot, B., Clapp, E. P., Corazza, G. E., Hennessey, B., Kaufman, J. C., Lebeda, I., & Lubart, T. (2020). Advancing creativity theory and research: A socio-cultural manifesto. *The Journal of Creative Behavior*, 54(3), 741–745. <https://doi.org/10.1002/jocb.395>
- Gray, H. (2016). Precarious diversity: Representation and demography. In M. Curtin & K. Sanson (Eds.), *Precarious creativity* (pp. 241–253). University of California Press.
- Guilford, J. P. (1950). *Fundamental statistics in psychology and education*. McGraw-Hill.

- Guilford, J. P. (1956). *Fundamental statistics in psychology and education* (3rd ed.). McGraw-Hill.
- Hennessey, B. A., & Amabile, T. M. (2010). Creativity. *Annual Review of Psychology*, 61(1), 569–598. <https://doi.org/10.1146/annurev.psych.093008.100416>
- 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>
- Hewitt, R. (2019). *SpinCycle project*.
- Hickey, M. (2001). An application of Amabile's consensual assessment technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education*, 49(3), 234–244. <https://doi.org/10.2307/3345709>
- Hickey, M., & Webster, P. (2001). Creative thinking in music. *Music Educators Journal*, 88(1), 19–23. <https://doi.org/10.2307/3399772>
- Hong, E., & Milgram, R. M. (2010). Creative thinking ability: Domain generality and specificity. *Creativity Research Journal*, 22(3), 272–287. <https://doi.org/10.1080/10400419.2010.503535>
- IBM. (2017). *IBM SPSS statistics*.
- Julmi, C., & Scherm, E. (2015). The domain-specificity of creativity: Insights from new phenomenology. *Creativity Research Journal*, 27(2), 151–159. <https://doi.org/10.1080/10400419.2015.1030310>
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342. <https://doi.org/10.1111/j.1745-3984.2001.tb01130.x>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Karweit, N. (1984). Time-on-task reconsidered: Synthesis of research on time and learning. *Educational Leadership*, 41(8), 32–35. <https://eric.ed.gov/?id=EJ299538>
- Kaufman, J. C., & Baer, J. (2004). Sure, I'm creative—But not in mathematics!: Self-reported creativity in diverse domains. *Empirical Studies of the Arts*, 22(2), 143–155. <https://doi.org/10.2190/26HQ-VHE8-GTLN-BJMM>
- Kaufman, J. C., & Baer, J. (2005). *Creativity across domains: Faces of the muse*. Psychology Press.
- 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. (2007). Resource review: Creativity. *Change: The Magazine of Higher Learning*, 39(4), 55–60. <https://doi.org/10.3200/CHNG.39.4.55-C4>
- Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into Practice*, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- Lou, S.-J., Chou, Y.-C., Shih, R.-C., & Chung, C.-C. (2017). A study of creativity in CaC2 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>
- Magerko, B., Freeman, J., Mcklin, T., Reilly, M., Livingston, E., Mccoid, S., & Crews-Brown, A. (2016). *EarSketch*: A steam-based approach for underrepresented populations in high school computer science education. *ACM Transactions on Computing Education*, 16(4), 1–25. <https://doi.org/10.1145/2886418>
- Marrone, R., Cropley, D. H., & Wang, Z. (2023). Automatic assessment of mathematical creativity using natural language processing. *Creativity Research Journal*, 35(4), 661–676. <https://doi.org/10.1080/10400419.2022.2131209>
- Mayasari, T., Kadarohman, A., Rusdiana, D., & Kaniawati, I. (2016). Exploration of student's creativity by integrating STEM knowledge into creative products. In *AIP Conference proceedings* (Vol. 1708, No. 1, p. 080005). AIP Publishing. <https://doi.org/10.1063/1.4941143>
- McKlin, T., Magerko, B., Lee, T., Wanzer, D., Edwards, D., & Freeman, J. (2018). Authenticity and personal creativity: How *EarSketch* affects student persistence. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (pp. 987–992).
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23. <https://doi.org/10.2307/1176219>
- Meusburger, P. D. I., Funke, J., & Wunder, E. (2009). *Milieu of creativity: An interdisciplinary approach to spatiality of creativity*. Springer.
- Mislevy, R. J., Almond, R. G., & Lukas, J. (2004). *A brief introduction to evidence-centered design* [CSE Technical Report 632]. The National Center for Research on Evaluation, Standards, Student Testing (CRESST).
- P21. (2019). *Framework for 21st-century learning*.
- Palmiero, M., Nori, R., Aloisi, V., Ferrara, M., & Piccardi, L. (2015). Domain-specificity of creativity: A study on the relationship between visual creativity and visual mental imagery. *Frontiers in Psychology*, 6, Article 1870. <https://doi.org/10.3389/fpsyg.2015.01870>
- Patterson, J. D., Barbot, B., Lloyd-Cox, J., & Beaty, R. E. (2023). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods*. Advance online publication. <https://doi.org/10.3758/s13428-023-02258-3>
- Patterson, J. D., Merseal, H. M., Johnson, D. R., Agnoli, S., Baas, M., Baker, B. S., Barbot, B., Benedek, M., Borhani, K., Chen, Q., Christensen, J. F., Corazza, G. E., Forthmann, B., Karwowski, M., Kazemian, N., Kreisberg-Nitzav, A., Kenett, Y. N., Link, A., Lubart, T., ... Beaty, R. E. (2023). Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts*, 17(4), 495–507. <https://doi.org/10.1037/aca0000618>
- Piffer, D. (2012). Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research. *Thinking Skills and Creativity*, 7(3), 258–264. <https://doi.org/10.1016/j.tsc.2012.04.009>
- Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist*, 39(2), 83–96. https://doi.org/10.1207/s15326985ep3902_1
- Qian, M., Plucker, J. A., & Yang, X. (2019). Is creativity domain specific or domain general? Evidence from multilevel explanatory item response theory models. *Thinking Skills and Creativity*, 33, Article 100571. <https://doi.org/10.1016/j.tsc.2019.100571>
- Rafner, J. (2021). Creativity assessment games and crowdsourcing. In *Creativity and cognition, C&C '21*. Association for Computing Machinery.
- Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. *Nature Human Behaviour*, 7(11), 1836–1838. <https://doi.org/10.1038/s41562-023-01751-1>
- Rafner, J., Wang, Q. J., Gadjacz, M., Badts, T., Baker, B., Bergenholtz, C., Biskjaer, M. M., Bui, T., Carugati, A., De Cibeins, M., & Noy, L. (2023). Towards game-based assessment of creative thinking. *Creativity Research Journal*, 35(4), 763–782. <https://doi.org/10.1080/10400419.2023.2198845>
- Rahimi, S. (2023). Going beyond the brick: Assessing and supporting creativity using AI-powered digital games. *Creativity Research Journal*. Advance online publication. <https://doi.org/10.1080/10400419.2023.2241779>
- 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., Fulwider, C., Jiang, S., & Shute, V. J. (2022). Predicting learning gains in an educational game using feature engineering and machine learning. In C. Chinn, C. Tan, C. Chan, & Y. Kali (Eds.), *ICLS Proceedings—international collaboration toward educational innovation for all* (pp. 2124–2125). International Society of the Learning Sciences.
- Rahimi, S., & Shute, V. J. (2021). First inspire, then instruct to improve students' creativity. *Computers & Education*, 174, Article 104312. <https://doi.org/10.1016/j.compedu.2021.104312>

- Rahimi, S., & Shute, V. J. (2023). Stealth assessment: A theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments. *Educational Technology Research and Development*. Advance online publication. <https://doi.org/10.1007/s11423-023-10232-1>
- Rahimi, S., Shute, V. J., & Zhang, Q. (2021). The effects of game and student characteristics on persistence in educational games: A hierarchical linear modeling approach. *International Journal of Technology in Education and Science*, 5(2), 141–165. <https://doi.org/10.46328/ijtes.118>
- Rahimi, S., Walker, J. T., Lin-Lipsmeyer, L., & Shin, J. (2023). Toward defining and assessing creativity in sandbox games. *Creativity Research Journal*. Advance online publication. <https://doi.org/10.1080/10400419.2022.2156477>
- Resnick, M. (2018). *Lifelong kindergarten: Cultivating creativity through projects, passion, peers, and play*. MIT Press. (Reprint edition).
- Resnick, M., Myers, B., Nakakoji, K., Shneiderman, B., Pausch, R., Selker, T., & Eisenberg, M. (2005). Design principles for tools to support creative thinking. In B. Schneiderman, G. Fischer, M. Czerwinski, B. Myers, & M. Resnick (Eds.), *Creativity support tools: A workshop sponsored by the National Science Foundation* (pp. 25–35). Carnegie Mellon University.
- Runco, M. A. (2003). Education for creative potential. *Scandinavian Journal of Educational Research*, 47(3), 317–324. <https://doi.org/10.1080/00313830308598>
- Runco, M. A. (2014). “Big C, Little C” creativity as a false dichotomy: Reality is not categorical. *Creativity Research Journal*, 26(1), 131–132. <https://doi.org/10.1080/10400419.2014.873676>
- Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, 24(1), 92–96. <https://doi.org/10.1080/10400419.2012.650092>
- Shaheen, R. (2010). Creativity and education. *Creative Education*, 1(03), 166–169. <https://doi.org/10.4236/ce.2010.13026>
- Sheard, J., Carbone, A., Lister, R., Simon, B., Thompson, E., & Whalley, J. L. (2008). Going SOLO to assess novice programmers. In *Proceedings of the 13th Annual Conference on Innovation and Technology in Computer Science Education—ITiCSE’08* (p. 209). ACM Press.
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Information Age Publishers.
- 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. (2017). Review of computer-based assessment for learning in elementary and secondary education: Computer-based assessment for learning. *Journal of Computer Assisted Learning*, 33(1), 1–19. <https://doi.org/10.1111/jcal.12172>
- Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, 116, Article 106647. <https://doi.org/10.1016/j.chb.2020.106647>
- Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., Kuba, R., Liu, Z., Yang, X., & Sun, C. (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>
- Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review*, 104(1), 66–89. <https://doi.org/10.1037/0033-295X.104.1.66>
- Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment*, 17(4), 396–408. <https://doi.org/10.1037/1040-3590.17.4.396>
- Smith, J. B., Jacob, M., Freeman, J., Magerko, B., & McKlin, T. (2019, December 4–6). *Combining collaborative and content filtering in a recommendation system for a web-based DAW*. Web Audio Conference WAC-2019, Trondheim, Norway.
- Smith, J. B., Truesdell, E. J. K., Freeman, J., Magerko, B., Boyer, K. E., & McKlin, T. (2020). Modeling music and code knowledge to support a co-creative AI agent for education. In J. Cumming, J. H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, & T. de Reuse (Eds.), *ISMIR* (pp. 134–141). International Society for Music Information Retrieval.
- Smith, J. B., Vinay, A., & Freeman, J. (2023). The impact of salient musical features in a hybrid recommendation system for a sound library. In A. Smith-Renner & P. Taele (Eds.), *Joint Proceedings of the ACM IUI Workshops*. CEUR Workshop Proceedings.
- Starr, C. W., Manaris, B., & Stalvey, R. H. (2008). Bloom’s taxonomy revisited: Specifying assessable learning objectives in computer science. *ACM SIGCSE Bulletin*, 40(1), 261–265. <https://doi.org/10.1145/1352322.1352227>
- Sternberg, R. J. (Ed.). (1988). *The nature of creativity: Contemporary psychological perspectives*. Cambridge University Press.
- Sternberg, R. J., & Lubart, T. I. (1996). Investing in creativity. *American Psychologist*, 51(7), 677–688. <https://doi.org/10.1037/0003-066X.51.7.677>
- Streich, S. (2006). *Music complexity: A multi-faceted description of audio content*. Universitat Pompeu Fabra.
- Thompson, E., Luxton-Reilly, A., Whalley, J. L., Hu, M., & Robbins, P. (2008). Bloom’s taxonomy for CS assessment. In *Proceedings of the Tenth Conference on Australasian Computing Education* (Vol. 78, pp. 155–161).
- Thul, E. (2008). *Measuring the complexity of musical rhythm*.
- Torrance, E. P. (1972). Can we teach children to think creatively? *The Journal of Creative Behavior*, 6(2), 114–143. <https://doi.org/10.1002/j.2162-6057.1972.tb00923.x>
- Torrance, E. P. (1974). *Torrance tests of creative thinking: Verbal tests, forms A and B, figural tests, forms A and B*. Xerox.
- Toussaint, G. T. (2004). A comparison of rhythmic similarity measures. In *Proceedings of the 5th International Conference on Music Information Retrieval*. ISMIR.
- Truax, B. (1992). Musical creativity and complexity at the threshold of the 21st century. *Journal of New Music Research*, 21(1), 29–42. <https://doi.org/10.1080/09298219208570598>
- Truesdell, E. J. K., Smith, J. B., Mathew, S., Katuka, G. A., Griffith, A. E., McKlin, T., Magerko, B., Freeman, J., & Boyer, K. E. (2021). Supporting computational music remixing with a co-creative learning companion. In *ICCC 2021* (pp. 113–121).
- Tubb, A. L., Cropley, D. H., Marrone, R. L., Patston, T., & Kaufman, J. C. (2020). The development of mathematical creativity across high school: Increasing, decreasing, or both? *Thinking Skills and Creativity*, 35, Article 100634. <https://doi.org/10.1016/j.tsc.2020.100634>
- Vinchon, F., Lubart, T., Bartolotta, S., Gironnay, V., Botella, M., Bourgeois-Bougrine, S., Burkhardt, J. M., Bonnardel, N., Corazza, G. E., Gläveanu, V., & Hanchett Hanson, M. (2023). Artificial intelligence & creativity: A manifesto for collaboration. *The Journal of Creative Behavior*, 57(4), 472–484. <https://doi.org/10.1002/jocb.597>
- Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity-intelligence distinction. Modes of thinking in young children: A study of the creativity-intelligence distinction*. Holt, Rinehart & Winston.
- Wardrip-Fruin, N. (2009). *Expressive processing: Digital fictions, computer games, and software studies*. MIT Press.
- Wingstrom, R., Hautala, J., & Lundman, R. (2023). Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists. *Creativity Research Journal*. Advance online publication. <https://doi.org/10.1080/10400419.2022.2107850>

Received July 23, 2023

Revision received January 16, 2024

Accepted January 21, 2024 ■