

Validity and Fairness of an Automated Assessment of Creativity in Computational Music Remixing

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

¹School of Teaching & Learning, College of Education, University of Florida, Gainesville, FL 32611, USA

²Georgia Tech Center for Music Technology, 840 McMillan Street NW, Atlanta 30308, GA, USA

³Expressive Machinery Lab, Georgia Institute of Technology, Atlanta, GA 30308, USA

⁴Computer & Information Science & Engineering, University of Florida, Gainesville, FL 32611, USA

⁵The Findings Group, Decatur, GA 30030, USA

Abstract

Creativity is one of the most crucial skills for success in life in the 21st century. However, assessing creativity in an automated, objective way is challenging. In this study, we designed and validated an automated assessment (an unobtrusive, formative assessment) of creativity in *EarSketch*, a computational music remixing platform where students write Python or JavaScript code to create pieces of music. Specifically, using an existing dataset of *EarSketch* projects ($n = 53$), we investigated the validity and fairness of an automated assessment of creativity. Our findings show that the automated assessment of creativity has reasonable convergent validity ($r = .50$) and discriminant validity; and this assessment is fair (i.e., no significant differences in terms of gender, grade, or race were found). The results of this research have the potential to inform the design of innovative educational programs and interventions that foster creativity and innovation in STEM education. 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 future challenges.

Keywords

Creativity, Automated Assessment, Computer Programming, EarSketch, Music, High School Students

1. Introduction and Background

As future jobs become more automated and replaced by AI-powered machines, 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 [1]. Moreover, creativity has been used in STEM education to broaden participation and engagement [2, 3, 4]. Integrating creativity in STEM education can improve students' STEM-related knowledge and skills as well as their creativity [4]. The current study aims to design and validate an automated assessment of creativity using an automated assessment technique [5] in a web-based, programming-learning environment called *EarSketch* [6]. Before talking about the details of this study, we briefly define creativity.

In general, creativity is the ability to come up with useful solutions for problems—or new and interesting ideas and objects—across a wide range of domains [7, 8, 9, 10]. According to Guilford [11], complexity is characterized as the degree to which a unified entity is formed by incorporating diverse and separate components. In her assessment of creativity, Amabile [12] incorporated *effort* as one of the factors that evaluators must consider when appraising a creative product. The model of creativity

AI in Education 2023 Conference: Automated Assessment and Guidance of Project Work, July 03-07, 2023, Tokyo, Japan

✉ srahimi@ufl.edu (S. Rahimi); jsmith775@gatech.edu (J. B. Smith); erinjkruesdell@gmail.com (E. J.K. Truesdell);

ashvala@gatech.edu (A. Vinay); keboyer@ufl.edu (K. E. Boyer); magerko@gatech.edu (B. Magerko);

jason.freeman@gatech.edu (J. Freeman); tom@thefindingsgroup.org (T. Mcklin)


🌐 <https://education.ufl.edu/faculty/rahimi-seyedahmad-ahmad/> (S. Rahimi)

🆔 0000-0001-9266-758X (S. Rahimi); 0000-0002-7075-6132 (J. B. Smith); 0000-0002-9421-8566 (E. J.K. Truesdell);

0000-0002-2487-2052 (A. Vinay); 0000-0003-3434-3450 (K. E. Boyer); 0000-0003-1900-4020 (B. Magerko); 0000-0003-3827-1060

(J. Freeman); 0000-0001-6120-7222 (T. Mcklin)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

that Guilford [13] introduced, for example, operationalized creativity as divergent thinking, with four sub-facets: *fluency* (the ability to produce a large number of ideas, solutions, 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). One of the frequently used assessment techniques is called the *Consensual Assessment Technique* [7]. In this technique, multiple experts (at least two) rate some 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. New assessment techniques that can objectively automate the assessment of a creative product are needed. Creativity has been assessed automatically in digital environments (e.g., digital games) in recent years. For instance, Shute and Rahimi [14] embedded and validated an automated assessment of creativity in a STEM game called *Physics Playground* [15]. Moreover, Rafner et al. [16] have reviewed game-based, automated assessments of creativity. Similarly, Yu et al. [17] used a maximum associative distance to automatically assess the creativity of the participants' responses. However, these studies usually base their assessment methods solely on divergent thinking operationalization of creativity. In this study, inspired by the traditional definitions of creativity in terms of creative, divergent thinking, and the aspects of a creative product, we operationalized creativity, designed an automated assessment of creativity based on that operationalization, and validated it using the consensual assessment technique in a project-based, music remixing environment called *EarSketch*.

EarSketch platform. *EarSketch* [6] is a web-based expressive computing learning environment designed to engage high school students in computer science through the use of Python or JavaScript code to remix music samples (from a library of over 4,000 sounds in a variety of musical genres and instruments) along a multi-track timeline (see Figure 1). It also includes a curriculum covering programming concepts in line with state and national standards (e.g., functions, loops, debugging, and peer review). *EarSketch* is designed to be accessible to students with little to no programming or music experience and has been used by over 1 million learners as of fall 2022.

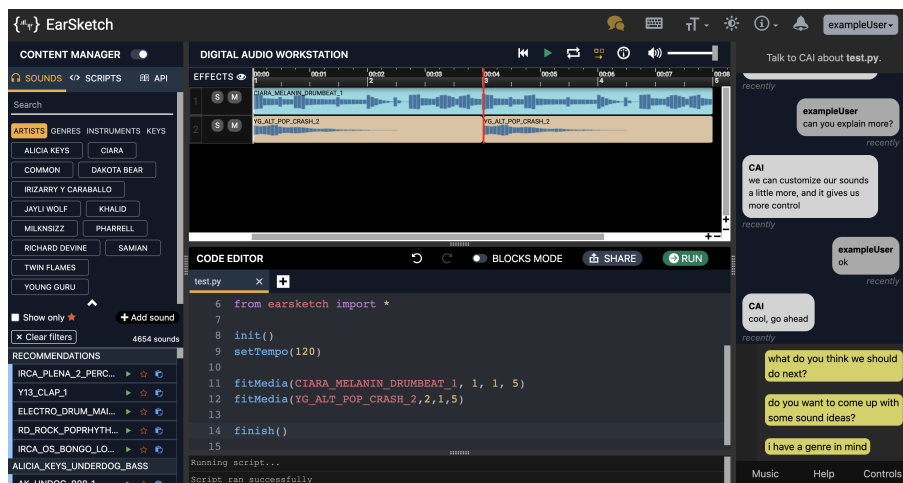


Figure 1: The *EarSketch* application's sound library (left), code editor (bottom), digital audio workstation (DAW) (top), and experimental CAI interface (right).

Automated Assessment. The type of automated assessment we are referring to in this paper has been introduced by Shute [5] and she called it *stealth assessment*. This assessment is a formative assessment (i.e., assessment for learning; see Shute and Rahimi [18]) technique that uses digital learning environments (historically, digital games) as vehicles for assessing *and* supporting various competencies such as basic knowledge and skills including physics [19]; and other hard-to-measure skills such as creativity [20, 14], persistence [21, 22], problem-solving [23]. The underlying framework of this automated assessment technique is the evidence-centered design (ECD) framework of assessment

[24, 25]. ECD has three main models: (1) the Competency Model (CM) which defines what is being assessed (i.e., unobservables); (2) the Evidence Model (ED) defines what indicators (i.e., observables) can provide evidence for the competency of interest. 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). And, (3) the Task Model (TM) defines what tasks (or learning environments) can elicit the evidence for the evidence model. Once these three models are designed and in place, the automated assessment computed estimates can be used (e.g., by making the learning experience adaptive or by providing targeted learning supports) to improve students' learning in real-time [26, 27]. In this study, we will address the following research question: *Is the automated assessment of creativity in EarSketch psychometrically sound?* That is, we will investigate the validity and fairness of this assessment.

2. Methods

The data was drawn from students' projects created in sessions held in five classrooms in the southeastern United States from our earlier studies investigating student use of CAI, a co-creative AI built for the *EarSketch* platform [28]. We used the data from 53 students (*Females* = 11; *Males* = 39; *Not Listed* = 3). Our sample was diverse with most of the students identifying as *White* ($n = 21$), *Hispanic/Latino* = 9, *Black/African American* = 11, and *Asian* = 8. Most of the students ($n = 27$) had not taken any courses related to programming; while some students had taken one other programming course ($n = 20$), and few students had taken two or more than two other programming courses ($n = 6$). Moreover, our sample included students from various grades (i.e., $9_{th} = 29$, $10_{th} = 10$, $11_{th} = 8$, and $12_{th} = 6$).

The students responded to a demographic and background questionnaire before starting working with *EarSketch*, 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, and four Likert-scale items about music enjoyment, coding enjoyment, making-music confidence, and learning-to-code confidence. The Task Model is the *EarSketch* environment. *EarSketch* already provides a lot of opportunities for students to show the evidence for the creativity competency model. Next, we explain the rules of evidence related to each indicator we used from *EarSketch* log data.

The competency model of creativity in this study includes divergent thinking and two other creative product qualities (i.e., complexity and effort). The gray boxes in Figure 2 show the observables for each CM sub-facet—i.e., the EM. Figure 2 shows the CM and EM of creativity in *EarSketch*.

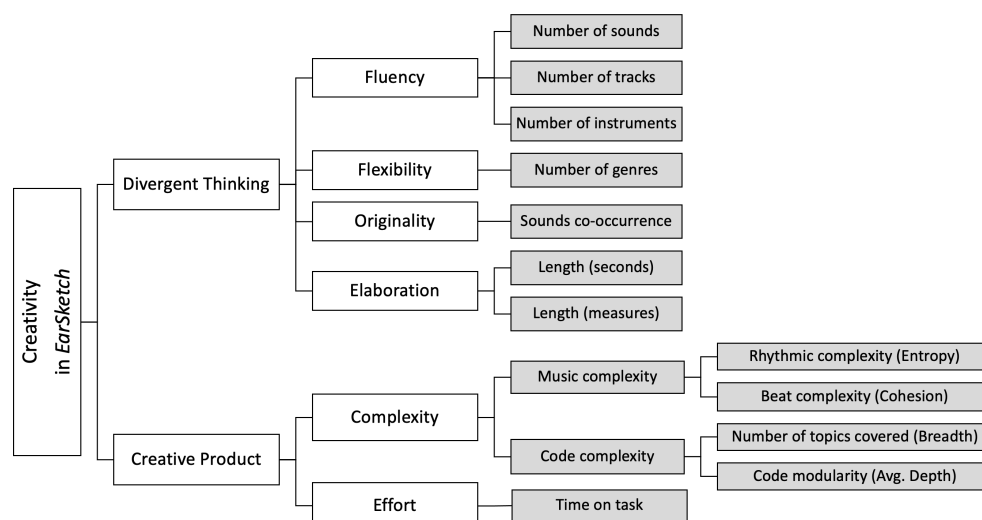


Figure 2: The creativity competency and evidence models in *EarSketch*

```

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: 1.0000

```

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: 1.7333

Figure 3: Three *EarSketch* scripts of increasing average depth with identical musical output.

Rules of evidence in ECD refer to a set of rules or 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 need to 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*. An updated version of *EarSketch*'s code complexity assessment tool was used to analyze the complexity of each project in the dataset. It estimates the user's knowledge of seventeen concepts covered in the *EarSketch* curriculum, assigning scores between 0 (if the concept does not appear in the user's code) and higher values between 1 and 3 indicating increasing levels of complexity of use.

The complexity calculator was expanded to include code *breadth* and *depth*. 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 by assigning each line of code a depth score. Code depth is calculated as the number of times a line of code is nested within a larger structure (such as a function definition or for loop), with a depth of 1 representing lines at the top level of the code. An increase in average depth indicates that the student has utilized 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 3.

In addition to the complexity calculator, *EarSketch* contains tools to model the musical contents of a project's musical structure [29], dividing it into sections and subsections in order to form a nested dictionary referred to as a Sound Profile [28]. 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 dataset. The values generated for each project included scores for each creativity sub-facet (see Figure 2), breadth and average depth scores, the number of unique sound samples used in the project, the number of the unique genre, and instrument tags present among the sounds in the project, the maximum number of sounds overlapping by being used simultaneously in separate tracks, the length of a project's output song in both seconds and measures, and average co-occurrence scores of sound samples in the project (statistically rarer combinations of

sounds found in an analysis of previous *EarSketch* projects [30]). Moreover, we investigate unique sound selections as a proxy for originality. Each pair of sounds is given a co-occurrence score (i.e., *sound co-occurrence*), based on the number of times the two sounds have been used at the same time in previous *EarSketch* projects [30]. Using these scores, we analyze the dataset of student projects to determine if they contain statistically rarer combinations of sounds—a proxy for originality.

Additionally, *Entropy*, the uncertainty in a given distribution, can be used as a measure of rhythmic complexity [31] is found to correlate with human perception of rhythm [32]. 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. Also, for every sound clip in *EarSketch*, we computed a “click track” representation, i.e., a binary vector where 1 indicates the presence of a beat at the timestamp and 0 indicates the absence of such a beat [33]. To compute *cohesion*, we computed the similarity (Hamming Distance) between every possible pair of successive sounds in a measure, as suggested by [34]. A lower distance indicates that the two tracks might be rhythmically cohesive.

Finally, as an indicator of mental *effort*, the *time-on-task* percentage has been used in a variety of user evaluations in human-computer interaction research and student evaluations [35], and has been measured alongside assessment of creativity [36]. It can be calculated as the ratio of time intervals (in an even division from the start to the end of a task) containing 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. Between its start and submission, actions (collected via keystroke logging) for each project are grouped on 10-second windows. Each project is given a time-on-task percentage, determined by the ratio of windows that contain actions by the user and the points that do not contain any actions logged.

To compute a final automated creativity score, we first standardized the low-level indicators (computing the z-scores for each observable shown in gray color in Figure 2). Then, we averaged those standardized estimates to compute the sub-facet variables fluency, flexibility, originality, elaboration, complexity, and effort. Next, we computed the divergent thinking score and creative product score by averaging the scores of their relevant sub-facets. Finally, we combined the averages of these two scores to obtain the overall automated assessment score.

2.1. External Assessment of Creativity

We used the consensual assessment technique [37] to externally assess the pieces of music ($n = 53$) and use the results from this assessment for validation purposes. Amabile (1982) asserts that if a group of experts in the domain at hand believe 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). These experts have been judging the quality of the projects created in *EarSketch* in the past several years. Thus, they were very well qualified as the experts who knew what was possible when it comes to creating music using *EarSketch* by high school students. 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.

3. Results

Generally, students created pieces of music that were considered somewhat creative on average ($M = 3.86, SD = .70$). The time on task (effort) was about 48% on average ($SD = 15.21$) across the 53 students. To investigate the validity of the automated assessment of creativity in *EarSketch*, we looked at the convergent and discriminant validity evidence. To examine the convergent validity of our assessment, we correlated the observables computed from the log data in *EarSketch* with the external

assessment of creativity (the expert ratings). Results show that fluency ($r = .37$), flexibility ($r = .30$), elaboration ($r = .21$), complexity ($r = .33$), and effort ($r = .25$) have reasonable, small to medium correlations with the external assessment of creativity. On a higher level, the divergent thinking score ($r = .39$) and the product score ($r = .40$) showed positive, medium correlations with the external creativity assessment. The overall automated assessment of creativity (the average score of the divergent thinking skills and product creativity) correlated positively ($r = .50$) with the external assessment of creativity. To investigate the discriminant validity of our assessment, we correlated the external assessment of creativity, the automated assessment of creativity, and its sub-facets with students' age, and the other self-report measures to investigate the discriminant validity of our assessment. Results show that there is no relationship between the self-report measures and the automated assessment of creativity *EarSketch* and its sub-facets. This finding indicates that the automated assessment of creativity in *EarSketch* is independent of those variables, thus, the assessment has good discriminant validity which is desirable.

To investigate the fairness of the automated assessment of creativity in *EarSketch*, we conducted several ANOVAs. Results show that there is no significant difference among different genders [$F_{gender}(2, 50) = .18, p = .84, Partial\eta^2 = .01$], races [$F_{race}(5, 47) = 1.46, p = .22, Partial\eta^2 = .14$], and grades [$F_{grades}(3, 49) = .40, p = .75, Partial\eta^2 = .02$]. These results indicate that the automated assessment of creativity in *EarSketch* is fair and is not performing differently for students from different genders, races, and grades.

4. Discussion and Conclusion

The validity of our automated assessment of creativity should be examined from various standpoints. First, 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; see Runco [38]) compared to traditional assessments of creativity (e.g., list as many alternative uses for a pen). Second, using ECD, consulting the literature on creativity, and consulting with music experts in *EarSketch*, we were able to establish the face validity of our competency model for creativity in *EarSketch*. In terms of convergent validity, our results showed that the automated assessment of creativity in *EarSketch* correlated (with a medium effect overall) with our external assessment of creativity both on an overall and sub-facet level. These results are promising as we only had one project per student and this was our first pass at designing and examining this assessment. 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 enjoyed coding). In terms of fairness, our findings indicated that there were no statistically significant differences among various groups present in our dataset (i.e., genders, races, grades). Overall, we can conclude that our automated assessment of creativity in *EarSketch* has a reasonable, promising level of validity and it shows to be fair based on the current data. In this study, we provided evidence for our automated assessment accuracy (i.e., convergent validity) or what Kane [39] 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 [39, 40]. Future studies can help us improve the validity and further investigate the fairness of our automated assessment using a larger sample size. One important aspect of automated assessment is to *use* the automated assessment estimates in real-time and provide personalized support (e.g., by informing the suggestions made by CAI or suggesting to the students to increase the variety of their sounds in their project) and some form of adaptation to maximize learning [e.g., 26]. One natural next step could be putting this automated assessment of creativity to use in *EarSketch*. In conclusion, the concept of automated assessment of creativity presents a promising approach to objectively and accurately measure creativity in a variety of settings. The results of this research have the potential to inform the design of innovative educational programs and interventions that foster creativity and innovation in the STEM workforce.

References

- [1] Education & Human Resources, STEM education for the future: A visioning report. National Science Foundation., Technical Report, 2020. URL: <https://www.nsf.gov/edu/Materials/STEM%20Education%20for%20the%20Future%20-%202020%20Visioning%20Report.pdf>.
- [2] B. Wang, P.-p. Li, Digital creativity in stem education: the impact of digital tools and pedagogical learning models on the students' creative thinking skills development, *Interactive Learning Environments* (2022) 1–14.
- [3] R. Miller, Integrating the arts and creativity in stem education: Emerging talent using steam, in: *STEM Education for High-Ability Learners*, Routledge, 2021, pp. 207–223.
- [4] A. Üret, R. Ceylan, Exploring the effectiveness of stem education on the creativity of 5-year-old kindergarten children, *European Early Childhood Education Research Journal* 29 (2021) 842–855.
- [5] V. J. Shute, Stealth assessment in computer-based games to support learning, *Computer games and instruction* 55 (2011) 503–524.
- [6] B. Magerko, J. Freeman, T. Mcklin, M. Reilly, E. Livingston, S. Mccoid, A. Crews-Brown, Earsketch: A steam-based approach for underrepresented populations in high school computer science education, *ACM Transactions on Computing Education* (2016).
- [7] T. M. Amabile, *Creativity in context*. Boulder.: Westview Press Harper Collins Publishers (1996).
- [8] G. R. Oldham, A. Cummings, Employee creativity: Personal and contextual factors at work, *Academy of management journal* 39 (1996) 607–634. ISBN: 0001-4273 Publisher: Academy of Management Briarcliff Manor, NY 10510.
- [9] J. Zhou, J. M. George, When job dissatisfaction leads to creativity: Encouraging the expression of voice, *Academy of Management journal* 44 (2001) 682–696. ISBN: 0001-4273 Publisher: Academy of Management Briarcliff Manor, NY 10510.
- [10] J. Zhou, J. M. George, Awakening employee creativity: The role of leader emotional intelligence, *The Leadership Quarterly* 14 (2003) 545–568. doi:10.1016/S1048-9843(03)00051-1, place: Netherlands Publisher: Elsevier Science.
- [11] J. P. Guilford, *Fundamental Statistics in Psychology and Education*, McGraw-Hill, 1950.
- [12] T. M. Amabile, The social psychology of creativity: A componential conceptualization., *Journal of personality and social psychology* 45 (1983) 357. doi:<https://psycnet.apa.org/doi/10.1037/0022-3514.45.2.357>, iISBN: 1939-1315 Publisher: American Psychological Association.
- [13] J. P. Guilford, *Fundamental statistics in psychology and education*, 3rd ed., *Fundamental statistics in psychology and education*, 3rd ed., McGraw-Hill, New York, NY, US, 1956. Pages: xi, 565.
- [14] V. J. Shute, S. Rahimi, Stealth assessment of creativity in a physics video game, *Computers in Human Behavior* 116 (2021). doi:10.1016/j.chb.2020.106647, place: Netherlands Publisher: Elsevier Science.
- [15] V. J. Shute, R. G. Almond, S. Rahimi, *Physics Playground (version 1.3)*[computer software], 2019.
- [16] J. Rafner, M. M. Biskjær, B. Zana, S. Langsfjord, C. Bergholtz, S. Rahimi, A. Carugati, L. Noy, J. Sherson, Digital games for creativity assessment: strengths, weaknesses and opportunities, *Creativity Research Journal* 0 (2021) 1–27. URL: <https://doi.org/10.1080/10400419.2021.1971447>. doi:10.1080/10400419.2021.1971447, publisher: Routledge _eprint: <https://doi.org/10.1080/10400419.2021.1971447>.
- [17] Y. Yu, R. E. Beaty, B. Forthmann, M. Beeman, J. H. Cruz, D. Johnson, A mad method to assess idea novelty: Improving validity of automatic scoring using maximum associative distance (mad), *Psychology of Aesthetics, Creativity, and the Arts* (2023).
- [18] V. Shute, S. Rahimi, Review of computer-based assessment for learning in elementary and secondary education: Computer-based assessment for learning, *Journal of Computer Assisted Learning* 33 (2017) 1–19. URL: <http://doi.wiley.com/10.1111/jcal.12172>. doi:10.1111/jcal.12172.
- [19] V. J. Shute, S. Rahimi, Stealth assessment of creativity in a physics video game, *Computers in Human Behavior* 116 (2020) 106647. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0747563220303940>. doi:10.1016/j.chb.2020.106647.

- [20] J. Rafner, Creativity assessment games and crowdsourcing, in: *Creativity and Cognition*, C&C '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: <https://doi.org/10.1145/3450741.3467465>. doi:10.1145/3450741.3467465, event-place: Virtual Event, Italy.
- [21] K. E. DiCerbo, Game-Based assessment of persistence, *Journal of Educational Technology & Society* 17 (2014) 17–28. URL: <https://www.jstor.org/stable/jeductechsoci.17.1.17>.
- [22] S. Rahimi, C. Fulwider, S. Jiang, V. J. Shute, 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*, 2022, pp. 2124–2125.
- [23] B. Akram, W. Min, E. Wiebe, B. Mott, K. E. Boyer, J. Lester, Improving stealth assessment in game-based learning with LSTM-based analytics, in: *International Conference on Educational Data Mining*, 2018. URL: <https://par.nsf.gov/biblio/10100664-improving-stealth-assessment-game-based-learning-lstm-based-analytics>.
- [24] R. G. Almond, R. J. Mislevy, L. S. Steinberg, D. Yan, D. M. Williamson, *Bayesian Networks in Educational Assessment*, Springer, 2015.
- [25] R. J. Mislevy, R. G. Almond, J. Lukas, A brief introduction to evidence-centered design, CSE Technical Report 632, The National Center for Research on Evaluation, Standards, Student Testing (CRESST), 2004. URL: <http://www.cresst.org/reports/r632.pdf>.
- [26] V. J. Shute, S. Rahimi, G. Smith, F. Ke, R. Almond, C. Dai, R. Kuba, Z. Liu, X. Yang, C. Sun, Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games, *Journal of Computer Assisted Learning* 37 (2020). URL: <http://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12473>. doi:<https://doi.org/10.1111/jcal.12473>, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.12473>.
- [27] S. Rahimi, R. G. Almond, V. J. Shute, Getting the first and second decimals right: Psychometrics of stealth assessment, in: M. P. McCreery, S. K. Krach (Eds.), *Games as Stealth Assessments*, DOI press, Lewes, DE, in press 2023, pp. 1–40.
- [28] E. J. K. Truesdell, J. B. Smith, S. Mathew, G. A. Katuka, A. E. Griffith, T. McKlin, B. Magerko, J. Freeman, K. E. Boyer, Supporting computational music remixing with a co-creative learning companion., in: *ICCC*, 2021, pp. 113–121.
- [29] J. Smith, E. J. K. Truesdell, J. Freeman, B. Magerko, K. E. Boyer, T. McKlin, Modeling music and code knowledge to support a co-creative ai agent for education., in: *ISMIR*, 2020, pp. 134–141.
- [30] J. Smith, M. Jacob, J. Freeman, B. Magerko, T. Mcklin, Combining collaborative and content filtering in a recommendation system for a web-based daw, in: *Proceedings of the International Web Audio Conference*, 2019.
- [31] E. Thul, *Measuring the complexity of musical rhythm* (2008).
- [32] R. de Fleurian, T. Blackwell, O. Ben-Tal, D. Müllensiefen, Information-theoretic measures predict the human judgment of rhythm complexity, *Cognitive Science* 41 (2017) 800–813.
- [33] J. B. Smith, A. Vinay, J. Freeman, The impact of salient musical features in a hybrid recommendation system for a sound library, in: *Joint Proceedings of the ACM IUI Workshops*, 2023.
- [34] G. T. Toussaint, A comparison of rhythmic similarity measures., in: *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR*, Barcelona, Spain, 2004. URL: <https://doi.org/10.5281/zenodo.1416812>. doi:10.5281/zenodo.1416812.
- [35] N. Karweit, Time-on-task reconsidered: Synthesis of research on time and learning., *Educational leadership* 41 (1984) 32–35.
- [36] M. Benedek, C. Mühlmann, E. Jauk, A. C. Neubauer, 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 (2013) 341.
- [37] T. M. Amabile, Social psychology of creativity: A consensual assessment technique, *Journal of Personality and Social Psychology* 43 (1982) 997–1013. doi:10.1037/0022-3514.43.5.997, place: US Publisher: American Psychological Association.
- [38] M. A. Runco, “Big C, little c” creativity as a false dichotomy: Reality is not categorical, *Creativity*

Research Journal 26 (2014) 131–132. ISBN: 1040-0419 Publisher: Taylor & Francis.

- [39] M. T. Kane, Current concerns in validity theory, *Journal of educational Measurement* 38 (2001) 319–342.
- [40] M. T. Kane, Validating the Interpretations and Uses of Test Scores, *Journal of Educational Measurement* 50 (2013) 1–73. Publisher: Wiley Online Library.