Melete: Validating the Creativity Support Index as a Metric for Evaluating the Integration of AI In Software Pipelines

Sokol Murturi[1,2,*,†], Matthew Yee-King[2], Joseph Walton-Rivers[1], Michael James Scott[1] and Marco Gillies[2]

[1]*Falmouth University, Faculty of Screen, Technology and Performance, Cornwall, United Kingdom*
[2]*Goldsmiths University of London, Department of Computing, London, United Kingdom*

Abstract

This paper evaluates the Creativity Support Index (CSI) for Mixed-Initiative Artificial Intelligence (MIAI) pipeline development. Determining the quality of the interaction between agents and users is valuable given the complex nature of MIAI pipelines. Within academic literature, the CSI is regarded as a practical measurement for assessing the usefulness of software. However, real-world validation of the CSI as a measurement tool for MIAI pipelines has not yet been established. We compared two undergraduate cohorts with 99 participants, who completed a level-design task using 'Melete' an MIAI pipeline, participants reported their experiences using CSI which resulted in 297 responses to the CSI. We then conducted a factor analysis to determine the validity of individual components of the CSI. Analysis of the responses indicates the CSI is a valuable measurement tool for testing MIAI pipelines, with limitations. We make recommendations to improve the CSI including; the disentanglement of measures and the development of a robust set of questionnaires.

Keywords

Creativity Support, Level Design, Development Tool, Games, Play, Human-Computer Interaction, Quantitative, Mixed-Initiative Artificial Intelligence, Procedural Content Generation, Validation

1. Introduction

The role of Artificial Intelligence (AI) has been expanding over the last ten years. Fields such as medical diagnostic systems, financial trading algorithms, driverless cars, customer engagement systems, linguistics, audio, games, and countless other areas have adapted or implemented AI algorithms. However, these systems require expert knowledge to interpret and operate. This limits the deployment of these systems to industry experts with this technology[1, 2]. There have been numerous efforts to provide user-friendly interfaces to these algorithms[3, 4]. This area of academic interest is known as Mixed-Initiative Artificial Intelligence (MIAI). To allow developers to create meaningful applications of MIAI pipelines it is crucial to establish measurement tools that effectively assess the usefulness of these software applications post-implementation.

Although much of the existing literature has examined the performance of AI algorithms, comparatively little work has gone into the evaluation of the user experience of these tools. Traditionally, much of the focus on verification of these pipelines focuses on the correctness of the output of these systems, rather than the user experience when using the pipelines. The earlier issues with the pipelines can be identified, the less costly it is to address these issues[5]. The focus of generative AI is to develop content rapidly, and the creative process itself is iterative. This highlights that our traditional approach

Synergy 2025 HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy.

sokol.murturi@falmouth.ac.uk (S. Murturi); m.yee-king@gold.ac.uk (M. Yee-King); Joseph.Walton-rivers@falmouth.ac.uk (J. Walton-Rivers); michael.scott@falmouth.ac.uk (M. J. Scott); M.Gillies@gold.ac.uk (M. Gillies)
https://www.falmouth.ac.uk/staff/sokol-murturi (S. Murturi); https://www.gold.ac.uk/computing/people/m-yee-king/ (M. Yee-King); https://www.falmouth.ac.uk/staff/joseph-walton-rivers (J. Walton-Rivers); https://www.falmouth.ac.uk/staff/dr-michael-scott (M. J. Scott); https://www.gold.ac.uk/computing/people/m-gillies/ (M. Gillies)
0000-0001-9466-8981 (S. Murturi); 0000-0001-6606-2448 (M. Yee-King); 0000-0002-3406-0584 (J. Walton-Rivers); 0000-0002-6803-1490 (M. J. Scott); 0000-0002-3100-9230 (M. Gillies)

to software development is inadequate for ensuring the effectiveness of MIAI pipelines for content generation.

Effective testing of the user experience for these applications requires suitable measures for identifying possible issues and addressing them. One such approach to doing this is through the use of survey-based measurement tools. Although these have found support within the existing literature[6], there are a range of possible limitations for these approaches which need to be considered. Unclear questions in measurement tools can hinder the evaluation process and lead to unusable data, rendering studies or analyses of software tools unserviceable. Measurement tools must delineate the variables that they assess. This ensures that the surveyor is assessing the correct metrics, and the respondent adequately responds to the survey.

Before the integration of AI into design tools the need to research user interaction with Creativity Support Tools(CSTs) was already identified by the Human Computer Interaction(HCI) community[7]. In his call for researchers to explore future developments in CSTs, Shneiderman[7] also emphasized the evolving nature of tool evaluation and the importance such evaluation plays in establishing the value of CSTs to end users. In addition to exploring the nature of CSTs, it is important to note the advancement of technology in this sector. The development of MIAI pipelines brings to the front the need to advance these measurement systems in line with the integration of AI based MIAI and co-creative artificial intelligence systems.

One notable contribution in this area is the Creativity Support Index (CSI)[8]. Drawing conceptual inspiration from the NASA Task Load Index, the CSI provides researchers with a straightforward, adaptable survey to assess the effectiveness of CSTs. The CSI evolved from an earlier pilot version, the Beta CSI, which was rooted in creativity theories and tested in three studies to assess its viability as a research metric. The CSI's questions are grounded in creativity research, yielding quantifiable and comparable results. A subsequent publication elaborated on the CSI, including a detailed case study[9].

Although the CSI has become a widely recognized and valuable tool, as seen in HCI and creativity research, the next step is to further validate it. The authors of the CSI, in Carol et al.[8] the authors called for its application in more studies to address usability issues identified in the Beta version. In this paper, we present our findings from such an analysis.

In this paper, we explore the evolution of creativity measurement and its limitations, particularly in the context of measuring MIAI systems. We differentiate MIAI systems from CSTs and provide examples of diverse MIAI systems developed in recent years to highlight their wide range of applications. Next, we explain how validity is assessed in measurement tools. We then describe the methodological procedure used to evaluate the CSI, followed by a demographic overview of the study participants. The results of the CSI factor analysis are presented, along with a discriminant validity test. A second factor analysis, excluding factors with discriminant validity issues, is also conducted, and we compare this modified model with the original CSI. Finally, we discuss the results of these analyses and conclude with recommendations for further research to improve measurement tools for both CSTs and MIAI systems.

2. Background

2.1. Measuring Creativity

Shneiderman[10] introduced a framework to aid in the development of digital interactive tools for creative problem-solving, a well-established area in creativity research (e.g.[11, 12]). In exploring the potential to enhance individual creativity through new tools, Shneiderman[10] considered both widely used general-purpose tools like text editors and spreadsheets and other specialized applications in architecture, graphic design, and engineering.

In the 1993 Creativity and Cognition symposium, Candy and Edmonds[13] emphasized the need to focus more on the study and development of CSTs to benefit "all people in any domain"[14]. However, to achieve this, the CHI community would "need to understand much more about the creative processes that we are trying to support"[14] In the years that followed, research interest in CSTs expanded, spurred by a 2006 U.S. National Science Foundation workshop that emphasized the need to make creative processes more efficient but also to foster innovation among users[15]. During this workshop, Shneiderman encouraged boldness in CST research and development, arguing that while the risks are high, "so are the payoffs for innovative developers, ambitious product managers, and bold researchers"[16]. Even described the creation of new CSTs as "a grand challenge for HCI researchers"[17].

While psychological creativity research boasts nearly seven decades of groundbreaking contributions, it is clear that HCI-oriented creativity research does not have as strong a tradition. Nevertheless, the field has made significant strides. A dedicated conference, the Creativity and Cognition, became an ACM SIGCHI conference in 1999, and the number of creativity-focused publications from the CHI community has surged since the late 1990s[18].

2.2. Human Computer Interaction

HCI-specific methods for measuring the impact of new CSTs have also been developed[9]. This evolution in HCI research has led to the notion of a potential of expanding creativity research in the field of HCI[18], which, while still in its early stages, may eventually parallel the more established research into traditional creativity task such as artistic expression and social interaction. This emerging HCI creativity research is said to be characterized by a focus on collaborative work and digitization, particularly the growing reliance on CSTs in creative processes, as well as a predominance of empirical research methodologies[18].

HCI researchers have implemented a broad spectrum of methodological approaches to validating pipelines. Qualitative methods can provide insights into the users' thought processes and provide nuanced information about how the user perceives the CST. The deployed techniques include structured interviews, speak-aloud interviews, expert interviews and grounded theory[19]. In contrast, the quantitative methods mainly focus on analysing the outputs of systems or conducting surveys[3]. These quantitative approaches allow for the rapid testing of multiple systems in a relatively short period of time. In addition to these benefits, the analysis of quantitative data can be useful for highlighting issues throughout the development life cycle of CST.

The criteria for evaluation similarly varied, encompassing both traditional creativity traits like flexibility and fluency, as well as classic usability principles[18]. There is a perceived conflict between the desire to measure the effectiveness of the creative pipeline and its outputs. It is difficult to assess creative outputs using objective measures, as creativity has a more subjective component. As a result, it can be difficult to disentangle the creative endeavour from using the system. This is not a new observation but rather a frequently discussed issue within the HCI community, as special interest groups have grappled with the complex challenge of evaluating research that extends beyond usability[20, 21].

Researchers have made progress toward establishing a standardized evaluation method for evaluating CSTs, most notably with the CSI[8]. The CSI is a psychometric survey developed to evaluate how effectively a digital CST aids users in their creative process. Its theoretical foundation is rooted in concepts from creativity and cognition. Support tools, drawing on Boden's work on creative exploration and play[22], formal theories of play[23], Csikszentmihalyi's concept of flow[24], and Shneiderman's design principles for creativity support tools[16]. The CSI consists of two sets of questionnaires, the first of which explores six different aspects of the CST: Collaboration, Enjoyment, Exploration, Expressiveness, Immersion, and Results worth the effort. A second survey consists of a paired-factor analysis which is meant to determine what the user values the most when using a CST for a particular task, with regard to the six factors presented above.

2.3. Mixed-Initiative Artificial Intelligence

Traditional CSTs allow for creative expression but do not make use of advances within the field of Artificial Intelligence - specifically generative artificial intelligence. The integration of these AI techniques into CSTs has led to the creation of a new range of pipelines where human authors can make use of AI techniques within their tools; these AI-augmented tools are referred to as MIAI[25]. MIAI systems are collaborative environments where users work with artificial intelligence agents, seamlessly blending their contributions to the creative process[26].

MIAI systems have been utilized in various creative fields, including art, music, dance, drawing, and game design[27]. In some MIAI systems, the computational agent directly manipulates a shared artifact, while in others, it offers suggestions to inspire users to develop new ideas. A key factor in how we can differentiate different MIAI systems is how the MIAI agent contributes to the creative process.

One of the MIAI interaction paradigms in design is the turn-taking action between a user and an AI agent in a shared artifact. Drawing Apprentice[28] is a web-based co-creative drawing system that analyzes the sketches and responds to the user's sketch. In the Drawing Apprentice, the user starts sketching on the canvas, then the AI agent generates and adds a sketch based on the user's sketch.

DuetDraw[29] is a MIAI drawing system that allows users and an AI agent to create pictures. DuetDraw assists users with various drawing tasks, such as completing an unfinished object, drawing the same object in a different style, suggesting complementary objects, identifying empty spaces on the canvas, and automatically colorizing sketches.

Cobbie[30] is a mobile robot with a recurrent neural network (RNN)–based MIAI approach and a mobile drawing system designed to support early-stage ideation. Cobbie generates inspirational sketches based on the designer's input.

While the MIAI interaction paradigms mentioned above demonstrate instances where an AI agent is directly engaged in a creative activity by performing actions similar to those of the user, another MIAI interaction paradigm involves the AI providing suggestions to the user.

The Sentient Sketchbook[19] and 3Buddy[31] are MIAI systems for game-level design. In both systems, the AI agent provides feedback and additional ideas to develop the game design. These systems use a turn-taking interaction but provide suggestions to the human designer rather than creating game levels directly.

Some recent studies investigated the role of AI and the impact of AI on ideation in human-AI collaboration. Liao et al.[32] presented three potential roles of AI-based inspirations in ideation that are closely related to the interaction paradigm providing suggestions described above: AI as representation creation (providing inspirations by suggesting texts or images), AI as an empathy trigger (supporting the designer's descriptive thinking), and AI as engagement (helping the designer avoid fossilization/stagnation and perform typical design actions).[32]

Figoli et al. claimed that the role of AI in human-AI collaboration relies on the capability of AI (i.e., AI as a teammate when AI performance is better than human performance and AI as external stimuli when AI performance is worse than human performance).[33] While the roles of AI emphasize the positive impact of AI in human-AI collaboration, Pandya et al. and Zhang et al. showed that human-AI collaboration can produce better outcomes when the AI contributes a better performance than the human. However, if the AI performs at the same level or worse than a human, it is likely to lead to less favorable results. These studies suggest that the roles and impacts of AI in human-AI collaboration require more investigation [34, 35].

2.4. Assessing Validity

According to Straub et al.[36], there are three types of validity and reliability in instrument development: content validity, construct validity, and internal reliability.

Content validity refers to the extent to which the items used to measure a construct accurately reflect its meaning and cover the full range of possible items that could represent the construct.

Construct validity concerns how well the items measure the theoretical construct they are intended to assess. This is often broken down into convergent validity and discriminant validity, both of which are necessary to establish construct validity[37]. After conducting our factor analysis we used discriminant validity to confirm our analysis.

Convergent validity measures the degree to which items that should theoretically be related are indeed related, while discriminant validity assesses whether items that should not be related are actually unrelated. Reliability refers to the consistency of responses to a set of related items that are intended to measure the same construct (i.e., internal consistency).

Concurrent validity is also considered when constructs are expected to be related; it examines whether a construct is related to or can predict another construct within the same instrument.

2.5. Melete

To deepen our understanding of MIAI pipelines, we developed Melete—a human-in-the-loop Creative Support Tool (CST) for generating asymmetrical 3D level topologies using the Wave Function Collapse (WFC) algorithm[38, 39]. Designers interact with Melete by selecting, editing, and exporting WFC-generated environments in JPEG or XML formats.

We evaluated Melete through four studies:

- Melete: Playtesting and 3D Environments for Mixed-Initiative Artificial Intelligence as a Method for Prototyping Video Game Levels[39]
- Melete: Exploring the Components of Mixed-Initiative Artificial Intelligence Pipelines for Level Design[40]
- Melete: The Importance of Kernel Interaction in Mixed-Initiative Artificial Intelligence Pipelines[41]
- Melete: Using Large Language Models (LLMs) as Co-Creators for Kernels in Mixed-Initiative Artificial Intelligence Pipelines[42]

The first two studies validated Melete: expert users found it enhanced engagement and supported iterative design, while novice users helped identify key MIAI themes—control, design, and play—informing a proposed interaction model. The final two studies used the CSI. The third examined Kernel interaction but yielded inconclusive CSI results. The fourth explored LLM-assisted design, revealing that most users preferred manual input, suggesting a need to re-evaluate CSI's applicability in MIAI contexts.

2.5.1. The structure of Melete

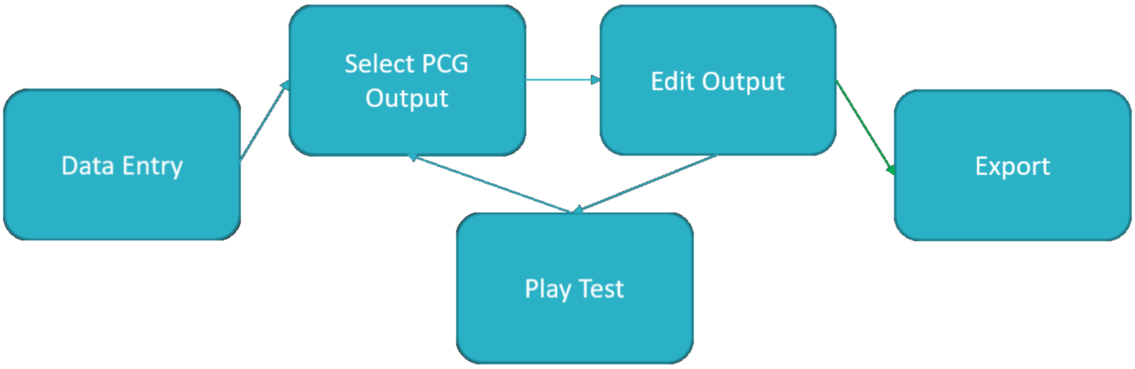
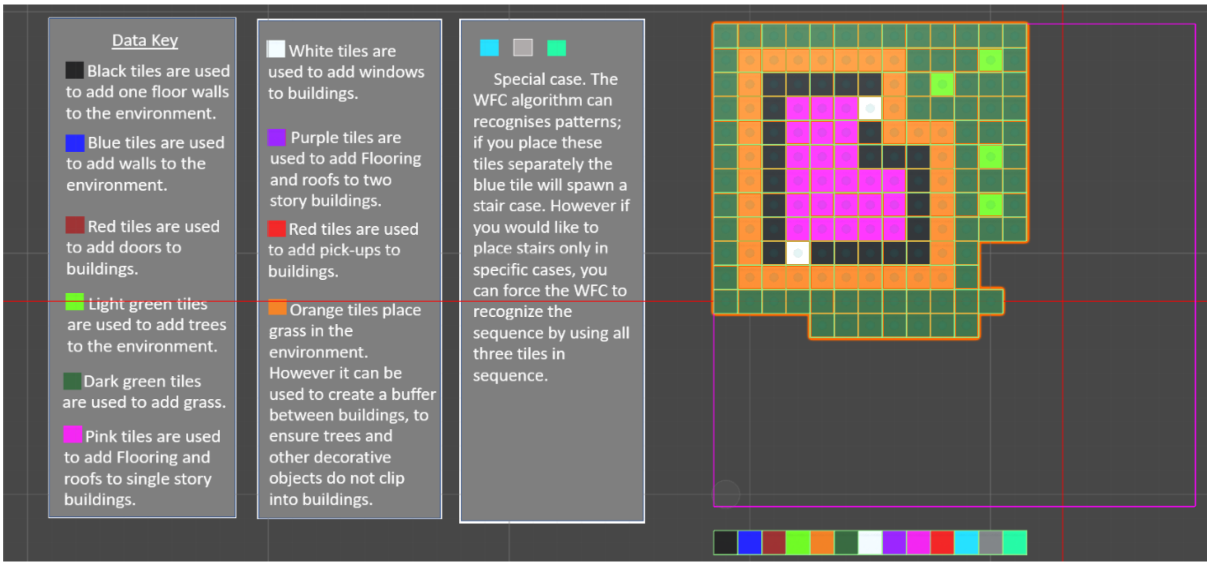


Figure 1: Melete Overview

The papers "Melete: Playtesting..."[39] and "Melete: Exploring the Components..."[40] detail Melete's implementation. Briefly, as shown in Figure 1, Melete operates in two phases: Kernel creation and an interaction loop.

In the Kernel creation phase (Figure 2), designers use the WFC algorithm to build a Kernel that guides the generation of stylistically consistent artefacts. Inspired by texture synthesis, WFC differs in that



Data Key

⬛ Black tiles are used to add one floor walls to the environment.

🟦 Blue tiles are used to add walls to the environment.

🟥 Red tiles are used to add doors to buildings.

🟩 Light green tiles are used to add trees to the environment.

🟩 Dark green tiles are used to add grass.

🟪 Pink tiles are used to add Flooring and roofs to single story buildings.

⬜ White tiles are used to add windows to buildings.

🟪 Purple tiles are used to add Flooring and roofs to two story buildings.

🟥 Red tiles are used to add pick-ups to buildings.

🟧 Orange tiles place grass in the environment. However it can be used to create a buffer between buildings, to ensure trees and other decorative objects do not clip into buildings.

🟦 ⬜ 🟩

Special case. The WFC algorithm can recognises patterns; if you place these tiles separately the blue tile will spawn a stair case. However if you would like to place stairs only in specific cases, you can force the WFC to recognize the sequence by using all three tiles in sequence.

Figure 2: WFC Kernel Interaction

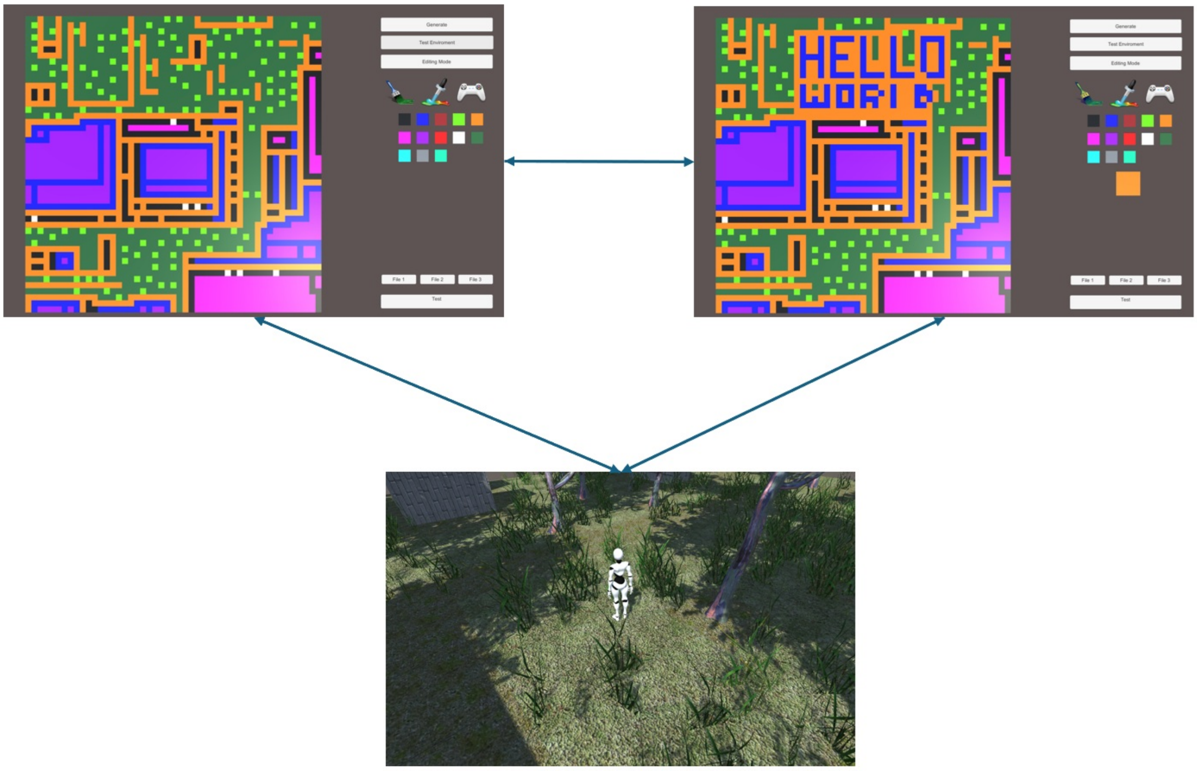


Figure 3: Interaction Loop

it avoids merging or averaging pixels, allowing for a palette-based approach that preserves gameplay semantics.

Designers interact with the WFC algorithm using a palette of game objects to build a small 10×10 input grid. This grid helps the algorithm generate possible 2×2 tile combinations, assign likelihood scores, and create a dictionary of overlapping tiles.

Although Melete focuses on 3D environment generation, abstracting game objects from the algorithm allows the palette to represent different gameplay or environmental elements. This enables designers to shape the environment's topology while the algorithm handles structural generation.

Once the WFC training input is finalized, designers enter the interaction loop (Figure 3).

Here, users select from multiple WFC-generated outputs, modify them as needed, and explore the environment in 3D using the game object palette. An avatar allows for first-person exploration. Users can move between selection, editing, and exploration freely, and export the final design in XML or JPEG format.

2.5.2. Melete: Playtesting and 3D Environments for Mixed-Initiative Artificial Intelligence as a Method for Prototyping Video Game Levels

While reviewing MIAI literature, we found that many systems lacked emphasis on exploring generated content—an essential part of iterative design. Melete addresses this by enabling users to explore environments through an avatar, supporting playtesting as a core component of its MIAI pipeline.

To validate Melete and highlight the value of playtesting, we conducted an expert analysis as detailed in "Melete: Playtesting and 3D Environments..."[39]. Five experts—spanning academia and industry in game design and AI—participated in two 20-minute sessions, designing a Battle Royale and an RTS environment after a brief tutorial.

Experts answered evaluation questions focused on usability, control, and the usefulness of playtesting. Their responses, analyzed thematically, confirmed Melete's utility for prototyping and emphasized the importance of genre-specific playtesting in MIAI systems. However, participants also noted that MIAI tools like Melete are best suited for prototyping, not full production.

2.5.3. Melete: Exploring the Components of Mixed-Initiative Artificial Intelligence Pipelines for Level Design

There is currently no consensus on the core components of MIAI pipelines, especially with the integration of systems like pathfinding, LLMs, and PCG. In "Melete: Exploring the Components of Mixed-Initiative Artificial Intelligence Pipelines for Level Design"[40], we investigated what novice designers consider essential in MIAI pipelines.

Twelve computing students (nine undergraduates, three postgraduates) participated in a qualitative study. Each designed a Battle Royale game using Melete under two conditions: one full version (Condition 5) and one of four partial versions (Conditions 1–4), each emphasizing a different stage of the pipeline. Conditions and order were randomized.

After each session, participants were interviewed using semi-structured questions and a think-aloud protocol. A thematic analysis of the responses revealed three main categories:

- Control themes included: personalization, understanding, adaptation, and tools.
- Design themes included: usability, aesthetics, and gameplay.
- Play themes included: exploration and enjoyment.

These insights informed the development of a Mixed-Initiative Interaction Model, offering guidance on features critical to effective MIAI pipeline design.

2.5.4. Melete: The Importance of Kernel Interaction in Mixed-Initiative Artificial Intelligence Pipelines

In the previous study, we identified the importance of Kernel interaction within MIAI pipelines. Overall, users reported feeling greater control over the output of the MIAI system. However, participants specifically encouraged further training focused on the WFC algorithm.

To address this, we developed and provided a short tutorial video explaining how the WFC algorithm generates content and how MIAI functions in theory. We then conducted a study aimed at examining the role of Kernel interaction in more detail.

One of the biggest challenges encountered in earlier studies was the analysis of qualitative data. To improve our evaluation, we explored alternative methods and identified the CSI as a useful tool for examining content generation systems. Based on the CSI framework, we designed an experiment involving three conditions:

- Condition 1: Participants interacted only with Melete's interface, serving as a baseline level editing tool.
- Condition 2: Participants engaged with Melete's interaction loop.
- Condition 3: Participants interacted with Melete's full system, including Kernel interaction.

In all conditions, participants were tasked with designing a Battle Royale-style game environment.

A G*Power analysis indicated that 45 participants were needed to perform an ANOVA on the CSI output across the three conditions. Students from an Introduction to Computing course were invited to participate voluntarily. This course aimed to provide students with experience working on large datasets, making it relevant to include an academic research project as part of their experience.

Out of 110 students who signed up for the study, 58 attended, though 16 participants did not complete the experiment. Additional students were randomly recruited to compensate for the shortfall. Ultimately, participants were randomly assigned the first condition. And completed all three conditions sequentially thereafter. Each session lasted 15 minutes, during which participants designed a Battle Royale-style map. After completing each condition, participants filled out a CSI questionnaire.

Upon completion of all three conditions and all three CSI questionnaires, the participants then performed the factor analysis component of the CSI. After data collection, a Mahalanobis distance analysis was conducted to identify outliers, resulting in the removal of three participants. This left 42 valid responses for analysis.

An ANOVA was performed on the CSI scores. All conditions scored between 61 and 63 out of 100 on the CSI, and the ANOVA results suggested that there were no significant differences between the conditions.

2.5.5. Melete: Using Large Language Models (LLMs) as Co-Creators for Kernels in Mixed-Initiative Artificial Intelligence Pipelines

Given the inconclusive results of the previous study, and the highlighted importance of Kernel interaction in both the expert analysis and component analysis, we sought to further explore ways to enhance Kernel interaction. The integration of LLMs into MIAI pipelines showed significant potential for streamlining the Kernel generation phase.

To enable LLMs to interact with Melete, we developed a custom webhook that allowed ChatGPT to design text-based level Kernels. We used ChatGPT 3.5 to generate these text files, which were then sent to Unity. Unity converted the text files into JPEG images that could be parsed by the WFC algorithm, thus fully integrating LLM-generated Kernels into the Melete pipeline.

In this participant study, we aimed to examine the differences in human-computer interaction with and without the assistance of LLMs. We created three experimental conditions:

- Condition 1: Participants provided a single prompt to ChatGPT to generate a Kernel for Melete's interaction loop.
- Condition 2: Participants engaged in a conversational interaction with ChatGPT to collaboratively design the Kernel.
- Condition 3: Participants used the original Melete system without LLM support, as described in previous studies.

In all conditions, participants were tasked with designing a Battle Royale-style game environment. The same tutorial video explaining the WFC algorithm and the theory behind MIAI, used in the previous experiment, was provided to participants in this study.

Participants were randomly assigned an initial condition and subsequently completed all three conditions sequentially. Each session lasted 15 minutes, during which participants created a Battle

Royale-style map. After completing each condition, participants filled out a CSI questionnaire, and after completing all three conditions, they completed the factor analysis component of the CSI.

A G*Power analysis suggested that 45 participants would be required to perform an ANOVA on the CSI results. Students from an Introduction to Computing course were invited to participate voluntarily. This course aimed to give students experience working with large datasets, making participation in an academic research project relevant to their studies. In total, 60 participants completed all three conditions. After applying a Mahalanobis distance function to identify and remove outliers, 57 valid results remained for analysis.

An ANOVA was then performed on the CSI scores. The results were unexpected: both LLM-supported conditions scored significantly lower than the original Melete system. Melete (without LLM assistance) achieved a mean CSI score of 70 out of 100, outperforming both LLM-based conditions.

The growth of GenAI and its integration into CSTs is an exciting and rapidly evolving area of research. However, as demonstrated in the studies above, analyzing CSTs—particularly MIAI pipelines—remains a challenging task.

Qualitative analysis methods, while valuable, are time-consuming and often open to subjective interpretation. Meanwhile, there is no standardized approach to quantitatively evaluating MIAI pipelines. In this context, the CSI offers a valuable starting point for building more structured methods of analysis.

Nevertheless, it is important to critically examine the role the CSI can play in adequately evaluating these systems. The following section of this paper will propose potential improvements to the CSI by identifying limitations in its current application within MIAI pipelines.

3. Method

To evaluate the CSI measurement tool, the CSI psychometric data gathered from: ”Melete: The Importance of Kernel Interaction in Mixed-Initiative Artificial Intelligence Pipelines”[41] and ”Melete: Using Large Language Models (LLMs) as Co-Creators for Kernels in Mixed-Initiative Artificial Intelligence Pipelines”[42] was used.

The responses to the surveys were then used to conduct a confirmatory factor analysis[37] and then grouped and based on their psychometric properties. Namely, based on the recommendations of Straub et al.[36], and other authors (e.g. [43]) reliability and validity need to be established to deem a measurement instrument adequate.

The data was then analyzed using SPSS and AMOS. Outliers from both studies were determined with a Manalobis distance function and chi-squared analysis. In ”Melete: The Importance of Kernel Interaction in Mixed-Initiative Artificial Intelligence Pipelines”[41] 4 sets of responses were removed based on that analysis as stated above and in ”Melete: Using Large Language Models (LLMs) as Co-Creators for Kernels in Mixed-Initiative Artificial Intelligence Pipelines”[42], 3 sets of responses were also removed as stated in the above section.

	Kernel Study	LLM Study
Condition 1	42 Responses	57 Responses
Condition 2	42 Responses	57 Responses
Condition 3	42 Responses	57 Responses
TOTAL		297 CSI Responses

To conduct the factor analysis, we first tested the CSI scores against a null model, which assumes no correlation between the factors (collaboration, enjoyment, exploration, expressiveness, immersion, and worthwhile results). The factor analysis was performed in AMOS using maximum likelihood estimation, estimated means, and intercepts.

To further validate these results and ensure there was no correlation or very low correlation between measures of unrelated constructs we conducted a discriminant validity analysis. Discriminant validity measures how well one variable differs from others in the same model. Testing discriminant validity ensures variables can explain more of their own data than errors or overlaps with other variables in the model. Scenarios in which discriminant validity issues can occur are when measurement errors or similar external, unmeasured influences affect the model. Alternatively, constructs within the conceptual framework are too similar. If this happens, the reliability of the measures is questionable[44]. Shared variance is the overlap between two variables, calculated by squaring their correlation[37].

3.1. Participants

Participants for this experiment were recruited from undergraduate games students at a British University. To ensure broad representation from disciplines across the game development sector, participants were drawn from various specialisms, including computer science, robotics, game design, game programming, game art and game writing. The experiment was advertised to first-year computing students in sessions and, more broadly, within the department to other disciplines.

4. Results and Data Analysis

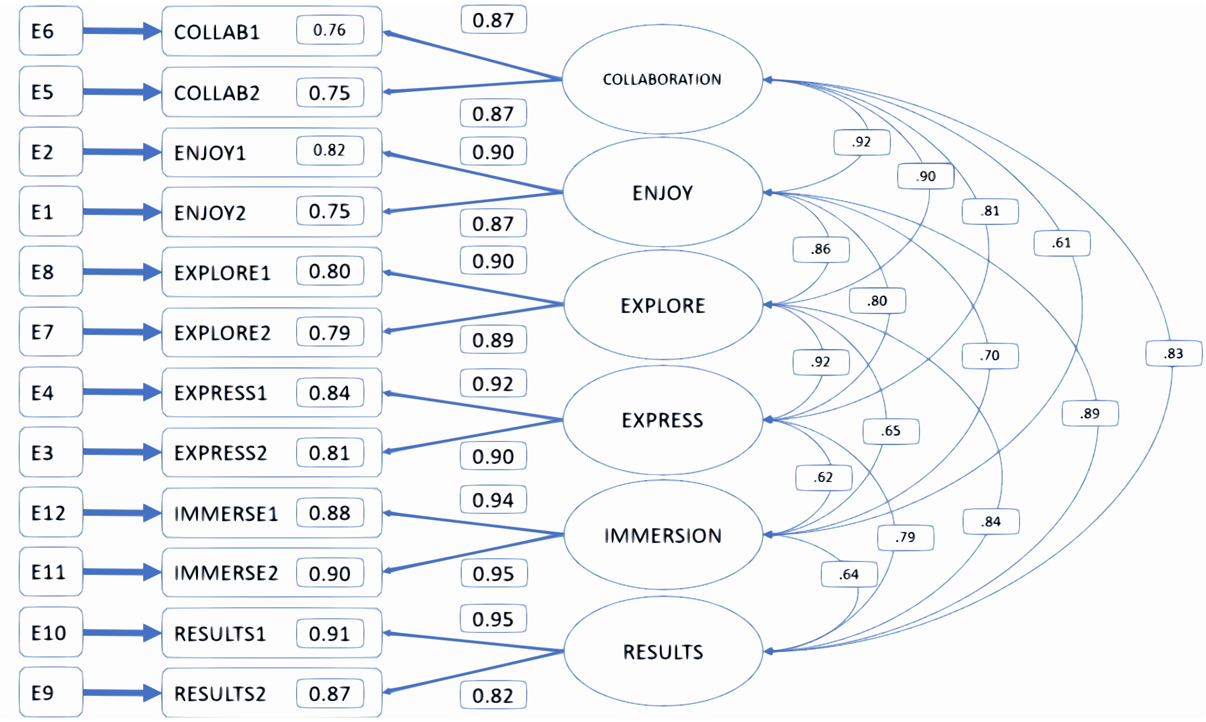
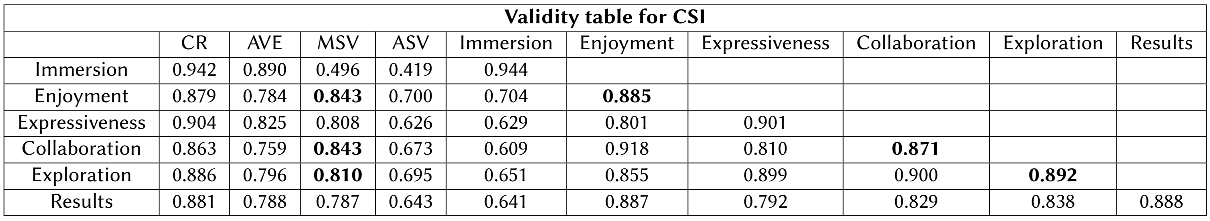


Figure 4: Creativity support index factor analysis

The diagram[Figure.4] is the factor analysis of the creativity support index. For the construct of Collaboration, the factor loadings for COLL1 and COLL2 are 0.87 and 0.87, respectively. In the Enjoy construct, ENJOY1 and ENJOY2 have loadings of 0.90 and 0.87. Similarly, for the Explore construct, EXPLORE1 and EXPLORE2 have loadings of 0.90 and 0.89. The Express construct shows loadings of 0.92 for EXPRESS1 and 0.90 for EXPRESS2. For Immersion, IMMERSE1 and IMMERSE2 exhibit of 0.94 and 0.95. Finally, for the Results construct, RESULTS1 and RESULTS2 have loadings of 0.95 and 0.82.

The latent constructs reveal correlations. For instance, the correlation between Collaboration and Enjoy is 0.92. Collaboration and Immersion are 0.61, Collaboration and Express 0.81 and Collaboration and Explore 0.90. Enjoy and Results 0.89. Whereas the correlation between Enjoy and Immersion is 0.70, Enjoy and Express is 0.80 and Enjoy and Explore is 0.86. With regards to Explore, its correlations with Express, Immersion and Results are 0.92, 0.65 and 0.84, respectively. Whilst Express correlates with Immersion and Results 0.62 and 0.79. Finally, Immersion and Results correlate only 0.64. These correlations suggest that the constructs are not independent of each other but are closely intertwined, and needed further exploration to ensure there is no discriminant validity.

Table 1
Validation table for CSI



Validity table for CSI										
	CR	AVE	MSV	ASV	Immersion	Enjoyment	Expressiveness	Collaboration	Exploration	Results
Immersion	0.942	0.890	0.496	0.419	0.944					
Enjoyment	0.879	0.784	**0.843**	0.700	0.704	**0.885**				
Expressiveness	0.904	0.825	0.808	0.626	0.629	0.801	0.901			
Collaboration	0.863	0.759	**0.843**	0.673	0.609	0.918	0.810	**0.871**		
Exploration	0.886	0.796	**0.810**	0.695	0.651	0.855	0.899	0.900	**0.892**	
Results	0.881	0.788	0.787	0.643	0.641	0.887	0.792	0.829	0.838	0.888

Some discriminant validity issues where identified by the validity analysis. Particularly Collaboration, Enjoyment, and Exploration, show high correlations, Collaboration (0.871), Exploration (0.892), and Enjoyment (0.885) indicating potential overlap between these constructs. To address the issues with discriminant validity we tried multiple models removing individual factors. However, other correlations started to appear, as a result, we just removed the three factors that where indicated in the CSI factor analysis.

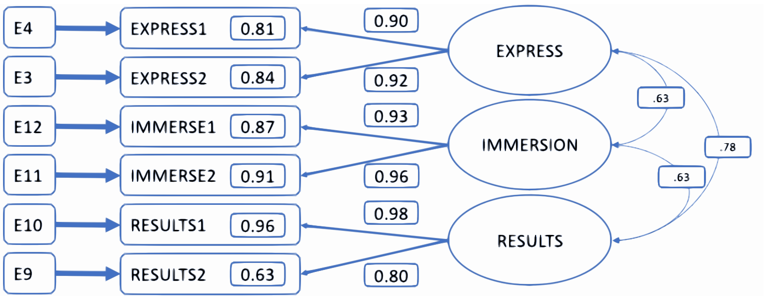
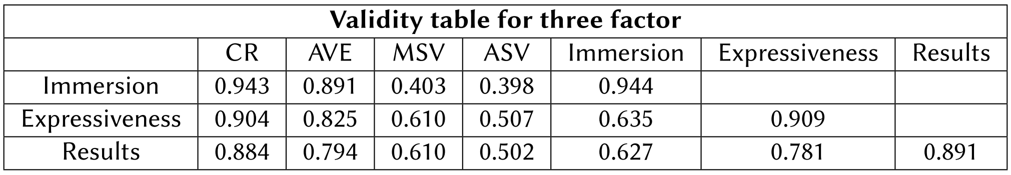


Figure 5: 3 Factor analysis, omitting Collaboration, Enjoy, Explore.

The factor analysis conducted on the latent variables: Express, Immersion, and Results. Express is linked to two observed variables, EXPRESS1 and EXPRESS2, with factor loadings of 0.90 and 0.92, respectively. Immersion is connected to IMMERSE1 and IMMERSE2, with loadings of 0.93 and 0.96. Results is tied to RESULTS1 and RESULTS2, with loadings of 0.98 and 0.80. The correlations among factors are, Express and Immersion are correlated with a coefficient of 0.63. Express and Results are correlated at 0.78. Immersion and Results have a correlation of 0.63. This model evaluates how well the observed variables measure their respective latent constructs and the relationships among the latent variables. The standardized loadings and correlations suggest a relatively good fit for most relationships.

Table 2
Validity table for three factor model



Validity table for three factor							
	CR	AVE	MSV	ASV	Immersion	Expressiveness	Results
Immersion	0.943	0.891	0.403	0.398	0.944		
Expressiveness	0.904	0.825	0.610	0.507	0.635	0.909	
Results	0.884	0.794	0.610	0.502	0.627	0.781	0.891

We did not find any discriminant validity issues with these factors.

The Model Indices Comparison Table compares the CSI model with a model corrected for discriminant validity issues. The corrected model shows a significant improvement in model fit, as indicated by a

Table 3
Model Indices Comparison table

Model Indices Comparison table		
	Creativity Support Index	Corrected for discriminant Validity Model
Chi-Square	104.114	14.150
Probability level	.000	.028
Root mean square error of approximation	.072	.065
Tucker-Lewis Index	0.971	.0.987
Standardised Root Mean Squared Residual	0.0191	.0148

much lower Chi-Square value (104.114 vs. 14.150) and a higher Tucker-Lewis Index (0.971 vs. 0.987), both suggesting a better fit for the corrected model. The Root Mean Square Error of Approximation (RMSEA) improves slightly from 0.072 to 0.065, moving closer to the ideal value below 0.06. Additionally, the probability level of the corrected model is still significant but less extreme (0.028), indicating improved validity.

5. Discussion

The factor analysis results indicate that the standardized estimates for the relationships between questions and factors exceeded the minimum threshold for a factor analysis, which is set at 0.7. The values ranged from 0.76 to 0.95, with a mean of 0.89, indicating that a substantial portion of variance was explained by the latent variables across all factors. Notably, RESULT2 showed the lowest variance explanation at 0.82 but still met the required threshold[Figure.4].

The CSI probability level was statistically significant ($p < 0.001$), which is expected given the large sample size and the chi-square value for the model was 104.114[Table.3], indicating an adequate fit. These findings support the overall validity of CSI framework. The validation of the CSI model is supported by the cut-off indices used for structural analysis.

However, the factor analysis also revealed correlations between several factors. For example, in the CSI model diagram, collaboration and exploration are highly correlated, as are worthwhile results with other factors. [Figure.4] Despite the conceptual separation of the CSI factors, this analysis indicates that participants may perceive some of these factors as interrelated4. This was further explored with a discriminant validity test[Table 1]. The model exhibited issues with discriminant validity, as presented in the discriminant validity tables. Notably, enjoyment, collaboration, and exploration. This was determined by the square root of the AVE being lower than the absolute value of the correlations with other factors.

It is also important to note that no validity concerns were identified with the three factors model which consisted of expressiveness, immersion, and results, and this model demonstrates a good fit with respect to these factors.[Table. 2]

The analysis suggests that participants may be responding to the items related to the three specific factors—enjoyment, exploration, and collaboration—in a similar manner, indicating that these concepts may be statistically indistinguishable from one another. This overlap suggests potential conceptual or empirical similarity between these factors, which may be contributing to the lack of discriminant validity observed in the model. This may be explained by external factors such as the participants opinion on AI. Alternatively, the phrasing, word choice, or number of questions presented in the psychometric survey may be more likely to be the cause of the issue. Although the survey is theoretically well grounded, the issues with correlation and discriminant validity suggest that there is room for improvement.

6. Conclusions

In conclusion, while the CSI demonstrates a generally strong model for assessing the usefulness of both CSTs and MIAI pipelines, there are issues with discriminant validity in the factors of enjoyment, exploration, and collaboration. We recommend expanding the latent variables within these factors and conducting further factor analysis to better understand and resolve the discriminant validity issues. Once these issues are addressed, further work can focus on improving the standardized RMR and RMSEA, ultimately leading to the development of an enhanced CSI model better suited for evaluating both CSTs and MIAI pipelines. With regards to the study presented there are some possible limitations in this work.

Firstly, the data collected draws primarily from undergraduate participants in a higher-education context, so these results might not be generalizable to a broader audience. In addition to this although we attempted to be as inclusive as possible the demographic sampling does skew towards the male gender which is representative of both industry sampling[45] and higher STEM education[46]. As such this could also be explored as future work.

Secondly, although this study uses a single tool, the design presents two possible configurations as part of the analysis. This can indicate some generalizability of the work, but both experiments were similar in nature, so the findings might not be generalizable to other pipelines, and ensuring that this is the case will require additional research. However, this is a good first step for verifying the validity of the CSI measurement tool.

Finally, another avenue for future work would be to compare the existing CSI to an improved version using the recommendations we have made in this paper. For example, a CSI with more latent variables, or one less factors tested using similar MIAI or Co-creative pipelines.

7. Declaration on Generative AI

During the preparation of this work, authors used ChatGPT and Grammarly for the following: **grammar and spellcheck**. After using this tool/service, the authors, reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Amershi, M. Cakmak, W. B. Knox, T. Kulesza, Power to the people: The role of humans in interactive machine learning, AI magazine 35 (2014) 105–120.

[2] T. Lubart, How can computers be partners in the creative process: classification and commentary on the special issue, International journal of human-computer studies 63 (2005) 365–369.

[3] N. Shaker, M. Shaker, J. Togelius, Ropossum: An authoring tool for designing, optimizing and solving cut the rope levels, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 9, 2013, pp. 215–216.

[4] H. Yu, Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching, Frontiers in Psychology 14 (2023) 1181712.

[5] J. Togelius, G. N. Yannakakis, K. O. Stanley, C. Browne, Search-based procedural content generation: A taxonomy and survey, IEEE Transactions on Computational Intelligence and AI in Games 3 (2011) 172–186.

[6] T. Punter, M. Ciolkowski, B. Freimut, I. John, Conducting on-line surveys in software engineering, in: 2003 International Symposium on Empirical Software Engineering, 2003. ISESE 2003. Proceedings., IEEE, 2003, pp. 80–88.

[7] B. Shneiderman, Creating creativity: user interfaces for supporting innovation, ACM Transactions on Computer-Human Interaction (TOCHI) 7 (2000) 114–138.

[8] E. A. Carroll, C. Latulipe, R. Fung, M. Terry, Creativity factor evaluation: towards a standardized survey metric for creativity support, in: Proceedings of the seventh ACM conference on Creativity and cognition, 2009, pp. 127–136.

[9] E. Cherry, C. Latulipe, Quantifying the creativity support of digital tools through the creativity support index, ACM Transactions on Computer-Human Interaction (TOCHI) 21 (2014) 1–25.

[10] B. Shneiderman, Supporting creativity with advanced information-abundant user interfaces, Springer, 2001.

[11] S. G. Isaksen, D. J. Treffinger, Celebrating 50 years of reflective practice: Versions of creative problem solving, The Journal of Creative Behavior 38 (2004) 75–101.

[12] S. G. Isaksen, D. J. Treffinger, Creative problem solving, The Basic Course. New York: Bearly Limited (1985).

[13] L. Candy, E. A. Edmonds, Proceedings of the international symposium creativity and cognition., 1993.

[14] L. Candy, K. Hori, The digital muse: Hci in support of creativity: " creativity and cognition" comes of age: towards a new discipline, interactions 10 (2003) 44–54.

[15] B. Shneiderman, G. Fischer, M. Czerwinski, M. Resnick, B. Myers, L. Candy, E. Edmonds, M. Eisenberg, E. Giaccardi, T. Hewett, et al., Creativity support tools: Report from a us national science foundation sponsored workshop, International Journal of Human-Computer Interaction 20 (2006) 61–77.

[16] B. Shneiderman, Creativity support tools: accelerating discovery and innovation, Communications of the ACM 50 (2007) 20–32.

[17] B. Shneiderman, Creativity support tools: A grand challenge for hci researchers, in: Engineering the user interface: From research to practice, Springer, 2008, pp. 1–9.

[18] J. Frich, M. Mose Biskjaer, P. Dalsgaard, Twenty years of creativity research in human-computer interaction: Current state and future directions, in: Proceedings of the 2018 Designing Interactive Systems Conference, 2018, pp. 1235–1257.

[19] A. Liapis, G. N. Yannakakis, J. Togelius, Designer modeling for sentient sketchbook, in: 2014 IEEE Conference on Computational Intelligence and Games, IEEE, 2014, pp. 1–8.

[20] J. Kaye, Evaluating experience-focused hci, in: CHI'07 extended abstracts on Human factors in computing systems, 2007, pp. 1661–1664.

[21] R. Mandryk, A. S. I. G. on Computer-Human Interaction, Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, 2018.

[22] M. A. Boden, The creative mind: Myths and mechanisms, 2004.

[23] J. C. Read, S. MacFarlane, C. Casey, Endurability, engagement and expectations: Measuring children's fun, in: Interaction design and children, volume 2, Citeseer, 2002, pp. 1–23.

[24] M. Csikszentmihalyi, Flow and the psychology of discovery and invention, HarperPerennial, New York 39 (1997) 1–16.

[25] S. Deterding, J. Hook, R. Fiebrink, M. Gillies, J. Gow, M. Akten, G. Smith, A. Liapis, K. Compton, Mixed-initiative creative interfaces, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2017, pp. 628–635.

[26] L. Mamykina, L. Candy, E. Edmonds, Collaborative creativity, Communications of the ACM 45 (2002) 96–99.

[27] S. J. Russell, P. Norvig, Artificial intelligence: a modern approach, Pearson, 2016.

[28] N. Davis, C.-P. Hsiao, K. Y. Singh, L. Li, S. Moningi, B. Magerko, Drawing apprentice: An enactive co-creative agent for artistic collaboration, in: Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition, 2015, pp. 185–186.

[29] C. Oh, J. Song, J. Choi, S. Kim, S. Lee, B. Suh, I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–13.

[30] Y. Lin, J. Guo, Y. Chen, C. Yao, F. Ying, It is your turn: Collaborative ideation with a co-creative robot through sketch, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–14.

[31] P. Lucas, C. Martinho, Stay awhile and listen to 3buddy, a co-creative level design support tool., in: ICCC, 2017, pp. 205–212.

[32] J. Liao, P. Hansen, C. Chai, A framework of artificial intelligence augmented design support, Human–Computer Interaction 35 (2020) 511–544.

[33] F. A. Figoli, F. Mattioli, L. Rampino, Artificial intelligence in the design process: The Impact on

Creativity and Team Collaboration, FrancoAngeli, 2022.

[34] R. Pandya, S. H. Huang, D. Hadfield-Menell, A. D. Dragan, Human-ai learning performance in multi-armed bandits, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 369–375.

[35] G. Zhang, A. Raina, J. Cagan, C. McComb, A cautionary tale about the impact of ai on human design teams, Design Studies 72 (2021) 100990.

[36] D. Straub, M.-C. Boudreau, D. Gefen, Validation guidelines for is positivist research, Communications of the Association for Information systems 13 (2004) 24.

[37] J. Hair, Multivariate data analysis, Exploratory factor analysis (2009).

[38] I. Karth, A. M. Smith, WaveFunctionCollapse is Constraint Solving in the Wild, Proceedings of the 12th International Conference on the Foundations of Digital Games (2017) 68:1–68:10. URL: http://doi.acm.org/10.1145/3102071.3110566. doi:10.1145/3102071.3110566.

[39] S. Murturi, J. Walton-Rivers, M. Scott, M. Yee-King, M. Gillies, Melete: Playtesting and 3d environments for mixed-initiative artificial intelligence as a method for prototyping video game levels, CHI Play NA (2025) NA.

[40] S. Murturi, T. Pellicone, M. Yee-King, M. Gillies, Melete: Exploring the components of mixed-initiative artificial intelligence pipelines for level design., In Review at SYNERGY, PhD Viva Passed, Unpublished NA (2025) NA.

[41] S. Murturi, M. Yee-King, M. Gillies, Melete: The importance of kernel interaction in mixed-initiative artificilal intelligence pipelines., in: Unpublished, PhD, viva passed, 2025, p. NA.

[42] S. Murturi, J. Walton-Rivers, M. Yee-King, M. Gillies, Melete: Using large language models (llms) as co-creators for kernels in mixed-initiative artificial intelligence pipelines, PhD NA (2025) NA.

[43] R. F. DeVellis, C. T. Thorpe, Scale development: Theory and applications, Sage publications, 2021.

[44] C. Fornell, D. F. Larcker, Evaluating structural equation models with unobservable variables and measurement error, Journal of marketing research 18 (1981) 39–50.

[45] T. Camp, 'computing, we have a problem…', acm inroads 3 (2012) 34–40.

[46] S. Kulturel-Konak, M. L. D'Allegro, S. Dickinson, Review of gender differences in learning styles: Suggestions for stem education., Contemporary Issues in Education Research 4 (2011) 9–18.