Behind the Scenes: Unpacking Students' Experience during a Collaborative AI Workshop using Multi-Modal Data

Isabella Possaghi
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
isabella.possaghi@ntnu.no

Feiran Zhang
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
School of Design
Hong Kong Polytechnic University
Hong Kong SAR, Hong Kong
feiran.zhang@ntnu.no

Kshitij Sharma
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
kshitij.sharma@ntnu.no

Sofia Papavlasopoulou
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
spapav@ntnu.no

Abstract

Artificial Intelligence (AI) is playing a growing role in K-12 education. However, curricula often lack structure and proper assessment when paired with collaborative approaches like Design Thinking (DT). Here, behavioral and affective dynamics are overlooked, even though they are indicators of both performance and quality of the learning experience, warranting a more in-depth exploration through Multi-Modal Learning Analytics. Therefore, we engaged 63 students, divided into 29 groups (aged 11 to 15) in a DT workshop on AI, analyzing their performance across each stage of their experience, including their behavioral and affective (i.e., emotional) states, using data collected from physiological sensors, audio, and video recordings. Our results show that certain conditions (e.g., joint visual attention, boredom, and high stress) consistently predicted positive or negative performance across all stages of the workshop, while affective states such as confusion, frustration, high engagement, and low stress fluctuate with implications on the learning experience.

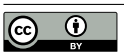
CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in interaction design.*

Keywords

Design Thinking; Collaboration; Multi-Modal Data; Learning Analytics; Education; Learning

ACM Reference Format:
Isabella Possaghi, Feiran Zhang, Kshitij Sharma, and Sofia Papavlasopoulou. 2025. Behind the Scenes: Unpacking Students' Experience during a Collaborative AI Workshop using Multi-Modal Data. In *Interaction Design and Children (IDC '25), June 23–26, 2025, Reykjavik, Iceland.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3713043.3728839



IDC '25, Reykjavik, Iceland

ACM ISBN 979-8-4007-1473-3/25/06
https://doi.org/10.1145/3713043.3728839

1 Introduction

Computer Science (CS) and close disciplines underwent considerable reframing in recent years. With an interdisciplinary deployment that constantly grows, CS education now includes broader considerations of social, societal, and ethical dimensions that parallel the diffusion of technology into everyday life [25, 101]. Similar progress is recommended in AI education as it is in CS education [38, 88, 101], especially because students are more and more engaged in AI-driven tools [25, 39, 98]. Grover recently described the profound transformation of society given digital advancements, positioning AI within CS curricula as necessary to navigate these evolving scenarios [25]. Yet, AI education for K-12 should not be a simple response to manage AI pervasiveness from a sole technical perspective but also as a proactive measure that equips young learners with agency and awareness on the topic to prevent inclusion disparities [39, 76]. Educational activities following open-ended blueprints close to constructionism are useful to this end. To prepare students for addressing the "wicked issues" of digitalization, curricula can incorporate hands-on strategies and scaffold learner-centered experiences for the construction of their own knowledge. These practices can diversify in affordances, subdividing them into project-based learning, maker education, Design Thinking (DT), and many more. Among these, DT is a human-centered, iterative approach that conventionally follows the stages of *Empathize*, *Define*, *Ideate*, *Prototype*, and *Test* [71]. Its popularity in CS stems from its alignment with computational thinking and problem-solving techniques, fostering creativity, collaboration, and out-of-the-box reasoning [61].

Conducted within teams, DT-related workshops supported by digital tools promote breaking down silos between technical and non-technical roles for collective engagement. Resulting dynamics are subjected to Self-Regulated Learning (SRL) [28] and Socially

Shared Regulation of Learning (SSRL) [62]. These important components of collaborative experiences merit further exploration using indicators extending beyond mere performance metrics [59, 90]. In fact, measuring performance alone, particularly when focused solely on learning gains assessed at the conclusion of an activity, provides limited insight into the dynamics that take place during collaboration, leaving these constructs impossible to analyze in practice. Nasir et al. draw attention to this challenge, which exacerbates in open-ended educational activities where solutions are not unidirectional like those frequently found in constructionist approaches [59]. In these contexts, understanding the in-group experience holistically requires consideration of real-time and process-oriented indicators of SSRL. Joint learning processes unfold in complex and dynamic interactions that are most often scrutinized using non-standardized and subjective rates, resulting in a lack of consistency [20]. The dual demands of evaluating collaborative learning processes and student-driven knowledge construction call for innovative assessment styles that are highly context-dependent [44]. In response, state-of-the-art sensors and instruments enable the use of Multi-Modal (MM) data, information collected and integrated from multiple channels, analyzed using Learning Analytics approaches [10, 11]. Here, we position our research, fulfilling the gap by combining learners' affective (i.e., emotional) and behavioral responses at the group level with performance scores from each DT stage, an area that remains highly unexplored. Chosen joint affective (i.e., engagement, stress, frustration) and behavioral (i.e., visual attention) measurements are flexible enough to capture the ebb and flow of collaboration while remaining rigorous enough to deliver actionable insights [81]. The following research questions reflect the objective of our study:

- **RQ1:** How are the different Design Thinking stages related in terms of performance outcomes?
- **RQ2:** What are the overall relationships between students' affective and behavioral responses and their performance outcomes across the Design Thinking stages?

To answer our questions, we carried out an intervention in authentic K-12 school settings during usual classroom schedules. The intervention involved a DT workshop to elaborate on AI and Machine Learning (ML) concepts collaboratively, with the scope of creating a playable game to raise awareness of the subject. In teams, students navigated problem-solving with different hands-on supports, such as shared paper-based sheets and a block-based programming environment. Here, MM data (students' physiological responses, facial expressions, and visual attention) were captured moment-to-moment along with team learning performance. Our results first map association between DT stages in terms of performance and then reveal which measurements are correlated with performance in each stage, reflecting factors that co-occur with higher or lower performance. This allows for a broader overview that indicates which states are more persistent than others and which are transient, occurring only during a single stage of the DT process. Our contributions are summarized as:

- *a)* We designed and implemented a workshop "in-the-wild", namely in a K-12 school setting. The empirical study presents insights from data collected from 63 students using multiple sources, including MM data.
- *b)* We show the relationship between the different DT stages, illustrating how performance outcome in one stage correlates, or does not, to subsequent stages. Moreover, we map the affective and behavioral states that are characteristic of performance for each DT stage.
- *c)* We provide implications for research and practice discussing the feasibility of curricula focused on AI, ML, and related concepts paired with a DT approach. Rich and innovative methods provide insights into the in-classroom orchestration of learning activities with adequate scaffolding, guidance, and attention to affective well-being.

2 Background and related work

2.1 Artificial Intelligence Education for K-12

In the introduction to his book on teaching AI, Zimmerman defines it as both a replicator of human intelligence and an effort to imbue machines with human-like qualities [107]. Among young populations, AI-driven tools and services proliferate, but most users remain unaware of the underlying mechanisms governing these systems, making them prone to misconceptions, misuse, and ethical dilemmas. These issues remain unresolved, with guidelines for AI integration into curricula that are still far from being solidified [25, 101]. Several factors underlie the complexity of this issue. Among others, [9], concerns about age appropriateness and accessibility can hinder a proper access of AI and ML concepts in education. As often applied in CS for programming [5], early exposure can rely on "unplugged" strategies [12, 53]. These approaches do not necessarily require direct interaction with AI or ML technologies but instead prioritize foundational principles such as database training, pattern recognition, and functional biases. These activities can gently introduce young learners to AI and ML concepts, building awareness and confidence in their understanding before breaking down more technological aspects [25]. Moreover, situating AI and ML learning in relatable contexts helps students apply these often abstract and distant concepts to their own lives, extending their engagement [12, 60], and encouraging them to think about the wider social and ethical consequences of AI and ML. [25]. Conversely, when these connections are absent or irrelevant, these educational activities diminish in effectiveness because disjointed from the world outside the classroom [57, 76, 98]. Multiple initiatives have contributed to bring AI in K-12 learning with guidelines and compendia [57]. For example, the Computer Science Teachers Association is working closely with the Association for the Advancement of Artificial Intelligence to revise the United States national standards for K-12 CS education by 2026 [98]. However, tailoring these resources to resonate with everyday in-school experiences remains an ongoing challenge. Despite the promise of these initiatives, recent reviews warn that the general lack of curriculum design significantly restricts the integration of AI and ML concepts into young students' educational journeys [9, 92, 101]. To address this, targeted efforts to draw insights from theoretical frameworks to inform the design of K-12 curricula are essential to bridging the gap between feasibility and meaningful learning experiences [76].

2.1.1 Design Thinking for AI and ML. Academic consensus supports engaging young learners in approaching AI and ML concepts guided by constructionist frameworks [1, 12, 26, 51]. Rooted in experiential learning pedagogy, constructionism is an approach that views students' personal interest and collaborative first-hand experience as cornerstones of knowledge production [31]. For instance, the symposium chaired by Morales-Navarro & Kafai compiles empirical research on constructionism to support learning AI and ML concepts in developmentally appropriate and personally relevant ways from an early age [58]. Additionally, UNESCO's 2022 global report on the state of AI education in K-12 settings supports the adoption of constructionist blueprints as a foundation for effective learning experiences [57], especially when integrated with DT and computational thinking modules. DT borrows from open-ended constructionism, emphasizing active creation while channeling it towards innovative solutions through user-centric models. It encourages open-mindedness, challenges assumptions, and allows for the iterative redefinition of problems [43, 63]. A partition of DT into five stages, *Empathize*, *Define*, *Ideate*, *Prototype*, and *Test*, is now widely proposed in K-12 educational programs. This framework was formalized and disseminated by the Stanford d.school to enhance its accessibility and applicability across disciplines and maintain references to authentic scenarios [71]. DT often partner with computational thinking given their complementing nature in fostering problem-solving with creativity [41] while handling the complexity of real-world issues [61], making it more and more popular in CS, including AI and ML learning scenarios. However, the overall design, orchestration, and evaluation of these experiences are challenging for the research community. Particularly on the latter, the review by Li et al. on task design for AI education in K-12 grades mapped a great availability of assessment methods. Yet, they suffer from limited efficacy, as conventionally used instruments are prone to high bias, undermining the reliability and validity of the results [51]. This position echoes Van Mechelen et al., who advocate for more formative assessments aimed at learners' journeys, enhanced through feedback and tailored guidance, rather than focusing exclusively on outcomes [101]. Consequently, curricula development and testing for emerging technologies, such as AI and ML, must investigate in-depth learners' responses to improve educational approaches that empower young learners to assert their agency in current and future digital landscapes confidently.

2.2 Affective and Behavioral Dynamics in Collaborative Learning Experiences

Multiple mechanisms characterize learning experiences and shape how individuals acquire knowledge. Within the Self-Regulated Learning (SRL) framework, the concept of being "socially situated" [28, p. 85] emphasizes that the process is both a personal endeavor and socially embedded. When collaboration happens on a shared task, bound by a collective commitment, learners form a dynamic social system with shared regulatory processes [79]. This phenomenon is called Shared Regulation of Learning (SSRL) [62] , and literature on the subject does not center on cognition only but also involves affective and behavioral dynamics. For example, the evolving nature of emotions (e.g., frustration, delight) during learning experiences and their impact [45]. Applicable at both individual and collaborative levels [103], D'Mello and Graesser proposed a model of affective dynamics to explore how emotional states impact and are influenced by cognitive processes [18]. They refer to *cognitive equilibrium* as a balanced state where two or more participants operating on the same task share similar cognitive states, such as common progress expectations, favoring conditions for learning with sustained communication and engagement [16]. However, the sophistication of learning activities (e.g., open-ended and/or collaborative ones) requires higher cognitive alignment between partners. In fact, tasks with greater complexity demand greater mental coordination, making the attainment of the *cognitive equilibrium* more critical [23, 97]. This condition is not static; rather, it oscillates between *cognitive equilibrium* and *cognitive disequilibrium*. The latter arises from a contradiction or mismatch between a given task and learners' pre-existing mental models, resulting in confusion. However, *cognitive disequilibrium* can also initiates knowledge growth, as it it compels learners to reassess, adapt, and expand their mental models, ultimately leading to the restoration of *cognitive equilibrium* through knowledge acquisition [19, 50].

Described dynamics interconnect with group-level phenomena: emotional exchange, cognitive appraisal, coping mechanisms, and feedback loops, among others. Van Kleef explains how emotions serve as relational signals that influence interpersonal attitudes [99]. Positive states can quickly spread among collaborative settings, fostering cohesion and reinforcing a shared sense of purpose. When one team member displays engagement, it can create a ripple effect, stimulating the morale and investment of the group. Negativity can also propagate within a team. Frustration or boredom exhibited by one individual can evoke similar feelings in others, resulting in tension, miscommunication, and a decline in group harmony [32]. These negative emotions may exacerbate conflicts and mine motivation, derailing performance if they are not acknowledged or managed. Even though emotional exchange is characterized by involuntary responses, such as automatic mimicry [66], its occurrence is neither uniform nor deterministic. Diverse factors, including the emotional awareness of team members [100], influence the degree of emotional contagion. Emotional awareness concerns the interpretation of one's own learning experience and the emotional states of others. A key component in developing such awareness is *cognitive appraisal*. Scherer and Moors define "appraisal" as the evaluation of a circumstance in relation to an individual's well-being (e.g., satisfaction or perceived harm) [77]. This evaluative process is central to emotional episodes, as it differentiates them into specific types, such as anger, sadness, or joy, and their direction towards a target [66]. As cognitive appraisal does not merely demarcate the nature of affective states but also subsequent behaviors, reframing appraisals can redirect actions to improve collaboration [91]. For instance, interpreting a team member's frustration as a call for help rather than a mistake promotes a more empathetic and supportive environment [37].

Similarly, *coping mechanisms* [66] manage social experiences, either by directly influencing the emotion itself or by changing the individual's response to it. In collaborative learning, coping mechanisms enable team members to adapt their emotional state and respond effectively to the feelings of others. An example is conceptualizing failures as learning opportunities instead of setbacks, possibly mitigating feelings of anger and defeat. The conscious

alteration of a negative emotional perspective into a more constructive view can lessen its impact [74]. Seeking support in the form of advice, reassurance, or clarification is another strategy to prevent feeling overwhelmed and to reestablish understanding. Such *coping mechanisms* are not exclusive to the individual level but also concern collective efforts to handle emotional climates. Teams may develop norms or practices over time to sustain a supportive atmosphere, such as regular check-ins or feedback sessions to address feelings related to the activity. Moreover, the circulation of emotions among team members can create *feedback loops* that can reciprocally influence cognitive load [102]. In fact, performance may suffer due to negative emotional contagion throughout the group. However, team members may adjust their emotional responses or behaviors when performance declines to restore emotional and cognitive equilibrium [68]. The study of SSRL and its dynamics applied to pedagogy and learning design looks at how affective, behavioral, and cognitive factors influence students' engagement with open-ended activities like DT and complex topics such as AI and ML. These factors are shaped by both the educational framework and the digital tools used, highlighting their significance in designing learning experiences.

2.3 Multi-Modal Approaches to Collaborative Learning

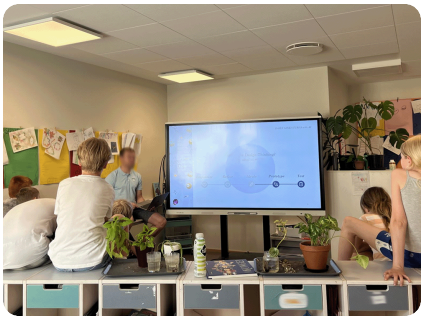
Despite being well-established, traditional data collection methods alone may fail to capture the complexity and richness of data generated during collaborative experiences. For instance, self-reports, observations, and standardized tests easily overlook the subtle shifts that occur throughout group interactions in affective, behavioral, and cognitive states [6, 106]. Adding to this, curricula with open-ended and hands-on modules introduce further complexity [93]. Collaborative, constructionist approaches like DT require assessment methods that can capture the nuances of this learning experience without disrupting the creative flow [44, 59, 90]. Technological advancements granted novel instruments and sensors to detect fine-tuned, real-time responses, tracking both behavioral and affective indicators with greater precision [35]. These assets also allow for further entry points from which data can be drawn, especially when the learning experience is digitally mediated or asks for digital interaction [90]. In this context, Multi-Modal (MM) data refers to the simultaneous collection and/or combination of different streams of information and takes the name MM Learning Analytics when leveraged to debrief the process of learning [11, 22, 81]. Over the years, the integration of behavioral data such as eye-gaze [48], gestures [89, 90], and speech [90] has become increasingly prevalent in research. These markers provide a window into how students interact with content, where their attention is directed, and how they are processing information. Spikol et al. used facial movements and orientations to analyze students' project-based group work in an LA approach. Their findings revealed collaborative features that predict artifact quality, providing data-driven insights for guiding constructionist pedagogies [90]. Additionally, emotional responses from facial expressions [87, 96], and physiological sensors [20, 95, 105] offer a deeper understanding of the learner's affective states, their levels, and fluctuations. Finally, MM data, when combined with metrics for tracking knowledge acquisition, has demonstrated efficacy in various studies [65, 72, 96]. While metrics like Relative Learning Gain (RLG) account for the proportional improvement in learning progress [75], they may miss a fuller picture of the learning experience by overlooking individual variations in affective and behavioral responses. For instance, electrodermal activity, measured via skin-contact sensors, can represent shared engagement, as physiological arousal tends to increase with sustained attention [13, 81, 95]. Similarly, heart rate and its variability can significantly reflect learners' ability to maintain balance and adapt to stress-related stimuli [72, 74, 81].

The rising interest in SSRL within Child-Computer Interaction (CCI) reflects the significance of understanding how collaborative dynamics take place in digitally rich educational environments. With a focus on learners' responses beyond cognition and their impact, the orchestration of collaborative educational experiences and the resultant performance can be optimized [52]. Moreover, because SSRL aligns closely with research in Computer-Supported Collaborative Learning (CSCL) [104], it can be harnessed in digitally rich educational environments, potentially leading to new insights into how technology-mediated collaboration impacts learning performance [34, 74]. Summarizing the field, Lämsä et al. report how future CSCL research would benefit from investigating group dynamics through time-sensitive analytics, offering a deeper understanding of how learning unfolds and is shaped by temporal factors [46]. With empirical validation, the contribution by Törmänen et al. combined observation, video recordings, and arousal state detection to explore affective conditions in emotion regulation at the team level. For a more comprehensive overview of students' emotional awareness and its support, the same authors advocate for further interventions centered on the interplay between cognitive and affective states during collaborative experiences [97]. By engaging middle schoolers in a coding workshop, Lee-Cultura et al. showcase the efficacy of combining visual attention data with emotional states retrieved at the group level to explore their causality and impact on coding production and quality [48]. Delving deeper on a theoretical level, Järvelä et al. conceptualized *"triggers"* as specific events or cues that prompt regulatory responses in CSCL with the introduction of a framework to contextualize adaptive and maladaptive SRL [33]. On this count, Li et al. draw from an MM dataset to investigate cognitive, behavioral, and emotional responses elicited by such triggers during a group project, providing a simulated yet authentic setting for social adaptation [52]. The findings highlight the diagnostic link between knowledge acquisition and emotional attitudes, which opens up possibilities for developing orchestration methods or support tools that provide scaffolded assistance to students. These insights can be expanded by looking at the collaborative process as a whole. Further research is needed to analyze hands-on experiences, such as DT for K-12, using objective metrics [55].

3 Method

3.1 Participants

In 2024, three middle schools in Norway participated in the research intervention. The intervention consisted of a six-hour DT workshop integrated into the regular school schedule (see *Figure* 1a). A total of 108 students aged 11 to 15 years (6th to 9th grades) took part in



(a) *Empathize* stage in the classroom

(b) Examples of personas

(c) Teams setting during the *Ideate* stage

Figure 1: Extracts from the Design Thinking workshop

the workshop. None had prior experience with the technologies used. Of these, 63 students (30 boys, 27 girls, 2 other, 4 undisclosed; Mean age = 12.79, SD = 1.18) joined the data collection. While all students participated in the workshop, inclusion in the data collection required parental consent and student willingness. The data collection adhered to GDPR regulations and received prior approval from SIKT [1].

3.2 Design Thinking Workshop for ML and AI Learning

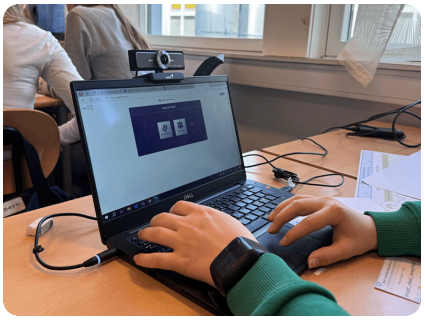
The workshop was split into two days of three hours each, including the regular recesses. It was led by a facilitator with a pre-service teaching background. Two researchers managed data collection and ensured the proper functioning of the technology. Students were subdivided into teams of both dyads and triads, with access to a shared laptop. This grouping was primarily based on logistical considerations, prioritizing dyads while also accounting for students' collaboration preferences. During the *prototyping* stage, the *SorBET (Sorting Based on Educational Technology)* website [2] was accessed through the shared device. *SorBET* is an online open-source platform where the learner can design, modify, and play sorting games by acting on a block-based coding interface [24]. The platform provides multiple customization options, such as embedding external resources, enabling the design of tailored activities that adapt flexibly to diverse contexts (AI and ML, in our case).

3.2.1 Design Thinking Workshop Steps. The DT workshop unfolded in five steps based on the Extend(DT)^2 activity plan [3], with the scope of creating a playable game and raising awareness on the topic. In the following lines, each step is detailed. **a. Empathize**: The facilitator introduced fundamental AI and ML concepts using videos (e.g., *How do machines learn?"*, *What is a database?"*), illustrating their everyday presence and potential ethical and societal biases. A discussion encouraged students to connect these concepts to their personal experiences and knowledge, fostering reflection and inspiration for the next stage. **b. Define**: Each team received a paper-based practice sheet to develop problem-solving skills by expanding the empathy exercise and defining a persona (example in *Figure* 1b) based on real-world references. Students analyzed user needs by identifying their personas' challenges and applied creative thinking to brainstorm AI-driven tools or systems to address them. An example outcome was: *This grandma wants new recipes for the Easter lunch to satisfy all her grandkids tastes! Perhaps a fancy AI helper could lend a hand!"*. **c. Ideate**: Teams received a new sheet to elaborate on the chosen challenges. This sheet guided students through a storyboard exercise, prompting them to expand on their persona's scenario (Figure 1c). They had to analyze AI applications and limitations by describing how the AI-driven tool or system would function, its successes, and potential shortcomings (e.g., *Oh no, all the recipes created by AI contain eggs, and one grandkid is allergic!"*). Overall, teams identified lessons learned about the design and limitations of AI and ML connected to their practical applications. Teams then engaged with *SorBET* to understand data organization for AI training. They ideated a fictitious database for their chosen AI-driven tool or system and applied categorization skills by selecting items and defining classifications. For example, a recipe suggestion system differentiated dishes into *"Kid's Favorites," "Very Popular," "Moderately Popular," "Occasionally Liked,"* and *"Unpopular."* This activity helped students grasp how AI systems rely on structured data to function effectively. **d. Prototype**: This stage aimed to develop computational thinking by transforming the designed database into an interactive format. Each team created gameplay within *SorBET*, integrating their database into a sorting game that tested their AI-driven categorization. Using *SorBET*'s block-based interface, students applied logical reasoning to conceptualize and implement gameplay mechanics, demonstrating how AI classifies information. **e. Test and Feedback**: After completing the game, teams exchanged their outcomes with another group to test the *SorBET* game. A feedback session encouraged reflection on AI decision-making and user experience design, serving as a closing discussion before the workshop concluded.

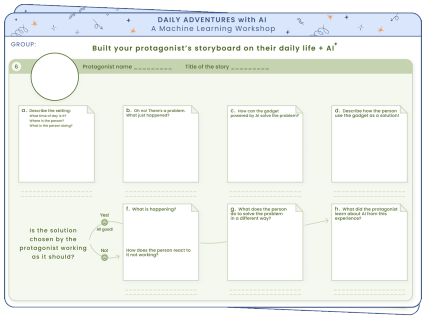
3.3 Data Collection

The 63 students participating in data collection were organized into 29 teams (24 dyads, 5 triads). The process incorporated three sources: individual pre- and post-knowledge tests, team artifacts, and MM data acquired in real-time. *Figure* 3 provides an overview of the data collection process. The pre-knowledge test was administered at the beginning of the workshop, while the post-knowledge

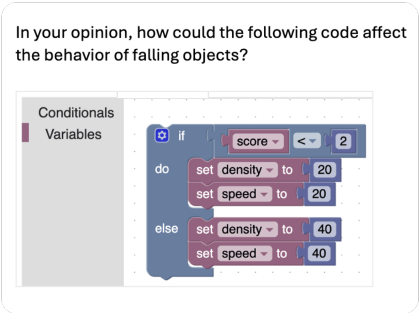
[1] The Norwegian Agency for Shared Services in Education and Research
[2] The *SorBET* platform: https://extendt2.com/widgets/sorbet/
[3] European Project Extend(DT)^2: https://extendt2.eu/



(a) Student wearing Empatica E4



(b) One of the two paper sheets



(c) Test excerpt for the *Prototyping*

Figure 2: Different data collection modalities

test was conducted at the end. Both consisted of multiple-choice questions divided into two sections. The first section assessed students' understanding of AI and ML concepts introduced during *Empathize*, while the second evaluated their ability to apply these concepts in a block-based coding environment, *SorBET* (*Figure* 2c), aligning with the *Prototyping* stage. Students' artifacts included Paper Sheet 1 from *Define*, Paper Sheet 2 (*Figure* 2b), and the *SorBET* database from the *Ideate* stage. Finally, MMD was gathered continuously throughout the DT workshop. A wide-angle camera on the laptop captured video and audio, including all team members (*Figure* 2c). Each team member wore an Empatica EmbracePlus (N=35) or Empatica E4 (N=28) wristband to collect biometric data continuously (*Figure* 2a). These devices measured key physiological parameters, considering students' Heart-Rate Variation (HRV) at 1Hz, blood volume pulse (64Hz), Electrodermal Activity (EDA) at 4Hz, and skin temperature at 4Hz. While this study did not use skin temperature or blood volume pulse, these data were retained for future research, particularly for forecasting applications. Video, audio, and biometric data were collected in a synchronized 30-second time window.

3.4 Data Pre-processing

3.4.1 Mounted Camera Data: Our video data was collected on-site in a school environment, presenting challenges typical of naturalistic research settings. While the raw footage was high quality and free from signal-based noise, variations in team arrangements (dyads or triads) during sessions posed difficulties for facial decoding. These changes occurred when students reached out to their classmates, left their teams temporarily, or unintentionally rotated the camera or device, resulting in other faces being captured in the frame. A deep learning-based face-tracking algorithm was used to monitor and trace team members consistently throughout the sessions, drawing on methodologies from similar studies [83, 86, 96]. This reduced biases related to team composition and unintended camera movements. Moreover, we performed video enhancement for facial feature detection to facilitate eye-tracking approximation via computer vision and deep learning [14]. To this end, we improved the resolution of key features in the eye area (e.g., eyebrows) to make them more distinguishable and enable the estimation of the location and movement of the gaze.

3.4.2 Wristband Data: We pre-processed the HRV and EDA data with a sliding window average. Namely, we used a moving window of 100 samples with overlapping 50 sample windows to remove any unwanted peak in the time series. Based on previous studies [21, 83, 84], the moving window approach allows for analyzing continuous data in smaller segments that overlap by 50%, preserving temporal continuity while enabling a more detailed analysis across different time scales. To account for potential subjective and contextual distortions (e.g., individual physical conditions), we considered the initial 30 seconds of each data stream to compute the mean and standard deviation, which were then used to normalize the subsequent time series for each student using z-scores. Previous studies have employed similar normalization techniques for HRV and EDA processing [8, 27, 47]. In this context, such a method ensures baseline control of physiological responses, thereby enhancing the reliability of group-level analyses.

3.5 Measurements

The following section details the measurements resulting from the data collection and employed in the data analysis. As depicted in *Figure* 3, the Group Relative Learning Gain (RLG) and Group artifact performance score rate teams' performance from each DT stage. These measurements served to answer RQ1. Affective aspects of SSRL were assessed through Joint Engagement (JEng), Joint Stress (JStr), and Joint Emotional States (JES), while Joint Visual Attention (JVA) represented the behavioral dimension. This subset of MM measurements was used with performance scores to address RQ2.

3.5.1 Group Relative Learning Gain (RLG). The Relative Learning Gain (RLG) quantitatively evaluates how much students had learned during the workshop. Specifically, it measures the relative improvement in students' knowledge (*Empathize*) and skills (*Prototyping*) from the pre-knowledge test to the post-test [65, 72, 96]. We calculated the individual RLG using the following formula, as proposed by Sangin et al., for pre- and post-knowledge tests using the questions that correspond to each of the mentioned stages [75]. We computed the RLG, which is considered a more meaningful measure than absolute learning, which does not account for prior knowledge. To compute the RLG at the team level, we aggregated the individual RLG scores and then calculated their mean average. Thus, the RLG was used to measure the *Empathize* performance

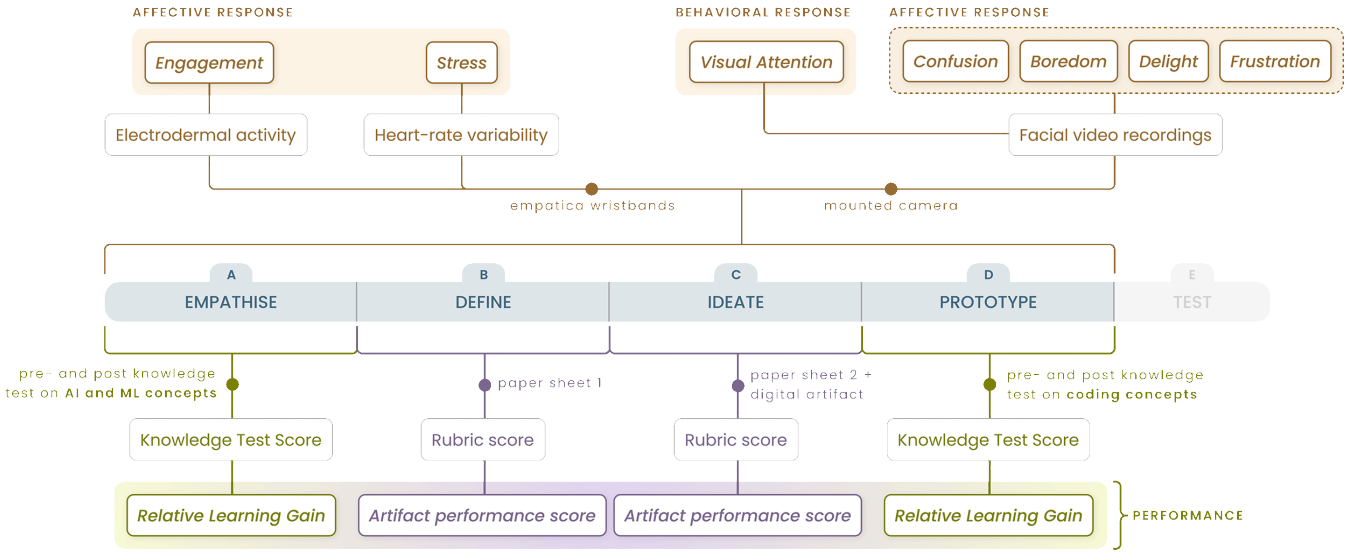


Figure 3: Workshop stages in connection with data collection instruments and measurements

score (range from 0-1) and the *Prototyping* performance score (range from 0-1).

$$RLG = \begin{cases} \frac{post\text{-}test - pre\text{-}test}{\text{Max } pre\text{-}test - pre\text{-}test}, & \text{if } post\text{-}test \geq pre\text{-}test \\ \frac{post\text{-}test - pre\text{-}test}{\text{pre-test}}, & \text{if } post\text{-}test < pre\text{-}test \end{cases}$$

3.5.2 Group artifact performance score: A quality score was assigned to three artifacts created by the teams during the workshop. Paper Sheet 1, produced in the *Define* stage, was the first artifact. The second and third artifacts, Paper Sheet 2 and the digital database in *SorBET*, were developed during the *Ideate* stage. Researchers with expertise in DT and computational thinking evaluated these artifacts using custom-designed rubrics (see Appendix *Tables* 7, 8, 9). These rubrics, developed in collaboration with four educational experts, ensured both relevance and accuracy. Key evaluation criteria included the elaboration, feasibility, and variety of solutions proposed by teams. The scores of each rubric, represent the performance scores of the *Define* (range from 0-15) and *Ideate* (range from 0-27) stages.

3.5.3 Joint Stress (JStr) High and Low: We derived physiological stress levels based on HRV recorded by the wristband. As a well-established physiological marker of stress, HRV reflects fluctuations in inter-beat intervals, with decreased variability commonly indicating heightened sympathetic activation and reduced parasympathetic regulation [36, 42, 78]. After determining stress levels for each team member, we sorted the data in ascending order and categorized the time series into high and low stress. Our data follows a normal distribution. We did not observe a bimodal distribution. Therefore, a median split was deemed unsuitable, as values near the 49th and 51st percentiles would be almost identical. To better distinguish high and low stress, the 33rd and 66th percentiles were used as thresholds. Time series above the 66th percentile were classified as high stress, while those below the 33rd percentile were categorized as low stress. To determine joint stress high (JStr-High) and joint stress low (JStr-Low), cross-recurrence analysis [86] was applied to the high- and low-stress time series for all team members (see *Figure* 4).

3.5.4 Joint Engagement (JEng) High and Low: We retrieved EDA data from wristbands to measure engagement, as it reflects sympathetic nervous system activity, with higher electrodermal responses indicating increased investment and enhanced information processing [7, 13, 30, 95]. This study focused on the mean phasic EDA signal component, whose rapid fluctuations indicate physiological arousal, following the method outlined by Di Lascio et al. [13]. As for stress, once we retrieved the individual engagement, we divided its time series into high- and low-engagement segments. Again, we set the 66th percentile as the threshold for high-engagement and the 33rd percentile for low-engagement. We then computed the cross-recurrence [86] of the high engagement time series for team members as a measure of joint engagement high (JEng-High). Similarly, the cross-recurrence [86] of the low engagement time series for team members served as a measure of low joint engagement (JEng-Low).

3.5.5 Joint Visual Attention (JVA):. JVA is defined as the proportion of time that students within the same team spend gazing at the same group of elements within a set time window. We follow these steps to compute the JVA from the recorded videos of students collaborating in dyads and triads: *a)* faces were detected in each frame of the video, *b)* for each detected face, its gaze direction was calculated using the facial image in the frame. This involved generating a 3D gaze vector pointing toward the screen (see visual representation in *Figure* 5a) with the software *OpenFace* [2, 4], *c)* the exact location of where the students are looking is the intersection point between the laptop screen plane and the 3D gaze vector. We referred to the location as "gaze-point", *d)* Once the gaze points for all team members were identified, we divided the laptop screen space into

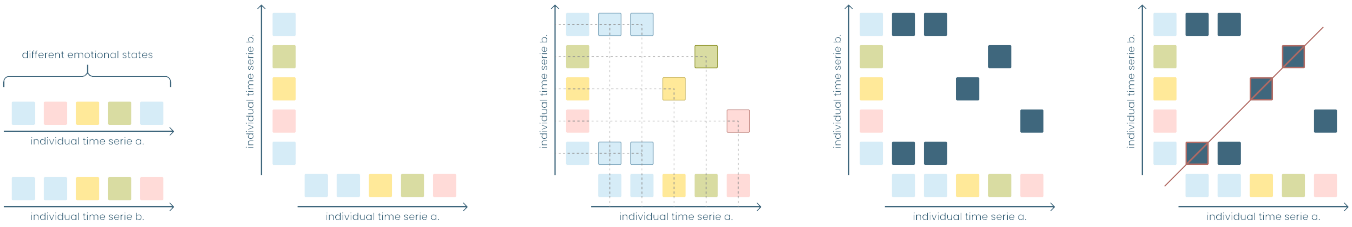


Figure 4: Cross-recurrent matrix to plot affective and behavioral synchrony. Adapted from Sharma et al. [86]. The two series are the two dimensions of the matrix. The red diagonal illustrates a perfect synchrony of states. This bi-dimensional matrix (for dyads) can be extended to three dimensions for triads.

20 rectangular regions, arranged in a four per five grid (*Figure* 5b). Each gaze-point was assigned to one region, and *e)* once again, we calculated the cross-recurrence [86] among students present in the video to determine their JVA. To ensure fair comparisons, we normalized the results based on the group size. We normalized using the chance level proportions of two people looking at the same region out of 20 at the same time and three people looking at the same region out of 20 at the same time. The probability of the first event is 1/20, while the probability of the second event is 1/320; therefore, the ratio of these values is used to normalize the group size.

3.5.6 Joint Emotional State (JES) Boredom, Confusion, Frustration, Delight: JES is defined as the proportion of time that students within the same team spend in a specific emotional state in the same set time window. We focused on four emotional states, frustration, boredom, confusion, and delight, which are identified as the most prominent across various studies [15, 17]. These states are also grounded in the affective subset outlined by D'Mello and Graesser in their "*Cognitive Disequilibrium Theory*" [18]. To identify emotions, we used Action Units (AUs) from the *Facial Action Coding System*, an anatomy-based framework [29] (*Figure* 4). The combination of AUs defines a specific emotion (*Table* 1). Moreover, the AUs' overall intensity reflects each detected emotion's prominence.

We describe the process to retrieve the JES for each team with these passages: *a)* faces were detected in each frame of the video, *b)* we followed the method described in Sharma et al. [87] to ensure consistency: we matched the same face in every frame of the video and assigned a unique ID to it, *c)* After assigning correct IDs to the faces, we used *OpenFace* [2, 4] to extract the AUs [56] for each frame, *d)* using the AUs, we calculated the proportions of the four emotions, delight, frustration, boredom, and confusion, over a fixed ten-second time window [86]. We used a generalized additive model to integrate the AUs and compute the expressions [54], and *e)* with the calculated proportions, we assessed how aligned the emotional states of the team members were by using the cosine similarity between the emotional probabilities of each student. A high cosine similarity indicates shared emotional experiences, while a low similarity suggests divergent emotional responses.

3.6 Data Analysis

3.6.1 Checking for Potential Confounds. As a first step, we check for potential confounding variables. An independent T-test (i.e., two-sample T-test) compared the means of the two kinds of teams,

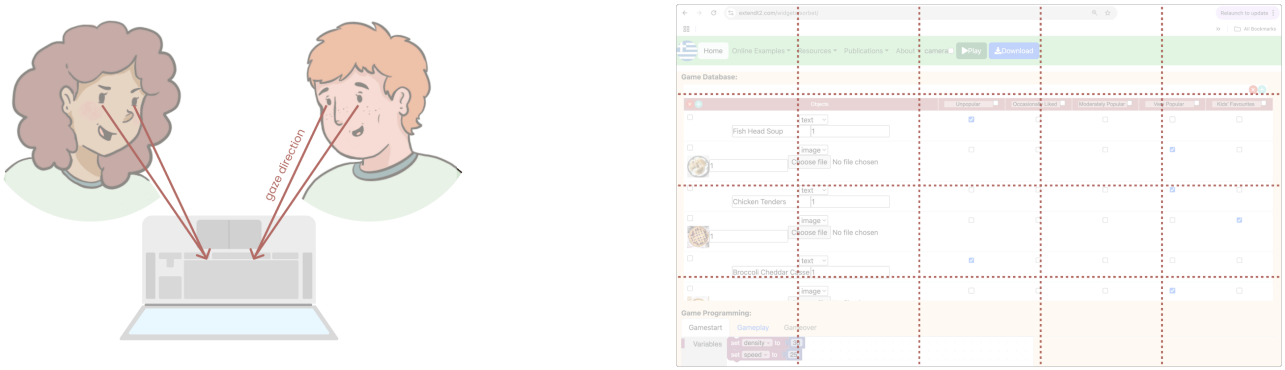
Emotions	Combination of AUs
Boredom	AU4, AU7, AU12
Frustration	AU12, AU43
Confusion	AU1, AU4, AU7, AU12
Delight	AU4, AU7, AU12, AU25, AU26

Table 1: AUs combinations grounded in the *"Cognitive Disequilibrium Theory"* [18]

namely dyads and triads, engaged in the *Empathize*, *Define*, *Ideate*, and *Prototype* stages of the DT workshop. This analysis was intended to eliminate group size (dyads vs. triads) as a confound in our further analyses, ensuring that any differences observed in performance were not due to the size of the group. Also, we conducted a T-test to check for any gender-related confounds, while age-related confounds were examined using a Pearson correlation analysis. Across all measurements, there was no evidence of gender or age-related confounds, either at the group or individual level.

3.6.2 Correlation Between Performances in Different DT Stages. Following this, a Pearson correlation analysis was conducted to examine the relationships between performance scores across the different stages of the DT workshop: *Empathize*, *Define*, *Ideate*, and *Prototyping*. This analysis measures the linear relationship between two variables, represented by a correlation coefficient (r) and a p-value. Positive values signify a direct correlation, while negative values suggest an inverse relationship. The p-value determines whether the observed relationship is statistically significant, with values greater than 0.05 indicating a lack of statistical significance. This analysis aimed to identify which stages were more closely connected in terms of performance and understand how progress in one stage may influence outcomes in the following ones.

3.6.3 Correlation of Performance and Affective/Behavioral Responses. Next, we performed another Pearson correlation analysis to explore the relationship between each joint MM affective (i.e., JEng, JStr, and JESs) and behavioral measurement (i.e., JVA) and the performance scores from each DT stage. This determined the patterns of association between affective and behavioral responses and students' performance during the workshop. Overall, we are using



(a) Representation of students' gaze direction

(b) Visual attention grid with 20 regions

Figure 5: Passages for Joint Visual Attention calculation.

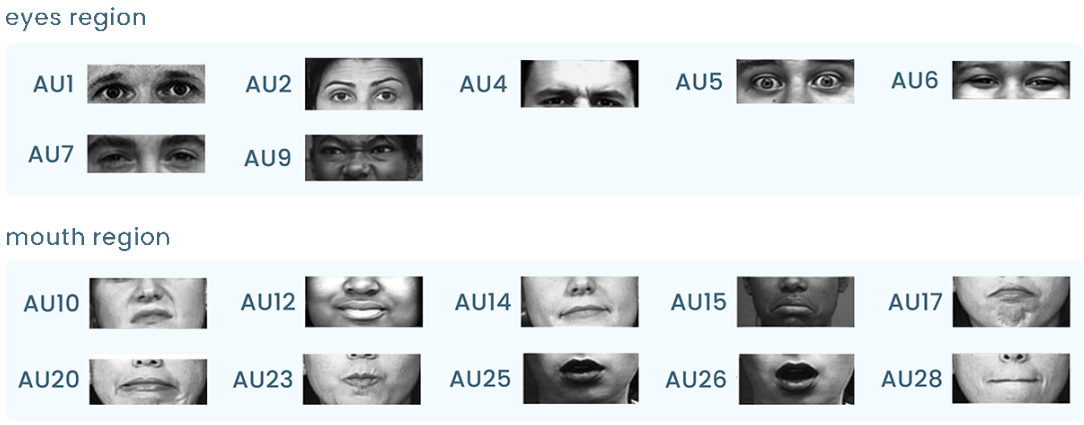


Figure 6: AUs from the *Facial Action Coding System* [29]

Pearson correlation for the following reasons: *a)* there is no theoretical or empirical basis for hypothesizing non-linear relationships between the MM measurements and performance variables in this study. *b)* MM measurements are temporal, while performance scores were collected at single points during each DT stage rather than continuously; this setup does not allow for cross-correlation or cross-recurrence analysis. *c)* our data follows a normal distribution (as confirmed by the Shapiro-Wilk test), the measurements are continuous rather than ordinal, and we did not observe outliers. The Bonferroni correction was applied to adjust p-values for multiple comparisons, accounting for the increased risk of false positives by modifying the threshold for statistical significance.

4 Results

4.1 Correlations Between Group Performances in DT Stages

Our analysis shows positive correlations between the *Empathize* stage and the *Ideate* (r = 0.51, p = 0.001) and *Prototyping* (r = 0.71, p = <0.0001) stages, both of which are statistically significant. The correlation coefficient between the *Define* stage and the *Ideate* stage performance scores suggests a positive correlation (r = 0.62, p-value = 0.0002), which also reaches statistical significance after the correction. Conversely, the correlations between the *Empathize* and *Define* stages (r = 0.24, p = 0.19), the *Define* and *Prototyping* stages (r = 0.17, p = 0.37), and the *Ideate* and *Prototyping* stages (r = 0.33, p = 0.07) are not statistically significant. *Table* 2 presents the correlation coefficients r and p-values between stages.

4.2 Correlations Between MM Measurements and DT Stages Group Performances

4.2.1 Multi-Modal Measurements and the Empathize stage. We started with computing correlations between the *Empathize* stage and the MM measurements. The results (see *Table* 3) indicate a significant positive correlation between the *Empathize* stage score and JVA (r = 0.47, p = 0.009). Conversely, significant negative correlations were observed with JES-Boredom (r = -0.53, p = 0.002) and JStr-High (r = -0.40, p = 0.02). No significant correlations (p-value < 0.05) were found between the *Empathize* stage score and the remaining affective states: JES-Confusion, JES-Frustration, JES-Delight, JEng-High, JEng-Low, and JStr-Low.

Performance score Mean (SD)	Empathize (r, *p*-value, *p*-uncorr)	Define (r, *p*-value, *p*-uncorr)	Ideate (r, *p*-value, *p*-uncorr)	Prototyping (r, *p*-value, *p*-uncorr)
Empathize 0.50 (0.15)	/	/	/	/
Define 9.86 (3.74)	0.24, 0.19, 0.03	/	/	/
Ideate 16.79 (5.03)	0.51, **0.001**, 0.0001	0.62, **0.0002**, 0.00003	/	/
Prototyping 0.60 (0.11)	0.71, **<0.0001**, <0.00001	0.17, 0.37, 0.07	0.30, 0.11, 0.01	/

Table 2: Correlation coefficients *r* and *p*-values between the DT stages, alongside performance scores (mean, SD). Significant correlations are in bold. The last numbers show the *p*-values before the Bonferroni correction. We used $\alpha_{critical} = 1 - (1 - \alpha_{altered})^k$ with k = 6.

Measurement	**Mean**	**SD**	**Coefficient *r***	***p*-value**	***p*-uncorr**
JVA	**0.49**	**0.33**	**0.47**	**0.009**	**0.001**
JES - Boredom	**0.41**	**0.28**	**-0.53**	**0.002**	**0.0002**
JES - Confusion	0.53	0.34	-0.22	0.24	0.03
JES - Frustration	0.42	0.26	-0.10	0.58	0.09
JES - Delight	0.57	0.25	0.05	0.77	0.15
JEng - High	0.50	0.27	-0.11	0.56	0.08
JEng - Low	0.45	0.29	-0.20	0.27	0.03
JStr - High	**0.52**	**0.33**	**-0.40**	**0.02**	**0.002**
JStr - Low	0.52	0.28	0.23	0.21	0.02

Table 3: Correlation of Multi-Modal Measurements with the *Empathize* stage scores. Significant correlations are in bold. The last column shows the p values before the Bonferroni correction. We used $\alpha_{critical} = 1 - (1 - \alpha_{altered})^k$ with k = 9.

4.2.2 Multi-Modal Measurements and the Define stage. Next, we computed the correlation between the MM measurements and the scores from the *Define* stage. From the Table 4, we observe a significantly positive correlation with JVA (r = 0.41, p = 0.02) and JES-Frustration (r = 0.39, p = 0.03). On the contrary, significantly negative correlations are found with JES-Boredom (r = -0.46, p = 0.01), JEng-Low (r = -0.58, p = 0.0009), and JStr-High (r = -0.54, p = 0.002). Finally, there is no correlation (p-value < 0.05) between JES-Confusion, JES-Delight, JEng-High, and JStr-Low and the performance score in the *Define* stage.

4.2.3 Multi-Modal Measurements and the Ideate stage. Proceeding in the same manner, we computed the correlation between the MM measurements and the performance scores from the *Ideate* stage. Referring to *Table* 5, we observed positive correlations between the *Ideate* stage score and JVA (r = 0.59, p = 0.0007), JES-Confusion (r = 0.51, p = 0.004), JEng-High (r = 0.46, p = 0.01), and JStr-Low (r = 0.45, p = 0.01). Conversely, the *Ideate* stage score was negatively correlated with JES-Boredom (r = -0.44, p = 0.01), JES-Frustration (r = -0.48, p = 0.007), and JStr-High (r = -0.65, p = <0.0001). Finally, no statistically meaningful correlation (p-value > 0.05) was found between the *Ideate* stage score and JES-Delight or JEng-Low.

4.2.4 Multi-modal Measurements and the Prototyping stage. Finally, following the same process, we computed the correlation between the MM measurements and the performance score from the *Prototyping* stage. Our results (*Table* 6) indicate that the performance score in the *Prototyping* stage is positively correlated with JVA (r = 0.58, p = 0.0008), JEng-High (r = 0.40, p = 0.03), and JStr-Low (r = 0.38, p = 0.04). Conversely, it is negatively correlated with JES-Boredom (r = -0.50, p = 0.005), JES-Frustration (r = -0.51, p = 0.002), and JStr-High (r = -0.55, p = 0.001). JES-Confusion, JES-Delight, and JEng-Low do not show any statistically significant correlation (a p-value > 0.05) with *Prototyping* performance score.

Measurement	Mean	SD	Coefficient *r*	*p*-value	*p*-uncorr
JVA	**0.48**	**0.25**	**0.41**	**0.02**	**0.002**
JES - Boredom	**0.51**	**0.29**	**-0.46**	**0.01**	**0.001**
JES - Confusion	0.43	0.28	-0.12	0.52	0.07
JES - Frustration	**0.38**	**0.25**	**0.39**	**0.03**	**0.003**
JES - Delight	0.51	0.28	-0.06	0.75	0.14
JEng - High	0.56	0.30	0.25	0.18	0.02
JEng - Low	**0.50**	**0.26**	**-0.58**	**0.0009**	**0.0001**
JStr - High	**0.52**	**0.31**	**-0.54**	**0.002**	**0.0002**
JStr - Low	0.50	0.31	0.25	0.18	0.02

Table 4: Correlation of Multi-Modal Measurements with the *Define* stage scores. Significant correlations are in bold. The last column shows the p values before the Bonferroni correction. We used $\alpha_{critical} = 1 - (1 - \alpha_{altered})^k$ with k = 9.

Measurement	Mean	SD	Coefficient *r*	*p*-value	*p*-uncorr
JVA	**0.50**	**0.30**	**0.59**	**0.0007**	**0.00007**
JES - Boredom	**0.49**	**0.29**	**-0.44**	**0.01**	**0.001**
JES - Confusion	**0.45**	**0.29**	**0.51**	**0.004**	**0.0004**
JES - Frustration	**0.48**	**0.28**	**-0.48**	**0.007**	**0.0007**
JES - Delight	0.40	0.29	0.29	0.12	0.01
JEng - High	**0.44**	**0.26**	**0.46**	**0.01**	**0.001**
JEng - Low	0.53	0.29	0.23	0.21	0.02
JStr - High	**0.58**	**0.30**	**-0.65**	**<0.0001**	**<0.00001**
JStr - Low	**0.52**	**0.28**	**0.45**	**0.01**	**0.001**

Table 5: Correlation of Multi-Modal Measurements with the *Ideate* stage scores. Significant correlations are in bold. The last column shows the p values before the Bonferroni correction. We used $\alpha_{critical} = 1 - (1 - \alpha_{altered})^k$ with k = 9.

4.2.5 Summarization of all correlations. Finally, we present the overall relationships between students' affective (i.e., JEng, JStr, and JESs) and behavioral (i.e., JVA) responses and their experience outcomes across the DT stages. *Figure* 7 provides a visual summary of the correlations between MM measurements and the performance scores across the four stages: *Empathize*, *Define*, *Ideate*, and *Prototyping*. Upward arrows (↑) represent positive correlations, indicating that an increase in the MM measurement is associated with higher performance scores. Downward arrows (↓) represent negative correlations, indicating that an increase in the measurement corresponds to lower performance scores. Blank cells indicate that these MM measurements showed no significant correlation. JVA remains consistently positive across all stages, while JES-Boredom and JStr-High consistently show negative correlations. JES-Confusion is specific to the *Ideate* stage, where it correlates positively, while JEng-Low appears only in the *Define* stage with a negative correlation. JEng-High and JStr-Low show positive correlations that become more evident in later stages. Finally, JES-Frustration correlates positively during *Define* but shifts to a negative correlation in the subsequent stages.

5 Discussion

In this paper, we presented a study conducted "in-the-wild" where students engaged in a DT workshop focused on AI and ML. Through tests, artifacts, and MM Learning Analytics data, we captured students' affective and behavioral responses to uncover the underlying team dynamics during the open-ended experience for a new perspective on the CCI field. Two main points emerged from our results: *a)* an analysis of the correlations between different DT stages and exhibited patterns with performance scores, and *b)* the relationship between students' performance scores and their observed behavioral and affective states. Beyond contextualizing these results, the discussion examines the feasibility of DT curricula for tackling "wicked problems", as the ones related to AI and ML for insights into orchestrating student-centered learning activities with appropriate scaffolding and guidance.

Measurement	Mean	SD	Coefficient *r*	*p*-value	*p*-uncorr
JVA	**0.48**	**0.29**	**0.58**	**0.0008**	**0.00008**
JES - Boredom	**0.49**	**0.31**	**-0.50**	**0.005**	**0.0005**
JES - Confusion	0.43	0.32	0.12	0.51	0.076
JES - Frustration	**0.47**	**0.30**	**-0.51**	**0.002**	**0.0002**
JES - Delight	0.53	0.31	0.21	0.27	0.03
JEng - High	**0.48**	**0.32**	**0.40**	**0.03**	**0.003**
JEng - Low	**0.47**	**0.32**	0.11	0.53	0.08
JStr - High	**0.54**	**0.32**	**-0.55**	**0.001**	**0.0001**
JStr - Low	**0.46**	**0.32**	**0.38**	**0.04**	**0.004**

Table 6: Correlation of Multi-Modal Measurements with the *Prototyping* stage scores. Significant correlations are in bold. The last column shows the p values before the Bonferroni correction. We used $\alpha_{critical} = 1 - (1 - \alpha_{altered})^k$ with k = 9.

5.1 Interpretation of results

5.1.1 In respect to RQ1: To answer our RQ1, we assessed the performance from each stage of DT (*Empathize*, *Define*, *Ideate*, and *Prototype*) with measurable outcomes to identify correlations in team performances across the DT workshop. The findings illustrate that early-stage performance, whether high or low, is related to patterns of success or difficulty in later stages. Notable correlations were observed between *Empathize* → *Ideate*, *Empathize* → *Prototype*, and *Define* → *Ideate*. Among these, the correlation between *Define* and *Ideate* is strong and positive, indicating that progress in the *Define* stage displays a relationship with achievement in the *Ideate* stage. This finding suggests that a clear problem definition in the *Define* stage may support more effective brainstorming and solution generation in the *Ideate* stage. In the context of high performance, we align with Sun et al., who state that the discussion of pertinent ideas can be leveraged to improve the downstream success [93]. In our DT workshop, this link may be attributed to the consistency of support materials, such as paper-based worksheets, which facilitated a smoother transition. These materials enable students to carry situational understandings developed during the *Define* stage into the *Ideate* stage. Specifically, the situational interpretation of the task, including challenges appraisal [66, 77], may have been formed during the *Define* stage and seamlessly applied in the subsequent *Ideate* stage. Finally, *Empathize*, strongly correlates with *Prototyping*, which could initially seem counterintuitive, given that these stages took place at opposite ends of the DT process. Nonetheless, during the *Empathize* stage, students engaged in brainstorming activities aimed at raising their peers' awareness of AI through the game. Therefore, a good conceptualization could facilitate block-based programming development, helping shape the decisions made during the gameplay design.

5.1.2 In respect to RQ2: Building on the performances observed for RQ1, we analyzed their connection with students' behavioral and affective states to address RQ2, showing that certain states exhibit more consistent correlations throughout the DT stages than others. More specifically, JVA, JES-Boredom, and JStr-High characterize conditions that, once experienced by group members collectively, tend to reoccur throughout the workshop. JVA presents a significant positive correlation throughout the workshop, displaying potential as a proxy for collaboration and shared understanding [85]. In our DT workshop, collaborative tools (that is, paper-based sheets and laptops) were not individualized but shared among group members. This may have encouraged the redirection of visual attention towards a common focus. In contrast, joint states of boredom and high stress among members are associated with low performance. The adverse effects of boredom on learning activities are well documented in the literature [16, 68, 69]. According to these theories, boredom is best kept as a short-lived "mood" ideally confined to a single task, as it can be difficult to reverse once it becomes prolonged [3]. In SSRL, boredom is an emotion with rapid contagion among learners working on the same task, especially because of distinctive externalization [32]. This phenomenon likely occurred in our case, where protracted boredom extended beyond individual DT stages, dampening the overall workshop experience and calling for reappraisal strategies [91]. Notably, boredom can persist even when interactions vary, such as through the eventual introduction of a digital tool. Concentrating on stress, while it can sometimes function as an activating emotion [64], tends to impair performance when it becomes chronic [83]. Our findings show that elevated stress levels are increasingly correlated with poor performance as the workshop advances, peaking in *Ideate*. This is close to similar entries, where stress is a disturbance in the learning process [64]. Stress disrupted the collaborative process by amplifying negative emotions within groups. The introduction of a digital tool, namely *SorBET*, coincided with increased negative group dynamics, suggesting a potential association with heightened stress levels that made recovery more difficult.

JES-Confusion, JEng-High, and JStr-Low show up as positive drivers during the last DT stages (i.e., *Ideate* and *Prototype*). In line with the *cognitive disequilibrium* theory, we report that confusion during ideation may stimulate deeper thinking or problem-solving efforts [19, 50]. Engaging in activities addressing AI and ML "wicked problems" frequently mismatches with students' existing knowledge. This difficulty is further intensified by potential differences

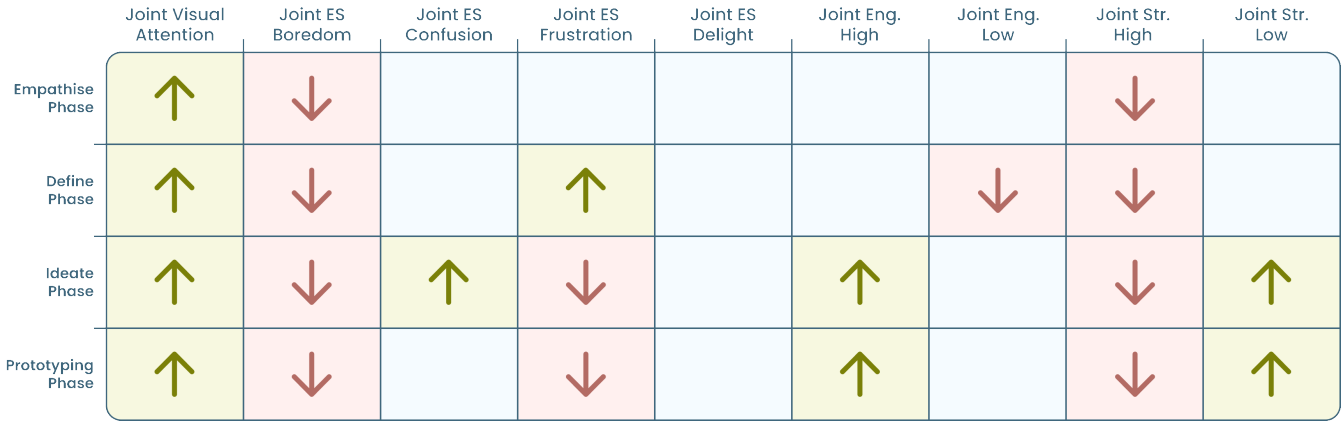


Figure 7: Positive and negative correlations between MM measurements and the *Empathize*, *Define*, *Ideate*, and *Prototyping* performance scores. Upward, green arrows (↑) represent positive correlations, and downward, red arrows (↓) represent negative correlations. Black cells indicate the absence of any correlation.

in background among team members: if confusion persists after several unsuccessful attempts to, for example, devise a solution, student engagement becomes at risk [19]. However, goal congruence among peers breaks through the impasse [87]. As the workshop progresses, increased engagement among group members is associated with higher productivity, supporting the notion that collaborative effort enhances idea development and its implementation within the programming platform [94]. Moreover, a state of low stress appears crucial during the latest stages. This reflects a positive team appraisal of the task, most likely as a manageable challenge [66, 77]. Previous familiarization with the workshop scope aligns with students framing their goals as achievable and reporting a sense of success in their addressing [91]. Finally, JEng-Low negatively correlates with *Define*, likely due to the need for sustained effort. Incomplete input from members is likely to hinder the *Define* stage, preventing the establishment of a solid foundation. Here, struggles with workload distribution could impede the team from advancing in the task. All the other stages show no correlation, suggesting less dependency on maintained engagement. The *Empathize* stage benefits from diverse contributions, even if they are sporadic or unstructured. Likewise, the creative and exploratory nature of *Ideate* and *Prototyping* allows for spontaneous insight and trial-and-error, making them more adaptable to varying levels of engagement.

Interestingly, we see JES-Frustration playing a dual role across the stages. In *Define*, frustration positively correlates with performance, indicating that mild manifestations of this state can be interpreted as triggers that push learners to clarify goals, refine problems, or persist through challenges at the early stages of the DT workshop [3, 33]. However, as the workshop progresses with *Ideate* and *Prototyping*, the presence of a joint state of frustration negatively correlates with production quality. Thus, frustration can serve a formative purpose in supporting the preliminary design of solutions in DT, particularly when sequenced with engagement. Nonetheless, it requires intervention in later stages if it disrupts the cognitive flow [3, 16], such as during an impasse caused by, for instance, prolonged disagreement over unsolved programming bugs [87]. Intelligent support systems or tailored supervision can facilitate the reappraisal of challenging situations and coping strategies to sustain performance throughout the process [49, 91]. For instance, early intervention in response to high stress experienced by one team member could help dissipate the sentiment and prevent escalation into complete disengagement and the spread of frustration among peers [19]. With the *cognitive disequilibrium* theory, we saw how activating negative emotions (i.e., stress, frustration) can sharpen analytical processes. However, if left unchecked, these emotions might induce more rigid communication, causing learners to become entrenched in their perspectives or even lead to avoidance behaviors, such as disengagement or reluctance to participate in further discussions [18, 91].

5.2 Implications

5.2.1 Implication for research. The process applied in the proposed study is built on theories that stress learning processes over the sole outcomes for a successful knowledge acquisition [18, 28, 62]. Against this background, the analysis of SSRL through MM data equipped a lens to examine the "hidden" mechanisms [11, 81, 105] that characterize students' experiences. The analysis dissects such experiences into finer, more understandable components, namely affective and behavioral factors, and outlines their impact. Our study serves as a feasibility example in reducing assessment approximation for collaborative, complex interactions, especially when they occur during open-ended activities for CS concepts (e.g., AI and ML) with diversified learning supports (i.e., paper-based and digital) [59]. Given the variation in tasks across stages, we also examined students' learning outcomes at each stage, incorporating a temporal dimension of sequentiality. This approach allowed us to explore how performance in one stage relates to the others. For example, we introduce computational thinking during the *Prototyping* stage but do not study it in isolation [72, 105]. Instead, we analyze it as an integral component of the entire learning experience. Even if our research focuses on DT, the lessons learned can resonate with

all the constructivist formats that value collaboration, problem-solving, and creativity [59]. Moreover, as students engaged in an activity hosted on a laptop with a programming module, our findings hold relevance in CSCL and CCI environments, where MM data collection can harness the existing educational technology infrastructure in classrooms [46, 104]. The intersection between SSRL and CSCL, when examined through the lens of MMLA, can push the boundaries of understanding learning in CCI. Through this, novel perspectives on students' journeys can be accounted for and can foster tailored interventions and support systems that are as flexible and dynamic as the constructionist activities they engage in [89].

Our study builds on existing theoretical frameworks of learners' responses, such as the D'Mello & Graesser model on affective dynamics [18]. While these models are foundational for understanding emotional drivers in learning environments, they may benefit from further expansion. This is quite relevant now, as MM tools have the potential to analyze a vast range of emotional, cognitive, and behavioral responses by tracking various signals. However, organizing these responses into predefined groups, as established by existing theories, may fail to capture learners' intricate and nuanced experiences exhaustively. Incorporating affective interdependencies [32, 72], or a time-sensitive map of collaborative phenomena (e.g., moments of appraisal, coping, etc.), could provide a more nuanced view of how emotions and behaviors interact in the context of learning, updating existing theoretical standards.

5.2.2 Implications for practice. Our results offer practical implications for curriculum development and the design of children's interactive learning experiences. We examined the correlations between performances across the different DT stages to understand their influence on one another. While some significant correlations were observed, the strongest relation was performance in the *Define* stage with performance in the *Ideate* stage. Contextualizing this within the proposed workshop, these two stages used the same materials, even though students were tasked with different objectives in each. This finding can orient the activity design, as it reports on the benefits of maintaining consistency in materials used throughout the process. At the same time, it highlights the cognitive load MM interactions impose on students. A shift in support tools, such as transitioning from paper-based resources to a block-based coding platform, may impact outcome efficiency. However, we observed that while introducing a new format can initially confuse, it may ultimately contribute to knowledge acquisition [19]. In the context of teaching computer science-related content like AI and ML, where concepts are often introduced unplugged and later approached with digital tools [53], this finding could help reconsider the facilitation style and scaffolding to manage negative sentiments [59].

Concerning the patterns of behavioral and affective states, JVA, JES-Boredom, and JStr-High are three variables that appear to be central to the performance and should be key targets for intervention planning to improve overall outcomes. For example, facilitating activities that further leverage team focus and coordination in introductory phases (e.g., structured brainstorming during *Empathize*) could equip teams with *"virtuous circles"* of feedback [102] later applicable as the experience unfolds [46]. The use of shared tools, or tools having collaborative entry points, likely amplified the group's collective focus, reinforcing mutual appraisal as group members worked together to achieve common outcomes with greater coherence [66, 91]. The persistence of boredom and high stress is an important marker for improving instructional design and facilitation techniques. Introducing variety, such as incorporating digital tools, may still fail to alleviate boredom if the root causes, such as a lack of challenges or repetitive tasks, are not addressed. Beyond thoughtful curriculum design and effective facilitation, auxiliary devices and the use of auxiliary devices (e.g., teacher-facing dashboards for classroom monitoring) can help detect the onset of these negative states. This enables timely interventions to mitigate their impact or, ideally, to predict and prevent them altogether [80]. While such states cannot be entirely eliminated and, in some cases, may even play a constructive role in knowledge acquisition, proper moderation is critical [73]. For example, high-stress, when combined with engagement and confusion, can signal a peak in task investment. Prolonged stress or boredom, however, can delay advancement. Strategies to maintain engagement, such as introducing variety, setting intermediate milestones, or providing stimulating tasks, can help counteract this. Stress management interventions, such as breaks, mindfulness exercises, or workload adjustments, should be integrated into the workflow. Moreover, facilitators should be equipped with real-time behavioral and affective indicators from students to address these states proactively, ensuring they do not persist in a way that impairs learning [80].

Moreover, affective and behavioral variables exhibit different relationships with success at each stage, indicating that phase-specific orchestration may be beneficial. Confusion has a positive correlation with performance in *Ideate*. Ideally, the presence of confusion in *Ideate* should be extended across phases to trigger students to question assumptions and generate innovative solutions consistently [19], especially in topics such as the one proposed, where teams faced open-ended tasks about AI. In fact, promoting uncertainty as something to embrace while also offering guidance to navigate through confusion might optimize performance from the early stages [3]. Low-stress states pair with high performance in the second half of the activity, where topics' understanding and goals are likely to be solidified. To sustain a more relaxed environment, support from the facilitator is important to mitigate a laid-back atmosphere and alleviate students' concerns. Interestingly, a high level of engagement is only linked to improved performance towards the end of the activity. This may indicate that having overly structured beginnings can reduce students' ability to make independent decisions, as they become overly reliant on guided instructional approaches. Moreover, the content to elaborate on should be diversified to match participants' interests well before the *Define* stage. Fluctuations in shared low engagement and performance appear more pronounced in tasks with a narrower focus [64]. In our case, this may be related to the instructions provided to redirect or refine team practice. Yet, its observed pattern with lower performance suggests that a rethinking of *Define* modules may be needed for the promotion of inclusive ideation and prototyping processes that invite contributions from all participants, regardless of their perceived levels of engagement.

Overall, the presence of specific affective and behavioral states shared at the group level, is significantly associated with performance outcomes. In the context of a DT workshop, emotional contagion [99] and appraisal [77] consider not only peers as influencing variables but also the *"object-focus"*, the created artifact, [66] toward which the emotion is directed. Artifacts play an important role in DT, and their design process can be used to act as mediators between group members, encouraging positive attitudes and discouraging disruptive ones. For instance, administered materials carefully designed for the scope can facilitate challenges re-appraisals (e.g., embedded prompts, tips) or joint agreement exercises, relieving negative sentiments spreading and/or escalation due to peer social conformity [66].

Lastly, using MM data to capture students' unfiltered and spontaneous states allows for accurate reports on regulation dynamics, such as emotional exchange within teams (e.g., contagion degree of boredom) and coping mechanisms (e.g., successful resolution of confusion) [66, 99]. This approach reduces the skew caused by social desirability bias (e.g., overstating positive experiences), which might occur when students are aware that their actions are being evaluated or when they are explicitly asked to reflect on their emotions and behaviors [49]. Wristbands and mounted cameras provided a non-intrusive method for collecting data that ensured equal coverage of all participants. This approach can be particularly beneficial for capturing information from students who might display non-verbal disposition or have difficulty actively articulating their emotions [49], thereby offering a more inclusive understanding and tailored assistance [81].

5.2.3 Ethical Considerations. Our study provides some reflections arising from using MM instruments with ethical integrity for "in-the-wild" studies with K-12 students. Considerations regarding those instruments brought to the classroom should be made in advance since they may impact its ecosystem [82]. For example, wristbands require direct contact with the students and can possibly distract from the experience or create discomfort as a wearable. Considering this, students should be informed, according to their age, about their use. Cameras, even if passive collectors of data, are still conspicuous in their presence and may alter student's natural behavior, potentially triggering a sense of being under constant "surveillance" (e.g., *'Does the teacher see what I am doing?"*) [11]. Practitioners in the field advocate to address these concerns [10, 11, 82]. Informed, we took steps starting from ensuring information privacy and transparency in the process to prevent negative responses [48, 87]. Moreover, students' individual preferences and needs should be accounted for by establishing dialogue and mutual trust. Recognizing students' willingness should involve a clear explanation of procedures and tools employed to alleviate concerns and indulge their curiosity [20]. Informing student about their data is another step to take. *"Where is my test ending up?"* and other questions must be followed up to correctly inform students on the scope of data collection, empowering them to participate confidently and with awareness of how their contributions will be used.

6 Limitations and Future Work

We discussed our findings, in light of the promise of MM data and LA as a window into students' collaborative experiences in hands-on contexts. However, to identify further opportunities for advancement, we must acknowledge the study's limitations. First and foremost, relying solely on quantitative analysis is a shortcoming, given the richness of the instruments used (e.g., paper-based sheets and game artifacts, video recordings). The data generated from these sources should be leveraged through qualitative approaches, such as thematic analysis, for more context-reliant implications. Additionally, while we employed a range of affective and behavioral responses to draw inferences about group-level dynamics, there is considerable room to expand this pool of data. For instance, measures that capture students' joint motivation [44], or joint mental effort [85] can be incorporated or explore different affective states from established frameworks (e.g., achievement-related emotions from the Control-Value Theory [67, 70]). Lastly, because performance is evaluated at discrete points and MM measurements are collected continuously, Pearson correlations cannot account for time-dependent influences or delayed effects between variables. Future work could address this by applying time-series analysis techniques, such as cross-correlation. We also report on the limitations in achieving full accuracy in emotion recognition. Given potential biases in the training libraries (e.g., cultural differences, neurodiversities), the algorithm may misidentify certain groups [40]. It is worth noting that the flexibility of DT can challenge the transferability of our results to larger scales. For instance, our performance assessment (i.e., quantitative analysis on tests and artifacts) may not align seamlessly with alternative DT deployments, which could impact the comparability of results. Future contributions should address a larger pool of students and introduce additional variables, such as fully digitalizing DT stages, to enhance the generalizability of results across different DT deployment styles.

7 Conclusion

We conducted an empirical intervention to unravel the "hidden" dynamics when students collaborate for a DT project centered on AI and ML. Therefore, data were collected from 63 students grouped into 29 teams. We assessed their performance in each DT stage and monitored their interaction responses with MM channels, adding a new layer of interpretation through behavioral and affective states. The results show close connections and notable divergences in performance across DT stages, suggesting that the way students engage with each stage of the process is dynamic and context-dependent. For each stage, students' joint states were mapped and correlated with performance to determine whether these responses displayed patterns with success or potential barriers to quality in their collaborative outcomes. Implications call for specificity in designing DT activities, ensuring that the most productive responses for each stage are harnessed (e.g., using confusion during the Ideate stage) to maximize learning outcomes. Our contribution hopes to leverage the intersection between educational psychology, constructionism, and CSCL to empower researchers, learners, and other stakeholders to make CCI experiences more impactful.

Selection and Participation of Children

Our research prioritized the willingness of children and made every effort to accommodate their preferences, placing their engagement and well-being at the forefront. The participation of children in the

study and data collection was regulated by an information letter and consent form signed by students' parents or guardians. These documents were designed in accordance with the requirements mandated by SIKT and approved by the same system, ensuring full compliance with national regulations and GDPR standards for the gathering, handling, anonymization, and protection of personal data. Detailed information about the research was shared in advance with children, their guardians, and other figures involved (e.g., the teacher and the institution) to facilitate an informed decision about joining the study. Children were able to opt out of the data collection process at any point.

Acknowledgments

We extend our heartfelt appreciation to the children, teachers, and schools whose participation made this research possible and truly meaningful. Funded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This research has received funding from the European Union's Horizon Europe Framework Programme for Research and Innovation under Grant Agreement No. 101060231, *Exten.D.T.2 - Extending Design Thinking with Emerging Digital Technologies*. The author also acknowledges the use of LLM (i.e., GPT-4 and Grammarly) for its assistance in grammar revision and language refinement throughout the writing process.

References

[1] Safinah Ali, Blakeley H Payne, Randi Williams, Hae Won Park, and Cynthia Breazeal. 2019. Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education. In *International workshop on education in artificial intelligence k-12 (eduai'19)*, Vol. 2. mit media lab Palo Alto, California, 1–4.

[2] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* 6 (2016).

[3] Ryan SJd Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (2010), 223–241.

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10. doi:10.1109/WACV.2016.7477553.

[5] Tim Bell and Jan Vahrenhold. 2018. CS unplugged—how is it used, and does it work? *Adventures between lower bounds and higher altitudes: essays dedicated to Juraj Hromkovič on the occasion of his 60th birthday* (2018), 497–521.

[6] Paulo Blikstein and Marcelo Worsley. 2016. Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics* 3, 2 (2016), 220–238.

[7] Ryan Cain and Victor R Lee. 2016. Measuring electrodermal activity to capture engagement in an afterschool maker program. In *Proceedings of the 6th Annual Conference on Creativity and Fabrication in Education*. 78–81.

[8] Alejandro Luis Callara, Laura Sebastiani, Nicola Vanello, Enzo Pasquale Scilingo, and Alberto Greco. 2021. Parasympathetic-sympathetic causal interactions assessed by time-varying multivariate autoregressive modeling of electrodermal activity and heart-rate-variability. *IEEE Transactions on Biomedical Engineering* 68, 10 (2021), 3019–3028.

[9] Lorena Casal-Otero, Alejandro Catala, Carmen Fernández-Morante, Maria Taboada, Beatriz Cebreiro, and Senén Barro. 2023. AI literacy in K-12: a systematic literature review. *International Journal of STEM Education* 10, 1 (2023), 29.

[10] Lucrezia Crescenzi-Lanna. 2020. Multimodal Learning Analytics research with young children: A systematic review. *British Journal of Educational Technology* 51, 5 (2020), 1485–1504.

[11] Mutlu Cukurova, Michail Giannakos, and Roberto Martinez-Maldonado. 2020. The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology* 51, 5 (2020), 1441–1449.

[12] Yun Dai. 2024. Integrating unplugged and plugged activities for holistic AI education: An embodied constructionist pedagogical approach. *Education and Information Technologies* (2024), 1–24.

[13] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement During Lectures Using Electrodermal Activity Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 103.

[14] Dandan Ding, Junchao Tong, and Lingyi Kong. 2020. A deep learning approach for quality enhancement of surveillance video. *Journal of Intelligent Transportation Systems* 24, 3 (2020), 304–314.

[15] Sidney D'Mello and Art Graesser. 2010. Modeling cognitive-affective dynamics with Hidden Markov Models. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 32.

[16] Sidney D'Mello and Art Graesser. 2011. The half-life of cognitive-affective states during complex learning. *Cognition & Emotion* 25, 7 (2011), 1299–1308.

[17] Jason E Dowd, Ives Araujo, and Eric Mazur. 2015. Making sense of confusion: Relating performance, confidence, and self-efficacy to expressions of confusion in an introductory physics class. *Physical Review Special Topics-Physics Education Research* 11, 1 (2015), 010107.

[18] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157.

[19] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction* 29 (2014), 153–170.

[20] Michail Giannakos, Mutlu Cukurova, and Sofia Papavlasopoulou. 2022. Sensor-based analytics in education: Lessons learned from research in multimodal learning analytics. In *The multimodal learning analytics handbook*. Springer, 329–358.

[21] Michail N Giannakos, Kshitij Sharma, Sofia Papavlasopoulou, Ilias O Pappas, and Vassilis Kostakos. 2020. Fitbit for learning: Towards capturing the learning experience using wearable sensing. *International Journal of Human-Computer Studies* 136 (2020), 102384.

[22] Michail N Giannakos, Kshitij Sharma, Ilias O Pappas, Vassilis Kostakos, and Eduardo Velloso. 2019. Multimodal data as a means to understand the learning experience. *International Journal of Information Management* 48 (2019), 108–119.

[23] Arthur C Graesser. 2020. Emotions are the experiential glue of learning environments in the 21st century. *Learning and Instruction* 70 (2020), 101212.

[24] Marianthi Grizioti and Chronis Kynigos. 2024. Integrating computational thinking and data science: The case of modding classification games. *Informatics in Education* 23, 1 (2024), 101–124.

[25] Shuchi Grover. 2024. Teaching AI to K-12 Learners: Lessons, Issues, and Guidance. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 422–428.

[26] Shuchi Grover, Brian Broll, and Derek Babb. 2023. Cybersecurity Education in the Age of AI: Integrating AI Learning into Cybersecurity High School Curricula. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 980–986.

[27] Ludovica Gualniera, Jatinder Singh, Federico Fiori, and Paramala Santosh. 2021. Emotional behavioural and autonomic dysregulation (EBAD) in Rett syndrome–EDA and HRV monitoring using wearable sensor technology. *Journal of psychiatric research* 138 (2021), 186–193.

[28] Allyson Hadwin, Sanna Järvelä, and Mariel Miller. 2017. Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In *Handbook of self-regulation of learning and performance*. Routledge, 83–106.

[29] P EKMAN-WV FRIESEN-JC HAGER. 2002. Facial Action Coding System. The Manual On CD ROM.

[30] Javier Hernandez, Ivan Riobo, Agata Rozga, Gregory D Abowd, and Rosalind W Picard. 2014. Using electrodermal activity to recognize ease of engagement in children during social interactions. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 307–317.

[31] Nathan Holbert, Matthew Berland, and Yasmin B Kafai. 2020. Introduction: fifty years of constructionism. (2020).

[32] Yasuhiro Imai. 2010. Emotions in SLA: New insights from collaborative learning for an EFL classroom. *The Modern Language Journal* 94, 2 (2010), 278–292.

[33] Sanna Järvelä and Allyson Hadwin. 2024. Triggers for self-regulated learning: A conceptual framework for advancing multimodal research about SRL. *Learning and Individual Differences* 115 (2024), 102526.

[34] Sanna Järvelä and Allyson F Hadwin. 2013. New frontiers: Regulating learning in CSCL. *Educational psychologist* 48, 1 (2013), 25–39.

[35] Sanna Järvelä, Hanna Järvenoja, and Jonna Malmberg. 2019. Capturing the dynamic and cyclical nature of regulation: Methodological Progress in understanding socially shared regulation in learning. *International journal of computer-supported collaborative learning* 14 (2019), 425–441.

[36] Susanna Järvelin-Pasanen, Sanna Sinikallio, and Mika P Tarvainen. 2018. Heart rate variability and occupational stress—systematic review. *Industrial health* 56, 6 (2018), 500–511.

[37] Hanna Järvenoja, Sanna Järvelä, and Jonna Malmberg. 2020. Supporting groups' emotion and motivation regulation during collaborative learning. *Learning and Instruction* 70 (2020), 101090.

[38] Yasmin B Kafai and Chris Proctor. 2022. A revaluation of computational thinking in K–12 education: Moving toward computational literacies. *Educational Researcher* 51, 2 (2022), 146–151.

[39] Yasmin B Kafai, Chris Proctor, Shuang Cai, Francisco Castro, Victoria Delaney, Kayla DesPortes, Christopher Hoadley, Victor R Lee, Duri Long, Brian Magerko, et al. 2024. What Does it Mean to be Literate in the Time of AI? Different Perspectives on Learning and Teaching AI Literacies in K-12 Education. In *Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024, pp. 1856-1862*. International Society of the Learning Sciences.

[40] Manmeet Kaur and Munish Kumar. 2024. Facial emotion recognition: A comprehensive review. *Expert Systems* 41, 10 (2024), e13670.

[41] Nick Kelly and John S Gero. 2021. Design thinking and computational thinking: A dual process model for addressing design problems. *Design Science* 7 (2021), e8.

[42] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. 2018. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry investigation* 15, 3 (2018), 235.

[43] Joyce Hwee Ling Koh, Ching Sing Chai, Benjamin Wong, Huang-Yao Hong, Joyce Hwee Ling Koh, Ching Sing Chai, Benjamin Wong, and Huang-Yao Hong. 2015. *Design thinking and education*. Springer.

[44] Madeleine Kröper, Doris Fay, Tilmann Lindberg, and Christoph Meinel. 2011. Interrelations between motivation, creativity and emotions in design thinking processes–an empirical study based on regulatory focus theory. In *Design creativity 2010*. Springer, 97–104.

[45] Peter Kuppens. 2015. It's about time: A special section on affect dynamics. *Emotion Review* 7, 4 (2015), 297–300.

[46] Joni Lämsä, Raija Hämäläinen, Pekka Koskinen, Jouni Viiri, and Emilia Lampi. 2021. What do we do when we analyse the temporal aspects of computer-supported collaborative learning? A systematic literature review. *Educational Research Review* 33 (2021), 100387.

[47] Jinhak Lee, Ho Bin Hwang, Seungjae Lee, Jayon Kim, Jeyeon Lee, Sanghag Kim, Jung Hee Ha, Yoojin Jang, Sejin Hwang, Hoon-Ki Park, et al. 2024. Analysis of Acute Stress Reactivity and Recovery in Autonomic Nervous System Considering Individual Characteristics of Stress using HRV and EDA. *IEEE Access* (2024).

[48] Serena Lee-Cultura, Kshitij Sharma, Giulia Cosentino, Sofia Papavlasopoulou, and Michail Giannakos. 2021. Children's play and problem solving in motion-based educational games: Synergies between human annotations and multimodal data. In *Interaction Design and Children*. 408–420.

[49] Serena Lee-Cultura, Kshitij Sharma, and Michail N Giannakos. 2023. Multimodal teacher dashboards: challenges and opportunities of enhancing teacher insights through a case study. *IEEE Transactions on Learning Technologies* 17 (2023), 181–201.

[50] Blair Lehman, Sidney D'Mello, and Art Graesser. 2012. Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education* 15, 3 (2012), 184–194.

[51] Li Li, Yu Fengchao, and Enting Zhang. 2024. A systematic review of learning task design for K-12 AI education: Trends, challenges, and opportunities. *Computers and Education: Artificial Intelligence* (2024), 100217.

[52] Yante Li, Yang Liu, Andy Nguyen, Henglin Shi, Eija Vuorenmaa, Sanna Järvelä, and Guoying Zhao. 2024. Interactions for socially shared regulation in collaborative learning: an interdisciplinary multimodal dataset. *ACM Transactions on Interactive Intelligent Systems* 14, 3 (2024), 1–34.

[53] Annabel Lindner, Stefan Seegerer, and Ralf Romeike. 2019. Unplugged Activities in the Context of AI. In *Informatics in Schools. New Ideas in School Informatics: 12th International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, ISSEP 2019, Larnaca, Cyprus, November 18–20, 2019, Proceedings 12*. Springer, 123–135.

[54] Huimin Liu. 2008. Generalized additive model. *Department of Mathematics and Statistics University of Minnesota Duluth: Duluth, MN, USA* 55812 (2008).

[55] Nikki G Lobczowski, Kayley Lyons, Jeffrey A Greene, and Jacqueline E McLaughlin. 2021. Socioemotional regulation strategies in a project-based learning environment. *Contemporary Educational Psychology* 65 (2021), 101968.

[56] Bethany McDaniel, Sidney D'Mello, Brandon King, Patrick Chipman, Kristy Tapp, and Art Graesser. 2007. Facial features for affective state detection in learning environments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29.

[57] F Miao and K Shiohira. 2022. K-12 AI curricula. A mapping of government-endorsed AI curricula. *UNESCO Publishing, URL https://unesdoc. unesco. org/ark:/48223/pf0000380602* 3 (2022), 1144399.

[58] Luis Morales-Navarro, Yasmin B Kafai, Ken Kahn, Ralf Romeike, Tilman Michaeli, Daniella DiPaola, Safinah Ali, Randi Williams, Cynthia Breazeal, Francisco Castro, et al. 2023. Constructionist Approaches to Learning Artificial Intelligence/Machine Learning: Past, Present, and Future. In *Proceedings of Constructionism 2023*.

[59] Jauwairia Nasir, Aditi Kothiyal, Barbara Bruno, and Pierre Dillenbourg. 2021. Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. *International Journal of Computer-Supported Collaborative Learning* 16, 4 (2021), 485–523.

[60] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2 (2021), 100041.

[61] Alannah Oleson, Brett Wortzman, and Amy J Ko. 2020. On the role of design in K-12 computing education. *ACM Transactions on Computing Education (TOCE)* 21, 1 (2020), 1–34.

[62] Ernesto Panadero and Sanna Järvelä. 2015. Socially shared regulation of learning: A review. *European Psychologist* (2015).

[63] Stefanie Panke. 2019. Design thinking in education: Perspectives, opportunities and challenges. *Open Education Studies* 1, 1 (2019), 281–306.

[64] Zacharoula Papamitsiou, Ilias O Pappas, Kshitij Sharma, and Michail N Giannakos. 2020. Utilizing multimodal data through fsQCA to explain engagement in adaptive learning. *IEEE Transactions on Learning Technologies* 13, 4 (2020), 689–703.

[65] Sofia Papavlasopoulou, Kshitij Sharma, and Michail N Giannakos. 2020. Coding activities for children: Coupling eye-tracking with qualitative data to investigate gender differences. *Computers in Human Behavior* 105 (2020), 105939.

[66] Brian Parkinson. 2011. Interpersonal emotion transfer: Contagion and social appraisal. *Social and Personality Psychology Compass* 5, 7 (2011), 428–439.

[67] Reinhard Pekrun. 2006. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review* 18 (2006), 315–341.

[68] Reinhard Pekrun. 2011. Emotions as drivers of learning and cognitive development. In *New perspectives on affect and learning technologies*. Springer, 23–39.

[69] Reinhard Pekrun, Thomas Goetz, Lia M Daniels, Robert H Stupnisky, and Raymond P Perry. 2010. Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of educational psychology* 102, 3 (2010), 531.

[70] Reinhard Pekrun and Elizabeth J Stephens. 2010. Achievement emotions: A control-value approach. *Social and Personality Psychology Compass* 4, 4 (2010), 238–255.

[71] Hasso Plattner, Christoph Meinel, and Larry Leifer. 2012. *Design thinking research*. Springer.

[72] Isabella Possaghi, Feiran Zhang, Kshitij Sharma, and Sofia Papavlasopoulou. 2024. Design Thinking Activities for K-12 Students: Multi-Modal Data Explanations on Coding Performance. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. 290–306.

[73] Joy R Rudland, Clinton Golding, and Tim J Wilkinson. 2020. The stress paradox: how stress can be good for learning. *Medical education* 54, 1 (2020), 40–45.

[74] Daniel Sánchez. 2024. Learning Under Stress: Enhancing Team-Based Simulation Training with Multimodal Data. In *17th International Conference on Computer-Supported Collaborative Learning (CSCL) 2024-CSCL Proceedings*, Vol. 2024.

[75] Mirweis Sangin, Gaëlle Molinari, Marc-Antoine Nüssli, and Pierre Dillenbourg. 2008. How learners use awareness cues about their peer's knowledge? Insights from synchronized eye-tracking data. (2008).

[76] Marie-Monique Schaper, Mariana Aki Tamashiro, Rachel Charlotte Smith, Maarten Van Mechelen, and Ole Sejer Iversen. 2023. Five design recommendations for teaching teenagers' about artificial intelligence and machine learning. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 298–309.

[77] Klaus R Scherer and Agnes Moors. 2019. The emotion process: Event appraisal and component differentiation. *Annual review of psychology* 70, 1 (2019), 719–745.

[78] Carmen Schiweck, Deborah Piette, Daniel Berckmans, Stephan Claes, and Elske Vrieze. 2019. Heart rate and high frequency heart rate variability during stress as biomarker for clinical depression. A systematic review. *Psychological medicine* 49, 2 (2019), 200–211.

[79] Dale H Schunk and Barry Zimmerman. 2011. *Handbook of self-regulation of learning and performance*. Taylor & Francis.

[80] Gayane Sedrakyan, Jonna Malmberg, Katrien Verbert, Sanna Järvelä, and Paul A Kirschner. 2020. Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior* 107 (2020), 105512.

[81] Kshitij Sharma and Michail Giannakos. 2020. Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology* 51, 5 (2020), 1450–1484.

[82] Kshitij Sharma and Michail Giannakos. 2021. Sensing technologies and child–computer interaction: Opportunities, challenges and ethical considerations. *International Journal of Child-Computer Interaction* 30 (2021), 100331.

[83] Kshitij Sharma, Serena Lee-Cultura, and Michail Giannakos. 2022. Keep calm and do not carry-forward: Toward sensor-data driven AI agent to enhance human learning. *Frontiers in Artificial Intelligence* 4 (2022), 713176.

[84] Kshitij Sharma, Evangelos Niforatos, Michail Giannakos, and Vassilis Kostakos. 2020. Assessing cognitive performance using physiological and facial features:

Generalizing across contexts. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–41.

[85] Kshitij Sharma and Jennifer K Olsen. 2022. What Brings Students Together?: Investigating the Causal Relationship Between Joint Mental Effort and Joint Visual Attention. In *Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning-CSCL 2022, pp. 131-138*. International Society of the Learning Sciences.

[86] Kshitij Sharma, Sofia Papavlasopoulou, and Michail Giannakos. 2019. Joint Emotional State of Children and Perceived Collaborative Experience in Coding Activities. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. ACM, 133–145.

[87] Kshitij Sharma, Sofia Papavlasopoulou, and Michail Giannakos. 2022. Children's facial expressions during collaborative coding: Objective versus subjective performances. *International Journal of Child-Computer Interaction* 34 (2022), 100536.

[88] Rachel Charlotte Smith, Marie-Monique Schaper, Mariana Aki Tamashiro, Maarten Van Mechelen, Marianne Graves Petersen, and Ole Sejer Iversen. 2023. A research agenda for computational empowerment for emerging technology education. *International Journal of Child-Computer Interaction* 38 (2023), 100616.

[89] Daniel Spikol, Emanuele Ruffaldi, and Mutlu Cukurova. 2017. Using multimodal learning analytics to identify aspects of collaboration in project-based learning. Philadelphia, PA: International Society of the Learning Sciences.

[90] Daniel Spikol, Emanuele Ruffaldi, Giacomo Dabisias, and Mutlu Cukurova. 2018. Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning* 34, 4 (2018), 366–377.

[91] Amber Chauncey Strain and Sidney K D'Mello. 2015. Affect regulation during learning: The enhancing effect of cognitive reappraisal. *Applied Cognitive Psychology* 29, 1 (2015), 1–19.

[92] Jiahong Su, Davy Tsz Kit Ng, and Samuel Kai Wah Chu. 2023. Artificial intelligence (AI) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence* 4 (2023), 100124.

[93] Chen Sun, Valerie J Shute, Angela EB Stewart, Quinton Beck-White, Caroline R Reinhardt, Guojing Zhou, Nicholas Duran, and Sidney K D'Mello. 2022. The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Computers in Human Behavior* 128 (2022), 107120.

[94] Jingjing Sun, Richard C Anderson, Tzu-Jung Lin, Joshua A Morris, Brian W Miller, Shufeng Ma, Kim Thi Nguyen-Jahiel, and Theresa Scott. 2022. Children's engagement during collaborative learning and direct instruction through the lens of participant structure. *Contemporary Educational Psychology* 69 (2022), 102061.

[95] Gahyun Sung, Harum Bhinder, Tianyi Feng, and Bertrand Schneider. 2023. Stressed or engaged? Addressing the mixed significance of physiological activity during constructivist learning. *Computers & Education* 199 (2023), 104784.

[96] Gabriella Tisza, Kshitij Sharma, Sofia Papavlasopoulou, Panos Markopoulos, and Michail Giannakos. 2022. Understanding Fun in Learning to Code: A Multi-Modal Data approach. In *Interaction Design and Children*. 274–287.

[97] Tiina Törmänen, Hanna Järvenoja, and Kristiina Mänty. 2021. All for one and one for all–How are students' affective states and group-level emotion regulation interconnected in collaborative learning? *International journal of educational research* 109 (2021), 101861.

[98] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for K-12: What should every child know about AI?. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9795–9799.

[99] Gerben A Van Kleef. 2009. How emotions regulate social life: The emotions as social information (EASI) model. *Current directions in psychological science* 18, 3 (2009), 184–188.

[100] Gerben A Van Kleef, Astrid C Homan, and Arik Cheshin. 2012. Emotional influence at work: Take it EASI. *Organizational Psychology Review* 2, 4 (2012), 311–339.

[101] Maarten Van Mechelen, Rachel Charlotte Smith, Marie-Monique Schaper, Mariana Tamashiro, Karl-Emil Bilstrup, Mille Lunding, Marianne Graves Petersen, and Ole Sejer Iversen. 2023. Emerging technologies in K–12 education: A future HCI research agenda. *ACM Transactions on Computer-Human Interaction* 30, 3 (2023), 1–40.

[102] Kristin Wäschle, Anne Allgaier, Andreas Lachner, Siegfried Fink, and Matthias Nückles. 2014. Procrastination and self-efficacy: Tracing vicious and virtuous circles in self-regulated learning. *Learning and instruction* 29 (2014), 103–114.

[103] Marcelo Worsley and Paulo Blikstein. 2015. Using learning analytics to study cognitive disequilibrium in a complex learning environment. In *Proceedings of the fifth international conference on learning analytics and knowledge*. 426–427.

[104] Kateryna Zabolotna, Jonna Malmberg, and Hanna Järvenoja. 2023. Examining the interplay of knowledge construction and group-level regulation in a computer-supported collaborative learning physics task. *Computers in Human Behavior* 138 (2023), 107494.

[105] Feiran Zhang, Isabella Possaghi, Kshitij Sharma, and Sofia Papavlasopoulou. 2024. High-performing Groups during Children's Collaborative Coding Activities: What Can Multimodal Data Tell Us?. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. 533–559.

[106] Leah Zhang-Kennedy and Sonia Chiasson. 2021. A systematic review of multimedia tools for cybersecurity awareness and education. *ACM Computing Surveys (CSUR)* 54, 1 (2021), 1–39.

[107] Michelle Zimmerman. 2018. *Teaching AI: Exploring new frontiers for learning*. International society for technology in education.

Appendix

Criteria	Description	Example	3 Points	2 Points	1 Point	0 Points
Topic Variety	How many topics does the group consider while exploring everyday issues pertaining to the chosen persona?	The group considered more than one initial topic. E.g., *"Creating the perfect menu"*	The group considered more than three topics	The group considered two to three topics	The group considered one topic	No topics were considered
Topic Elaboration	To what extent does the group elaborate on the chosen personas' challenges to explore?	The group elaborates and details challenges for the chosen persona. E.g., *"Grandma's problem is creating the perfect menu for all the grandkids for Easter lunch"*	The group presented detailed and varied challenges	The group presented challenges, but with limited elaboration.	The group mentioned one challenge with little detail	No challenges were identified
Topic Relevance	To what extent does the group elaborate solutions to explore directly addressing AI & ML?	The group listed topics with elaboration from videos' information and discussions among peers. E.g., *"We can create an AI-driven recipe database to inspire her"*	The group presented a well-developed AI-driven solution	The group presented a basic AI-driven solution, but with limited elaboration.	The group mentioned one solution with little detail	No solutions were identified
Topic Awareness	To what extent does the group elaborate solutions to explore addressing trust in AI & ML?	The group elaborated on the system reliability: *"It seems trustable because a lot of people use AI, but the Grandma should check the outcome in any case"*	The group presented solutions discussing trust and reliability	The group addressed trust but lacked depth	The group mentioned trust with minimal elaboration	No discussion of trust was present
Topic Consistency	To what extent does the worksheet show the usefulness of the proposed solution with respect to AI & ML?	The connection between sheet points 3, 4, and 5 shows consistency and relevance in proposing a solution	The solution is consistent and relevant to AI & ML	The solution is somewhat relevant to AI & ML	The solution shows limited relevance to AI & ML	No relevance to AI & ML was demonstrated

Table 7: Rubric A, the rubric to assess the Paper Sheet 1

Criteria	Description	Example	3 Points	2 Points	1 Point	0 Points
Topic Relevance	To what extent does a scenario belonging to an ML & AI topic unfold?	The scenario built presents elements related to ML & AI. E.g., the description of an AI-driven digital tool or service	The scenario is consistent and relevant to AI & ML	The scenario is relevant to AI & ML	The scenario shows limited relevance to AI & ML	No relevant consistent was demonstrated
Topic Consistency	To what extent does the scenario maintain consistency with the chosen persona and their challenges?	The scenario is built around the chosen persona. For example, at least one of the Grandma's challenges is addressed	The scenario fully aligns with challenges and expands on their context	The scenario partially addresses challenges, with some misalignment or gaps	The scenario only loosely relates to challenges and/or lacks detail	The scenario does not relate to challenges
Topic Communication	To what extent does the scenario hold communicative significance (message clarity) in the domain of ML & AI?	The scenario built presents elements able to instruct peers about ML & AI-driven tools and services, how they can solve everyday problems	The scenario explains concepts in a way peers can easily understand.	The scenario conveys concepts but with some gaps or lack of clarity	The scenario includes basic ML & AI elements but lacks clarity or detail	The scenario does not communicate concepts clearly
Critical Thinking	Does the scenario address issues (or how to avoid/prevent them) related to AI & ML usage?	The scenario discusses challenges. E.g., *"The grandma should ask every year to update the recipe collection to avoid assumptions on grandchildren's taste"*	Challenges are thoroughly addressd with suggestions to mitigate them	Some challenges are identified, with limited or general solutions	The scenario mentions challenges but does not provide meaningful solutions	The scenario does not address potential limitations

Table 8: Rubric B.1, the rubric to assess the Paper Sheet 2

Criteria	Description	Example	3 Points	2 Points	1 Point	0 Points
Database design complexity	To what extent does the artifact present complexity in items' categorization (number of categories)?	The database presents four categorization options for items. E.g., *"My Grandchildren's Favourites"*, *"Moderately tasty"*, etc.	More than 5 categories included	5 to 4 categories are included	3 to 2 categories included	1 or no categories included
Database design complexity	To what extent does the artifact present complexity in items' categorization (number of items)?	The database presents fifteen items that can be categorized. E.g., *"Blueberry pie"*, *"Fish soup"*, *"Falafels"*, *"Grilled vegetables"*, etc.	More than 6 items included	6 to 5 items included	5 to 4 items included	Less than 3 items included
External media embeddment	To what external resources are employed to develop the chosen scenario and enrich the experience?	Each item is described with a picture. The background has been changed to match the theme	More than 4 pictures included	4 to 3 pictures included	2 to 1 pictures included	No pictures included
Topic Relevance A	To what extent does the SorBET database include a built scenario belonging to an ML & AI topic?	The scenario built presents elements (items, classification descriptions) related to ML & AI	The scenario is thoroughly aligned with an ML & AI topic and integrates relevant examples	The scenario aligns with an ML & AI topic but lacks full detail or depth	The scenario loosely connects to an ML & AI topic, with significant gaps or inaccuracies	The scenario does not relate to an ML & AI topic
Topic Relevance B	To what extent does the *SorBET* database include a built scenario belonging to the *chosen* AI & ML topic?	The scenario built presents elements (items, classification descriptions) related to the specific ML & AI topic chosen by the group	The scenario fully incorporates the chosen topic with clear descriptions	The scenario partially incorporates the chosen topic, but with gaps in relevance	The scenario includes elements of the topic but with misalignment	The scenario does not reflect the chosen topic

Table 9: Rubric B.2, the rubric to assess the *SorBET* database artifact