Accepted at the 42nd International Conference
on Software Maintenance and Evolution (ICSME 2026)
Replication and Negative Results track, to appear

Emotion Self-Monitoring in Agile Retrospectives: A Replication Study

Daniela Grassi[1], Filippo Lanubile[1], Nicole Novielli[1], Alexander Serebrenik[2]
[1]University of Bari, Bari, Italy, Email: {daniela.grassi,filippo.lanubile, nicole.novielli}@uniba.it
[2]Eindhoven University of Technology, Eindhoven, The Netherlands, Email: a.serebrenik@tue.nl

***Abstract*—Emotion-awareness has been recently investigated in software engineering in relation to developers' well-being and productivity. A former study showed that self-monitoring of emotions through visualization tools combining biometric and self-reported data can enhance agile retrospective meetings, with newly formed novice student teams benefiting more than consolidated professional ones. The original authors hypothesized that the practice is most valuable for newly formed teams whose roles and collaboration dynamics are still being defined, but their sample did not include teams combining novelty with member competence and authentic project pressure. We present a partial replication in a third, intermediate context: a professional training program in which trainees form temporary teams to deliver real software products to industrial stakeholders. We conducted a within-subjects study with nine trainees organized in three teams over four agile sprints, comparing control and experimental conditions, and collected qualitative and quantitative data through interviews and focus groups. Our results partially support the original findings: trainees reported individual benefits in emotional awareness, but no improvements during retrospective meetings, owing to their reluctance to share emotional insights with teammates in newly formed teams. We refine the original hypothesis accordingly: emotion-monitoring tools support individual emotional self-awareness, but their team-level benefits during retrospectives appear to require a baseline of psychological safety that newly formed teams under external pressure may not yet have established.**

***Index Terms*—Replication study, emotional awareness, agile retrospective meetings, biometric sensors, self-report, professional training**

I. Introduction

In a collaborative development environment, emotions shape communication, influence conflict resolution, and affect how decisions are made [1], [2], [3]. Emotional awareness, the ability to recognize and interpret one's own emotions and those of others [4], is a critical competence for effective teamwork in software engineering, as it may help developers not only manage negative emotions but also enhance teamwork effectiveness by supporting effective conflict resolution and strengthening collaboration [5]. Studies show that positive emotions like joy and excitement foster engagement and productivity, while negative emotions such as frustration and confusion can significantly hinder individual performance and team collaboration [6], [7], [8], [9], [10], [11].

Researchers have proposed various tools for supporting emotional awareness in software teams: Emotimonitor, a Trello power-up for emoji-based emotion reporting [12], later extended into the AI-powered EmoReflex [13]; TimeMarbles, which shows that holistic self-reflection improves workday perception [14]; and structured self-reflection practices that foster developer productivity and well-being [15].

More recent work investigated the integration of emotional feedback into agile retrospective meetings using EmoVizPhy [16], an emotion-visualization tool designed to enable self-assessment of emotions through self-report and sensor-based measurements. This approach has been empirically evaluated in a controlled study involving two student teams and one professional development team [17] (i.e., the *original study* replicated in this article). The original authors found a notable distinction between the two populations: novice student teams perceived the practice of emotion self-monitoring as beneficial during retrospectives, while professional developers, working in a consolidated team, perceived limited benefits. Based on this evidence, the original authors hypothesized that emotion self-monitoring may play a valuable role *for novice developers with little-to-no experience in agile practices, as well as in facilitating the consolidation of new agile teams for which roles and collaboration dynamics are still being defined*. However, the sample of the original study did not include any team that simultaneously satisfied the conditions implied by this hypothesis, since the two factors—individual inexperience with agile and team novelty—were inherently confounded in the novice student teams. As a result, the proposed explanation remained an open conjecture rather than an empirically verified boundary condition. Testing this hypothesis directly requires a context that bridges the gap between undergraduate capstone teams and consolidated professional teams: a setting in which team members are skilled and engaged in real software development work, yet have not yet established shared roles, mutual trust, or collaboration routines. An ideal context for validating this hypothesis is offered by professional training programs, such as Engineering Doctorate (EngD) programs[1] which often simulate real-world conditions more closely than traditional academic settings, with participants working on industry-sponsored projects under authentic constraints, deadlines, and stakeholder expectations [18]. These programs face unique challenges due to participants typically being strangers with diverse disciplinary backgrounds, who must rapidly form effective teams and deliver meaningful outcomes in limited time frames. This combination provides

[1]See, e.g., https://engd.cs.st-andrews.ac.uk/engd-overview/, https://www.tue.nl/en/education/engd-programs.

an ideal setting to probe the boundary of the original findings and assess the extent to which the team-novelty hypothesis holds when other contextual factors are controlled. Given this context, we investigate the following research question:

RQ: *Does self-monitoring of emotions enhance agile retrospective meetings in newly-formed interdisciplinary teams operating under authentic project constraints?*

To answer this research question, we replicated the original study within an EngD program, where interdisciplinary teams develop solutions for real stakeholders using agile practices. We reused EmoVizPhy and applied the same experimental protocol of the original study, collecting both quantitative and qualitative data through interviews and focus groups.

Our findings partially support the original findings and contribute novel insights on the contextual factors influencing the effectiveness of emotion-monitoring tools. Specifically, *similarly* to the original study, participants found emotion-monitoring practices useful for individual self-reflection and emotional awareness. However, *differently* from the perception of the novice developers in the original study, trainees did not report any significant benefits during retrospectives, mostly due to their reluctance in sharing emotional insights with teammates in newly formed teams. Our results also suggest the importance of combining multiple data sources to support effective emotional monitoring practices. Finally, privacy concerns emerged as associated with the integration of emotion monitoring, particularly regarding surveillance.

II. Background

A. Emotional Awareness in Software Engineering

Emotions have gained increasing relevance in software engineering, where their impact on developers' productivity, emotional triggers, and well-being is a growing area of interest [19], [20], [21]. A consolidated body of research has shown that positive moods and states of flow correlate with higher performance, while frustration and negative affect hinder focus and progress [7], [8], and that self-reflection practices can increase developers' awareness of their work habits and foster productivity and well-being [15]. Building on these foundations, Girardi et al. [22] have demonstrated the feasibility of using physiological signals to detect emotions during programming with a follow-up study confirming that positive emotional valence correlates with perceived productivity in industry settings [23]. Recent work has further refined this approach: clustering developers by their physiological profiles improves the accuracy of sensor-based classifiers [24], while speech-based emotion recognition on retrospective recordings captures emotional dynamics that correlate with self-assessed team performance [25].

Beyond detection, a growing line of work has examined the human factors shaping developers' experience. Madampe et al. proposed a framework to support emotional intelligence and team goals when handling requirements changes [26], and characterized how negative emotions emerge in agile contexts and how managers can address them [27]. Cerqueira et al. [28] highlighted empathy as a key driver of practitioners' well-being and mental health. Martinez Montes et al. [29] identified the individual, social, and organizational factors that influence software engineers' well-being, complementing the affect-centric perspective.

Several tools have been proposed to enhance emotional awareness and reflection. McDuff et al. [30] explored personal reflection through *AffectAura*, *TimeMarbles* [14] for more holistic self-reflection in knowledge work, Emotimonitor [12] for emoji-based reactions in Trello, and its AI-powered extension EmoReflex [13]. Among the tools explicitly designed for agile retrospective meetings, *EmoVizPhy* [17] combines self-reported emotions, biometric signals, and team activity data in a unified visualization that supports reflection at both the individual and team level. The tool was empirically evaluated in a controlled study [17] (see Section III), which we replicate.

B. The Engineering Doctorate in Software Technology

The context of this study is the Engineering Doctorate (EngD) program in Software Technology of Eindhoven University of Technology, a context that represents an ideal experimental setting for addressing our research question. The program is a professional training initiative that integrates academic training with practical industrial experience, making it particularly well-suited for investigating the phenomena under study. Specifically, the hands-on project phase, which forms the basis of our investigation, engages trainees in team-based collaboration on complex, open-ended problems that mirror real industry challenges, with team composition rotating across blocks to form groups of up to seven members.

The EngD is a two-year, salaried program focused on the design and development of software. The first year consists of in-house training activities at the University, while the second year involves a ten-month capstone design project carried out as a full-time employee within a partner company. Throughout both years, the program employs agile development methodologies and project-based learning frameworks. Trainees engage with industry-sponsored projects while adhering to professional software development practices such as sprint planning, regular daily stand-up meetings, systematic retrospectives, and iterative delivery cycles.

During the first year, trainees participate in foundation-building activities organized as four sequential blocks of approximately 11 weeks each. Each block consists of a training period—during which trainees acquire technical, business, and professional competencies—and a hands-on project. All trainees enrolled in the EngD program receive joint supervision combining academic supervisors with industry mentors.

III. The Original Study

A. Study Design

The original study investigated to what extent self-monitoring of emotions can enhance agile retrospective meetings by improving emotion awareness of participants. The original authors implemented self-monitoring of emotions through the collection of biometrics and self-reported information about emotions, which were then visualized using EmoVizPhy [16] before the retrospective meetings to inform discussion during the meeting.

The study involved two distinct populations, both adopting the Mad/Sad/Glad retrospective format [31]. The first consisted of ten Italian undergraduate computer-science students (seven male, three female, aged 20–25) with limited to no professional experience in agile development, organized into two teams of five (Teams A and B) working on a capstone project within an undergraduate Software Engineering course, with two-week sprints. The second was a team of three Italian developers from an industry partner, already familiar with agile practices and working as a consolidated team on a full-time project with one-week sprints.

The instruments included: (i) the Empatica E4 wristband to collect electrodermal activity (EDA) and heart rate (HR) signals, known to be reliable proxies for emotions [22], [32], [24], [8]; (ii) a popup application based on the Self-Assessment Manikin (SAM) technique [33] for the self-report of emotional valence (positive vs. negative), arousal (low vs. high activation), dominance (low vs. high), perceived productivity, current activity, and free-text notes; (iii) the EmoVizPhy visualization tool [16], which displays EDA peaks, HR over time, and self-reports as interactive tooltips.

The study adopted a *crossover design* for student teams: in Sprint 1, Team A operated in the experimental condition (with EmoVizPhy) and Team B in the control condition; the roles were reversed in Sprint 2. The professional team participated for two weeks, both in the experimental condition. In the experimental condition, participants individually reviewed EmoVizPhy data for up to 15 minutes before entering the retrospective meeting, to recall emotion-triggering episodes occurred during the sprint.

Data collection comprised: (i) the cards contributed and corrective actions proposed during the retrospectives; (ii) semi-structured individual interviews with all participants, conducted at the end of the experimental phase, including six items on a 4-point ordinal scale; (iii) focus groups, one per team, conducted at the end of the experiment. The qualitative analysis was carried out by means of open coding followed by focused coding, organizing the codes into four themes: *Developer's attitude*, *Emotion representation*, *Perceived usefulness for retrospective meetings*, and *Instrumentation UX*.

B. Summary of Results

The authors observed a slight increase in the experimental condition for both the number of cards contributed and the corrective actions proposed across both populations. However, the most striking difference concerned the *type* of actions proposed: student participants nearly doubled their corrective actions addressing social and interpersonal dynamics (e.g., taking time to listen when others share their problems), whereas professional developers proposed no corrective actions in this category during the experimental sprint.

The qualitative analysis of participants' answers shows that students reported predominantly positive feedback regarding *improved card writing*, *improved actions proposal*, *improved emotional awareness*, and *promising use cases* for EmoVizPhy. Professional developers, on the other hand, reported only limited benefits, which the original authors attributed to three factors: the participants' familiarity with agile practices, the consolidated nature of their team, and the short, one-week duration of their sprints. Based on Tuckman's model [34], they hypothesized that self-monitoring emotions through EmoVizPhy may be particularly valuable in supporting the consolidation of new agile teams, where roles and collaboration dynamics are still being defined.

IV. Our Replication

A. Motivation for Conducting the Replication

We conducted this replication because the original authors explicitly hypothesized that emotion self-monitoring through EmoVizPhy is most beneficial for newly formed teams, but their sample did not allow this hypothesis to be tested in isolation. In the original study, team novelty was confounded with individual inexperience with agile practices: the only newly formed teams were also teams of novices (i.e., undergraduate students), while the only competent team was already consolidated. As a result, the team-novelty hypothesis remained an open conjecture requiring further investigation.

Our context provides the missing configuration. Interdisciplinary EngD trainees form temporary teams to deliver real software products to industrial stakeholders. Trainees come from different backgrounds and have not previously collaborated, yet they are more familiar with software development compared to undergraduate students involved in the original study. Furthermore, to strengthen the external validity of the original study, the current replication setting involves trainees operating under authentic project constraints, also involving pressure from external stakeholders defining the requirements in a way that the original sample did not.

B. Changes to the Original Study

This replication makes the following explicit changes to the original study, as also summarized in Table I.

Population and context. To address our research question, we recruited our participants in a population of newly-formed teams of trainees from the EngD program (cf. Section II-B). This will enable comparison with findings observed with the newly-formed teams of students involved in the original study.

Biometric devices. We replaced the Empatica E4 used in the original study with the Empatica EmbracePlus wristband, since the E4 has been discontinued by the manufacturer and is no longer commercially available. Both devices are produced by the same manufacturer and record the same signals.

Retrospective format and quantitative comparison. Retrospective meetings are an integral part of the EngD curriculum, and the choice of format is left to each team as part of their training. We could therefore not impose a single retrospective format, as was done in the original study with Mad/Sad/Glad.

TABLE I
SETTINGS COMPARISON BETWEEN THE ORIGINAL STUDY AND THIS REPLICATION.

Setting	**Original study**	**This replication**
Experiment site	University of Bari (IT), industry partner	Eindhoven University of Technology (NL), EngD Program
Sample population type	Undergrad students; consolidated pro team	Trainees in professional training program
# Participants	10 students + 3 professionals	9 trainees
# Teams	2 student teams + 1 professional team	3 teams
Team size	5 (students) / 3 (professionals)	2–4
Team status	Just formed (students) / consolidated (pros)	Newly formed, interdisciplinary
Project type	Capstone (students) / industry project (pros)	Industry-sponsored project
Familiarity with agile	None (students) / High (professionals)	Mixed, in training
# Sprints / Sprint duration	2 sprints / 2 weeks (students), 2 Sprints 1 week each (pros)	4 sprints / 1 week each
Experimental design	Crossover (students), no control (pros)	Within-subjects (2 control + 2 experimental)
Retrospective format	Mad/Sad/Glad	Mad/Sad/Glad, Dream Team, Happiness Radar, Start/Stop/Continue
Data sources	Cards & actions, interviews, focus groups	Interviews, focus groups
Codebook	Original (4 themes, 14 codes)	Original codebook + 2 emerging codes
Inter-rater agreement statistic	Cohen's κ (2 raters)	Fleiss' κ (3 raters)

As a consequence of this constraint, the teams adopted heterogeneous formats both across teams and across sprints, making the notion of "card" and "action" inconsistent and not directly comparable. We therefore focus our comparison on the qualitative themes and on the ordinal-scale interview items, which are directly comparable across studies.

C. *Participants*

Participants were the first year trainees of the EngD program described in Section II-B during the hands-on project phase.

Our study involved three teams comprising two, three, and four members, respectively. The nine participants (six men, three women) were aged from 25 to 30 years. Each team was assigned a distinct industrial project sponsored by different stakeholders. Each team participated in the study for four weeks, for a total of four agile sprints. Throughout the study period, teams were responsible for maintaining comprehensive project backlogs and delivering incremental results. Their workflow included regular sprint planning sessions, daily stand-ups, sprint reviews, and retrospectives. The study culminated with teams presenting final demonstrations of their software products to stakeholders.

Participants were recruited on a voluntary basis. To minimize selection bias, we emphasized that participation would not affect their program outcomes and that they could withdraw at any time. Informed consent was obtained from all participants, and the study protocol was approved by our ethical review board.[2] To ensure the privacy of participants, we anonymized all biometric and self-report data.

D. *Instrumentation*

To collect data, we employed two main instruments: a wearable device to collect physiological signals and a self-report tool to collect participants' emotions. The emotion data are then visualized by participants through a visualization tool.

Data Collection through Emotion Self-monitoring Tools. We collect Electrodermal Activity (EDA) using the Embrace Plus wristband,[3] (sampling rate: 4 Hz). We focus on the EDA signal as its peaks are recognized indicators of stress responses [35] and high emotional arousal [36].

Recordings from wearable sensors may include noise due to extraneous factors such as sensor displacement, movement artifacts, or temporary electrode contact loss. Therefore, preprocessing is essential to ensure meaningful interpretation of the signals in relation to the participants' emotional and cognitive states. To this aim, we adopt the tool by Taylor et al. [37] for artifact detection and removal in the EDA signal, and we compute EDA peaks using the same method, which relies on the first derivative of the curve.

To capture participants' subjective experiences, we enabled self-report of emotions by replicating the experience-sampling methodology used in the original study and previously adopted in software engineering research [38], [23], [16]. Specifically, we reused the publicly available pop-up application by Girardi et al. [22] and used in the original study [17]. In line with the original study, the notification frequency was set to one prompt per hour, with the option for participants to postpone or dismiss the notification within a defined interval.

Emotion Visualization. To support awareness and reflection on the collected data during agile retrospectives, we reused EmoVizPhy [16], an open-source tool that integrates self-reported emotions with biometric data. The panel plots the EDA signal over time, with vertical lines denoting peaks indicative of potential stress episodes [35]. Interactive plots support zooming, panning, and hovering to reveal contextual tooltips with self-reported ratings, activity, and notes. This integrated representation enables participants to jointly interpret physiological and self-reported data, fostering emotion awareness and discussion during retrospective meetings. As in the original study, participants accessed their individual self-reports through the EmoVizPhy interface before retrospective meetings, to facilitate recall of relevant emotional episodes.

E. *Empirical Protocol*

Study Setup The day before the experiment, the experimenter had a preliminary briefing with the participants, who signed the informed consent forms, acknowledging the purpose, procedures, and data handling processes. The experi-

[2] ERB n. ERB2023MCS46
[3] https://www.empatica.com/en-eu/embraceplus/

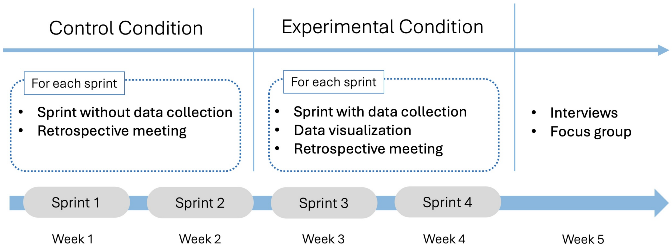

Fig. 1. Timeline of the experiment for each team.

menter also demonstrated how to wear the wristband and how to install the pop-up application and the EmoVizPhy tool.

Study Duration and Timeline All teams participated in the study over four consecutive one-week sprints using a within-subjects design (cf. Figure 1). Each team began with a two-week control condition (Sprints 1 and 2), during which they conducted retrospective meetings using their standard practices without any intervention from EmoVizPhy. Then, teams transitioned to a two-week experimental condition (Sprints 3 and 4). During this phase, participants actively engaged with the self-monitoring tools by (a) using the self-report tool to log their emotions and activities, and (b) wearing the wristband to collect biometrics throughout their working hours. In the experimental condition (Sprints 3 and 4), participants individually accessed the EmoVizPhy visualization tool upon conclusion of the sprint and right before entering the retrospective meeting. They were instructed to individually review (up to 15 minutes) and analyze their combined biometric and self-reported emotion data before starting the retrospectives to gain insights into possible triggers for visualized emotions and recall episodes that could potentially inform the discussion of the retrospective meeting. After this individual phase, the regular meeting starts with the entire team.

Retrospective Meetings Retrospective meetings were held at the end of each sprint. Since these retrospectives were an integral part of the course curriculum, we could not impose a specific format on the teams; instead, each team was free to adopt the retrospective format of their choice. Each team adopted retrospective formats of their liking: Team 1 used the *Mad/Sad/Glad* format [31] in the experimental condition and the *Dream Team* retrospective [39] in the control condition, where participants reflected on ideal team practices, collaborative values, and behaviors to avoid. Team 2 employed the *Happiness Radar* approach [40] in both conditions, where team members rated their satisfaction and shared brief reflections on their last agile iteration. Team 3 adopted the *Start-Stop-Continue* method in both conditions, guiding participants to discuss behaviors or practices to start, stop, or continue [41].

Interviews and Focus Groups To understand the perceived usefulness of emotion self-monitoring, we conducted individual, semi-structured interviews with all study participants. The questionnaires for the individual interview and focus groups are replicated from the original study. Six of the interview questions (see Figure 2) required participants to provide answers on a 4-point ordinal scale and to elaborate on their responses, while the remaining ones are open-ended questions. The experimenter conducted all interviews immediately at the end of the experimental phase, with each interview lasting between 15 and 30 minutes. At the end of the study, we carried out three focus groups, one with each team. The focus groups were conducted by the experimenter and lasted between 30 and 45 minutes. Both during the interviews and the focus groups, the experimenter took notes and made an audio recording. Afterwards, we transcribed the audio recordings and coded the transcripts for qualitative analysis.

F. Analysis

We analyzed the interview and focus-group transcripts following the thematic analysis approach proposed by Braun and Clarke [42], reusing the codebook from the original study. The codebook, including codes with their definitions, is included in the replication package[4]. One author manually extracted 96 excerpts from the interview and focus-group transcripts, discarding irrelevant contextual information (e.g., details about participants' backgrounds). Two rounds of coding were carried out. During the first round, three coders manually annotated half of the excerpts using the original codebook. The coders used the code `OTHER` whenever they did not identify a suitable code from the codebook. Moreover, if they felt an excerpt had been included by mistake (e.g., because it was not actually relevant to the study's research goals), they used the code `DISCARD`. After the first coding round, the three raters participated in a plenary meeting to discuss and resolve the disagreement cases. Furthermore, two new codes emerged as a result of the use of `OTHER`: *Self-report association* and *Importance of multimodality*. Then, the three coders engaged in a second round of coding, incorporating the two new codes. The three coders then convened to discuss and resolve the few remaining disagreements. As a result of this second stage, no new codes were included, thus reaching saturation. The final list of tags is reported in Table II.

V. Results

A. Qualitative analysis

In Table II, column T, we present the results of the qualitative coding on the transcripts of individual interviews and focus groups conducted with the participants. To enable comparison with the original study, we also report the findings from the original study (Students, Professional developers in the columns S and P, respectively) As a result of the coding (cf. Section IV-F), 15 excerpts were discarded as not being relevant for the research question, resulting in 81 excerpts. We assess the inter-annotator agreement using Fleiss' κ [43], obtaining a value of 0.85, which indicates almost perfect agreement [44]. The original study, conducted by two raters, used Cohen's κ and reported a value of 0.76.

Developer's Attitude This theme captures participants' attitudes toward using the self-monitoring tools and sharing emotions during retrospective meetings. The codes under this

[4]https://figshare.com/s/26fdf67c4aeed71cc894

TABLE II
TAXONOMY OF THEMES AND CODES FROM THE QUALITATIVE ANALYSIS OF INTERVIEW AND FOCUS GROUP TRANSCRIPTS. THE TABLE LISTS THE PARTICIPANTS GROUPED INTO STUDENTS (S) AND PROFESSIONAL DEVELOPERS (P) (REPRODUCED FROM THE ORIGINAL STUDY [17]), AND TRAINEES (T) FROM THIS REPLICATION. CODES IN ITALIC (*Self-report association*, *Importance of multimodality*) EMERGED IN THIS REPLICATION.

Themes	**Codes**		**S**	**P**	**T**
Developer's attitude	Willingness to share emotions	positive feedback	S1, S2, S3, S4, S7, S9	P1, P2, P3	T7, T8
		negative feedback	S5, S6, S8, S10	-	T5, T6, T9
	Privacy	no concerns	ALL (S1–S10)	P2	ALL (T1–T9)
		concerns	-	-	T1, T2, T3, T5, T7, T8
Emotion representation	Interpretability of biometrics	positive feedback	S2, S3, S4, S5, S8	-	T8
		negative feedback	S1, S3, S6, S7, S10	P3	-
	Self-report association	feasible	-	-	T4, T5, T8, T9
		challenging	-	-	-
	Peak-event association	feasible	S2, S3, S5, S6, S7, S8, S9, S10	P1, P3	T4, T5, T6, T8, T9
		challenging	S6	P1, P2	-
	Importance of notes		S2, S3	P3	T3, T6, T7
	Importance of multimodality		-	-	T9
Perceived usefulness for retrospective meetings	Focus and productivity boost		S1, S10	-	-
	Improved card writing	positive feedback	S4, S5, S6	P3	-
		negative feedback	S2, S4	P1, P2	ALL (T1–T9)
	Improved actions proposal	positive feedback	S4, S5	-	-
		negative feedback	S2	P3	-
	Improved emotional awareness	positive feedback	S1, S3, S4, S5, S7, S9, S10	P3	T3, T4, T6, T7, T9
		negative feedback	S4, S6, S7	-	T5, T8
	Promising use cases		-	P1, P2, P3	T1, T2, T4, T5, T7, T9
	Desire to keep monitoring emotions		ALL (S1–S10)	-	T9
Instrumentation UX	UX of the Wristband	positive feedback	ALL (S1–S10)	-	T5, T6, T7
		negative feedback	-	P2, P3	T7, T8, T9
	UX of the Self-reporting Tool	positive feedback	S1, S4, S6, S8, S9	P2	T4, T5, T7
		negative feedback	S1, S2, S3, S4, S6, S10	P1, P2, P3	T1, T2, T4, T5, T6, T7, T8
	UX of EmoVizPhy	positive feedback	S2, S3, S5, S6, S7, S10	-	-
		negative feedback	S2, S6, S8, S10	P3	T8, T9
	Recommendations for improvement		S1, S2, S3, S4, S6, S8, S9	P2	T7, T8, T9

theme include *Willingness to Share Emotions* and *Privacy*. In the original study, most participants (students and professionals) expressed a positive attitude toward the *willingness to share emotions* during retrospectives, emphasizing the importance of honesty and openness in team dynamics. In our replication, two trainees expressed willingness to share emotions, particularly when they felt it was relevant to the team's progress: *"I always express my emotions. I always tell everything because it helps the team to progress and become better."* (T8). However, three trainees reported being selective in sharing emotions, especially when they felt the issues were not critical or could disrupt team harmony. For example, one participant reported: *"Yes, I chose not to disclose some emotions. I was worried that sharing these feelings might worsen the situation and make things more complicated for the team. I thought it was better to keep those feelings to myself to avoid potential conflicts."* (T5).

As for *privacy* concerns, the original study reported they were minimal, with all students and one professional developer expressing no concerns about the collection of biometric and emotional data. Similarly, in our replication, trainees reported no concerns about how their data was handled during the study; however, they raised concerns about potential application in real-world scenarios where peers or managers could access the data. This suggests that, despite trusting the research process, participants were skeptical about sharing their data in real-world usage of the tool: *"If I'm the team leader, I want to get the data, but as a member, I'm not sure about my team leader seeing it. It shows activities outside work, so I don't think it's necessary to show it to the PM."* (T8).

Emotion Representation This theme addresses participants' perceptions of how their emotions were represented through biometric signals and self-reported data. The original study included three codes under this theme, i.e. *Interpretability of biometrics*, *Peak-event association*, and *Importance of notes*, which we extended with the new codes *Self-report association* and *Importance of multimodality*.

For the code *Interpretability of biometrics*, participants in the original study had mixed experiences: some found the biometric visualizations easy to understand, others struggled due to a lack of familiarity with the data. In our replication, only one participant reported positive feedback, recognizing a connection between EDA peaks and physical movement or potentially engaging activities: *"Most of the time, my visualizations show a steady line, except during physical activity or during that presentation. Also, I've come to realize that I might remain too calm during regular work periods."* (T8).

Consistent with the original study, regarding the code *Peak-event association*, most of the trainees were able to link EDA peaks to stressful events, such as meetings or deadlines. For example, T4 stated: *"Yes, there was a time when we had a meeting, and we ended up having a heated argument. I remember that clearly because it was a very important meeting. We were discussing some critical issues, and the stress level was very high, which showed up as a peak in the data."*. None among the trainees reported difficulties in associating peaks with specific events.

The new code *Self-report association* refers to participants' ability to link self-reported emotions with specific events or activities during the sprint. *While Peak-event association* concerns the recall of events triggered by EDA peaks, Self-report association captures the analogous process starting from self-reported emotions, shown as yellow dots in EmoVizPhy. Trainees found it feasible to associate the self-reports with a specific event, helping them to recall explicit details about what had happened. For example, one trainee noted: *"There was an event that was very intense, and the self-report matched the high-stress peak I saw in the data."* (T4). This code did not appear in the original codebook, which focused on the association between biometric peaks and events; its spontaneous emergence in our replication indicates that participants actively engaged with self-reports as an independent channel for reconstructing emotional episodes, beyond their role as contextual annotations of biometric signals.

Participants also emphasized the *importance of notes* in recalling emotional episodes, reinforcing the original finding that contextual information enhances the interpretability of biometric signals: *"I would suggest making the notes in the pop-up mandatory instead of optional. It would be helpful to correlate the events with the data."* (T7).

Although discussed by only one participant, the new code *Importance of Multimodality* emphasizes the value of combining biometric data and self-reported emotions. The trainee noted that, while self-reports were useful for recalling specific events, biometrics provided objective insights into emotional states: *"I think [biometrics and self-reported emotions] are useful in different ways. For recollecting events that happened, the self-report is more useful because it's more explicit about what's been happening. But for understanding myself and my responses to stress better, I think the bracelet is more effective. The self-report is something you write yourself, while what's registered by the bracelet just happens to you. So it's less biased in that way."* (T9).

Perceived usefulness for retrospectives This theme captures participants' views on the usefulness of the emotion self-monitoring practice in improving the effectiveness of retrospective meetings. It includes codes that highlight the specific benefits of the proposed tools (e.g., *Focus and productivity boost*, *Improved card writing*, *Improved action proposals*, and *Improved emotional awareness*). Additionally, it features a code outlining potential *Promising use cases* for EmoVizPhy and another reflecting participants' *Desire to continue monitoring emotions* in the future.

This is the theme where the most striking divergence from the original study emerges. As for *Improved Card Writing*, while the original study reported predominantly positive feedback from students, all nine trainees in our replication reported that using the tool did not improve their card writing during retrospectives nor the action proposal. Instead, they considered the tool more useful for personal reflection rather than team discussions. One trainee explained: *"It's good to have that as a support tool that tells me, 'you were feeling this way or you might have been stressed this much during the week'. However, I didn't really think of it when I was talking about or participating in the retrospective. It was more of a personal reflection tool for me. I relied on it to validate my feelings and stress levels, but it didn't heavily influence my contributions or feedback during team discussions. Having that external validation was reassuring, but I still based my retrospective input primarily on my personal recollection and interpretation of events."* (T5).

As for the code *Improved Emotional Awareness*, several trainees reported positive feedback, stating that the tool helped them better understand their emotional patterns and stress levels: *"I didn't know that the effect of, for example, bad time management in my life, in my mood is too much. [. . .] It's more important than I was thinking about it. So I should consider the timing that I'm going to bed and I'm waking up."* (T6). This aligns with the original finding that self-monitoring tools can enhance emotional awareness, helping participants gain new insights about themselves. However, few trainees reported negative feedback regarding *Improved Emotional Awareness*, in particular due to difficulties interpreting the biometric data: *"No, partly because I don't know how to evaluate the severity of what the data is telling me. Like if I see a big, spike on the EDA data, okay, I was stressed, but I don't know how I was stressed or if I should do something about it. "* (T5).

Most of the trainees identified several *Promising use cases* for EmoVizPhy, particularly for self-awareness and onboarding new team members: *"I think individuals who are already self-aware may not need the tool to express their feelings. However, for those who aren't naturally inclined towards self-reflection, the tool could be valuable across various software development activities, including sprint planning and retrospectives."* (T9). This reinforces the finding that EmoVizPhy supports emotional self-reflection and team-wide sharing among novices less familiar with agile practices.

Finally, similar to the original study—in which all students expressed the *Desire to keep monitoring emotions*—one trainee in our replication also expressed interest in continuing to monitor their emotions. This suggests that, despite the challenges, participants see value in using self-monitoring tools for personal growth and emotional awareness: *"I can see how it might be personally useful for self-reflection."* (T9).

Instrumentation UX. In line with the original study, we investigated themes concerning the user experience (UX) of the instrumentation, including the EmbracePlus wristband, the self-reporting tool, and EmoVizPhy.

Feedback on the *wristband* was mixed, in line with the discomfort reported in the original study. One trainee noted: *"It was hard to get it tight enough that it would stay in place. But also not be annoying or too tight. Getting that balance right was hard."* (T9).

As for the *self-reporting tool*, the original study found that while some participants appreciated its journaling aspect, many found the frequent pop-ups disruptive. Similarly, trainees valued the tool for prompting reflection, while criticized it for interrupting their work. One trainee commented: *"The pop-up application was quite annoying for me because it interrupted my work. If I was in the middle of something or* *if I was pair programming and someone else was watching my screen, I felt a bit awkward."* (T7). Another offered a counterpoint: *"Sometimes the interruption was bad, but sometimes it was good because it made me take a break, which I normally don't really have."* (T5). This reinforces the original finding that self-reports' intrusiveness can negatively impact workflow, suggesting future tools may prompt users during natural breaks.

Regarding the *UX of EmoVizPhy*, the original study reported mixed experiences, with some struggling to interpret biometric signals. In our replication, two trainees similarly found it hard to interpret, and one suggested interface improvements. This aligns with the original findings on the need for better visualization guidelines. Participants also offered further *Recommendations for Improvement*. Echoing the original study's call for more contextualized insights, one trainee suggested filtering out data from physical activities: *"For me, I think you should remove the physical activities from the data so I can see better. [. . .] It can help if this is your first experiment."* (T8). Another recommended making notes mandatory in the self-reporting tool: *"Making the notes optional is probably not good. Because then the person will not enter any description, so they might not be able to correlate better with their data because it's not easy to remember it."* (T7). These recommendations highlight the importance of balancing contextual richness with usability.

B. Quantitative analysis

During the semi-structured interviews, we also posed a series of questions to the participants to triangulate and complement our qualitative analysis with their insights. The questions and answers from the nine participants are reported in Figure 2. While the wristband did not bother participants, the pop-up questionnaire was perceived as more disruptive: two participants reported being bothered "Some of the time" and one "Most of the time," with some admitting to occasionally skipping the prompts. This aligns with the original study, where the wristband was likewise seen as unobtrusive by students, and opinions on the pop-up were divided. It contrasts, however, with what was reported by the professional developers, who were all bothered by the wristband but found the pop-up less intrusive.

Regarding the participants' ability to link specific sprint events with the biometrics and self-reports displayed in the system, the answers showed mixed effectiveness across different data types. For associating EDA peaks with specific events, participants provided mostly positive feedback, with two responding "Seldom", two "Some of the time", and five "Most of the time". Self-report associations with specific events showed the most positive results, with only one participant responding "Seldom", four "Some of the time", and four "Most of the time". These results are in line with the students in the original study and more positive than those of the professional developers for EDA peaks, while self-report associations were comparably positive across all groups.

Regarding the helpfulness of the data visualization for writing notes during the retrospective, seven participants reported that it was "Never" helpful, while only one found it helpful "Some of the time" and another one "Most of the time". This finding diverges from the predominantly positive feedback reported by students in the original study, while it aligns with the limited benefit reported by the professional developers. For recalling emotionally intense episodes, the visualization showed moderate effectiveness, with two participants finding it helpful "Seldom", three "Some of the time", and two "Most of the time". Finally, regarding the disclosure of emotions during retrospective meetings, participants reported varied comfort levels, with two responding "Never", one "Seldom", five "Some of the time", and one "Most of the time". This pattern is consistent with the responses of the students in the original study, whereas all professional developers expressed willingness to share their emotions.

VI. Discussion

Comparison with the Original Study. The results of our analysis align with the original study for the *individual* dimension. In particular, we confirm the feasibility of associating biometrics and self-reports with specific events, the perception of improved emotional awareness in the experimental condition including the tool, the existence of promising use cases (especially onboarding and self-reflection scenarios). Also, we confirm the comfort with wearing the wristband reported by students participating in the original study, as well as the participants' mixed reception of the pop-up self-report tool.

In contrast, we could not confirm the *team-level* findings of the original study, in particular, the perceived improvement of card writing, the proposal of more (and more relational) corrective actions, and the high willingness to share emotions during the retrospective. The most notable divergence concerns the *Improved card writing* code: in the original study, students reported predominantly positive feedback, whereas in our replication, all nine trainees reported negative feedback. Unlike in the original study, privacy concerns emerged: six of nine participants raised explicit reservations about real-world use of the tool and self-monitoring protocol, though none had concerns about the study itself.

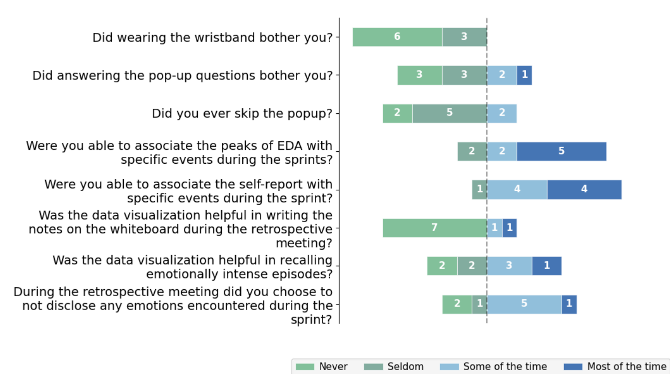

Fig. 2. Responses to the questionnaire administered during the interviews.

Emotional Awareness and Self-Reflection in Newly-Formed Teams. The trainees expressed strong appreciation for biometric data as a tool for personal reflection and emotional awareness: five of nine participants indicated improved emotional awareness when implementing self-monitoring of emotions (see Table II). Participants reported that the biometric visualizations helped them gain new insights into their emotional patterns, particularly in relation to stress and time management. This reinforces findings from the original study, showing that self-monitoring tools can be beneficial for novice developers or those new to agile practices, as they provide a structured way to reflect on emotional experiences [17], [45]. However, despite reporting the tool as useful for improved emotion awareness, the trainees did not report any beneficial impact on their contributions during retrospective meetings in terms of card writing or action proposal. This finding is in contrast with the original study, in which both students and professional developers benefited from emotion-monitoring tools in team retrospectives.

These results suggest that the effectiveness of such tools may be highly dependent on team maturity. Psychological safety is a plausible explanatory mechanism: our empirical evidence may be explained by the fact that the trainees were unfamiliar with one another and grouped by the course supervisor. At the moment of the study, the teams were still in the early stages of development, which is often characterized by uncertainty and limited trust [34]. In the experimental context, a climate of psychological safety had likely not yet developed. As highlighted by Alami et al. [46], openness and trust are crucial antecedents of psychological safety, and these elements may not have been fully in place in this setting. Indeed, the temporary nature of the teams (reforming every 11 weeks), cultural diversity, and lack of previous relationships created additional barriers to the open emotional sharing that would maximize the team benefits of emotion-monitoring tools. Some participants (T5, T6, T9) explicitly mentioned discomfort with sharing insights gained from their emotional data with the group, supporting this interpretation. This contrasts with the original study, where existing relationships and consolidated collaboration patterns appeared to facilitate emotional transparency.

Despite limited team-level impact, the use of EmoVizPhy and the implementation of self-monitoring practices demonstrated clear value for individual reflection. This suggests that, even when the group context is not ideal for open sharing, self-monitoring tools can still support personal learning and awareness. This finding suggests that emotional awareness tools may serve different functions at different stages of team development.

We therefore propose to refine the original hypothesis as follows: *emotion self-monitoring through EmoVizPhy supports retrospective meetings most strongly in newly-formed teams that have already developed a baseline of psychological safety*, a situation likely to be present in capstone student teams, where members are classmates from the same course, with similar backgrounds and no external project deadlines, but not necessarily in time-bounded professional training cohorts.

To bridge the gap between personal reflection and team discussions, future tools could incorporate features that encourage users to share insights from their self-monitoring data in a structured and non-intrusive way during retrospectives. Anonymized or aggregated data could be used to facilitate discussions about team-wide emotional patterns without compromising individual privacy [47].

Balancing Contextual Richness and Biometrics: Our empirical findings reveal the critical importance of combining multiple data sources for effective emotional awareness. The trainees consistently reported the feasibility of associating biometric peaks, typically indicating stress [35], with the corresponding emotional triggers. This extends the original findings by highlighting a new theme on the *importance of multimodality* in emotion-monitoring systems, with trainees reporting on the importance of self-reports and notes for recalling specific events associated with biometrics. In particular, the trainees emphasized that, while biometric data provided valuable insights into their stress levels and emotional states, self-reports were crucial for linking these biometric signals to specific activities or events.

Combining both data sources holds potential for enhancing emotional awareness, towards improving the quality of retrospective discussions. One participant explicitly articulated this distinction: *"For recollecting events that happened, the self-report is more useful because it's more explicit about what's been happening. But for understanding myself and my responses to stress better, I think the bracelet is more effective. [. . .] it's less biased in that way."* (T9). This observation suggests that different data modalities serve complementary purposes in emotional awareness development—self-reports for contextual understanding and biometrics for objective self-assessment. The emphasis on notes also suggests that future tools should explicitly support users in interpreting multimodal emotional data, not just visualize it.

Participants also discussed how self-reporting could be perceived as intrusive and time-consuming, suggesting that future tools should aim to minimize disruptions while still capturing meaningful contextual information. Future studies might include consideration of approaches for integrating calendar information [48] or automatically recognizing emotional states [22] in order to reduce the need for self-reporting through the pop-up application for emotion self-assessment.

Privacy concerns: While we implemented an empirical protocol involving minimal privacy issues within our research context, participants raised concerns about potential broader applications in professional settings. These concerns take on particular importance in professional training contexts, where power dynamics between supervisors and trainees, assessment pressures, and institutional data policies create additional complexity. Specifically, the trainees expressed concerns about the broader implications of sharing biometric and emotional data with peers or managers. One trainee (T8) explicitly worried about access by the team leader, suggesting how this concern extends to instructor-trainee relationships, where emotional data could potentially influence grading or team assignments.

These findings contrast with the original study, which reported that both students and professional developers were generally comfortable with emotion self-monitoring through self-report and biometrics. This might be due to the professional training context introducing unique ethical considerations not present in purely academic settings. Our protocol mitigates these concerns by design: data access remains strictly individual, with users privately reviewing their emotional feedback before any voluntary sharing. Nevertheless, future studies in educational or professional settings should explicitly communicate how privacy is addressed in the protocol.

A. Threats to Validity

In the following, we discuss threats to validity [49]. For each validity dimension, we present the trade-offs of corresponding decision points, the alternatives considered, the rationale for our choice, and its implications [50].

External validity is limited by the single training program and the relatively small sample (nine participants across three teams). Multi-site recruitment would have broadened diversity but introduced site-level variation confounding the experimental effect. The rationale was that within-program homogeneity isolates the configuration missing from the original study, i.e., newly-formed teams. The implication is that generalization is limited to similar training contexts and the small sample precludes statistical generalization. However, the authentic project constraints strengthen ecological validity over laboratory or course-based studies.

Construct validity is challenged by emotion-measurement issues and a decision point here concerns how to operationalize developers' emotions. We combined self-reports (SAM-based pop-up) and biometrics (EmbracePlus wristband) rather than relying on either alone. Self-reports are susceptible to social desirability bias, while biometrics suffer from movement artifacts and lack contextual grounding. The rationale was that triangulation addresses both weaknesses while preserving comparability with the original study. The implication is residual measurement noise in both channels; the emergence of the *Importance of multimodality* code (Section V) provides post-hoc support for this choice.

Internal validity may be affected by our choice to observe teams only during early formation (four one-week sprints), the phase targeted by the original hypothesis. As such, benefits in mature teams remain unobserved. Furthermore, we adopted a within-subjects design rather than between-subjects (infeasible with three teams) or crossover; the implication is that learning effects are not isolated. Finally, participation was voluntary to avoid coercion under supervisor power asymmetry. While voluntary participation may introduce a selection bias, the diversity of responses, including critical ones, suggests broad attitudes were captured.

Conclusion validity concerns the robustness of qualitative analysis. However, this was conducted by three independent coders reporting with Fleiss' κ = 0.85, indicating almost perfect agreement [44]. The quantitative analysis is descriptive, given the sample size. Unfortunately, we could not replicate the original quantitative analysis of cards and corrective actions due to the heterogeneity of retrospective formats and small team sizes, as discussed in Section IV-B, which poses a need for future investigations in this direction.

VII. Conclusion

This paper presented a replication of a previous study on the use of self-monitoring of emotions to support agile retrospective meetings. We replicated the original protocol with a population that, in the original authors' own framing, should have most strongly benefited from the proposed approach: newly-formed, interdisciplinary teams operating under authentic project constraints. Specifically, we ran the study with three teams of EngD trainees (nine participants in total) over four one-week sprints in a within-subjects design, using the same instrumentation as the original study, with the only technical change of replacing the discontinued Empatica E4 with the more recent EmbracePlus wristband.

Our findings reveal that, while emotional awareness practices provided individual benefits, they did not enhance team collaboration during retrospectives. Participants reported difficulties in sharing emotional insights with teammates in newly-formed teams lacking psychological safety, suggesting that individual and team benefits are distinct outcomes requiring different approaches. Taken together, the original study and this replication outline a more nuanced picture of when emotion self-monitoring supports agile retrospectives. The original authors hypothesized that EmoVizPhy is most beneficial for newly-formed teams, based on the contrast between novice student teams (positive results) and a consolidated professional team (limited benefits). Our replication, conducted on *newly-formed teams under authentic project pressure*, fails to replicate the team-level findings while confirming the individual-level ones, suggesting that team novelty alone is not sufficient for team-level benefits to materialize. Drawing conclusions across the original study and this replication, we proposed to refine the original hypothesis: emotion self-monitoring through EmoVizPhy supports agile retrospectives most strongly in newly-formed teams that have already developed baseline psychological safety, whereas benefits diminish when psychological safety is still absent (early forming/storming under external pressure) or no longer needed (mature performing teams). Our analysis also yielded two new codes extending the original taxonomy (*Self-report association* and *Importance of multimodality*) and revealed a rise in privacy concerns regarding real-world deployment.

Future replications should investigate the impact of emotional awareness tools in teams with established trust, further explore multimodal data-collection approaches, and develop stronger privacy-preserving approaches for team emotional discussions. Longitudinal studies examining the transition from individual to team benefits would provide valuable insights for future deployments of emotion-monitoring tools in software engineering contexts.

Acknowledgments

This research was co-funded by the Complementary National Plan PNC-I.1 – Research initiatives for innovative technologies and pathways in the health and welfare sector – D.D. 931 of 06/06/2022 ("DARE – DigitAl lifelong pRevEntion initiative", code PNC0000002, CUP B53C22006420001), and by the European Union – NextGenerationEU through the Italian Ministry of University and Research.

References

[1] D. Gachechiladze, F. Lanubile, N. Novielli, and A. Serebrenik, "Anger and its direction in collaborative software development," in *2017 IEEE/ACM 39th International Conference on Software Engineering: New Ideas and Emerging Technologies Results Track (ICSE-NIER)*, 2017, pp. 11–14.

[2] K. Madampe, R. Hoda, and J. Grundy, "A framework for emotion-oriented requirements change handling in agile software engineering," *IEEE Transactions on Software Engineering*, vol. 49, no. 5, pp. 3325–3343, 2023.

[3] M. M. Imran, P. Chatterjee, and K. Damevski, "Uncovering the causes of emotions in software developer communication using zero-shot llms," in *Proceedings of the IEEE/ACM 46th international conference on software engineering*, 2024, pp. 1–13.

[4] R. D. Lane and G. E. Schwartz, "Levels of emotional awareness: a cognitive-developmental theory and its application to psychopathology." *The American journal of psychiatry*, vol. 144, no. 2, pp. 133–143, 1987.

[5] A. Fountaine and B. Sharif, "Emotional awareness in software development: Theory and measurement," in *2017 IEEE/ACM 2nd International Workshop on Emotion Awareness in Software Engineering (SEmotion)*, 2017, pp. 28–31.

[6] M. Boekaerts and R. Pekrun, "Emotions and emotion regulation in academic settings," in *Handbook of educational psychology*. Routledge, 2015, pp. 90–104.

[7] D. Graziotin, X. Wang, and P. Abrahamsson, "Happy software developers solve problems better: psychological measurements in empirical software engineering," *PeerJ*, vol. 2, p. e289, 2014.

[8] S. C. Müller and T. Fritz, "Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress," in *37th IEEE/ACM International Conference on Software Engineering, ICSE 2015, Florence, Italy, May 16-24, 2015, Volume 1*. IEEE Computer Society, 2015, pp. 688–699.

[9] R. Pekrun and L. Linnenbrink-Garcia, "Academic emotions and student engagement," in *Handbook of research on student engagement*. Springer, 2012, pp. 259–282.

[10] A. Agrawal and T. Menzies, "Is "better data" better than "better data miners"?: on the benefits of tuning smote for defect prediction," 05 2018, pp. 1050–1061.

[11] F. Ebert, F. Castor, N. Novielli, and A. Serebrenik, "An exploratory study on confusion in code reviews," *Empir. Softw. Eng.*, vol. 26, no. 1, p. 12, 2021. [Online]. Available: https://doi.org/10.1007/s10664-020-09909-5

[12] M.-A. A. El-Migid, D. Cai, T. Niven, J. Vo, K. Madampe, J. Grundy, and R. Hoda, "Emotimonitor: A trello power-up to capture and monitor emotions of agile teams," *Journal of Systems and Software*, vol. 186, p. 111206, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016412122100279X

[13] K. Madampe, J. Grundy, M. Nguyen, E. Welstead-Cloud, V. T. Huynh, L. Doan, W. Lay, and S. Hashim, "Emoreflex: an ai-powered emotion-centric developer insights platform," *Automated Software Engineering*, vol. 32, no. 1, p. 22, 2025.

[14] A. Ruvimova, J. McGrenere, and T. Fritz, "Timemarbles: A more holistic approach to self-reflecting on focus in the knowledge workplace," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI 2026, Barcelona, Spain, April 13-17, 2026*, N. Oliver, D. A. Shamma, H. Candello, P. César, P. Lopes, A. Bozzon, T. Kosch, V. Q. Liao, X. Ma, V. Artizzu, F. Draxler, G. López, A. V. Reinschluessel, X. Tong, and P. O. T. Dugas, Eds. ACM, 2026, pp. 1532:1–1532:16. [Online]. Available: https://doi.org/10.1145/3772318.3791793

[15] A. N. Meyer, G. C. Murphy, T. Zimmermann, and T. Fritz, "Enabling good work habits in software developers through reflective goal-setting," *IEEE Transactions on Software Engineering*, vol. 47, no. 9, pp. 1872–1885, 2019.

[16] D. Grassi, F. Lanubile, N. Novielli, and A. Serebrenik, "Towards supporting emotion awareness in retrospective meetings," in *2023 IEEE/ACM 45th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*. IEEE Computer Society, 2023, pp. 101–105.

[17] D. Grassi, F. Lanubile, N. Novielli, L. Quaranta, and A. Serebrenik, "Self-monitoring of developers' emotions: the case of agile retrospective meetings," *ACM Trans. Softw. Eng. Methodol.*, Sep. 2025, just Accepted. [Online]. Available: https://doi.org/10.1145/3766064

[18] A. T. Aerts and Y. Dajsuren, "The pdeng program on software technology: Experience report on a doctorate level architecture training program," in *European Conference on Software Architecture*. Springer, 2020, pp. 247–262.

[19] N. Novielli and A. Serebrenik, "Sentiment and emotion in software engineering," *IEEE Software*, vol. 36, no. 5, pp. 6–23, 2019.

[20] P. Godliauskas and D. Šmite, "The well-being of software engineers: a systematic literature review and a theory," *Empirical Software Engineering*, vol. 30, no. 1, p. 35, 2025.

[21] N. Novielli and A. Serebrenik, *Emotion Analysis in Software Ecosystems*. Cham: Springer International Publishing, 2023, pp. 105–127. [Online]. Available: https://doi.org/10.1007/978-3-031-36060-2_5

[22] D. Girardi, N. Novielli, D. Fucci, and F. Lanubile, "Recognizing developers' emotions while programming," in *42nd Int. Conf. on Software Engineering (ICSE '20), July, 2020, virtual conference*, 2020, pp. 666–677.

[23] D. Girardi, F. Lanubile, N. Novielli, and A. Serebrenik, "Emotions and perceived productivity of software developers at the workplace," *IEEE Transactions on Software Engineering*, pp. 1–1, 2021.

[24] D. Grassi, F. Lanubile, A. Motca-Schnabel, and N. Novielli, "A cluster-based approach for emotion recognition in software development," in *2025 IEEE/ACM 18th International Conference on Cooperative and Human Aspects of Software Engineering (CHASE)*. IEEE, 2025, pp. 239–247.

[25] W. N. Awan, M. Paasivaara, and P. A. Gloor, "When emotions matter: Speech-based emotion detection in capstone project retrospective meetings while working hybrid," in *International Conference on Software Engineering (ICSE-SEET'26), April*, 2026, pp. 12–18.

[26] K. Madampe, R. Hoda, and J. Grundy, "Supporting emotional intelligence, productivity and team goals while handling software requirements changes," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 6, pp. 1–38, 2024.

[27] ——, "Addressing bad feelings in agile software project contexts: Considering team welfare and developer mental health," *IEEE Software*, vol. 41, no. 4, pp. 44–50, 2024.

[28] L. Cerqueira, S. Freire, D. F. Neves, J. P. S. Bastos, B. Santana, R. Spínola, M. Mendonça, and J. A. M. Santos, "Empathy and its effects on software practitioners' well-being and mental health," *IEEE Software*, vol. 41, no. 4, pp. 95–104, 2024.

[29] C. M. Montes, B. Penzenstadler, and R. Feldt, "The factors influencing well-being in software engineers: A mixed-method study," *ACM Transactions on Software Engineering and Methodology*, 2025.

[30] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "Affectaura: An intelligent system for emotional memory," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '12)*. ACM, May 2012, pp. 849–858. [Online]. Available: https://www.microsoft.com/en-us/research/publication/affectaura-an-intelligent-system-for-emotional-memory/

[31] E. Derby, D. Larsen, and K. Schwaber, *Agile retrospectives: Making good teams great*. Pragmatic Bookshelf, 2006.

[32] F. Fagerholm and T. Fritz, *Biometric Measurement in Software Engineering*, 2020, pp. 151–172.

[33] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy & Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[34] B. W. Tuckman, "Developmental sequence in small groups." *Psychological Bulletin*, vol. 63, no. 6, pp. 384—-399., 1965. [Online]. Available: https://doi.org/10.1037/h0022100

[35] J. H. Westerink, R. J. Rajae-Joordens, M. Ouwerkerk, M. van Dooren, S. Jelfs, A. J. Denissen, E. Penning de Vries, and R. van Ee, "Deriving a cortisol-related stress indicator from wearable skin conductance measurements: Quantitative model & experimental validation," *Frontiers in Computer Science*, vol. 2, p. 39, 2020.

[36] R. Kocielnik, N. Sidorova, F. M. Maggi, M. Ouwerkerk, and J. H. D. M. Westerink, "Smart technologies for long-term stress monitoring at work," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, 2013, pp. 53–58.

[37] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard, "Automatic identification of artifacts in electrodermal activity data," vol. 2015, 08 2015, pp. 1934–1937.

[38] D. Graziotin, X. Wang, and P. Abrahamsson, "Do feelings matter? on the correlation of affects and the self-assessed productivity in software engineering," *Journal of Software: Evolution and Process*, vol. 27, no. 7, pp. 467–487, 2015. [Online]. Available: https://doi.org/10.1002/smr.1673

[39] Parabol, Inc., "Dream team retrospective template," 2025, accessed: 2025-09-25. [Online]. Available: https://www.parabol.co/templates/sprint-retrospectives/dream-team/

[40] Teemood, "Mood-related retrospectives: How to use team mood data to plan your next sprint retrospective," 2025, accessed: 2025-09-25. [Online]. Available: https://blog.teammood.com/mood-related-retrospectives

[41] A. Tariq, P. Conrad, C. Hundhausen, A. Yu, and O. Adesope, "Improving agile retrospectives through metacognitive scaffolding," in *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, 2025, pp. 1106–1112.

[42] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, p. 77, 2006.

[43] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971. [Online]. Available: https://doi.org/10.1037/h0031619

[44] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: http://www.jstor.org/stable/2529310

[45] S. D. Kauer, *Emotional self-awareness and depressive symptoms: an investigation of an early intervention mobile phone self-monitoring program for adolescents PhD thesis School of Psychological Science*. The University of Melbourne, 2012. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1fe328cceb9caa6b829c826be6250a3ab58cf988

[46] A. Alami, M. Zahedi, and O. Krancher, "Antecedents of psychological safety in agile software development teams," *Information and Software Technology*, vol. 162, p. 107267, 2023.

[47] E. K. Choe, B. Lee, H. Zhu, N. H. Riche, and D. Baur, "Understanding self-reflection: how people reflect on personal data through visual data exploration," in *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, ser. PervasiveHealth '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 173–182. [Online]. Available: https://doi.org/10.1145/3154862.3154881

[48] M. Züger, S. Müller, A. Meyer, and T. Fritz, "Sensing interruptibility in the office: A field study on the use of biometric and computer interaction sensors," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, 2018, pp. 1–14.

[49] C. Wohlin, P. Runeson, M. Hst, M. C. Ohlsson, B. Regnell, and A. Wessln, *Experimentation in Software Engineering*. Springer Publishing Company, Incorporated, 2012.

[50] M. P. Robillard, D. M. Arya, N. A. Ernst, J. L. C. Guo, M. Lamothe, M. Nassif, N. Novielli, A. Serebrenik, I. Steinmacher, and K.-J. Stol, "Communicating study design trade-offs in software engineering," *ACM Trans. Softw. Eng. Methodol.*, vol. 33, no. 5, Jun. 2024. [Online]. Available: https://doi.org/10.1145/3649598