2022 IEEE/ACM 44th International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)

2022 IEEE/ACM 44th International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET) | 978-1-6654-9592-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSE-SEET55299.2022.9794177

Seeking Flow from Fine-Grained Log Data

Benjamin Ultan Cowley
University of Helsinki
Helsinki, Finland
ben.cowley@helsinki.fi

Arto Hellas
Aalto University
Espoo, Finland
arto.hellas@aalto.fi

Petri Ihantola
University of Helsinki
Helsinki, Finland
petri.ihantola@helsinki.fi

Juho Leinonen
University of Helsinki
Helsinki, Finland
juho.leinonen@helsinki.fi

Michiel Spape
University of Helsinki
Helsinki, Finland
michiel.spape@helsinki.fi

ABSTRACT

Flow is the experience of deep absorption in a demanding, intrinsically-motivating task conducted with skill. We consider how to measure behavioural correlates of flow from fine-grained process data extracted from programming environments. Specifically, we propose measuring affective factors related to flow non-intrusively based on log data. Presently, such affective factors are typically measured intrusively (by self-report), which naturally will break the flow. We evaluate our approach in a pilot study, where we use log data and survey data collected from an introductory programming course. The log data is fine-grained, containing timestamped actions at the keystroke level from the process of solving programming assignments, while the survey data has been collected at the end of every completed assignment. The survey data in the pilot study comprises of Likert-like items measuring perceived educational value, perceived difficulty, and students' self-reported focus when solving the assignments. We study raw and derived log data metrics, by looking for relationships between the metrics and the survey data. We discuss the results of the pilot study and provide suggestions for future work related to non-intrusive measures of programmer affect.

CCS CONCEPTS

• **Social and professional topics** → *Computing education.*

KEYWORDS

flow, stress, log data, survey data, difficulty, educational value, focus

ACM Reference Format:
Benjamin Ultan Cowley, Arto Hellas, Petri Ihantola, Juho Leinonen, and Michiel Spape. 2022. Seeking Flow from Fine-Grained Log Data. In *44nd International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET '22), May 21–29, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3510456.3514138

ICSE-SEET '22, May 21–29, 2022, Pittsburgh, PA, USA

ACMISBN978-1-6654-9592-9/22/05...$15.00
https://doi.org/10.1145/3510456.3514138

1 INTRODUCTION AND BACKGROUND

Having achieved the experiential state of *flow* – an intrinsically-rewarding state of deep attentional involvement in a challenging task [31] – just to have it interrupted by an incoming call, email, notification from an open web application, or by someone knocking on the door asking "can I ask you a question?", you might skip the preferred action, for example answering "you already did". Regardless of the action you choose, the flow is gone and it will take time to regain focus. It might even be the case that the interruption led you to focus on an unnecessary but seemingly important task, which you cannot really do anything about, causing frustration and stress.

While the previous backdrop to our work comes from a situation that could have happened whether working remotely or at an office, the global challenge of the COVID-19 pandemic has made remote work and online tools with a wide variety of reminder and notification functionality central to our knowledge economy. However, this may have come at a cost: the pressures of balancing work and personal life can elicit acute stress in knowledge workers and students. Furthermore, the diminished social visibility has reduced the potential for organizations and universities to successfully adapt their practices to signs of stress, or for peer-groups to provide support, thereby preventing mental health problems from escalating. Therefore, it has become critical to actively monitor remote workers and students for signs of stress and proactively foster healthy task-engagement.

1.1 Flow and stress among (novice) developers

For novice programming students, Bosch et al. found that flow and confusion are the most common affective states students experience in their first programming learning session [2]. However, students also experience negative states such as boredom and frustration. Experiencing flow during programming was positively correlated while confusion was mostly negatively correlated with performance in Bosch et al.'s study. Bosch et al. did not study stress directly, and found that anxiety – which perhaps most closely relates to stress out of the affective states they studied – was rare. However, Leppink et al. found that college students in general had high levels of stress and that being stressed tends to lead to worse academic performance [27].

The state of the art in measuring the flow experience and stress in software engineering is based on intrusive methods. Stress level and flow state are estimated with questionnaires, sometimes augmented

with physiologic indicators such as heart rate variance, electrodermal activity, muscular tension, and salivary cortisol analysis [38]. These methods are good for research, but not for long-term monitoring. Moreover, physiological data are typically difficult to interpret without surveys. For example, valence can be measured by facial observation (e.g., with EMG – electromyography, or camera), but that alone is unreliable as, for example, smile muscles are also activated by grimace. Thus, there is a strong need to study non-intrusive flow and stress detection.

One option for non-intrusive affective state detection is analyzing typing patterns [9, 19]. Typing patterns are keyboard movement traces of a user that describe latencies between moving from one key to another. They can be used to accurately identify who is typing, also when the user is programming [25, 29]. Previously, Epp et al. [9] have suggested that keystroke dynamics can be used to identify emotional states, including relaxation and tiredness. Recently, stress has also been related to fine-grained log data such as mouse usage [39] and typing force [12]. Perhaps because of certain physiological and psychological similarities of stress and flow, the current nonintrusive prediction approaches rarely make distinction between these two. However, flow depends on the perceived balance between difficulty of the task and skills [17], and previous work in the context of computing education has used typing-pattern changes 1) to detect difficulty in software development [14], 2) to detect programmers' existing knowledge and performance [8, 22, 26, 40], 3) derived student time-on-task from typing [23], and 4) found out how nonlinear movement in the code while solving the task is often related to the experienced difficulty of the task [10]. Thus, it might be possible to estimate the challenge-skill antecedent of flow with similar log data as well.

1.2 Flow and stress as mental and physiological states

Flow and stress are strongly related mental and physiological states. Flow is defined as the mental state occurring when a person is so much concentrated on an activity that they lose track of time and awareness of the self. This can occur when the difficulty of a task matches or slightly exceeds the individual skills [31]. However, as illustrated in Figure 1 (left side), if skills exceed demands, anxiety or stress (c.f. [21]) may also occur. Physiologically, flow and stress are also similar, characterised by high physiological arousal, as indicated by increased sympathetic nervous system activation [32]. However, although flow is generally experienced as positive, accumulated stress-system activity caused by daily challenges is well known to have severe negative health effects [30].

Though similar task demands can therefore equally elicit flow and stress, these two states result from very different appraisals. Appraisals are high level summaries of cognitive load and affective states as either within one's capability and volition or outside (see the right side of Figure 1). In other words, appraisals determine whether a task is seen as 'challenging, not overwhelming, and a positive experience', thereby increasing the likelihood of 'flow' experiences. These therefore require self-regulation and executive control, so as to facilitate continuous engagement by allocating cognitive resources (e.g. working memory, operation span) and inhibiting external interference [6, 33, 36]. Appraisals are, however, also determined by affective and motivational states, that can support or disrupt flow. For example, emotional dimension theory (e.g. [37]) suggests flow results from limited positive arousal and valence, stress from high arousal and low valence, and boredom from low arousal and negative valence. These, in turn, may undermine pre-existing approach motivations [34], potentially leading to distractibility and task disengagement.

Although appraisals in themselves are impossible to measure directly or continuously, the cognitive and affective mental states giving rise to stress and flow can be indexed using parallel psychophysiological recording and self reporting. Physiological measures can include, for example, recording brain activity using EEG (electroencephalography), skin conductivity (arousal) with EDA (electrodermal activity), heart rate (arousal) with ECG (electrocardiography), facial muscle activity (emotions) with EMG. EEG has been used to measure cognitive states (e.g. ERPs – event-related potential – detected spare capacity by detecting sentence predictability [13]), emotional states (e.g. frontal asymmetry indexes approach/avoidance motivation and anger [18]), and appraisal (e.g. theta oscillations in relation to appraisals [1]). Arousal, however, is better quantified using autonomous nervous system activity measurements provided by skin conductance (EDA) and heart-rate (ECG) measurements [16, 20]. Emotions are more readily detected from facial muscle activity (EMG), as even without awareness, discrete emotions elicit subcutaneous muscle activity from areas related to their expression on the face. Webcams may be used as a low-cost substitute, as they have been shown to offer a feasible, remote solution to detect micro-expressions (MEs; [28]), a substantial part of the same signal as EMG [5]. Furthermore, webcams have been used to detect flow in games [4], and mental workload [3]. Finally, to establish the diagnostic quality of the diverse range of measurements, one can use experiential self-report measurements (ESMs) to determine the degree to which flow and stress are experienced.

1.3 Our proposal: non-intrusive log based flow and stress detection

In this idea paper, we propose the development of non-intrusive methods for automatically inferring programmers' stress and flow from their working process. Our hypothesis is that appraisals of cognitive resources and affective states moderate the relationship between task demands and stress or flow outcomes. By obtaining a wide variety of psychophysiological and survey data in programming tasks with gradually increasing ecological validity, it could be possible to predict cognitive and affective states resulting in flow or stress. Simultaneously, we propose the collection of multimodal, fine-grained log data that is to be used as features from which to train machine learning models that can detect flow and stress unobtrusively. We believe one can develop a system that automatically recognises appraisals of psychophysiological states, so as to adapt task environments towards optimising flow and reducing stress. Such a system would benefit the future society by facilitating a healthy online work and study environment.

Our overarching hypothesis is that fine-grained log data can reveal the flow experience and stress related to software development. Examples of research objectives we propose in this idea paper are:

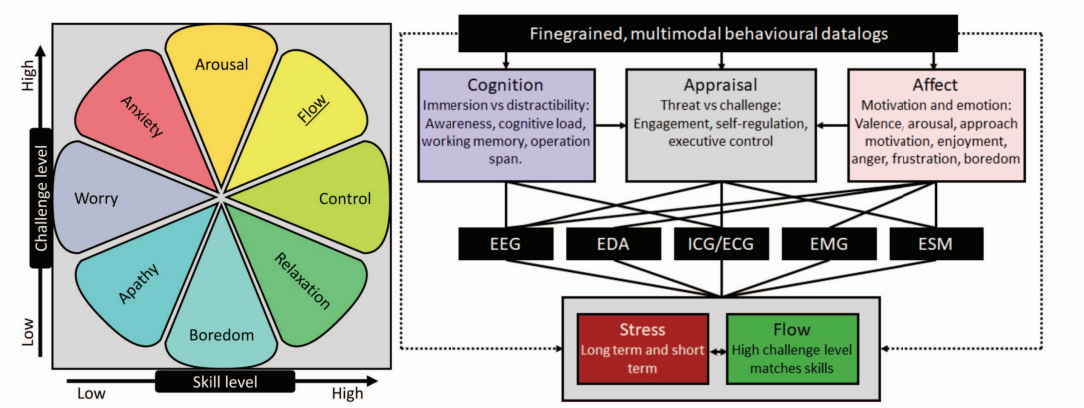

Figure 1: LEFT: Connection between skill level and difficulty (adapted from [7]). RIGHT: We propose using behavioural logs to detect cognitive and affective states that determine how perceived challenges can result in stress and flow, and validating the models using physiological measurements.

O1: How to model flow and acute stress? This question asks what cognitive and affective processes cause flow, and what differentiates flow from stressful high-intensity states in programming work? The model includes cognitive states (attention, control potential, cognitive load) and affective/motivational states (positive/negative affect, withdraw/approach motivation, discrete emotional states, and problem appraisals) to investigate their diagnostic quality as measures of flow and acute stress. The objective is to establish a model of causes and consequences of flow and stress, enhancing flow theory by embedding it in a more quantifiable biological framework.

O2: How to model behaviour using log data? The second question focuses on: What types of log data can be collected non-intrusively from learning environments and other environments used by novice programmers and how are the log data intertwined with each other? Potential data sources include, but are not limited to typing patterns, mouse data, active software, webcam-based psychophysiology, etc. The objective is to understand how different data sources complement each other to provide an illustrative data stream with minimal exposure to sensitive data.

O3: How to use log data and physiology to predict acute stress and flow? Given suitable models of flow and acute stress (after O1), this objective focus on the minimum-viable measurement properties of flow and acute stress and how well log data predicts it in the context of learning programming? The key focus is behavioural data, which is already often collected [15], although when creating such model, physiological data would also be needed.

O4: how to apply predictive models? Maintaining flow and reducing (negative) stress can yield higher productivity, improve workplace well-being, and reduce stress-related burnouts and learner dropouts. Objective 4 asks: how to use information on programmers' flow and stress to improve their day-to-day work life? The key focus is on constructing an interface that can be used in combination with other tools to aid programmers' work (e.g. silencing low-priority notifications in the case of flow).

Overall, the work on O1 could create more insight into the connection between flow and stress, as well as their fluctuations, which then could be used to adjust inputs used for flow and stress for O3. Work on O2 would lead more data for O3, which could increase our understanding of the link between various types of data collected from working environments and flow and stress. Work on O3, including transforming log data into derived metrics that better allow predicting flow and stress, could ultimately lead to applicable predictive models, which would be explored in O4. Together, the understanding formed through the research objectives could lead to a model that could be used to predict flow and stress relying solely on log data collected from the working environment and working process.

1.4 Pilot study

To start this line of work, we conducted a pilot study analyzing data collected from an introductory programming course. Our objective in the pilot study is to explore the feasibility of identifying flow from programming log data. Our research questions for the pilot study are:

RQ1. What is the relation between log data based features and student-reported affective metrics including focus during the task?

RQ2. What is the relationship between educational value, assignment difficulty, and focus in work?

Out of the larger scale research objectives, this pilot study relates to both O2 and O3. Related to O2, we explore different feature engineering options for log data based features, and related to O3, we examine the relationship between log data based features and student reported affective metrics: focus during the task, perceived educational value and assignment difficulty.

2 DATA AND METHODS

2.1 Data

For the pilot study, using an augmented IDE [41], we collected log data and self-reported data from an introductory programming course offered at the University of Helsinki. In the IDE, log data is collected from different student actions within the IDE when

they are working on course assignments. Logs are collected when students insert code (both typing at the keystroke level and pasting code into the editor), when they delete code, when they run their program, and when the editor window focus changes (editor window becomes active/inactive). In the course, when submitting a programming assignment that is assessed as correct by an automated assessment server, students are asked about the educational value and difficulty of the assignment, as well as how focused they were when working on the assignment. The questions were formed as Likert-like questions ranging from not at all (1) to very much so (5). In addition, as students were solving the programming assignments, log data was collected from the students as they were solving programming assignments.

From the log data for each assignment, we extracted the following raw log data based features for each student and assignment pair.

(1) Number of code inserts (keystrokes)
(2) Number of code pastes
(3) Number of code removes
(4) Number of project runs
(5) Number of editor window focus changes (i.e. switching from tab to another in the editor)

In addition to the log data features based directly on the raw log data, we process the raw log data into session related features. We derive session indicators that include the following log data based features for coding sessions with pauses that last no longer than one minute, five minutes, ten minutes, and thirty minutes. To determine a single session, we start counting backwards from the moment the student submits a working assignment which is the point when the survey data related to the affective factors is collected until the first pause that is longer than N minutes (where N is one, five, ten or thirty).

(1) Number of submitted assignments
(2) Number of worked on assignments
(3) Total duration of session

We only included data from students who were active in the course, i.e. had some log events and had completed at least one programming assignment. The data used for this study contained 167,668 completed assignments with answers to the Likert-like questions from a total of 2958 students. No background information on the students was available for this study.

The data was collected and used in accordance with the ethical processes outlined by the University of Helsinki.

2.2 Analysis

To answer the first research question, we performed correlation analyses using Spearman's rank correlation to study the relationship between the extracted log data based features and the self-reported affective features (educational value of assignment, difficulty of assignment, and focus when working on assignment). In the analyses, Bonferroni correction for multiple testing was performed.

To answer the second question, we studied both the correlations for the first research question as well as visually analyzed the relationship of the affective features.

3 RESULTS AND DISCUSSION

3.1 Relationship between log features and student-reported features

The relationship between the log features and student-reported features is described in Table 1. The correlations were calculated using Spearman's rank correlation, and all the included correlations are statistically significant ($p < 0.001$) after Bonferroni correction.

The correlations, although statistically significant due to the large number of samples, are mostly weak to negligible. The strongest correlations between the log features and the student-reported features are observed between the assignment difficulty and the raw metrics, which include statistics calculated from the events collected for the particular assignment. Here, the strongest correlation is observed between the number of window focus events and the assignment difficulty ($r = 0.36$), which suggests that more difficult assignments require more jumping back and forth between windows. Similarly, in general, it seems that the more difficult assignments require more work from the students, which is also understandable.

On the other hand, the raw metrics correlate relatively poorly with the educational value of the assignment and to what extent the student focused on the work. Correlations here are, in general, also positive – e.g. larger numbers of code inserts was linked with higher educational value.

When looking into the derived features that represent sessions, we observe that many of the correlations (despite being small to negligible) are negative. As an example, the metric "submitted assignments (5)" counts the number of unique assignments that the student has submitted without a break of at least 5 minutes for a specific student assignment pair (i.e., work on previous assignments is also counted to these features, if the work happens so that no pause of at least 5 minutes is observed in the log data). These observations in general suggest that the more assignments the student works on within a session, the lower the educational value, perceived difficulty, and focus is. On the other hand, the derived metric that describes the session length has a weak to negligible positive correlation with the affective metrics. In practice, this indicates that longer sessions suggest better educational value, assignment difficulty, and focus on solving the assignments.

This finding indicates that the sessions where the affective factors were rated the highest – most educational, difficult, and with the greatest focus – were longer sessions where students only worked on few assignments. One possible explanation is that students who are able to complete multiple assignments in quick succession are more experienced in programming beforehand, and thus both do not need to focus on the assignments as much, and do not perceive them as difficult or educational as their less experienced peers. Another potential explanation is that students who concentrate more during the assignment get more out of the assignment, thus rating it more highly regarding educational value and their focus during the assignment.

In addition to correlations, we also tested if multiple linear regression based on the log data could predict the self-reported focus. Even after step-wise feature selection the adjusted r-squared remained as low as 0.04. Using logistic regression to predict between low (values 1-3) and high (values 4-5) did not lead to meaningful

	EV	AD	FW
CODE INSERTS	0.19	0.31	0.14
CODE PASTES	0.11	0.23	0.06
CODE REMOVES	0.18	0.33	0.11
PROJECT RUNS	0.15	0.25	0.10
WINDOW FOCUS EVENTS	0.20	0.36	0.14
SUBMITTED ASSIGNMENTS (1)	-0.14	-0.13	-0.11
WORKED ON ASSIGNMENTS (1)	-0.08	-0.09	-0.04
TOTAL DURATION (1)	0.05	0.07	0.04
SUBMITTED ASSIGNMENTS (5)	-0.17	-0.16	-0.14
WORKED ON ASSIGNMENTS (5)	-0.12	-0.11	-0.09
TOTAL DURATION (5)	0.08	0.18	0.03
SUBMITTED ASSIGNMENTS (10)	-0.16	-0.15	-0.13
WORKED ON ASSIGNMENTS (10)	-0.12	-0.12	-0.09
TOTAL DURATION (10)	0.08	0.17	0.04
SUBMITTED ASSIGNMENTS (30)	-0.14	-0.14	-0.11
WORKED ON ASSIGNMENTS (30)	-0.12	-0.13	-0.08
TOTAL DURATION (30)	0.10	0.15	0.06
EDUCATIONAL VALUE	1.00	0.54	0.47
ASSIGNMENT DIFFICULTY	0.54	1.00	0.29
FOCUSED ON WORK	0.47	0.29	1.00

Table 1: Spearman correlations between the log derived features and student-reported features (EV = educational value, AD = assignment difficulty, FW = focus in work). The features with a number in parenthesis describe session features, where the extracted values represent data from a continuous session with no breaks lasting more than the number of minutes in the parentheses. The correlations are statistically significant (p-value < 0.001, Bonferroni corrected).

results. When 70% of the data was used for creating the model and 30% to test the model, overall accuracy of the model was 0.61. Accuracy of the majority vote baseline classifier was 0.584, signaling that the logistic regression performed poorly.

3.2 Relationship between student-reported features

To study the relationship between the student-reported affective features – perceived assignment difficulty and educational value and focus during work, we both look at the correlations (outlined in Table 1) and plot the features (shown in Figure 2). In general, the correlations between the student-reported features are stronger than the correlations between the student-reported features and log data features. These correlations are mostly moderate. The lowest correlation is between focus and difficulty ($r = 0.29$), while the strongest correlation is between educational value and difficulty ($r = 0.54$). The correlation between educational value and focus is moderate ($r = 0.47$).

We plotted the relationship of the three values to further study how they are related. The plot outlining the relationship of assignment difficulty, educational value, and focus is show in Figure 2. In general, we observe a positive correlation between all the variables. At the same time, with very difficult assignments (4-5 out of 5), we observe that the educational value does not increase regardless of focus. In fact, with lower focus (1-2 out of 5), the educational value of the assignment decreases if the assignment is considered very difficult as evidenced by the lower value for educational value for focus levels 1 and 2 when going from difficulty 4 to 5.

We do not know, however, whether this is a correlation or causation; it is possible, for example, that the students are distracted by something which then causes them to both lose focus and to perceive the assignments as having less educational value.

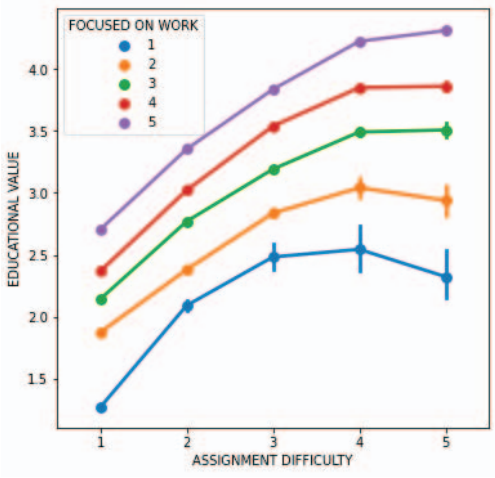

Figure 2: Students' self-reported affective factors: assignment difficulty on the x-axis, educational value on the y-axis, and focus on work in the plot. All three affective factors were measured on a 1-5 Likert scale and are averaged here over all assignments of the course.

4 CONCLUSION AND FUTURE WORK

In this idea paper, we proposed the development of non-intrusive methods of automatically detecting programmers' stress and flow from log data. As a pilot study on this topic, we explored data collected in an introductory programming course. In the pilot study, we examined the correlation between student self-reported affective factors and log data features. Our results suggest that the correlation between simple log features and student self-reported affective metrics is weak, which is in line with prior work by Henrie et al. who explored measuring engagement from log data [11]. In contrast with their work which did not find statistically significant relationships, however, we found the relationship between log data and student affective states statistically significant; possibly due to the larger amount of data available in this study. Our results also support earlier findings where it was suggested that certain types of data (e.g. IDE log data or survey data) correlate well within their own type (e.g. some IDE log data feature with another IDE log data feature) and worse across different types of data (e.g. IDE log data feature with a survey data feature) [24].

There are a few possible explanations for the finding that the correlation between self-reported focus and log data was weak. First and foremost, the data was collected from a programming course, where the participants likely do not know the topic yet and need

to switch back and forth between the programming environment and the learning materials, potentially also looking for help from other sources. In the pilot, we relied on data that was collected from the programming environment. Thus, we do not – for example – know about how and whether the students used other resources or programs. It is possible that to accurately measure flow and stress, one needs to collect multimodal data; for example, in addition to the log data from the programming environment, one could collect data related to information seeking behavior or collect observational data from classrooms with students solving the assignments [35]. Another potential avenue for further exploration is that there are student and/or assignment specific factors that we did not consider: in the pilot study, we averaged data over all students and all assignments. For example, some prior work has found that in the context of modeling student learning, there might be a sub-population of students who produce high quality data that would work better than using all student data [42]. Additionally, it is possible that with further feature engineering, i.e. transforming the raw log data into better features, we could have found stronger links between the affective factors and the log data derived features. Lastly, one possible explanation of course is that flow and stress cannot be accurately measured from non-intrusively collected data, such as log data as was the case in our pilot study.

One aspect of our future work is using physiological data to measure students' stress and flow in addition to self-reported data to possibly gain a more accurate view of the ground truth related to students' affective states. Additionally, we are interested in taking prior programming experience into account in the analysis as it is possible that flow and stress detection work differently for experienced and novice programmers. Lastly, regarding the analysis, we are going to explore more advanced data analyses than those conducted in the reported pilot study. For example, instead of looking at self-reported affect values directly, it is worthwhile to consider deviations in these, essentially taking students' basic affect levels into account, and similarly taking assignment specific factors such as difficulty into account in the analysis.

REFERENCES

[1] Ljubomir I Aftanas, Sergey V Pavlov, Natalia V Reva, and Anton A Varlamov. 2003. Trait anxiety impact on the EEG theta band power changes during appraisal of threatening and pleasant visual stimuli. *International Journal of Psychophysiology* 50, 3 (2003), 205–212.

[2] Nigel Bosch, Sidney D'Mello, and Caitlin Mills. 2013. What emotions do novices experience during their first computer programming learning session?. In *International Conference on Artificial Intelligence in Education*. Springer, 11–20.

[3] Frédéric Bousefsaf, Choubeila Maaoui, and Alain Pruski. 2014. Remote detection of mental workload changes using cardiac parameters assessed with a low-cost webcam. *Computers in biology and medicine* 53 (2014), 154–163.

[4] Andrew Burns and James Tulip. 2017. Detecting flow in games using facial expressions. In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 45–52.

[5] Eva Cerezo, Isabelle Hupont, Cristina Manresa-Yee, Javier Varona, Sandra Baldassarri, Francisco J Perales, and Francisco J Seron. 2007. Real-time facial expression recognition for natural interaction. In *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 40–47.

[6] Maurizio Corbetta and Gordon L Shulman. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience* 3, 3 (2002), 201–215.

[7] Mihaly Csikszentmihalyi. 1997. *Finding flow: The psychology of engagement with everyday life*. Basic Books.

[8] John Edwards, Juho Leinonen, and Arto Hellas. 2020. A study of keystroke data in two contexts: Written language and programming language influence predictability of learning outcomes. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 413–419.

[9] Clayton Epp, Michael Lippold, and Regan L Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the sigchi conference on human factors in computing systems*. 715–724.

[10] Ilenia Fronza, Arto Hellas, Petri Ihantola, and Tommi Mikkonen. 2019. An Exploration of Cognitive Shifting in Writing Code. In *Proceedings of the ACM Conference on Global Computing Education*. 65–71.

[11] Curtis R Henrie, Robert Bodily, Ross Larsen, and Charles R Graham. 2018. Exploring the potential of LMS log data as a proxy measure of student engagement. *Journal of Computing in Higher Education* 30, 2 (2018), 344–362.

[12] Javier Hernandez, Pablo Paredes, Asta Roseway, and Mary Czerwinski. 2014. Under pressure: sensing stress of computer users. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 51–60.

[13] Cynthia R Hunter. 2020. Tracking cognitive spare capacity during speech perception with EEG/ERP: Effects of cognitive load and sentence predictability. *Ear and hearing* 41, 5 (2020), 1144–1157.

[14] Petri Ihantola, Juha Sorva, and Arto Vihavainen. 2014. Automatically detectable indicators of programming assignment difficulty. In *Proceedings of the 15th Annual Conference on Information technology education*. 33–38.

[15] Petri Ihantola, Arto Vihavainen, Alireza Ahadi, Matthew Butler, Jürgen Börstler, Stephen H Edwards, Essi Isohanni, Ari Korhonen, Andrew Petersen, Kelly Rivers, et al. 2015. Educational data mining and learning analytics in programming: Literature review and case studies. *Proceedings of the 2015 ITiCSE on Working Group Reports* (2015), 41–63.

[16] Johannes Keller, Herbert Bless, Frederik Blomann, and Dieter Kleinböhl. 2011. Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology* 47, 4 (2011), 849–852.

[17] Johannes Keller and Anne Landhäu\s ser. 2012. The Flow Model Revisited. In *Advances in Flow Research*, Stefan Engeser (Ed.). Springer New York, New York, NY, 51–64. https://doi.org/10.1007/978-1-4614-2359-1_3

[18] Nicholas J Kelley, Ruud Hortensius, Dennis JLG Schutter, and Eddie Harmon-Jones. 2017. The relationship of approach/avoidance motivation and asymmetric frontal cortical activity: A review of studies manipulating frontal asymmetry. *International Journal of Psychophysiology* 119 (2017), 19–30.

[19] Agata Kołakowska. 2013. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 6th international conference on human system interactions (HSI)*. IEEE, 548–555.

[20] Elise Labbé, Nicholas Schmidt, Jonathan Babin, and Martha Pharr. 2007. Coping with stress: the effectiveness of different types of music. *Applied psychophysiology and biofeedback* 32, 3 (2007), 163–168.

[21] Richard S Lazarus and Susan Folkman. 1984. *Stress, appraisal, and coping*. Springer publishing company.

[22] Juho Leinonen. 2019. *Keystroke Data in Programming Courses*. Ph. D. Dissertation. University of Helsinki.

[23] Juho Leinonen, Francisco Enrique Vicente Castro, Arto Hellas, et al. 2021. Fine-Grained Versus Coarse-Grained Data for Estimating Time-on-Task in Learning Programming. In *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021)*. The International Educational Data Mining Society.

[24] Juho Leinonen, Leo Leppänen, Petri Ihantola, and Arto Hellas. 2017. Comparison of time metrics in programming. In *Proceedings of the 2017 acm conference on international computing education research*. 200–208.

[25] Juho Leinonen, Krista Longi, Arto Klami, Alireza Ahadi, and Arto Vihavainen. 2016. Typing patterns and authentication in practical programming exams. In *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*. 160–165.

[26] Juho Leinonen, Krista Longi, Arto Klami, and Arto Vihavainen. 2016. Automatic inference of programming performance and experience from typing patterns. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*. 132–137.

[27] Eric W Leppink, Brian L Odlaug, Katherine Lust, Gary Christenson, and Jon E Grant. 2016. The young and the stressed: Stress, impulse control, and health in college students. *The Journal of nervous and mental disease* 204, 12 (2016), 931–938.

[28] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. 2017. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE transactions on affective computing* 9, 4 (2017), 563–577.

[29] Krista Longi, Juho Leinonen, Henrik Nygren, Joni Salmi, Arto Klami, and Arto Vihavainen. 2015. Identification of programmers from typing patterns. In *Proceedings of the 15th Koli Calling conference on computing education research*. 60–67.

[30] Bruce S McEwen. 1998. Protective and damaging effects of stress mediators. *New England journal of medicine* 338, 3 (1998), 171–179.

[31] Jeanne Nakamura and Mihaly Csikszentmihalyi. 2002. The concept of flow. In *Handbook of positive psychology*, C. R. Snyder and Shane J. Lopez (Eds.). Oxford University Press, New York (NY), 89–105.

[32] Corinna Peifer, André Schulz, Hartmut Schächinger, Nicola Baumann, and Conny H Antoni. 2014. The relation of flow-experience and physiological arousal under stress—can u shape it? *Journal of Experimental Social Psychology* 53 (2014),

62–69.

[33] Reinhard Pekrun. 2006. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review* 18, 4 (2006), 315–341.

[34] Falko Rheinberg and Stefan Engeser. 2018. Intrinsic motivation and flow. In *Motivation and action*. Springer, 579–622.

[35] Ma Mercedes T Rodrigo and Ryan SJd Baker. 2009. Coarse-grained detection of student frustration in an introductory programming course. In *Proceedings of the fifth international workshop on Computing education research workshop*. 75–80.

[36] Ira J Roseman. 2013. Appraisal in the emotion system: Coherence in strategies for coping. *Emotion Review* 5, 2 (2013), 141–149.

[37] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[38] Nandita Sharma and Tom Gedeon. 2012. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer methods and programs in biomedicine* 108, 3 (2012), 1287–1301.

[39] David Sun, Pablo Paredes, and John Canny. 2014. MouStress: detecting stress from mouse motion. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 61–70.

[40] Richard C Thomas, Amela Karahasanovic, and Gregor E Kennedy. 2005. An investigation into keystroke latency metrics as an indicator of programming performance. In *Proceedings of the 7th Australasian conference on Computing education-Volume 42*. 127–134.

[41] Arto Vihavainen, Thomas Vikberg, Matti Luukkainen, and Martin Pärtel. 2013. Scaffolding students' learning using test my code. In *Proceedings of the 18th ACM conference on Innovation and technology in computer science education*. 117–122.

[42] Michael Yudelson, Steve Fancsali, Steve Ritter, Susan Berman, Tristan Nixon, and Ambarish Joshi. 2014. Better data beats big data. In *Educational Data Mining 2014*. Citeseer.