Analyzing Emotions in Conceptual Models Verification Tasks performed in Online Contests

Angela Mayhua-Quispe
Universidad Nacional de San Agustín de Arequipa
Arequipa, Peru
amayhuaq@unsa.edu.pe

Franci Suni-Lopez
Universidad Nacional de San Agustín de Arequipa
Arequipa, Peru
fsunilo@unsa.edu.pe

Nelly Condori-Fernandez
Computer Science Department
Universidade da Coruña
A Coruña, Spain
n.condori.fernandez@udc.es

Maria Fernanda Granda
Computer Science Department
University of Cuenca
Cuenca, Ecuador
fernanda.granda@ucuenca.edu.ec

***Abstract*—Emotion research in the area of software engineering has gained significant attention. Mostly this research has been focused on understanding the role of emotions in software programming carried out within collaborative software development environments. With the purpose of providing more evidence on emotion research in early stages of the software life cycle, in this paper, we report results of a live study conducted in competitive conditions. The main objective of the study is to analyze the emotions expressed by competitors, when perform verification tasks with the support of CoSTest, a model-driven testing tool. Our results show that participants tend to experience more positive emotions (*e.g.,* attentive, alert, active) than negative emotions (upset, hostile, afraid) when verification tasks are performed in an online contest.**

***Index Terms*—emotion, stress, verification, conceptual models, perceived usefulness**

I. Introduction

It is well known that stress has a negative impact on the quality of products, as it increases workers error rates [1]. Stress can also have a negative effect on a person's mood, which can influence on his/her working mode [2]. Consequently, it is reasonable to investigate the influence of emotions on this particular group of professionals related to software engineering. Several studies have been conducted to try to find the relationship between the software developer's emotions and productivity (*e.g.,* [3], [4], [5]). Most of the studies focus on understanding the role of emotions, by applying sentiment analysis to the textual developer-generated content in development environments [6]. However, the study of emotions in early stages of a software development process has not been yet well investigated. In [7], we proposed the design of a live study for analyzing emotions experienced by competitors, playing the role of software analysts, when verify conceptual models (CM) with the support of a testing tool, named CosTest [8]. Although the live study was accepted to run in SEmotion 2020[1], it had to be adapted since the event went virtual. For this purpose, we reuse and adapt the original live study proposal [7] regarding research questions and experiment procedure. In this paper, we report the first results of running our adapted live study run in two online contests. Also, we discuss some limitations that we have experienced during both competitions in virtual mode.

The remainder of this paper is organized as follows. In Section II, we describe the adaptation of our original live study design. Results are provided in Section III, where we report the flow of emotional states experienced by our competitors as well as their perceptions on usefulness of the testing tool used for detecting defects in CMs. In Section IV, we discuss the limitations in addressing validity threats of our study. Finally, conclusions and further work are discussed in Section V.

[1]https://semotion.github.io/2020/program.html

II. Study (Re-)Design

The present study is based on the design of the live study proposed for SEmotion [7]. However, as it could not be conducted by following the original plan, we had to reuse and adapt this proposal in terms of the objective, research questions and data collection procedure. This was because of the coronavirus disease 2019 (COVID-19) pandemic, which forced us to move from an In-Person workshop to a fully online event. In this section, we present the actual design used for running our experiment in competitive conditions.

A. Goal, Research Questions, Variables and Metrics

The objective of the experimental study is double: first, we aim to *analyze* the emotional states experienced by competitors when verify conceptual models with the support of a testing tool. Secondly, *evaluate* the perceived usefulness of a testing tool. From our objectives, the following research questions are derived:

> **RQ**$_1$: *How was the flow of emotional states during the correction of defects tasks?*
> **RQ**$_2$: *Is the testing tool perceived as useful to support the correction of conceptual models?*

From the research questions, the following variables were identified: **independent variables:** (i) the *CoSTest tool* that is used to automatically detect defects in conceptual models. (ii) The selected *conceptual models* (CM), and (iii) the *defects injected* into the CMs. As **dependent variables:** we identified the following variables: 1) *user emotional state* that is measured by self-report questionnaires based on the

International Positive and Negative Affect Schedule Short Form (I-PANAS-SF) [9], [10], the Visual Analogue Scale (VAS) for stress, adapted from [11]; and the Subjective Units of Distress Scale (SUDS) [12] (see Section II-C1 for more details); 2) *perceived usefulness* defined as the individual's perception to use the CoSTest tool for enhancing or improving her/his performance in correcting defects.

B. Participants and Experimental Context

The experiment involved 16 volunteer subjects (14 men and two women), who accepted an invitation to participate in a live study. 69% of the participants belong to a age range between 21 and 28 years old.

The live study was conducted in the context of a verification contest to correct defects of a set of conceptual schemes by using a testing tool, named CoSTest [8] to detect defects, and an UML editor to correct the conceptual schemas (*i.e.,* UML class diagram). Thus, a CM is considered *correct* if the model is absent of defects. For the contest, the winner was determined by the total number of corrected defects in the CMs in less time. The subjects have not prior domain knowledge of the artifacts (CM with defects), which were created by the researchers. But, a prior knowledge and experience on modeling UML class diagrams using tools or editors (*e.g.,* UML2Tools editor[2]) was highly required.

We run two contests: The first one was carried out as part of the SEmotion 2020 live study track[3], which included 2 participants (S15 and S16 in Table I). Due to this low number of participants, we decided to repeat the study in running the contest as part of an online mini-course. An invitation was sent to undergraduate students from Computer Science of the Universidad Nacional de San Agustin (Peru) and Universidad de Cuenca (Ecuador). The duration of the course was two hours[4]. Although the course counted with 33 attendees, only 14 participated in the competition (S1 - S14 in Table I).

C. Instrumentation

1) Questionnaire: We implement a web-based survey using Google Forms, which was composed by three sets of questions regarding:

- **Demographic data**; we ask about sex, range of age, educational background and domain expertise.
- **Emotion state**; where we use the following instruments: the I-PANAS-SF questionnaire that is a list of 10 adjectives used to describe different emotional states: 5 states of Positive Affect (PA) and 5 states of Negative Affect (NA). The PA scale measures activity and pleasure, while the NA scale relates to fear and stress [13]. The Visual Analogue Scale (VAS) is a measurement instrument for assessing anxiety/stress level [11] and was used at different moments during the contest with a 6-point scale. This instrument uses commonly a horizontal line to represent a range of values, from the minimum to the maximum value, so that subject marks a point on the range where he/she perceives his/her anxiety/stress state. Lastly, the Subjective Units of Distress Scale (SUDS) [12] used to measure the intensity of distress in the subject, this tool is rated on a scale from 0 (totally relaxed) to 10 (highest anxiety/distress that you have ever felt).
- **Experiment feedback**, a post-questionnaire that includes open questions about the instrumentation, the timing allocated for each phase, and complexity of the verification task.

2) Verification tasks: During the contest, the participants have to verify and correct defects in two conceptual models: CM1 which models a Super Stationery system, and CM2 that represents a Photography Agency system. Following the verification task proposed in [7], each CM has associated six test cases to be run using the CoSTest tool, each test case is used to verify the presence of one defect in the corresponding CM. After running the test cases, CoSTest displays a list of defects that can be corrected in any order. Participants have to re-run the test cases to verify the correctness of the changes. The participant will be focused first on CM1, having the possibility to upload his/her solution or skip to continue with CM2. For carrying out the contest, we provide the participants with the following material: (i) a virtual machine for VirtualBox[5] with all the software required for this contest, (ii) a brief description of each information system modeled in CM1 and CM2, and (iii) an example test suite (each one with six test cases) to be used during the training phase.

Regarding other materials, only the consent form was slightly modified with respect to the data collected since originally we had considered to use a wearable device for sensing physiological data and measuring stress in real-time.

D. Procedure

The adapted procedure is shown in Fig. 1 and its three main stages are explained below.

- *Preparation*: as the first stage, the details about the experiment were explained and the informed consent form was read and signed by the participants. Next, as the CoSTest could be a new tool for the participants, we give training for about 30 minutes that included instructions to configure the virtual machine and the execution of the verification process using a test CM. After training, we need to uniform the emotional state of all participants (*e.g.,* someone could come to the experiment already stressed) to avoid the influence of previous emotions; for that reason, participants are asked to stay quiet and watch a video to get relaxed, then they reported their current level of stress in LS_{t1} (see the *Preparation* module in Fig. 1).

[2]https://www.eclipse.org/modeling/mdt/?project=uml2tools
[3]https://kusisqa-project.github.io/semotion2020/
[4]https://kusisqa-project.github.io/costest2020/
[5]https://www.virtualbox.org/

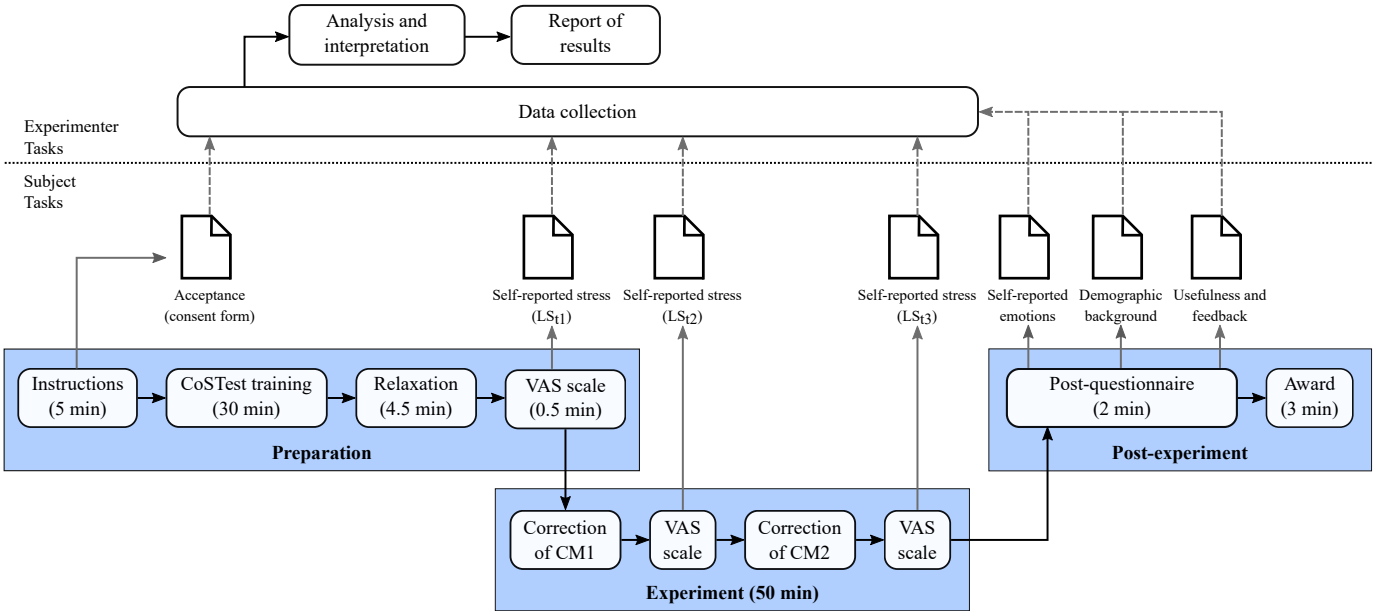

Fig. 1. Procedure carried out within a virtual contest. Adapted from [7].

- *Correction contest*: this stage takes about 50 minutes and its activities can be seen in the *Experiment* module of Fig. 1. All participants begin verifying and correcting defects of the CM1 using the provided resources (*e.g.,* test cases, description of the CM) and the CoSTest tool to list the defects (see Subsection II-C2 for more details regarding the verification tasks). It is important to remark that participants are able to submit their solution at any time as they consider within this stage. Once the solution is uploaded, the participant reports his perceived level of stress at that moment (LS_{t2}) and he has the option to continue with the CM2 executing the same process or skip it.
- *Post-experiment*: after finish the contest, participants are asked to complete a brief demographic questionnaire, self-response emotional questionnaires to report their perceived emotions during the contest, and give information on the perceived usefulness of CoSTest and feedback for improving the experiment (see Subsection II-C1 for more details about these questionnaires). After processing all submissions of the participants, we reward the participant who has corrected more defects during the contest.

The list of times that were originally allocated for each stage of the study is shown in Fig. 1, which was used only with participants from the online mini-course. For the live study run in SEmotion, the duration of the contest (experiment stage) was 30 min and the time of the CoSTest training was reduced to 15 minutes.

TABLE I
LEVEL OF STRESS (LS) REPORTED AFTER THREE STAGES BASED ON THE VAS SCALE AND THEIR CHANGES (CH) IN RELATION TO THE PREVIOUS STAGE. ALSO, THE LS REPORTED USING THE SUDS SCALE (LS_{su}).

ID	LS_{t1}	LS_{t2}	Ch_1	LS_{t3}	Ch_2	LS_{su}
S1	6	6	0	4	-2	1
S2	2	3	+1	4	+1	5
S3	2	4	+2	N/R	-	3
S4	3	3	0	3	0	7
S5	2	4	+2	N/R	-	7
S6	3	3	0	3	0	5
S7	2	3	+1	N/R	-	6
S8	2	2	0	1	-1	0
S9	3	3	0	N/R	-	5
S10	2	1	-1	N/R	-	1
S11	1	3	+2	N/R	-	10
S12	2	2	0	2	0	7
S13	3	5	+2	6	+1	10
S14	1	3	+2	5	+2	7
S15	1	3	+2	N/R	-	0
S16	1	4	+3	N/R	-	7
Avg	2.3	3.2		3.5		5
SD	1.2	1.2		1.6		3.2
Mode	2	3	0,+2	4	0	7

III. RESULTS

A. RQ1: Flow of emotional states

As can be seen in Fig. 1, there are three stages where subjects reported their level of perceived stress (LS): the first one after watching the relaxing video (LS_{t1}); and the last two ones after analyzing/correcting defects in CM1 and CM2 (LS_{t2} and LS_{t3}, respectively). Table I presents the data reported by the sixteen subjects (Sx). In the column LS_{t1}, we can observe that the 94% of subjects achieved a relaxed state (≤ 3), which can indicate us *i.e.,* the selected video was helpful for the relaxation stage. Then, focusing on the

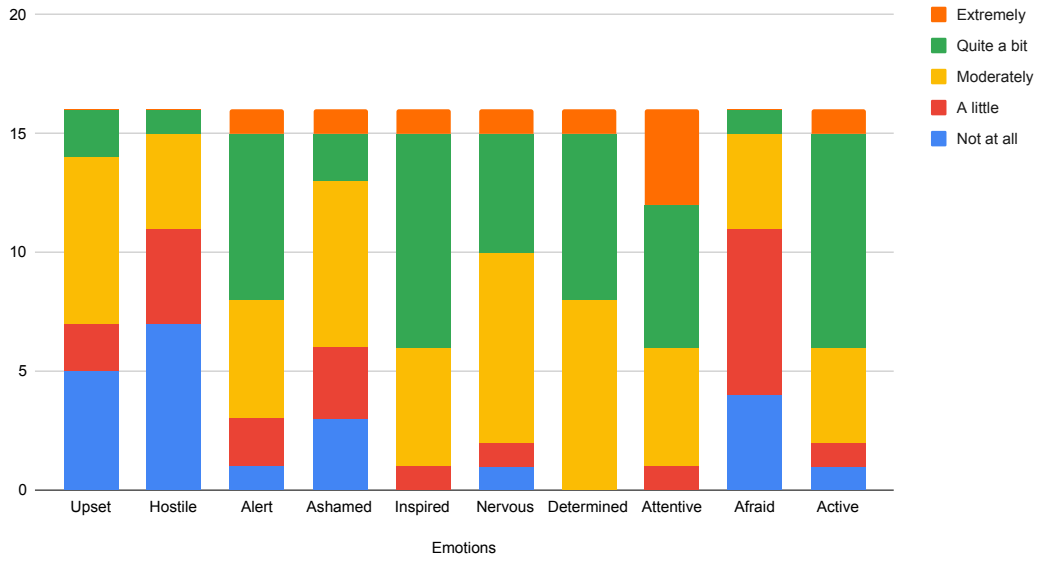

Fig. 2. Overview of different emotions perceived by the subjects (in a 5-point scale, based on the I-PANAS-SF [10]) for participating in the contest correcting defects.

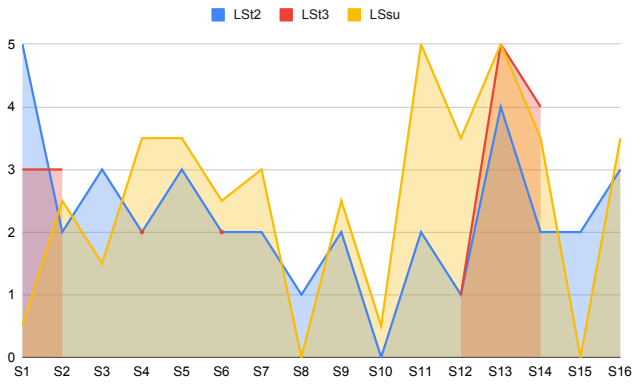

Fig. 3. Comparison of levels of stress (redimensioned to the same scale [0 - 5]) reported in LS_{t2}, LS_{t3}, and LS_{su} per each subject.

first change (Ch_1) from LS_{t1} to LS_{t2}, we found that most of the participants (except S10 who decreased one level) maintain (38% have 0 as value) or increase their level of stress (*e.g.,* 12% of participants increased their stress in one level (+1), and 38% did in two levels (+2)). This change could due to that analyzing and correcting defects in CMs for first time using a new testing tool can be a hard task. In column LS_{t3}, we note that half of participants did not provide a response (*i.e.,* N/R) because they decided to omit the correction of CM2 during the contest. This decision was mainly because of the lack of time for completing the task. However, for some of the participants who analyze/correct defects of CM2, their level of stress was maintained (0 is the value in Ch_2) in comparison with LS_{t2}.

Regarding the data collected after the experiment (see Fig. 1), we first examined the responses gave by the subjects using the I-PANAS-SF questionnaire [10]. It can be seen from the Fig. 2 that some negative affects (*e.g.,* upset, hostile, afraid) were not experienced by most of the participants during the contest (*i.e.,* answers are “a little" or “not at all"). However, the affect *nervous* was experienced very intensely because of the nature of the experimental context (*i.e.,* a competition). About positive affects (*i.e.,* inspired, determined, attentive, active, alert), all of them were experienced by the participants (*i.e.,* answers are “moderately", “quite a bit" or “extremely"), being the emotions *inspired* and *active* more frequent in the participants. This tendency to feel more positive affects than negative ones might be due to the type of task (correcting defects in the CMs).

Regarding the data collected through the question based on the SUDS scale [12], the scores assigned by the subjects can be seen in the last column of Table I. The question was formulated for analyzing the anxiety/distress perceived when the corrections were not successful after applying modifications in the CM. In order to compare the LS values gathered in each stage (during and after the contest), the corresponding scales were redimensioned to 5 points scale. Fig. 3 shows this comparison between LS_{t2}, LS_{t3}, and LS_{su}. From the figure, we can infer that some subjects (*e.g.,* S2, S5, S6, S9, S10, S13, S14, S16) perceived distress (LS_{t2} and LS_{t3}) due to that their attempts for correcting defects in the CM were not successful. We can also see that only four subjects (S4, S7, S11, S12) reported higher levels of distress than those that were reported during the contest (LS_{t2} and LS_{t3}). These observations suggest that most of the participants were more relaxed after uploading their solutions during the competition. Finally, S3, S8 and S15

experienced low anxiety/distress for bad solutions. Given that S3 and S8 were part of the organized online mini-course, we think that their low distress/anxiety might be due to that their motivation to participate in the contest was more on learning the testing tool than winning the contest. And in the case of subject S15, we think that it could be due to that the contest was carried out only with S16 as part of the SEmotion 2020 live study. Moreover, we realized that both were coworkers of the same institution.

B. RQ2: Perceived usefulness of a testing tool for verifying CMs

During the post-experiment stage, subjects answered six questions formulated in 5-point Likert scale (*i.e.,* from *strongly disagree* to *strongly agree*). These questions are related to whether the testing tool (*i.e.,* CoSTest) would be useful in their jobs. We first conducted a reliability analysis on these six questions. The reliability was conducted using the Chronbach alpha. The generic value obtained was 0.82, indicating that the items are reliable.

To answer RQ2, firstly the scores of each subject were averaged over the different items relevant for measuring PU. This way, we obtained a mean value for each subject. Then, for verifying whether these scores assigned by the competitors were significantly better than the middle score on the Likert scale for an item, we used the one-tailed sample t-test [6], which was applied with a significance level of 5 %, i.e., alpha = 0.05. According to the results shown in Table III, we corroborated empirically that the CoSTest tool is perceived as useful for performing their task (CM verification).

Finally, we calculated the frequency distribution on the scores given by the subjects to each item (see Fig. 4). From this figure, we observed that around 55% of subjects responded “agree" or “strongly agree". This means that participants are highly interested in using the tool in future activities. For instance, 69% of subjects agreed with the item I3: *Using CoSTest in my job would increase my productivity*. However, we also note that some participants were neutral, by answering "neither agree neither disagree". This neutral answer from this group of subjects could be due to a full interaction with the CoSTest tool might not have been completed. In this respect, it is important to remark that eight participants did not manage to finish the correction task of any CM.

IV. Threats to validity

A. Internal validity

This validity is related to factors in the experiments (*e.g.,* place, settings) that could affect the observed variables. We identified two possible threats: *experiment settings*; we mitigate this threat by performing both experiments in similar conditions for each participant (*i.e.,* material, verification tasks, rules of contest). For example, the settings of the required tools (*i.e.,* CoSTest and UML2Tools Editor) were pre-defined on the virtual machine for simplifying the software installation. However, despite we notified participants two weeks before the experiment, 19 subjects did not download the virtual machine, as consequence, they did not participate in the experiment. The other threat is about *emotions of participants before starting the experiment*; to mitigate this threat, we have planned a relaxing phase (*i.e.,* participants watched a video [7]) to uniform the emotions.

TABLE II
Inter-Item Correlation Matrix.

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**
Q1	1.000	.179	.256	.179	.434	.399
Q2	.179	1.000	.524	.574	.165	.382
Q3	.256	.524	1.000	.359	.398	.254
Q4	.179	.574	.359	1.000	.655	.676
Q5	.434	.165	.398	.655	1.000	.868
Q6	.399	.382	.254	.676	.868	1.000

TABLE III
One-Sample Test.

	Test Value = 3					
	t	**df**	**Sig. (2-tailed)**	**Mean Difference**	**95% Confidence Interval of the Difference**	
					Lower	Upper
PU	4.770	15	.000	.61458	.3400	.8892

B. Construct validity

The used instruments in the experiments are based on questionnaires with self-reported responses and consequently, participants could hide information about their emotional states or personal information; nevertheless, this threat is mitigated through our privacy and confidentiality terms that specify their information and responses are going to be anonymous. Furthermore, the selected instruments are well known and have been used in other works to measure emotions [14], [15]. Additionally, we used an inter-item correlation analysis to evaluate the construct validity of our response variable based on user perceptions (see Table II).

C. External validity

This issue is about the generalization of our results; a possible threat could be the selection of participants. However, we think this threat was mitigated by inviting participants (*i.e.,* attendees from SEmotion and mini-course) that have different personalities, experiences and educational backgrounds, such as master/PhD students, senior researchers, and practitioners from the Software Engineering community.

V. Conclusions and future work

The aim of the present research was to analyze the emotional states of competitors when correct defects in

[6]This statistical test was used because the data distribution was normal

[7]https://youtu.be/1La4QzGeaaQ

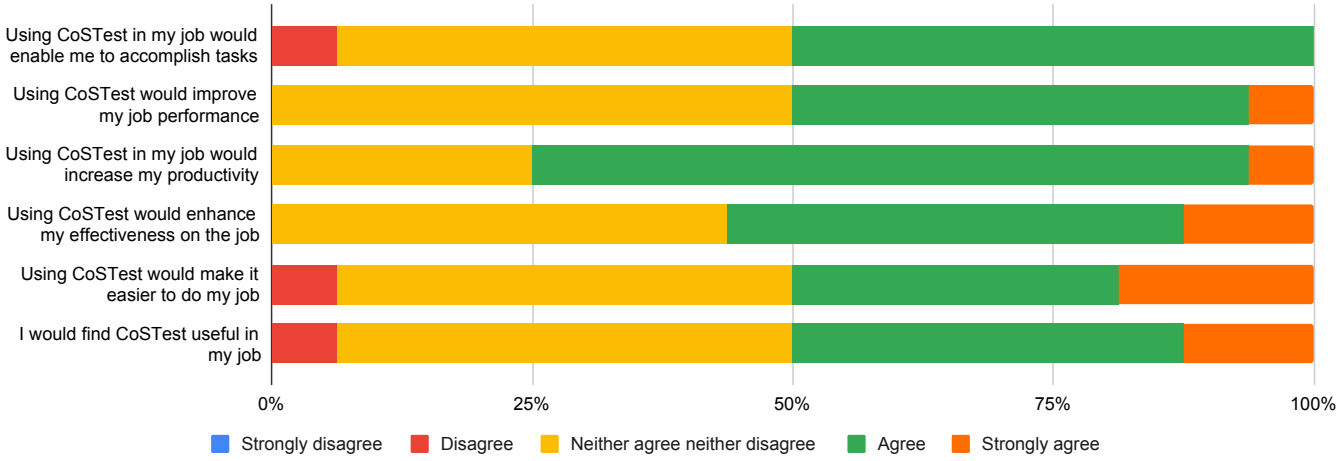

Fig. 4. Frequency distribution on the perceived usefulness by the subjects after using the CoSTest tool.

conceptual models, which can be detected with the support of a testing tool. This study has found that generally, the competitors can increase their levels of stress when a new testing tool is used for first time. Furthermore, the nervousness is also experienced very intensively due to the competitive nature of the contest. However, we have noted that the presence of positive emotions (*e.g.,* attentive, alert, active) is higher than negative ones in this type of contest because the task of correcting defects in CMs involves problem-solving skills.

Regarding the perceived usefulness, we corroborated empirically that the CoSTest tool was perceived as useful for performing CM verification. As one of the major limitations of this study was that emotions were measured only through a set of self-response questionnaires, further empirical research is needed to understand the influence of emotions on efficiency for verifying conceptual models within a Model-driven development context. We plan to replicate the study, by using wearable sensors not only for validating our stress detector [2], but also for creating our own public physiological dataset.

References

[1] J. Ostberg, D. Graziotin, S. Wagner, and B. Derntl, "A methodology for psycho-biological assessment of stress in software engineering," *PeerJ Comput. Sci.*, vol. 6, p. e286, 2020.

[2] F. S. Lopez, N. Condori-Fernandez, and A. Catala, "Towards real-time automatic stress detection for office workplaces," in *Information Management and Big Data*. Springer International Publishing, 2019, pp. 273–288.

[3] M. R. Wrobel, "Emotions in the software development process," in *2013 6th International Conference on Human System Interactions (HSI)*, Jun. 2013, pp. 518–523.

[4] D. Graziotin, F. Fagerholm, X. Wang, and P. Abrahamsson, "On the unhappiness of software developers," in *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, EASE 2017, Karlskrona, Sweden, June 15-16, 2017*, E. Mendes, S. Counsell, and K. Petersen, Eds. ACM, 2017, pp. 324–333.

[5] A. Serebrenik, "Emotional labor of software engineers," in *Proceedings of the 16th edition of the BElgian-NEtherlands software eVOLution symposium, Antwerp, Belgium, December 4-5, 2017*, ser. CEUR Workshop Proceedings, S. Demeyer, A. Parsai, G. Laghari, and B. van Bladel, Eds., vol. 2047. CEUR-WS.org, 2017, pp. 1–6.

[6] N. Novielli, D. Girardi, and F. Lanubile, "A benchmark study on sentiment analysis for software engineering research," in *Proceedings of the 15th International Conference on Mining Software Repositories, MSR 2018, Gothenburg, Sweden, May 28-29, 2018*, A. Zaidman, Y. Kamei, and E. Hill, Eds. ACM, 2018, pp. 364–375.

[7] A. Mayhua-Quispe, F. Suni-Lopez, M. F. Granda, and N. Condori-Fernandez, "How do negative emotions influence on the conceptual models verification? a live study proposal," in *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, ser. ICSEW'20. NY, USA: ACM, 2020, p. 581–585.

[8] M. F. Granda, N. Condori-Fernández, T. E. J. Vos, and O. Pastor, "Costest: A tool for validation of requirements at model level," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*, Sep. 2017, pp. 464–467.

[9] J. Karim, R. Weisz, and S. U. Rehman, "International positive and negative affect schedule short-form (i-panas-sf): Testing for factorial invariance across cultures," *Procedia - Social and Behavioral Sciences*, vol. 15, pp. 2016 – 2022, 2011.

[10] E. R. Thompson, "Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas)," *Journal of Cross-Cultural Psychology*, vol. 38, no. 2, pp. 227–242, 2007.

[11] V. S. Williams, R. J. Morlock, and D. Feltner, "Psychometric evaluation of a visual analog scale for the assessment of anxiety," *Health and Quality of Life Outcomes*, vol. 8, no. 1, p. 57, 2010.

[12] C. L. Benjamin, K. A. O'Neil, S. A. Crawley, R. S. Beidas, M. Coles, and P. C. Kendall, "Patterns and predictors of subjective units of distress in anxious youth," *Behavioural and Cognitive Psychotherapy*, vol. 38, no. 4, p. 497–504, 2010.

[13] U. Engelen, S. D. Peuter, A. Victoir, I. V. Diest, and O. Van den Bergh, "Verdere validering van de positive and negative affect schedule (panas) en vergelijking van twee nederlandstalige versies," *gedrag en gezondheid*, vol. 34, no. 2, pp. 61–70, Apr 2006.

[14] M. Wróbel, M. Finogenow, P. Szymańska, and J. Laurent, "Measuring positive and negative affect in a school-based sample: A polish version of the PANAS-c," *Journal of Psychopathology and Behavioral Assessment*, vol. 41, no. 4, pp. 598–611, Feb. 2019.

[15] B. A. Karanian, A. Parlier, V. Taajamaa, and G. Monaghan, "Engineering emotion : Students tell stories about the costs of being innovative," in *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, Oct. 2018.