Happy Software Developers Solve Problems Better: a Psychometrics-based Empirical Study

For several years in the context of software development, it has been claimed that the most significant quality and productivity gains are achieved by focusing on people. However, this claim has been rarely verified by Software Engineering research. Software Engineering is a young discipline, which is responsible to study the aspects of the software construction process. However, the research in the field faces the challenges that the software construction process is difficult to be studied under a purely Engineering viewpoint. Software construction is heavily dependent on human aspects and cognitive abilities. Lately, there has been an emerging awareness that a multidisciplinary viewpoint has to be adopted. More specifically, the psychometrics have been proposed in empirical Software Engineering research to study human aspects but still little research exists. It has been established that software developers solve problems through cognitive processing abilities. According to the research done in the field of Psychology, affective states - emotions, moods, feelings - deeply influence cognitive processing abilities and the performance of workers. In Software Engineering research, the affective states of software developers have been rarely investigated, in spite that affective states are a subject of several studies in Human Computer Interaction and in Computational Intelligence studies. The impact of affective states on the creativity and analytical problem-solving skills of software developers in general has not been investigated. Contending that the role of affective states on developers has been ignored in Software Engineering, there is a call for further research . In this paper, we report an experiment with 42 participants to investigate the impact of affective states on creativity and analytical problem-solving skills of software developers. The results offer support for the claim that happy developers are indeed better problem solvers in terms of their analytical abilities. The contributions of this study are (1) in understanding the impact of affective states on problem-solving skills of developers, (2) and in introducing psychometrics to be used in empirical Software Engineering studies.

Authors and Affiliations

Daniel Graziotin, Free University of Bozen-Bolzano, Faculty of Computer Science, Piazza Domenicani 3, 39100 Bolzano, Italy

Xiaofeng Wang, Free University of Bozen-Bolzano, Faculty of Computer Science, Piazza Domenicani 3, 39100 Bolzano, Italy

Pekka Abrahamsson, Free University of Bozen-Bolzano, Faculty of Computer Science, Piazza Domenicani 3, 39100 Bolzano, Italy

Introduction

For several years now, it has been claimed that a way to improve software developers' productivity and software quality is to focus on people (Boehm, 1990). The advocates of Agile software development stress this to the point that "If the people on the project are good enough, they can use almost any process and accomplish their assignment. If they are not good enough, no process will repair their inadequacy – 'people trump process' is one way to say this." (Cockburn & Highsmith, 2001). However, it is difficult to verify this claim empirically.

Software Engineering faces the challenging, yet beautiful difficulty that software development activities are rather different from industrial processes carried on by machines. Human aspects play a fascinating, complex role in the construction of software products, and the process itself is mainly intellectual (Fischer, 1987; Graziotin, Wang, & Abrahamsson, 2013; Khan, Brinkman, & Hierons, 2010). Lately, the discipline of Software Engineering is beginning to adopt a multidisciplinary view and to borrow theories from more established disciplines, like Cognitive Psychology, Human-Computer Interaction, and Management.

A proposal to satisfy the need to study human aspects in empirical Software Engineering research is using psychometrics (Feldt, Torkar, Angelis, & Samuelsson, 2008). While this call has been mildly echoed, the authors of this article note a limited research on the role of emotions and moods of software developers as humans.

As individuals, we act based on the emotions as we encounter the world through a mood. Affects enable the "mattering" of things. They are the medium within which acting towards the world takes place (Ciborra, 2002, pp. 159–165). Practically speaking, the ability to sense moods and emotions of software developers may be essential for the success of an Information Technology firm. (Denning, 2012). Thus, it is necessary to study the role of affective states on software developers. Affective states – emotions, moods, and feelings – are an important field of research in Psychology and Management studies, which are reference fields for this study.

Every day, software developers are required to solve problems. They need to plan strategies in order to find the only possible solution to a given problem or to generate multiple creative, innovative ideas. Among many skills required for software development, developers must possess high analytical problem-solving skills and creativity. Both of them are cognitive processing activities. Indeed, software development activities are accomplished through cognitive processing abilities (Khan et al., 2010).

Cognitive processing abilities are deeply linked with the affective states of individuals (Ilies & Judge, 2002), Research in Psychology, Cognitive Science, and Management studies the correlation between the affective states of individuals, their work-related performance, and their cognitive processing abilities, among them creativity and analytical problem-solving skills. No consensus has been reached on these linkages yet; however, there is a tendency to consider unhappy individuals to perform better in terms of analytical problem-solving abilities.

In Software Engineering research, the affective states of software developers have been rarely investigated, in spite that affective states are a subject of several studies in Human Computer Interaction and Computational Intelligence studies (e.g. (Lewis, Dontcheva, & Gerber, 2011; Tsonos, Ikospentaki, & Kouroupetrolgou, 2008)). The impact of affective states on the creativity and analytical problem-solving skills of software developers in general has not been investigated. Contending that the role of affective states on developers has been ignored in Software Engineering, there is a call for further research (Khan et al., 2010; Shaw, 2004).

This article serves for multiple purposes. (1) It provides a better understanding on the impact of the affective states on the creativity and analytical problem-solving capacities of developers. (2) It introduces and validates psychometrics on affective states, creativity and analytical-problem-solving skills in empirical Software Engineering. It (3) raises the need to study human aspects in Software Engineering by employing a multidisciplinary viewpoint.

The research objective of this study is: analyze analytical and creative problem-solving skills of software developers, with respect to their affective states, from the point of view of the researcher, in the context of a controlled environment (i.e., a computer laboratory) where M.Sc. and B.Sc. students in Computer Science perform tasks related to the object of study.

The results of this study indicate that the happiest software developers are also those who possess the highest analytical problem-solving skills. Affective states may not influence the creativity of software developers, but the data inspires future research avenues.

This article adheres to suggested guidelines when reporting controlled experiments in Software Engineering[1] (Jedlitschka & Ciolkowski, 2008; Wohlin et al., 2000) and it is organized as follows. In the remaining part of this section, the background research and the related work of this study are reviewed. An overview is provided on the theories behind affective states and skills for software development in Psychology and Cognitive Science. The constructs, psychometrics and the high-level hypotheses are reported. The Materials and Methods section presents the experimental design of the experiment. That is, the context, the participants' characteristics, the required instruments, the tasks of the experiment, and the analysis methods are presented. The Results and Discussion section describes the execution and the results of the experiment, including the descriptive statistics, the hypothesis testing, the interpretation of the results, their implications, and the lessons learned. The last section concludes the paper and suggests future research opportunities.

Related Work

This sub-section contains the related work on affective states of software developers and the literature review on the impact of affective states on problem-solving skills for software development. The constructs and the employed psychometrics are presented after the formulated hypotheses.

1 PeerJ requires different sections than those suggested by the literature. Therefore, the ordering and the position of the sections of this article are different from those of the followed guidelines. However, this article adheres to the guidelines in terms of content.

Affective States

It is difficult to differentiate terms like affective state, emotion, and mood. Emotions are states of mind that are raised by external stimuli and are directed toward the stimulus in the environment by which they are raised (Plutchik & Kellerman, 1980). Moods are affective states in which the individuals feels good or bad, and either likes or dislikes what is happening around them. Moods are believed to last longer than emotions (Parkinson, Briner, Reynolds S., & Totterdell, 1996). There is no clear agreement on a differentiation of the terms emotion and mood. Many authors consider mood and emotions as the same entity (Dow, 1992; Kaufmann & Vosburg, 1997; Parkinson et al., 1996). Therefore, this article adopts the term *affective states* as a general umbrella for emotion and mood. It has been previously adopted by the authors of this manuscript (Graziotin et al., 2013).

There are two main approaches to categorize affective states. One is called the discrete approach and seeks a set of basic affective states that can be uniquely distinguished (Plutchik & Kellerman, 1980). Examples include “interested”, “excited”, and “guilty”. The other approach groups affective states in major dimensions, which enable clear distinction among them (Russell, 1980). With this approach, affective states are characterized by their valence (pleasure), arousal (excitement, reacting from a stimulus, sensation of mental awakening), and dominance (control of a stimulus, over-learning). (Lewin, 1935; Morris, 1995; Russell, 1980).

The dimensional approach is mainly employed to assess affective states response triggered by immediate stimulus (Lewis et al., 2011; Morris, 1995). In other words, this methodology is suitable to measure how the individuals feel with respect to a stimulus (e.g., a graphical user interface) rather than assessing their general affective states. In this the discrete approach is employed as it is usually adopted when studying the impact of affective states on a participant’s characteristics, when the participant emotional reactions to external stimuli is not a concern - e.g., (Brand, Reimer, & Opwis, 2007; Forgeard, 2011).

Psychology studies often classify affective states as negative, neutral (less often), or positive according to the treatments to induce affective states or other psychometrics brought in by the measurement instruments, usually questionnaires. Several techniques exist to induce affective states on participants. The procedures are variegated: showing films to participants, playing some kind of music, showing pictures and photographs, let participants remember happy and sad events in their lives (Lewis et al., 2011; Westermann & Spies, 1996). That is, if the participants enjoy the stimuli, they will receive a positive mood induction effect. Recent studies question the effects of mood-induction techniques, especially when studying pre-existing affective states of the participants (e.g., (Forgeard, 2011)).

Affective States and Software Developers

There is limited research on the impact of affective states on software developers. To these authors’ knowledge, only two publications exist, which employ psychometrics to study the affective states of software developers.

Shaw (2004) observes that the role of emotions in the workplace is a subject of study in the management literature. Information systems research focuses on job outcomes such as stress, turnover, burnout, and satisfaction. However, little or no attention has been given to emotions of IT professionals. Shaw's work-in-progress paper reports a framework for studying affects in the workplace to investigate the fluctuation of moods of twelve software developers. The study shows that the affective states of a software developer may dramatically change during a period of 48 hours. However, the study is a work-in-progress paper and no continuation is known. The paper calls for research on the affective states of software developers.

This call is echoed by Khan et al. (2010). In the study, a correlation with cognitive processing abilities and software development is theoretically demonstrated. The authors construct a theoretical two-dimensional mapping framework in two steps. In the first step, programming tasks are linked to cognitive tasks. For example, the process of constructing a program – e.g. modeling and implementation – is mapped to the cognitive tasks of memory, reasoning, and induction. In the second step, the same cognitive tasks are linked to affective states. Two empirical studies on affective states and software development are then reported, relating a developer's debugging performance to the affective states. In the first study, affective states were induced to software developers, who were then asked to complete a quiz on software debugging. The second study was a controlled experiment. The participants were asked to write a trace on paper of the execution of algorithms implemented in Java. The study provides empirical evidence for a positive correlation with the affective states of software developers and their debugging performance. This study recommends more research on the topic.

Problem-solving and the Impact of Affective States

Among many theories, two ways of thinking for problem-solving have been distinguished in Psychology and Cognitive Science, namely divergent thinking and convergent thinking, which are strictly related to creativity and analytical problem-solving.

Divergent thinking leads to no agreed upon solutions and involves the ability to generate a large quantity of not necessarily correlated ideas (Csikszentmihalyi, 1997). Convergent thinking involves solving well-defined, rational problems that often have a unique, correct answer. It emphasizes speed and working out from what is already known. It leaves little room for creativity as the answers are either right or wrong (Cropley, 2006; Csikszentmihalyi, 1997). While divergent and convergent thinking are not proven to be synonyms of creativity and analytical problem-solving capacities - there are not clear definitions for these terms (Amabile, 1982) - these are the two dimensions that most studies analyze while still claiming to study creativity and analytical problem-solving (Csikszentmihalyi, 1997).

A creative performance can be either defined in terms of creative outcome or by the processes that lead to the creation of the creative results (Amabile, 1982; Davis, 2009). In this paper, creativity is studied in terms of *creative outcome* as it is the most explored in the referenced disciplines.

According to a recent meta-analysis on the impact of affective states on creativity (i.e., creative outcomes) in terms of the quality of generated ideas, positive affective states lead to higher creativity than neutral affective states, but there are not significant differences between negative and neutral affective states or between positive and negative affective states (Baas, De Dreu, & Nijstad, 2008). Another recent meta-analysis agrees that positive affective states have moderate effects on creativity with respect to neutral affective states. However, it shows that positive affective states also have small, non-zero effects on creativity with respect to negative affective states (Davis, 2009). Lewis et al. (Lewis et al., 2011) provide evidence for higher creativity under induced positive and negative affective states with respect to non-induced affective states. However, participants low in depression have higher creativity with induced negative affective states but no benefits have been found for individuals with induced positive affective states (Forgeard, 2011). Sowden & Dawson (2011) find that the quantity of generated ideas is boosted under positive affective states, but no differences in terms of quality is found.

Few studies have assessed the impact of affective states on the analytical problem-solving skills of individuals, and the understanding of the impact is still limited even in Psychology studies. Analytical problem-solving skills - thus, related to convergent thinking - are believed to be higher under the influence of negative affective states than positive affective states (Abele-Brehm, 1992; Melton, 1995). Somewhat contradictory to these studies, Kaufmann and Vosburgh (Kaufmann & Vosburg, 1997) show that negative affective states are significantly negatively correlated with analytical problem-solving skills, while positive and neutral affective states are not significantly correlated with analytical problem-solving. On the other hand, the processes of transferring, learning, and solving analytical problems deteriorate with negative emotions (Brand et al., 2007).

The literature review produced two high-level research hypotheses, which are further refined after the presentation of the constructs and the employed psychometrics:

- A difference is expected in the creative work produced by developers with respect to their affective states.
- A difference is expected in the analytical problem-solving skills of developers with respect to their affective states.

Constructs and Psychometrics

The most notable survey for the discrete approach is the Positive and Negative Affect Schedule (PANAS) (D. Watson, Clark, & Tellegen, 1988; David Watson, Clark, & Tellegan, 1988). It is a 20-item survey that represents positive affects (PA) and negative affects (NA). The scale received critiques in recent studies. Among them, it does not always represent real feelings and tends to capture only high-arousal feelings in general (Diener et al., 2009). There are recent, modern proposals to further reduce the PANAS scale and improve it, bringing the discrete theory closer to the dimensional approach.

To measure the affective states of the participants, this study opted for the Scale of Positive and Negative Experience (SPANE). SPANE is a modern proposal, which has been validated to converge to the previous affective states measurement instruments (Diener et al., 2009). SPANE is composed of 12-items, balanced in terms of negative experiences and positive experiences during the past 4 weeks. Among the other psychometrics derived from the survey, the Affect Balance Score (SPANE-B) is an indicator of the pleasant and unpleasant emotions caused by how often experienced happened to the participant. SPANE-B ranges from -24 (completely negative) to +24 (completely positive). Even if the SPANE-B score is a fuzzy indication of the affective states felt by individuals, it could be employed to split participants in groups using a median split. It is common to adopt the split technique on affective states psychometrics (e.g., (Forgeard, 2011)). Regression analysis is also possible if the data is suitable for it.

To the knowledge of the authors of this article, there are no studies in Software Engineering research, which define software development tasks suitable to measure the creativity and the analytical problem-solving skills of software developers. Although strict development tasks could be prepared, there would be several threats to validity. Participants with different backgrounds and skills are expected. It is almost impossible to develop a software development task suitable and equally challenging for first year BSc students and second year MSc students. For example, asking a first year BSc student to implement a design pattern may result in a negative performance no matter the affective states. For this reason, this study stays at a higher level of abstraction and measures their creativity and analytical problem-solving skills with validated tasks from Psychology and Cognitive Science.

There are different tasks that support a measurement of the creativity of the participants. The most adopted task is to generate creative ideas - e.g., the generation of captions for ambiguous photographs - or short solutions for uncommon and bizarre problems (e.g., (Forgeard, 2011; Kaufman, Lee, Baer, & Lee, 2007; Lewis et al., 2011; Sowden & Dawson, 2011)).

In order to measure the creativity, according to the Consensual Assessment Technique (Amabile, 1982), independent judges expert in the field of creativity score the captions using a Likert-item related to the creativity of the artifact to be evaluated. The judges have to use their own definition of creativity (Amabile, 1982; Kaufman et al., 2007). The Likert-item is represented by the sentence “This caption is creative”. The value associated to the item ranges from 1 to 7, where 7 is “I strongly agree” and 1 is “I strongly disagree”.

The scores for measuring creativity are the average of the scores of all the generated ideas (ACR), the best score for the ideas written by each participant (BCR) (Forgeard, 2011) (creators are often judged by their best work rather than the average of all their works (Kaufman et al., 2007)), and the number of generated ideas (NCR) (Sowden & Dawson, 2011).

Measuring analytical problem-solving skills is less problematic than measuring creativity. There is only one solution to a given problem (Cropley, 2006). The common approach is to assign points to analytical problem-solving tasks (Abele-Brehm, 1992; Melton, 1995). The approach employed in this study is to combine quality and quantity of results by assigning points to the achievements of analytical tasks and measure the time spent on planning the solution. The Tower of London game (a.k.a. Shallice’s test) is a game aimed to determine impairments in planning and executing solutions to analytical problems (Shallice, 1982). It is similar to the more famous Tower of Hanoi game in its execution. Figure 2 provides a screenshot of the game. The rationale when choosing this task is straightforward: the participants mentally prepare and execute a high-level algorithm in order to solve each presented problem.

The Analytical Problem Solving (APS) score is defined as the ratio between the progress score achieved in each trial of the Tower of London Game (TOLSS), and the seconds needed to plan the solution to solve each trial (PTS). TOLSS ranges from 0 to 36 because there are 12 problems to be solved and each one can be solved in maximum 3 trials. PTS is the number of milliseconds passed from the presentation of the problem to the first mouse click in the program. In order to have comparable results, a function to map the APS ratio to a range from 0.00 to 1.00 is employed.

Hypotheses

The following are the research hypotheses of this paper, which derive from the high-level hypotheses from the literature review. The alternative hypotheses are two tailed because the literature review does not provide clear indications. The average value of the constructs is denoted with the Greek symbol μ.

Hypothesis H1, H2, and H3 are on a difference in the best creative score (BCR), the average creativity score (ACR) and the number of generated creative solutions (NCR) produced by software developers feeling positive affective states and by software developers feeling non-positive affective states.

- $H1_0$: $\mu_{BCR_{N-POS}} = \mu_{BCR_{POS}}$ vs. $H1_A$: $\mu_{BCR_{N-POS}} \neq \mu_{BCR_{POS}}$
- $H2_0$: $\mu_{ACR_{N-POS}} = \mu_{ACR_{POS}}$ vs. $H2_A$: $\mu_{ACR_{N-POS}} \neq \mu_{ACR_{POS}}$
- $H3_0$: $\mu_{NCR_{N-POS}} = \mu_{NCR_{POS}}$ vs. $H3_A$: $\mu_{NCR_{N-POS}} \neq \mu_{NCR_{POS}}$

Hypothesis H4 is on a difference in the analytical problem-solving scores (APS) between software developers feeling positive affective states and software developers feeling non-positive affective states.

- $H4_0$: $\mu_{APS_{N-POS}} = \mu_{APS_{POS}}$ vs. $H4_A$: $\mu_{APS_{N-POS}} \neq \mu_{APS_{POS}}$

Materials and Methods

This section carefully described the empirical experiment. From the context and participants, the objects and instrumentation, the procedure and the analysis methods, the reader can evaluate the study and replicate it.

Context and Participants

The formal experiment is run in a controlled environment (a computer laboratory). The participants are obtained from the students of Computer Science. There are no restrictions in gender, age, nationality, or level of studies. The participation is voluntary and in exchange of credit points. The affective states are entirely random between participants.

An IRB approval for conducting empirical studies on human participants is not required by the institution. However, written consent was obtained from all subjects. Participants were advised (both informally and on the consensus form) about the data retained and that anonymity was fully ensured. No sensitive data has been collected in this study. Participants were assigned a random participant code in order to link the gathered data. The code is by now way linked to information, which would reveal the participant's identity.

Objects and Instrumentation

The objects required for the creativity task are six color photographs with ambiguous meanings. Figure 1 is one of the six photographs. They are available from the authors upon request. The material required for the analytical problem-solving task is provided by the open-source PEBL software (Piper et al., 2011). One computer per participant is required. The PEBL software automatically collects all the required data for measurement purposes.

Procedure

The participants arrive to the controlled environment only knowing that they participate to an experiment. As soon as they sit on their workstation, they read a reference sheet, which is included in Article S1. The sheet summarizes all steps of the experiment. The researchers also assist the participants during each stage of the experiment.

Participants are divided in two groups, namely N-POS and POS, by using a median split of the SPANE-B score only in the analysis phase of the experiment. Therefore, they are not aware of the group existence. They participate as single individuals to the experiment.

The experiment procedure is composed of four activities: (1) Affective states measurement (2) Creativity task (3) Affective states measurement (4) Analytical problem-solving task.

In the *creative generation task*, the participants receive two random photographs from the set of the six available ones, one at a time. Participants imagine participating to the “Best Caption of the Year” contest and try to win this contest by writing the best captions possible for these two photographs. They write as many captions as they want for the pictures. The full instructions given to the participants are the same employed by Forgeard (Forgeard, 2011).

In the *analytical problem-solving task*, the participants open the PEBL software. The test battery is set up to automatically display the Tower of London game, namely Shallice test ([1,2,3] pile heights, 3 disks, Shallice’s 12 problems). The PEBL software displays specific instructions before the start of the task. The instructions state how the game works and that the participants have to think about the solution before starting the task - i.e., doing the first mouse click. In Fig. 2, provides a screenshot of the first level of the game.

Although the participants do not have strict time restrictions to complete the tasks, they are advised on the time usually required to complete each task and that the second task begins only after every participant finishes the first task. This apparent freedom is in reality based on the typical time limits known from previous studies.

Analysis Procedure and Validity Evaluation

In this experiment, two groups are compared. Therefore, it requires typical statistical tests such as the t-test and the Wilcox test. Depending on the distribution of the data, there is also the possibility to perform a regression analysis of the task scores vs. the SPANE-B score, without considering the two groups.

An agreement measurement such as the Cronbach’s Alpha (Cronbach, 1951) provides validation of the two SPANE affective states measurements. Cronbach’s alpha is a coefficient of internal consistency and interrelatedness especially designed for psychometric tests. It considers the variance specific to individual items. The value of Cronbach's alpha ranges from 0.00 to 1.00, where values near 1.00 mean excellent consistency (Cortina, 1993; Cronbach, 1951).

For the creativity task, the captions are collected in a spreadsheet file to be distributed to the judges in charge to score the creativity. For the analytical problem-solving task, the PEBL software instructs the participants, provides the task itself, and collects several metrics, including those exploited to build the APS score.

Except the collection of the written captions, all steps of the experiment are therefore automatized and require the use of a computer. For this reason, a third person double-checks the spreadsheet containing the transcribed captions. The data is then aggregated and analyzed using the open-source R software.

It is important to report how psychometric tests are described and administered to the participants. In addition to the general instructions contained in the reference sheet (Article S1), each of the four experiment activities has related specific instructions. The SPANE questionnaire instructions provided to the participants are available in the paper Diener et al. (2009), but are currently freely accessible on one author's academic website (Ed Diener & Biswas-Diener, 2009). The creativity task instructions and the specific instructions for the judges are available as appendix in the study by Forgeard (2011). The Tower of London game displays specific instructions before the beginning of the game. As PEBL is open-source software, the reader is advised to obtain it to read the instructions.

Two supervisors are present during the experiment in order to check the progresses of the participants and to answer to their questions.

The participants are blind to the conditions of the experiment. They do not know about the existence of the groups.

The judges are blind to the conditions of the experiment. That is, they receive the six pictures with all the participants' captions grouped per picture. The judges are not aware of the presence of other judges and rate the captions independently.

Results and Discussion

This section merges the proposed "Execution" and "Analysis" sections of the guidelines. It describes the implementation of the research design; it provides descriptive statistics, and the hypothesis testing.

The experiment was run in January 2012. The designed data collection process and the validity process were fully followed. No deviations occurred. Both tasks required thirty minutes to be completed, and the participants completed the two surveys in ten minutes each time. No participants dropped from the study.

Participants

A total of 42 participants were sampled from the Faculty of Computer Science students of the Free University of Bozen- Bolzano. Of the 42 participants, 33 were male and 9 female. The sample had a mean age of 21.50 years old (standard deviation=3.01 years) and was diverse in provenience country: Italy 74%, Lithuania 10%, Germany 5%, and Ghana, Nigeria, Moldavia, Peru, U.S.A. all with a 2.2%

frequency. Additionally, the number of years of study has been recorded, with a mean of 2.26 years (SD=1.38 years). The working result of one participant from the creative generation task and another from the analytical problem-solving task have been excluded. The two participants did not follow the instructions and submitted non-complete data. No participants reported previous knowledge of the tasks.

Five independent judges have been recruited, to rate the captions produced in the first task. Two judges are internationally recognized professors of Design & Arts, two judges are local professors of Humanistic Studies (one is also an expert in Cinema & Theatre Arts), and one judge is a professor in Creative Writing.

Descriptive Statistics

Interesting observations arise by inspecting the two SPANE-B sessions on the affective states of the participants. Table 1 contains the measures of central tendencies and dispersion of the SPANE-B score before the creativity task and before the analytical problem-solving task. Figure 6 and Fig. 7 illustrate the distribution of the SPANE-B scores in the two measurement sessions. As both the statistics and the histograms illustrate, the sample of software developers can be considered as happy people.

The data shows a slight change in the group composition. Before the creativity task, 20 students were classified as N-POS while 21 students were part of the POS group. The average value of SPANE-B was $\mu_{\text{SPANE-B}_{CR}} = 7.58$ (SD=7.06).

Before the analytical problem-solving task, 19 students composed the N-POS group and 22 students belonged to the POS group. The average value of SPANE-B was $\mu_{\text{SPANE-B}_{APS}} = 8.70$ (SD=6.68).

The first SPANE questionnaire session happened as soon as the participants arrived to the laboratory. The SPANE-B value obtained from this measurement session let us estimate the SPANE-B population mean for software developers $\mu_{\text{SPANE-B}_{DEV}} = 7.58$ (SD=7.06). This estimate of the population mean has a .95 confidence interval of (5.29, 9.85).

Table 2 summarizes the task scores of two groups. Interestingly, the two creativity scores ACR and BCR have many commonalities. The box-plots of the ACR and BCR scores between the two groups (BCR score in Fig. 3) do not indicate noticeable differences. The two outliers of the N-POS groups for the BCR score are only at the graphical level: their commitment to the task was high and provided many, different captions for the photographs.

The scatterplots of the ACR and BCR scores versus the SPANE-B score (Fig. 8) may indicate a weak trend of higher creativity when the SPANE-B value reaches extreme values (-24 and +24).

The boxplots for the NCR score (Fig. 4) indicate different median values for the two groups. The median for the number of generated captions is indeed four for the N-POS group and six for the POS group. However, the lower quartiles of the two groups are almost the same while there is a tiny difference between the two upper quartiles.

The boxplots for the APS score of the two groups (Fig. 5) show a difference between the two groups. The distribution of the N-POS group appears as negatively skewed while the POS group is likely positively skewed in its distribution. While the two medians of the groups are not far away from each other, the two distributions seem to overlap only between the upper quartile of the N-POS group and the lower quartile of the POS group.

The scatterplot of the APS score in Fig. 9 is appealing: the APS points for the N-POS group look linearly, negatively correlated with the SPANE-B score. The APS scores for the POS do not present an evident correlation with the SPANE-B score. It seems that excellent APS score are only achieved in the POS group.

Hypothesis Testing

As the data could not permit insightful regression analysis, the research hypotheses have been tested with statistical tests for comparing two groups. Table 3 summarizes them.

Research hypotheses H1 and H2 were tested using unpaired, two-tailed t-tests. On the other hand, hypothesis H3 required a Mann-Whitney U Test because the assumptions of normality were not met (Shapiro-wilk test for normality, W = 0.8834, p-value = 0.02036 for N-POS and W = 0.8742, p-value = 0.0114 for POS).

There is no empirical evidence for rejecting the null hypotheses $H1_0$, $H2_0$, $H3_0$.

H4 was tested with unpaired, two tailed t-tests with Welch's correction, because of a significant difference in the variances of the two groups (F-test for differences in variances, F = 3.3209, degrees of freedom = 21/18, p-value = 0.01277).

With a p-value = 0.0079, reject $H4_0$ is rejected in favor of $H4_A$. There is significant evidence on a difference of analytical problem-solving skills of the participants with respect to their affective states. A two-sample permutation test confirms the results (T = 168, p-value = 0.0097).

Evaluation of Results and Implications

The empirical data does not support a difference in the creativity with respect to the affective states of the participants. This study results agree with those of Sowden & Dawson (2011), who did not find differences in the creativity of the generated ideas between the groups. The results regarding the number of produced ideas are similar with those of the study, but not significant. This study results deviate from those by Forgeard (2011), where non-depressed participants provided more creative captions under negative induced affective states. Nevertheless, it must be noted that the depression factor has not been controlled in this study.

Figure 8 suggests a tendency of higher ACR when the SPANE-B approaches extreme values such as -24 or +24. One interpretation is that higher creativity is achieved when individuals feel extreme affective states, either in the positive or negative directions. The tendency could be even stronger for the POS group. However, any specific relationships cannot be claimed at this stage.

There is significant evidence that the study participants feeling the most positive affective states possess higher analytical problem-solving skills. This study results are in contradiction with those of Abele-Brehm (1992) and Melton, (1995), where the higher analytical problem-solving skills were found in participants feeling negative affective states. The data of this research supports the theorized hypothesis of Kaufmann & Vosburg (1997), where participants feeling positive affective states performed non-significantly better in problem-solving tasks.

Figure 9 suggests a V-shaped relationship of the analytical problem-solving skills and the SPANE-B, having a vertex in proximity of the value 0 for the SPANE-B score. This should be addressed in future studies with a larger amount of participants.

Interestingly, the average value of the SPANE-B score for the sample as soon as the participants arrived to the laboratory was $\mu_{SPANE\text{-}B_{DEV}} = 7.58$ (SD=7.06). The .95 confidence interval of (5.29, 9.85), estimated for the true mean score, suggests that software developers are happy on the average.

Additionally, it has been reported how the participants enjoyed the creativity task and happily committed to it. This observations was reflected by the data: they generated 220 captions - averaging 5.24 captions per participant - and the SPANE-B value showed a +1.12 increase. A paired t-test proved that the increase in the affective states of the participants was significant (p-value = 0.0045). The Cronbach's alpha reliability measurement for the two SPANE questionnaire sessions was 0.975 over 1.000, indicating an excellent reliability of the data. Therefore, there is evidence that the participants provided stable and consistent data in the two SPANE sessions.

This validates the capabilities of the adopted measurement instrument for the affective states measurements and shows that even simple and short activities may raise the affective states of software developers.

The theoretical implications of this study are that positive affective states of software developers are indicators of higher analytical problem-solving skills. While the same is not shown for creativity, the data trends inspire prosecution of research in this field. This study introduces and validates psychometrics related to affective states to creativity and to analytical problem-solving skills in empirical Software Engineering research. This research empirically demonstrates that the affective states measurement instrument (SPANE) is capable to detect mood induction effects in a short run timespan.

Validity of the Study

A limitation of this study lies in the sample. The participants are all Computer Science students. Even though there is diversity in nationality and experience in years of study, they have a limited software development experience with respect to the professionals. However, Kitchenham et al. (2002) & Tichy (2000) argue that students are the next generation of software professionals. Thus, they are remarkably close to the interested population and may even be more updated on the new technologies (Kitchenham et al., 2002; Tichy, 2000). Höst, Regnell, & Wohlin (2000) are along the same line as they find a non-significant difference in the performance of professional software developers and students on the factors affecting the lead-time of projects. There is awareness that not all the universities offer the same curricula and teaching methods and that students may have different levels of knowledge and skills (Berander, 2004). Still, given the high level of abstraction provided by the tasks in this study, a hypothetical difference between this study participants and software professionals would likely be in the magnitude, not in the direction of the results (Tichy, 2000).

Although the number of the participants is acceptable, a coverage of the SPANE-B range in the negative direction could not be obtained. On the other hand, there is no evidence that the distribution of SPANE-B for the population of software developers should cover the full range of [-24, +24]. Additionally, studies estimating the SPANE-B mean for any population are not known. For this reason, an estimation of the affective states population mean for software developers is offered: $\mu_{\text{SPANE-B}_{\text{DEV}}} = 7.58$ with a .95 confidence interval of (5.29, 9.85). Thus, it may be that the population true mean for SPANE-B is above +7 and significantly different from the central value of the measurement instrument, which is 0.

All the measurements and instruments in this study came from a literature review of the reference fields. Although the APS metric was not formally defined elsewhere, it is built up upon standard psychometric measurements, and it refers to the background research in analytical problem-solving tasks where the results are unique, mathematically defined and calculated.

Lessons Learned

The most valuable lesson learned is on the distribution of the affective states of the participants. Even if 42 participants have been recruited, the SPANE-B score has never fallen below the value of -9 and its average value was always greater than +7 on a scale of [-24,+24]. Before the run of the experiment, a more homogeneous distribution of the participants with respect to the SPANE-B score was expected. Still, the distribution of the SPANE-B score was acceptable and let us split the participants in two groups. While this had no particular implications for this experiment, successive studies employing different tasks and measurement but employing affective states measurement should consider this study’s SPANE-B results. For future studies, the suggestion is to adopt strong mood induction techniques on randomized groups of participants and measure how the mood induction effect was effective.

Conclusions

For decades, it has been claimed without strong empirical research that a way to improve software developers’ productivity and software quality is to focus on people. A proposal to effectively study the human aspect of developers in empirical Software Engineering research is to adopt psychometrics.

By observing the reference fields (Psychology, Cognitive Science, Human-Computer Interaction, Management research), it has been understood that software developers solve problems in creative and analytical ways, through cognitive processing abilities. Cognitive processing abilities are deeply linked with the affective states of individuals, i.e., emotions, moods and feelings. This study calls for further research on the affective states of software developers.

This paper, reported a formal experiment on the impact of affective states on important software development skills and capacities, particularly analytical problem-solving skills (convergent thinking) and creativity (divergent thinking). The understanding provided by this study should be part of basic science (i.e., essence) in Software Engineering rather than leading to direct, applicable results. This work

(1) provides a better understanding of the affective states and their impact on problem-solving capacities of software developers; it (2) introduces and validates psychometrics on affective states, creativity and analytical-problem-solving skills in empirical Software Engineering; it (3) continues to raise the awareness on the need to effectively study the human aspect of software developers, as unique individuals, under a multidisciplinary viewpoint. Although the claim 'people trump process' is far from being empirically validated yet, this study provides tools, evidence, and an attitude towards its validation.

Future research should provide additional details to the claims reported in this article. The weak, yet promising trends in the creativity data inspire a prosecution of research in this area. A replication of this experiment with an order of magnitude larger may provide significant data and even enable regression analyses to verify how the intensity of affective states may matter in the creativity of software developers. It is necessary to study the affective states of software developers from a process-oriented view, to observe possible correlation with work-related achievements, and productivity while developing software. There is the need to focus on practical implications, as well. Qualitative research should answer how the creativity of software developers influences design artifacts and the source-code of a software system. Research must be done on how mood induction effects affect the quality of a software system and the productivity of developer.

Software developers are unique human beings. By embracing a multidisciplinary view, human aspect can be effectively studied. By inspecting how cognitive activities influence their performance, research will open up a totally new angle and a better understanding of the creative activity of software development.

Acknowledgments

The authors would like to thank the students participating to the experiment. The authors would also like to acknowledge Elena Borgogno, Cristiano Cumer, Federica Cumer, Kyriaki Kalimeri, Paolo Massa, Matteo Moretti, Maurizio Napolitano, Nattakarn Phaphoom, and Juha Rikkilä for their kind help during this study.

References

Abele-Brehm A. 1992. Positive versus negative mood influences on problem solving: A review. *Polish Psychological Bulletin 23*:203–221.

Amabile TM. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology 43*:997–1013. doi:10.1037//0022-3514.43.5.997

Baas M, De Dreu C, Nijstad B. 2008. A meta-analysis of 25 years of mood-creativity research: hedonic tone, activation, or regulatory focus? *Psychological Bulletin 134*:779–806.

Berander P. 2004. Using students as subjects in requirements prioritization. In *Proceedings 2004 International Symposium on Empirical Software Engineering 2004 ISESE 04* (pp. 167–176). Ieee. doi:10.1109/ISESE.2004.1334904

Boehm B. 1990. Understanding and Controlling Software Costs. *IEEE Transactions on Software Engineering 14*:1462–1477.

Brand S, Reimer T, Opwis K. 2007. How do we learn in a negative mood? Effects of a negative mood on transfer and learning. *Learning and Instruction 17*:1–16. doi:10.1016/j.learninstruc.2006.11.002

Ciborra C. 2002. *The Labyrinths of Information: Challenging the Wisdom of Systems* (1st ed.). New York, New York, USA: Oxford University Press, USA.

Cockburn A, Highsmith J. 2001. Agile software development, the people factor. *Computer, IEEE 34*:131–133. doi:10.1109/2.963450

Cortina JM. 1993. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology 78*:98–104. doi:10.1037/0021-9010.78.1.98

Cronbach L. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika 16*:297–334. doi:10.1007/BF02310555

Cropley A. 2006. In Praise of Convergent Thinking. *Creativity Research Journal 18*:391–404. doi:10.1207/s15326934crj1803_13

Csikszentmihalyi M. 1997. *Creativity: Flow and the Psychology of Discovery and Invention*. *Personnel* (p. 464). Harper Perennial.

Davis M. 2009. Understanding the relationship between mood and creativity: A meta-analysis. *Organizational Behavior and Human Decision Processes 108*:25–38. doi:10.1016/j.obhdp.2008.04.001

Denning PJ. 2012. Moods. *Communications of the ACM 55*:33. doi:10.1145/2380656.2380668

Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi D, Oishi S, Biswas-Diener R. 2009. New Well-being Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings. *Social Indicators Research 97*:143–156. doi:10.1007/s11205-009-9493-y

Dow J. 1992. External and Internal Approaches to Emotion. *Psycoloquy 3*:

Ed Diener, Biswas-Diener R. 2009. Scale of Positive and Negative Experience (SPANE). Retrieved March 19, 2013, from http://internal.psychology.illinois.edu/~ediener/SPANE.html

Feldt R, Torkar R, Angelis L, Samuelsson M. 2008. Towards individualized software engineering: empirical studies should collect psychometrics. In *International Workshop on Cooperative and human aspects of software engineering* (pp. 49–52). ACM. doi:10.1145/1370114.1370127

Fischer G. 1987. Cognitive View of Reuse and Redesign. *IEEE Software 4*:60–72. doi:10.1109/MS.1987.231065

Forgeard MJ. 2011. Happy people thrive on adversity: Pre-existing mood moderates the effect of emotion inductions on creative thinking. *Personality and Individual Differences 51*:904–909. doi:10.1016/j.paid.2011.07.015

Graziotin D, Wang X, Abrahamsson P. 2013. Are Happy Developers More Productive? In Jens Heidrich, M. Oivo, A. Jedlitschka, & M. T. Baldassarre (Eds.), *14th International Conference on Product-Focused Software Process Improvement (PROFES 2013)* (pp. 50–64). Paphos, Cyprus: Springer Berlin Heidelberg. doi:10.1007/978-3-642-39259-7_7

Höst M, Regnell B, Wohlin C. 2000. Using Students as Subjects — A Comparative Study of Students and Professionals in Lead-Time Impact Assessment. *Empirical Software Engineering 5*:201–214. doi:10.1023/A:1026586415054

Ilies R, Judge T. 2002. Understanding the dynamic relationships among personality, mood, and job satisfaction: A field experience sampling study. *Organizational Behavior and Human Decision Processes 89*:1119–1139. doi:10.1016/S0749-5978(02)00018-3

Jedlitschka A, Ciolkowski M. 2008. Reporting experiments in software engineering. *Guide to Advanced Empirical Software Engineering* Section II, 201–228.

Kaufman JC, Lee J, Baer J, Lee S. 2007. Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity 2*:96–106. doi:10.1016/j.tsc.2007.04.002

Kaufmann G, Vosburg J. 1997. “Paradoxical” Mood Effects on Creative Problem-solving. *Cognition & Emotion 11*:151–170. doi:10.1080/026999397379971

Khan IA, Brinkman W, Hierons RM. 2010. Do moods affect programmers’ debug performance? *Cognition, Technology & Work 13*:245–258. doi:10.1007/s10111-010-0164-1

Kitchenham BA, Pfleeger SL, Pickard LM, Jones PW, Hoaglin DC, Emam K El, Rosenberg J. 2002. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering 28*:721–734. doi:10.1109/TSE.2002.1027796

Lewin K. 1935. *A dynamic theory of personality*. *Trans by Adams F and Zener K New York McGrawHill* (Vol. 2008, p. 77). New York, New York, USA: McGraw-Hill. doi:10.1109/IEMBS.2008.4649489

Lewis S, Dontcheva M, Gerber E. 2011. Affective computational priming and creativity. *Proceedings of the 2011 annual conference on Human factors in computing systems* 735–744. doi:10.1145/1978942.1979048

Melton RJ. 1995. The Role of Positive Affect in Syllogism Performance. *Personality and Social Psychology Bulletin 21*:788–794. doi:10.1177/0146167295218001

Morris JD. 1995. SAM: The Self-Assessment Maniking - An Efficient Cross-Cultural Measurement of Emotional Response. *Adversitising Research 35*:

Parkinson B, Briner R, Reynolds S., Totterdell P. 1996. *Changing moods: The psychology of mood and mood regulation*. *Longman* (1st ed., p. 4). London: Addison-Wesley Longman.

Piper BJ, Li V, Eiwaz MA, Kobel Y V., Benice TS, Chu AM, … Mueller ST. 2011. Executive function on the Psychology Experiment Building Language tests. *Behavior Research Methods*. doi:10.3758/s13428-011-0096-6

Plutchik R, Kellerman H. 1980. *Emotion, theory, research, and experience*. *Theories of Emotion* (Vol. 1). London: Academic Press.

Russell J. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology 39*:1161–1178.

Shallice T. 1982. Specific Impairments of Planning. *Phil Trans R Soc B 298*:199–209. doi:10.1098/rstb.1982.0082

Shaw T. 2004. The emotions of systems developers. In *Proceedings of the 2004 conference on Computer personnel research Careers, culture, and ethics in a networked environment - SIGMIS CPR ’04* (p. 124). New York, New York, USA: ACM Press. doi:10.1145/982372.982403

Sowden PT, Dawson L. 2011. Creative feelings. In *Proceedings of the 8th ACM conference on Creativity and cognition - C&C '11* (p. 393). New York, New York, USA: ACM Press. doi:10.1145/2069618.2069712

Tichy W. 2000. Hints for reviewing empirical work in software engineering. *Empirical Software Engineering 5*:309–312. Retrieved from http://www.springerlink.com/index/rr70j282h2k01960.pdf

Tsonos D, Ikospentaki K, Kouroupetrolgou G. 2008. Towards modeling of Readers' Emotional State response for the automated annotation of documents. *2008 IEEE International Joint Conference on Neural Networks IEEE World Congress on Computational Intelligence* 3253–3260. doi:10.1109/IJCNN.2008.4634260

Watson D., Clark LA, Tellegen A. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology 54*:1063–70.

Watson David, Clark LA, Tellegan A. 1988. The Positive and Negative Affect Schedule. *Journal of Personality 8*:1988.

Westermann R, Spies K. 1996. Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology 26*:557–580. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1099-0992(199607)26:4<557::AID-EJSP769>3.0.CO;2-4/abstract

Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A. 2000. *Experimentation in software engineering: an introduction*. (V. R. Basili, Ed.)*Springer Netherlands* (Vol. 15, p. 228). Kluwer Academic Publishers. Retrieved from http://www.amazon.com/Experimentation-Software-Engineering-Introduction-International/dp/0792386825

Figure 1

A photograph for the creativity task.

Figure 2

The first level of the Tower of London game.

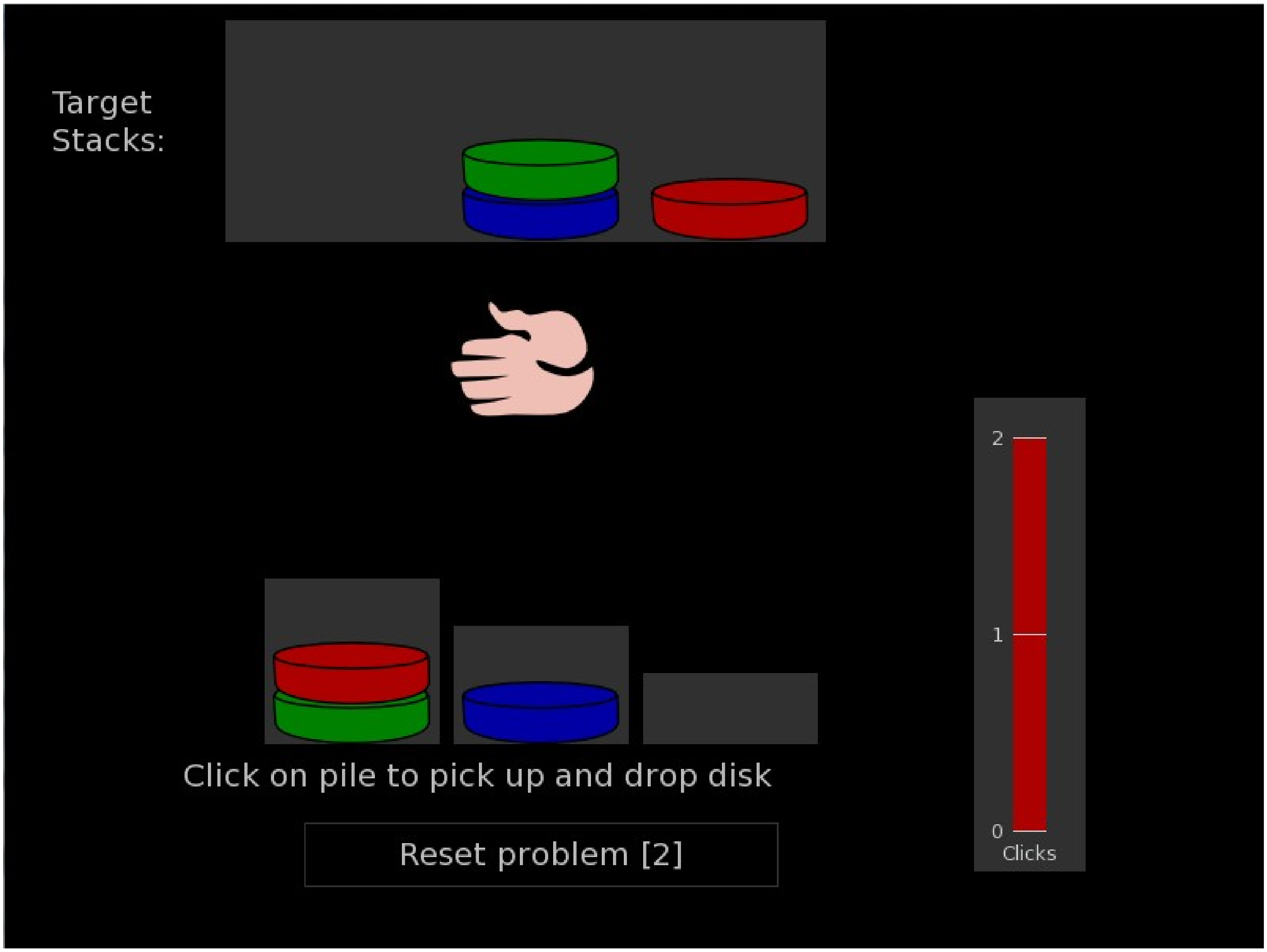

Figure 3

Best Creativity Score (BCR) Boxplots of the two groups.

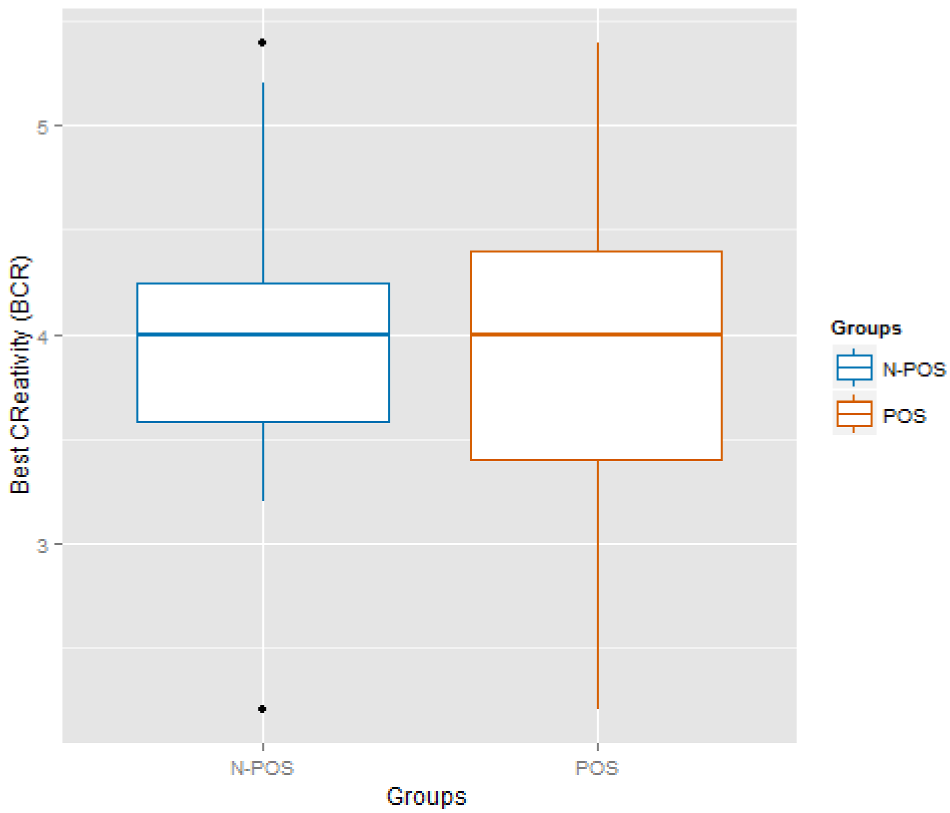

Figure 4

Number of Creative Ideas (NCR) Boxplots of the two groups.

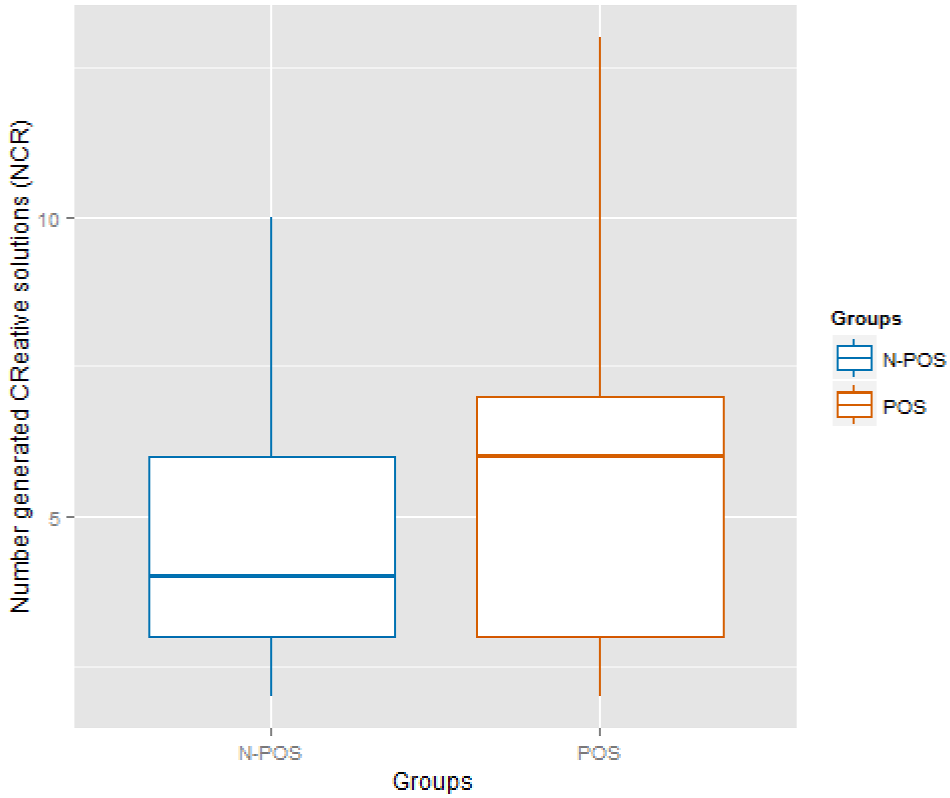

Figure 5

Analytical Problem-Solving score (APS) boxplots of the two groups.

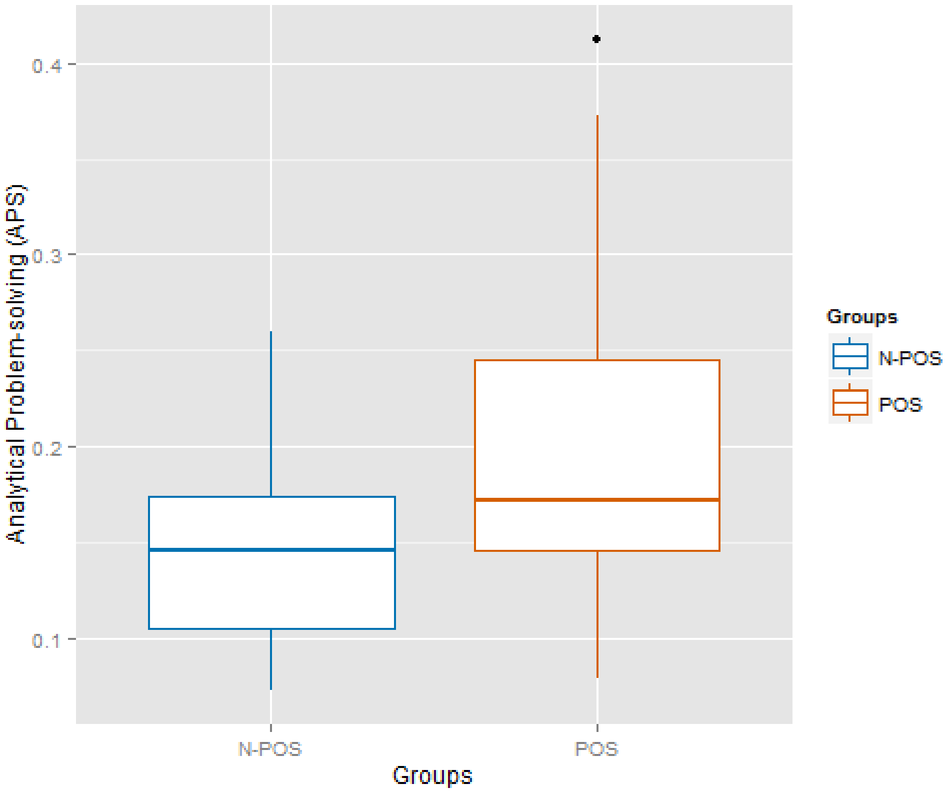

Figure 6

Affect Balance (SPANE-B) distribution before the creativity task.

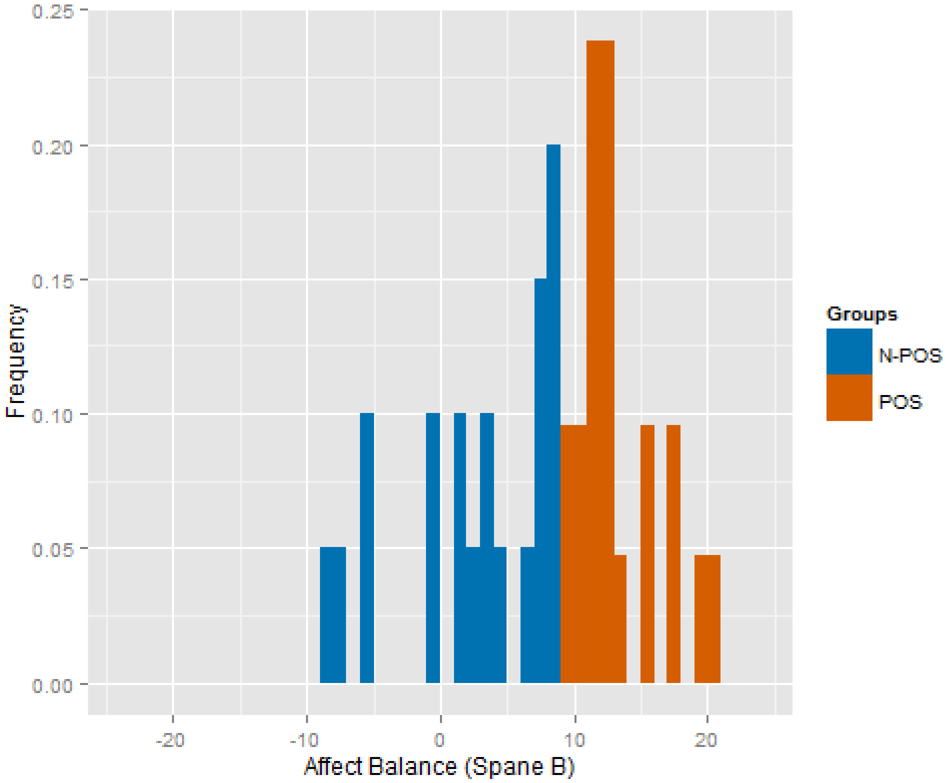

Figure 7

Affect Balance (SPANE -B) distribution before the analytical problem-solving task.

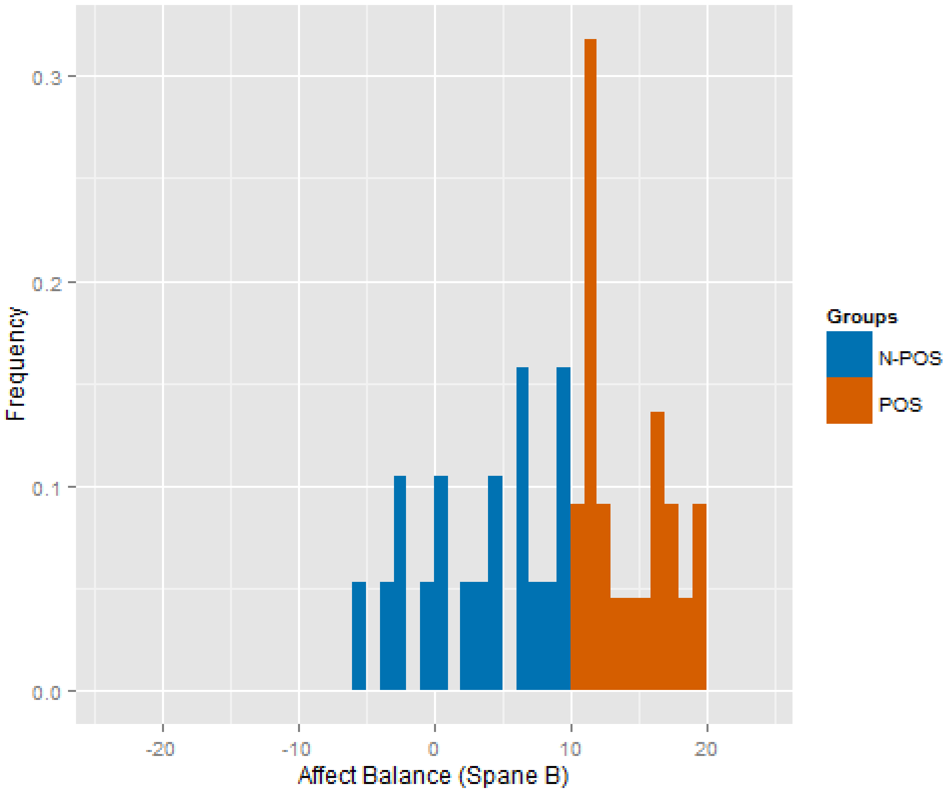

Figure 8

Average Creativity score (ACR) vs. Affect Balance (SPANE-B) scatterplot.

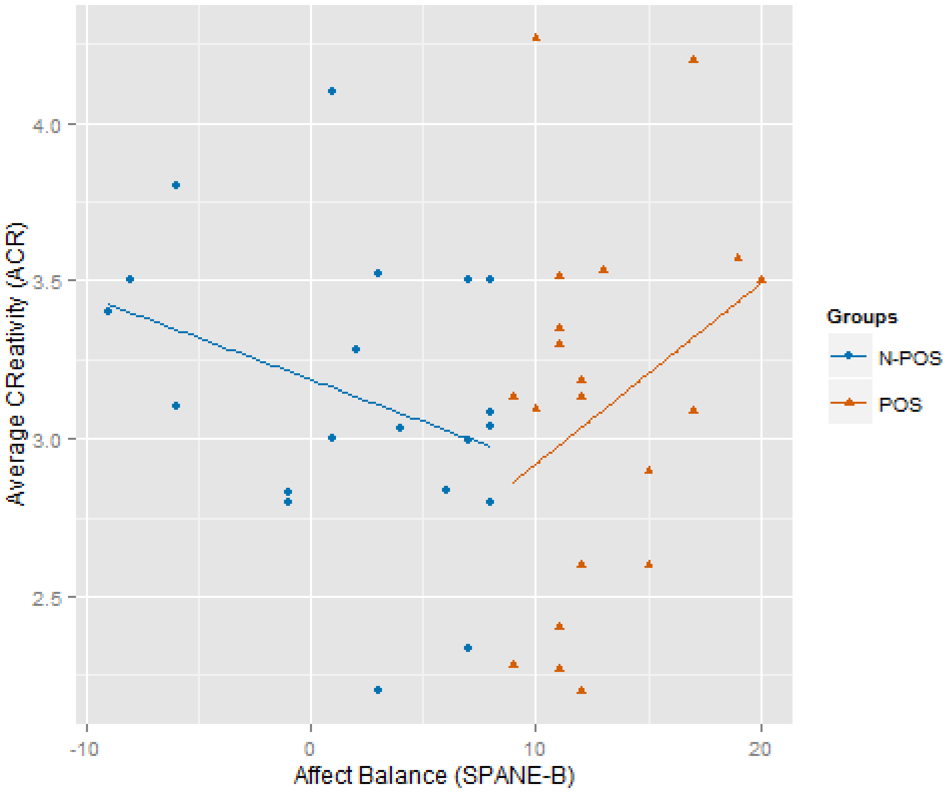

Figure 9

Analytical Problem-Solving score (ACR) vs. Affect Balance (SPANE -B) scatterplot.

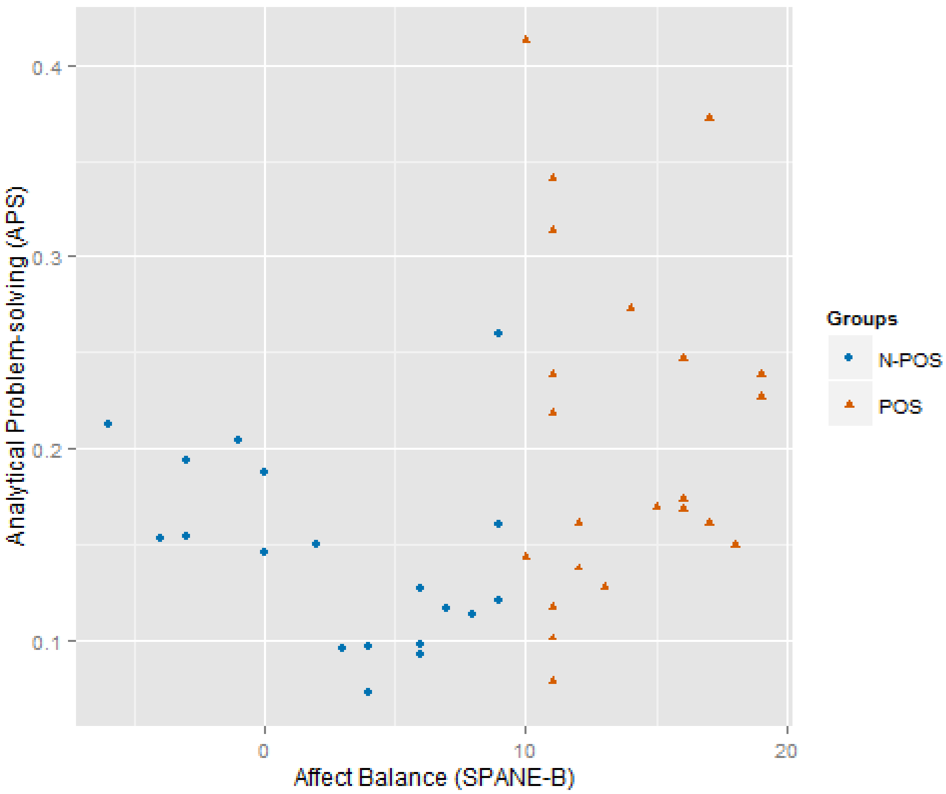

Table 1 (on next page)

Mean, Standard Deviation, and Median of Affect Balance (SPANE-B) before each task

	SPANE-B		
Measurement Time	Mean	Standard Deviation	Median
Before CR Task	7.58	7.04	9
Before APS Task	8.70	6.68	10

Table 2(on next page)

Mean and Standard Deviation of the Task Scores divided by the groups.

	Average Score (Std. Deviation)	
	N-POS Group	POS Group
ACR	3.13 (0.45)	3.08 (0.58)
BCR	4.02 (0.76)	3.98 (0.76)
NCR	4.70 (2.34)	5.90 (3.46)
APS	0.14 (0.04)	0.20 (0.08)

Table 3(on next page)

Hypothesis Testing.

RH	Mean Diff.	D.F.	T-value	W-value	p-value
H1	0.048	39	-0.206	-	0.8379
H2	0.051	39	-0.310	-	0.7577
H3	-1.204	-	-	167.50	0.2678
H4	-0.062	33.45	2.821	-	0.0079