A quasi-experiment to evaluate the impact of mental fatigue on study selection process

Ricardo Britto, Muhammad Usman, Nasir Mehmood Minhas

1. Introduction

Systematic literature reviews (SLR) and mapping studies (SMS) have become very popular in the software engineering research community. In this type of research, researchers identify, select and extract data from existing literature. Depending on the investigated research topic, researchers will have to evaluate and decide about the inclusion/exclusion of several studies. This scenario would demand the researchers to carry out a monotonous auditory detection task that may take several hours/days.

Existing literature indicates that loss of alertness associated with mental fatigue is highly correlated with fluctuations in the performance of people carrying out auditory detection tasks [1-4]. Loss of alertness has been successful measured through analyzing eye activity [2][4] and electroencephalogram frequency [1][3].

It is reasonable to assume that loss of alertness due to mental fatigue may also affect the correctness of SLR/SMS results, wherein researchers are faced with auditory detection tasks. However, to the best of our knowledge, there is no empirical evidence supporting this hypothesis. Thus, we conducted a quasi-experiment to evaluate the impact of mental fatigue on the study selection process of systematic secondary studies (SLR and SMS). The focus is on the level-1 selection, wherein researchers base the selection process on reading the title and abstracts of candidate studies.

The remainder of this report is organized as follows: Section 2 presents some related work. Section 3 details the employed experimental design. The results are presented in Section 4 and discussed in Section 5. Section 6 discusses the validity threats associated with this work, followed by our conclusions and vision on future work in Section 7.

2. Related Work

In this section we discuss some related works on the impact of mental fatigue on task performance and study selection process.

2.1 Mental fatigue and task performance

Mental fatigue is a well-known phenomenon in modern society, and its impact on task performance has been investigated in many studies. Mizuno and Watanabe [14] investigated the impact of mental fatigue on work efficiency employing an attention and working memory test with 14 healthy subjects over a four-hour period. The results showed an increase in the error count with time spent on the test task, i.e. the task performance deteriorated with increase in the mental fatigue. Later in another similar study involving the same attention and working memory test, Shigihara et al. [15] noted the increase in the error counts after the subjects performed a fatigue-inducing mental task. Boksem et al. [16] also noted that the three-hour task performance increased the fatigue level, and negatively impacted the

performance efficiency of the fatigued participants. The reaction time, errors and false alarms increased with time spent on task for these participants. Mental fatigue decreases the "goal-directed" attention, which makes subjects prone to shift their attention due to irrelevant stimuli [16]. These results show that the mental fatigue leads to the decrease in the task performance and attention levels.

2.2 Study selection process

As stated above the impact of mental fatigue on the study selection process of secondary studies has not been investigated. However, few attempts have been made to investigate study selection process with respect to the other issues. Ali and Petersen [6] evaluated different study selection strategies in systematic literature reviews. They observed that the use of more inclusive study selection strategy lead to the selection of additional relevant articles. However, they also noted, during their systematic review, that the use of most inclusive strategy needs more time and effort with little gain. Octaviano et al. [7] conducted an experiment wherein they tested the effectiveness and efficiency of a semi automatic strategy (SCAS: Score Citation Automatic Selection) for study selection in SLRs. They concluded that the use of SCAS results in considerable effort reduction (22.33%) at the cost of little loss of evidence (1.6 evidence per SLR) i.e. it saves lot of time, but may result in excluding few relevant articles.

3. Experimental Design and Operation

In this section, we describe the research questions addressed herein, the design of the experiment and its operation. We followed the guidelines proposed by Wohlin et al. [8] to carry out the quasi-experiment reported in this report.

3.1 Research questions

The experiment will be conducted to addres the following research question:

- RQ1 - Does mental fatigue affects the correctness of the decisions during study selection process in secondary studies?
- RQ2 - Does mental fatigue affects the researchers' confidence in their decisions during study selection process in secondary studies?

3.2 Experiment definition

The experiment is defined as follows [8]:

- Analyze the **impact of mental fatigue on study selection process**,
- With respect to the **correctness and confidence** of the selection decisions,
- From the point of view of **researchers**,
- In the context of **software engineering master's students** conducting secondary studies.

3.3 Preparation and planning

We developed an experiment plan, which is composed by the sample selection and commitment, description of the experimental package, definition of variables, the statement of hypotheses and the description of employed design principles.

3.3.1 Sample selection and commitment

The subjects of the experiment were selected using convenience sampling. The sample consists of students of the Masters in Software Engineering program at a public sector University in Pakistan. All students are aware of the systematic literature review and mapping studies, and the accompanying processes described in the relevant guidelines [9,10].

2.3.2 Experimental package

The experimental package's[1] object is the task of selecting primary studies for a mapping study about taxonomies in software engineering. This object is based on a mapping study that we conducted recently to characterize the state of the art on taxonomy research in software engineering discipline [12]. The documents of the experimental package are:

- **Consent form –** A form that describes the objectives of the experiment and how the collected data will be used.
- **Description of the object** – It contains the description of the systematic mapping study whose candidate primary studies' titles and abstracts the subjects will read during the experiment.
- **Instructions** – A document describing the instructions that the subjects have to follow during the experiment.
- **Pre-experiment questionnaire** – A paper-based questionnaire asking information about the subjects' background (education level, experience in software engineering and experience in conducting secondary studies) and a questionnaire about the subjects' state of mind using PANAS questionnaire [13].
- **Titles and abstracts of 100 candidate primary studies** – It contains titles and abstracts of 100 candidate primary studies randomly selected from the set of the original mapping study that the subjects have to read during the selection process. Each study has the associated selection result (included or excluded), as per the Usman et al [12].
- **Selection criteria** – It describes the selection criteria the subjects have to apply during the selection process.
- **Data collection form** – A Google form that contains the titles and abstracts and measurement instruments (selection result, mental fatigue and confidence scales)[2].
- **Pos-experiment questionnaire** – A paper-based questionnaire composed by general questions about the experiment.

3.3.2 Variables

The independent variable (intervention) is the mental fatigue. We measure mental fatigue using a self-reported mental fatigue instrument. Each subject had to provide his/her mental fatigue level after reading set of 10 abstracts on the following 4-point Likert scale:

1. I can continue this experiment without any problem. My ability for sustained mental effort is not reduced

[1] The experimental package is available at goo.gl/znOZpk

[2] The data collection form is available at goo.gl/forms/4Jw3yqKziU3JHLbK2

2. I am a bit fatigued, but am still able to make the required mental effort without any break
3. I am fatigued, and need to take a short break before moving on
4. I am highly fatigued and cannot continue with this experiment today

The main dependent variable affected by mental fatigue is the correctness of the subjects' decision during study selection process, which is measured by comparing the result generated by each subject with the results of Usman et al [12].

The secondary dependent variable affected by mental fatigue is the subjects' confidence level in their include/exclude decisions during the study selection process. The confidence level is measured using a 3-point Likert scale (low, medium and high).

3.4 Data analysis

We used Bayesian statistics to analyze the collected data. Bayesian statistics is based on the Bayes Theorem. It differs from the Frequentist statistics in two ways: interpretation of results and use of prior information. In Frequentist statistic, based on a cut point (p value), one hypothesis is selected over the other. In Bayesian statistics on the other hand, analysis of results is based on probability values that are used to compare one hypothesis with the other.

Bayes theorem is defined as:

$$\text{posterior probability} = \frac{\text{prior probability}_i \times \text{likelihood}_i}{\text{prior probability}_i \times \text{likelihood}_i + \text{prior probability}_j \times \text{likelihood}_j}$$

Or if stated in terms of A and B:

$$p(A \mid B) = \frac{p(A) \times p(B \mid A)}{p(A) \times p(B \mid A) + p(A') \times p(B \mid A')} \quad \ldots\ldots (1)$$

Where in A is the parameter (or hypothesis) under study and B is the data [11]. The probability of the hypothesis A can be revised with the availability of more/new data (B).

To address the formulated research questions, we evaluated the following hypotheses using Bayes theorem:

- H1 – The null hypothesis assumes that mental fatigue does not affect the correctness of the selection decisions during study selection process.
- H2 – The null hypothesis assumes that mental fatigue does not affect the confidence level of subjects in their selection decisions during study selection process

3.5 Data preparation

Before analyzing the data collected during the experiment session, we have to conduct some preparation. In order to simplify the analysis, we joined the values of the variables mental fatigue and confidence levels. Mental fatigue had four scale values initially (see Section 3.3.2), which were grouped into two values: fatigued and non-fatigued states. The transformation was possible due to the fact that the first two values in the fatigue scale are

related to a state that allows the subject to continue the study selection process without taking any break, while the last two are related to the fatigue states that prevent the subjects from continuing the selection process.

Likewise, we grouped the confidence-related results into two groups: high confidence or not having high confidence. This means that we grouped the low/medium confidence results into one group (i.e., not having high confidence).

All the preparation described above was carried out by means of R scripts. As a result, a centralized spreadsheet was created and used as input for the data analysis.

3.6 Operation of the experiment

We operationalized the experiment as follows:

1. We provided a short presentation to the potential subjects in which we explained the following:
   - Goal of the experiment.
   - Task(s) that each subject has to perform.
   - Experimental material.
   - Data collection instruments.
   - Estimate of the experiment duration.
2. Once the potential subjects provided their consent to participate in the experiment, we requested them to fill a pre-experiment questionnaire.
3. We handed over to the subjects the instructions to apply the inclusion and exclusion criteria.
4. We shared the google form URL with the subjects. The subjects had to read the titles and abstracts online, and also to provide their decision, i.e. to include or to exclude a study. Besides their decision, the subjects had to express their confidence level (Low, Medium and High) with each decision. Moreover, after every 10 abstracts the subjects were asked to report their fatigue level.
5. The subjects were told to complete the screening of all 100 titles and abstracts in one session.
6. At the end, the subjects were requested to fill the post-experiment questionnaire.

4. Results

In this section we present the results corresponding to each research question and associated hypotheses.

4.1 Subjects' information

The subjects of our experiment are students of Masters in Software Engineering program at a public University of Pakistan. A total of 18 students agreed to participate in the experiment. Out of the 18 students, five students have defended their research thesis successfully, eight were working on their theses and five students were starting their research work. Table 1 shows that only two subjects had never conducted any secondary study, while 10 had designed and conducted secondary studies.

Out of 18, only six subjects were included in our analysis. We excluded 12 subjects due to the following reasons (see Figure 1):

- Two subjects violated the instructions of the experiment.
- Four subjects were excluded because they were unable to complete experiment.
- Six did not reached a mental fatigue state, as per their own self-assessment.

Note that the subjects answered the PANAS questionnaire in the pre-experiment. The goal was to identify whether the subjects were in any state of mind that could bias the data analysis. However, none of the subjects were excluded due to the results of this questionnaire, since no extreme state of mind was reported.

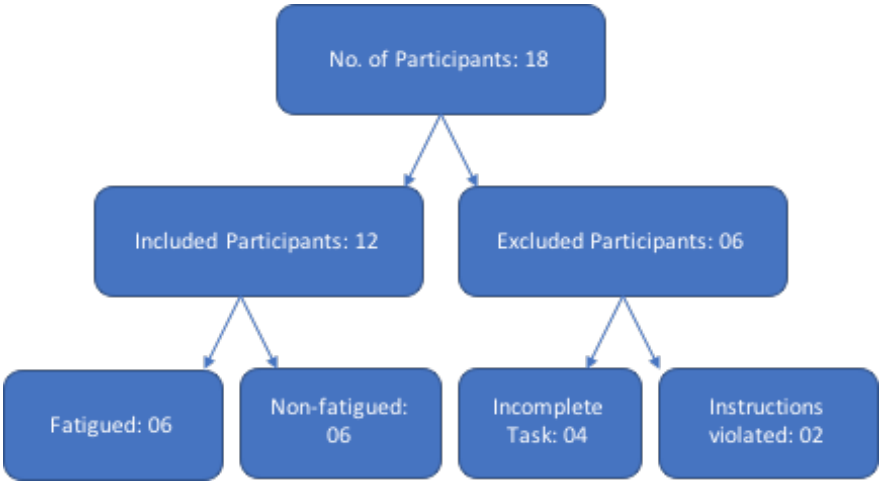

Figure 1: Details about the excluded subjects.

Table1: Information about the subjects.

Participant ID	Education Level	Research Experience	Secondary Study Design	Secondary Studies Conduction
P1	MASTERS	No Experience	0	1
P2	MASTERS	No Experience	0	1
P3	MASTERS	1 Year	2	3
P4	MASTERS	1 Year	0	1
P5	MASTERS	No Experience	0	0
P6	MASTERS	1 Year	1	2
P7	MASTERS	1 – 3 Year	2	3
P8	MASTERS	1 Year	1	2
P9	MASTERS	1 Year	0	1
P10	MASTERS	1 – 3 Year	1	2
P11	MASTERS	1 – 3 Year	1	2
P12	MASTERS	1 Year	0	0
P13	MASTERS	1 – 3 Year	1	1
P14	MASTERS	No Experience	0	0
P15	MASTERS	1 Year	1	1
P16	MASTERS	1 Year	2	2
P17	MASTERS	No Experience	0	0
P18	MASTERS	1 – 3 Year	3	3

4.2 RQ1: Impact of mental fatigue on correctness of selection decisions

This question aims to see if mental fatigue has an impact on the correctness of the selection decisions. The participants' decisions were compared with the benchmark's decisions. We assumed that the benchmark decisions were indeed correct. We did so because the benchmark was the result of consensus between at least two researchers who developed reasonable knowledge about the domain of the mapping study corresponding to the benchmark.

We used the Bayes theorem to investigate the hypothesis that the mental fatigue impacts the correctness of the selection decisions. The following variables are used in the analysis:

- Correctness of decision
  - C = Decision is correct.
  - I = Decision is incorrect.
- Mental Fatigue
  - F= Presence of fatigue during the selection process.
  - NF = Absence of Fatigue during the selection process.

Using the formula in equation (1), we have the following probabilities:

- P (C | F) = 72%, P (I given F) = 28%

These results show that in fatigued state the probability of correct decisions is bigger than incorrect decision. We also computed the following probabilities corresponding to the non-fatigued state:

- P (C |NF) = 76%, P (I given NF) = 24%

It shows that in non-fatigued state, the probability of arriving at correct decision is 76%. It is slightly larger than the corresponding probability (72%) in the fatigued state. It shows that the probability of correct decisions is higher in non-fatigued state as compared to the fatigued state.

4.3 RQ2: Impact of mental fatigue on subjects' confidence level in their selection decisions

This question aims to investigate if mental fatigue impacts the confidence level of researchers in their selection decisions during study selection process. We used the Bayes theorem to investigate the following variables:

- Confidence level
  - NH = Confidence is not high in selection decisions
- Mental Fatigue
  - F = Presence of fatigue during the selection process

Using the formula in equation (1), we calculated the following probabilities:

- P (NH | F) = 52%, P (H given F) = 48%

It shows that the probability of not having high confidence in decisions in fatigued state is 52%. As it is slightly above the half way mark, it is hard to infer anything from it.

We also checked for the probability of not having high confidence in selection decisions in non-fatigued state. We calculated the following probabilities:

- P (NH |NF) = 62%, P (H given NF) = 38%

These results show that there is a 62% probability of not having high confidence in selection decisions in non-fatigued state. These two set of probabilities highlight an interesting point i.e., the confidence in selection decisions is relatively higher in fatigued state (48%) as compared to non-fatigued state (38%). However, in both cases, the probability of high confidence is below 50%.

5. Discussion

In large mapping studies, researchers often have to screen hundreds of titles and abstracts of candidate primary studies. Out of the 12 participants who completed the experiment task, 6 (50%) reached a fatigue state during the experiment. However, the preliminary results presented herein do not support our hypotheses: fatigued subjects were to have higher probability of making mistakes and lower probability of having high confidence associated with their selection results.

Apart from minor comparative differences between fatigued and non-fatigued states' impact on the correctness and confidence levels, the results for RQ1 and RQ2 did not indicate any significant impact of fatigue on correctness and confidence level. This may be due to a combination of the following reasons:

a) **Small number of subjects** - Only six subjects were accounted for in the data analysis.
b) **Homogenous sample of participants** - All participants were Masters student of the same university, and have similar backgrounds.
c) **The self-reported fatigue instrument** - Self-reporting based instruments have their own shortcomings. It is possible that the participants were not able to properly classify their fatigue levels.
d) **Simple experiment object** - The experiment object was very simple. The subjects were asked to decide whether or not studies report taxonomies in software engineering. The corresponding inclusion and exclusion criteria were simple, which facilitate the task, not demanding high mental effort by the subjects.

6. Validity Threats

Conclusion validity is concerned "with issues that affect the ability to draw the correct conclusions about relations between the treatment and the outcome of an experiment" [8]. The main threats of this category related to our investigation are the low reliability of measures (the amount of noise related to a measure) and low statistical power. Regarding the reliability of the measures, we based our measurement approaches on instruments used by other researchers. However, since they are based on self-assessment, it might be the case that the subjects did not provide completely accurate data. Regarding statistical power, the size of the experiment sample is too small, which is probably related to the results did not support the investigated hypotheses.

Threats to **internal validity** are concerned with ”influences affecting the independent variable with respect to causality without the researcher’s knowledge” [8]. The main threat of this category related to our investigation is the mortality threat (subjects dropping off the experiment). Considering the nature of the experiment, that required the subjects to get fatigued, there was a high chance that people would either not accept to participate or would abandon/not finalize the experiment. To mitigate that, we provided a very comprehensive explanation about the experiment operation and asked for the subjects to join only if they were able to finalize the tasks, even after knowing the effort required to do so.

Construct validity is the ability to ”generalize the result of the experiment to the concept or theory behind the experiment” [8]. The main threats of this category related to our investigation are mono-method bias (only one method to measure a construct), hypothesis guessing (the subjects try to guess the real purpose of the study), evaluation apprehension (the subjects get anxious about being evaluated), and researcher expectancies (when a researcher communicates the desirable result of a research). Regarding mono-method bias, there is just one measurement method per construct, which is a limitation of the study. We piloted the measurement methods to identify other the measures behaved in the expected way. Nevertheless, the measures are obtained through self-assessment, which gives room for inaccurate results. To mitigate hypothesis guessing, we avoided giving any hint about the main purpose of the investigation. To mitigate the last two threats, we made an effort to make the subjects comfortable and also clarified that there was no right and wrong answers for the experiment tasks.

External validity is the ability to ”generalize the results of an experiment to industrial practice” [8]. Considering the small number of subjects and the low variety, it is not possible to generalize the results presented herein. Furthermore, the place where the experiment took place was a lecture room and all the students were together. This setting does not necessarily resemble the environment where researchers would conduct study selection.

7. Conclusions

We reported the design and preliminary results of an experiment to investigate the impact of mental fatigue on study selection process of secondary studies.

Eighteen subjects participated in the experiment session, but only six were accounted for in the data analysis. Due to the small number of subjects, in addition to other issues (e.g. simplicity of the experiment task), it was not possible to identify any significant impact of mental fatigue on the correctness and confidence of selection results.

As future work, we plan to replicate this experiment with more participants, and with a more complicated experimental object involving relatively complex selection choices. We would also investigate other available mechanisms for eliciting mental fatigue more effectively.

References

[1] S. Makeig and M. Inlow, “Lapse in alertness: coherence of fluctuations in performance and EEG spectrum,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 86, no. 1, pp. 23–35, 1993.

[2] K. F. Van Orden, T. P. Jung, and S. Makeig, “Combined eye activity measures accurately estimate changes in sustained visual task performance,” *Biol. Psychol.*, vol. 52, no. 3, pp. 221–240, 2000.
[3] S. Makeig and T.-P. Jung, “Tonic, phasic, and transient {EEG} correlates of auditory awareness in drowsiness,” *Cogn. Brain Res.*, vol. 4, no. 1, pp. 15–25, 1996.
[4] R. Martin and J. M. Carvalho, “Eye blinking as an indicator of fatigue and mental load—a systematic review,” *Occup. Saf. Hyg. III*, no. October 2016, pp. 231–235, 2015.
[5] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer, *Eye Tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.
[6] Ali, Nauman Bin, and Kai Petersen. "Evaluating strategies for study selection in systematic literature studies." In Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 45:1-45:4, 2014.
[7] Fábio Octaviano, Cleiton Silva, and Sandra Fabbri. "Using the SCAS strategy to perform the initial selection of studies in systematic reviews: an experimental study", In Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering , pp. 25:1-25:10, 2016.
[8] Wohlin, Claes, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. Experimentation in software engineering. Springer Science & Business Media, 2012.
[9] Keele, Staffs. "Guidelines for performing systematic literature reviews in software engineering." Technical report, Ver. 2.3 EBSE Technical Report. EBSE. sn, 2007.
[10] Petersen K, Feldt R, Mujtaba S, Mattsson M. Systematic Mapping Studies in Software Engineering. In Proceedings on International Conference on Evaluation and Assessment in Software Engineering, pp. 68-77, 2008.
[11] Eidswick, J., *A Bayesian alternative to null hypothesis significance testing.* Shiken Research Bulletin, 2012. **16**(1).
[12] Muhammad Usman, Ricardo Britto, Jürgen Börstler, Emilia Mendes, Taxonomies in software engineering: A Systematic mapping study and a revised taxonomy development method, Information and Software Technology, Volume 85, Pages 43-59, 2017.
[13] Watson, David, Lee A. Clark, and Auke Tellegen. "Development and validation of brief measures of positive and negative affect: the PANAS scales." Journal of personality and social psychology 54, no. 6 (1988): 1063.
[14] Mizuno, Kei, and Yasuyoshi Watanabe. "Utility of an advanced trail making test as a neuropsychological tool for an objective evaluation of work efficiency during mental fatigue." In Fatigue science for human health, pp. 47-54. Springer Japan, 2008.
[15] Shigihara, Yoshihito, Masaaki Tanaka, Akira Ishii, Seiki Tajima, Etsuko Kanai, Masami Funakura, and Yasuyoshi Watanabe. "Two different types of mental fatigue produce different styles of task performance." *neurology, psychiatry and brain research* 19, no. 1 (2013): 5-11.
[16] Boksem, Maarten AS, Theo F. Meijman, and Monicque M. Lorist. "Effects of mental fatigue on attention: an ERP study." *Cognitive brain research* 25, no. 1 (2005): 107-116.