Effect of Time-pressure on Perceived and Actual Performance in Functional Software Testing

Iflaah Salman
M3S Group, University of Oulu
Oulu, Finland
iflaah.salman@oulu.fi

Burak Turhan
Department of Computer Science
Brunel University London, UK
burak.turhan@brunel.ac.uk

ABSTRACT

Background: Time-pressure is an inevitable reality of software industry that influences the performance of software engineers. It may result in adverse effects on software quality or distort the perception of performance on executed tasks to differ from actual performance. **Objective:** We aim to investigate the effect of time-pressure on perceived and actual performance of software testers in the context of functional software testing. **Method:** We performed two controlled experiments with 87 graduate students in two academic terms. We assessed actual performance in terms of coverage (i.e. percentage of test cases correctly identified) and perceived performance using NASA-TLX. We have an independent factorial design for our experimental study. **Results:** The results reveal a significant effect of time-pressure on actual performance. However, we could not observe a significant effect of time-pressure on the perceived performance of the participants for the task undertaken. We also observed a significant negative correlation between actual and perceived performance when controlled for time-pressure and experimental session factors. **Conclusion:** Time-pressure affects the actual performance in a testing task but the perception of accomplishment by the testers is sustained irrespective of time-pressure, indicating an over-estimation issue. Perception of performance should be adjusted to align with reality to account for the effect of time pressure. This will lead to better self estimates of performance.

CCS CONCEPTS

• **Software and its engineering → Software testing and debugging**;

KEYWORDS

Software Quality, Software Testing, Time Pressure, Experiment, Performance, Self-assessment

ACM Reference Format:
Iflaah Salman and Burak Turhan. 2018. Effect of Time-pressure on Perceived and Actual Performance in Functional Software Testing. In *ICSSP '18: International Conference on the Software and Systems Process 2018 (ICSSP '18), May 26–27, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3202710.3203148

ICSSP '18, May 26–27, 2018, Gothenburg, Sweden

1 INTRODUCTION

Time-pressure, studied also as schedule pressure and deadline pressure, is reported to have adverse effects on performance in many fields, e.g., auditing, projects work, and health and safety [12]. The fast paced software industry is also enduring the challenge of time-pressure, which is often considered as a negative factor [15]. For example, Baddo and Hall [2] found time pressure as a leading demotivating factor for software process improvement. Another study by Wilson and Hall [27], which investigates software engineers' views of quality, revealed time-pressure as a considerable factor to have a negative impact on the development cycle. However, some empirical studies also report the positive effects of time-pressure, e.g., time-pressure improved efficiency in requirements review [15] and multiple testers under time-pressure delivered better defect rates compared to individuals under no time pressure [14]. A recent literature review by Kuutila et al. [12] found that the most investigated phases of software development, from the perspective of time-pressure, are testing and code quality, yet the authors recommend conducting more studies on this topic since the empirical evidence is still scarce [12], [15].

Other than the field of Software Engineering (SE), time-pressure is found to affect perceived performance or self-judgement of performance in marketing and education domains. For example, Papamitsiou and Economides suggest that time-pressure and stress may affect the self judgement about performance of the students taking tests. However, the authors' investigation of the postulated phenomenon did not reveal a significant difference between pre and post tests' perception of performance [19]. Although, a significant difference was found between the actual performance and pre and post test's perception of performance [19]. In the field of marketing, Andre and Smith [1] found a negative relationship between the perceived creativity of marketing ideas and the perceived time-pressure by the marketing professionals. Despite of time-pressure, self-assessment of performance or skills is a non-trivial task, due to which self-perceptions of knowledge and skills are often in conflict with the reality [5]. Accurate perception of performance, of a self-executed task, is of vital importance as it helps in improving actual performance and life-long learning [4], [16]. Accurate perceptions or self-assessments help in identifying strengths and weaknesses, which also lead towards improved actual performance [16].

There is a limited to no literature available in SE, to the best of our knowledge, that compares actual performance with the perceived or self-assessed performance. However, the domain of software project management compares estimated efforts (expert estimations) with actual efforts invested in the projects, e.g., [11] and [17]. Comparison of estimated and actual efforts of project management is in contrast with the objective of our study because 1) we focus

on perceived and actual performance 2) we compare actual performance with the performance perceived after the execution of the task. Since it is non-viable to alleviate time-pressure from software industry, and software testing being the most impacted phase of this reality, we scope our investigation to functional testing to study this gap. We therefore, aim to *investigate the effect of time-pressure on actual and perceived performance in software testing*.

In order to meet our objective, we performed two controlled experiments with graduate students across two academic terms. We employed independent factorial design with two levels of time pressure and two levels of experimental sessions. The actual performance is assessed in terms of percentage of identified functional test-cases, and perceived performance is measured using NASA-TLX (workload assessment procedure). The results show a significant impact of time pressure on actual performance but no effect on perceived performance. Moreover, actual performance negatively correlates with perceived performance. Our study contributes by performing an experimental study exploring the effect of time pressure not only on actual performance but also, how perception of performance varies under time pressure, which is novel in SE.

The next section presents related work which is followed by Section 3 elaborating on research methodology. Section 4 details experimental execution, which is followed by Section 5 presenting the results. The results are further discussed in Section 6 with threats to validity in Section 7. The study is concluded in Section 8.

Table 1: Time-Pressure (tp) Studies in SE

Study	Context & Aim	Results
[27]	Software Quality; Investigation of the software engineer's views of the quality	tp is suspected to impact negatively overall in the development cycle.
[2]	Software Process; Issues demotivating the software practitioners for Software Process Improvement (SPI)	tp identified as one of the demotivating factors for SPI.
[25]	Software Development (Database Query); investigation of the relation between time availability and task complexity	Time availability did not affect task performance. However, task complexity influenced performance at all levels of time availability.
[18]	Software Development; Effects of schedule pressure and budget on software cycle time and effort	Schedule pressure did not significantly impact the outcomes of development.
[14]	Software Testing; Effect of multiple individuals working on the same task and tp on the effectiveness and efficiency of manual testing	Multiple time-pressured testers delivered better defect detection compared to individual testers under no-time-pressure.
[23]	Global Software Testing (GST); perception of testing and deadline pressure	tp is perceived negatively by the test-engineers in terms of quality. However, team configurations in GST do effect the perception and experience of tp.
[15]	Software Testing; Effects of tp on effectiveness, efficiency. Effect of knowledge on tp and the perception of tp	tp did not cause negative effects on effectiveness. No effect of knowledge was observed on tp. tp improved the efficiency in requirements review and test-case development.

2 RELATED WORK

Table 1 presents the studies conducted in SE either directly on the topic of time-pressure (tp) or that observed tp as a considerable factor in the studied context. The table is adopted from Mäntylä et al.[15] but has been revised to fit to our context. The three columns in the table show; the study reference, the context and aim of the study and the results of the respective study from the perspective of time-pressure. We can see from the table that three studies - Baddoo and Hall [2], Shah et al. [23] and Wilson and Hall et al. [27] - report negative effects of tp. Two studies, Topi et al. [25] and Mäntylä et al. [15] could not observe the effects of time-pressure on the outcomes of development and effectiveness respectively.

Several studies in the area of software project management have focused on comparing expert estimations (expert judgement) with the actual effort invested in the projects, in comparison with tools and methods based estimation methods. We summarise these studies only from the perspective of the accuracy of expert estimations without arguing over their efficacy compared to tools or models based estimations. Because, it is in more alignment with our study objectives to consider accuracy and performance of expert estimators (humans). We therefore, consulted the most recent literature review by Jørgensen [10], on the topic of expert estimations, to choose the relevant studies that date back to 1999 only.

Myrtveit and Stensrud [17] found that expert estimators (professionals) provided more accurate estimates when they used the analogy tool and regression model compared to the estimates based solely on dataset of past projects. Another study based on analogy (case-based reasoning strategy) based software effort estimations revealed expert estimators to be more accurate in selecting analogues and estimating compared to the tool's selection of analogues and regression models' estimations [26]. Kitchenham et al. [11] based on the results of their study on maintenance and development estimation accuracy, concluded that human-mediated estimation process can result in quite accurate estimates. Overall, the results of these studies are mixed from the perspective of accuracy of expert judgement.

In other fields, e.g., education, health and marketing it is common to use self-assessments and assess their accuracy in order to improve performance and behaviour, e.g., [1], [4] and [16]. In SE, Mäntylä et al. [15] used NASA-TLX for assessing perceived time-pressure and workload assessment for the tasks performed by the participants. The authors, however, did not use NASA-TLX to compare the perceived and actual performance of their participants.

In contrast to the presented earlier work from SE domain, making estimations and comparing it with the historical data to know its accuracy, we assess the performance of the task after its execution instead. We measure the self-assessment of performance and compare it with the actual performance as a response to the effect of time-pressure. We scope our study to software testing domain and the designing of functional test-cases. We share the same study domain of software testing with [14], [15] and [27]. Moreover, we do not detect defect rate as a measure of performance and the development of test-cases on a high level as done by Mäntylä et al. in [14] and [15]. Our required level of detail for a designed test-case is explained in next section.

3 RESEARCH METHODOLOGY

We follow the guidelines by Wohlin et al. [28] for the reporting of experimental definition and planning stages.

3.1 Goal

Our study aims to examine the effect of time-pressure on the actual performance for a given task compared with the perception of accomplishment formed for the same task. We explore this in the context of functional software testing. Functional software testing exercises software testing on the basis of requirements specifications and is also referred as black-box testing [13]. Investigation scope of our functional software testing is limited to the designing of test-cases rather than their execution. According to Goal-Question-Metric, we define the goal of our study as [3]:
Analyse *the designed suite of functional test-cases*
For the purpose of *examining the effects of time-pressure*
With respect to *actual and perceived performance*
From the point of view *of researchers*
In the context of *an experiment run with graduate students (as proxies for novice professionals) in an academic setting.*

Our research question therefore, is defined as:
RQ: How does time-pressure affect the actual and perceived performance of software testers in designing functional test-cases?

3.2 Design

We have a between groups factorial design - Independent Factorial Design, depicted in Table 2. *TP* and *NTP* stand for time-pressure and no-time-pressure, i.e., treatment and control groups respectively. The two experimental sessions are indicated as *ES*1 and *ES*2. Our design has a single object that limits the effect of task-treatment interaction as a confounding factor, which we discuss further in Section 7. However, this allows us to study the effect of time-pressure with more control. Further details on time-pressure manipulation and experimental sessions are presented in subsequent sections.

Table 2: Experimental Design

	TP	NTP	Object
ES 1	Group 1	Group 2	MusicFone
ES 2	Group 3	Group 4	

3.3 Experimental Materials

All materials including pre and post experimental data collection, object of the study, experimental execution guidance scripts are available online [22].

3.3.1 Pre/post-questionnaires. Background data related to the experience of the participants in software development and testing both in academia and industry was collected using an online survey. Industrial experience relates to working in different roles, e.g., developer, tester, whereas, academic experience relates to the applied learning of development and testing as a part of coursework.

We used *NASA Task Load Index (TLX)* for post-experimental data collection. NASA-TLX is developed by NASA Ames Research Center to assess the workload for a task experienced by the task-performer [8]. The workload assessment in this procedure is based on the rating of six attributes; demands imposed on the participant (mental, physical and temporal demand) and interaction of a participant with the task (performance, effort and frustration) [8]. Therefore, we used NASA-TLX's *performance* attribute as a self-assessment measure for perceived performances of the participants in our experiment. Another major reason for using NASA-TLX was to use the ratings of *temporal demand* attribute. The assessment of the perceived temporal demands by the participants informs us how effectively we manipulated time-pressure in the controlled and experimental groups. Thus, adding towards the validity of our experiment. NASA-TLX also leverages us collecting feedback from the participants in a more standardised manner.

3.3.2 Experimental object. The object of our study is the requirements specification document of MusicFone application [22]. MusicFone is a GPS based mobile-phone music playing application which helps an app-user to prepare an itinerary for attending concerts based on the selection of artists, which are suggested to the user based on the currently played artist by MusicFone application. The participants of the study are required to design functional test-cases for MusicFone application.

MusicFone has also been used by other reported experiments, e.g., [7] and [21], as a realistic task - task that simulates realistic programming. The selection of this task hence, also reduces the non-realism aspect of our study, per Sjøberg et al. [24]. For the purpose of our experiment, we modified the original application's specifications in a way that they can serve as a requirements specifications for designing test-cases. In addition to the specifications, we also provided a screenshot of the implemented MusicFone application for the participants to develop a better understanding of the task. We provided a screenshot of an implementation that is consistent with the specifications.

3.3.3 Test-case design template. We provided the participants a template for designing the test-cases, to ensure consistency in data collection. Table 3 presents the provided template consisting of four columns; *ID*, *description* of the test-case, *input/pre-condition* for mentioning state/conditions that should pre-exist for the test-case and *expected output* for the state/conditions that should be met respective to the test-case execution. The last two columns also aid towards developing a better understanding of the designed test-case for the researcher during data-extraction phase. The template also consisted of an example test-case for the participants to know the level of detail required for designing a test-case. The template was provided in paper form, i.e., test-case design task for the object of the study was a pen and paper activity.

Table 3: Test-case Design Template

Test-Cases			
ID	Description	Input/pre-condition	Expected Output/Post-condition
1	Application displays 20 artists to the AppUser.	Get Recommendations Clicked; Artists from Last.Fm website	20 artists are displayed in the recommendation's section.

3.4 Participants

Convenience sampling was used to draw a sample from the graduate level software engineering students at the University of Oulu, as proxies for novice professionals. The participants are students enrolled in the International Master's Degree programme of 2015

and 2016 academic years for the software quality and testing course, forming groups for experimental sessions *ES*1 and *ES*2 respectively. The students of both batches participated to the experiment voluntarily providing a consent form. Students who participated in the experiment were offered bonus points in their final grading. Four students from ES1 and three from ES2 did not give consent to participate - their data is not used in this paper. All students, however, irrespective of their consent, performed the experimental task as a class activity. The final count of participants in 2015 session is 43 (22 in TP, 21 in NTP) and 45 (24 in TP and 21 in NTP) in 2016 experimental session.

Figures 1 and 2 present the experience data collected as background information of $ES1-2015$ and $ES2-2016$ participants. ADE stands for academic development experience, ATE for academic testing experience, and *I* is for *industrial* in IDE and ITE respectively. In the legends, *m* stands for *months* and *y* for *year(s)*. The comparison of figures inform us that participants of ES1 are relatively more experienced, particularly considering ADE and ATE. Though participants in the category of $>= 1y$ *and* $< 3y$ of ADE of ES2 are more than the similar category of ES1 but the rest of the categories especially, $>= 3y$ of ES1 is 37% and of ES2 is 18%. Similarly, there are 84% participants in $< 6m$ category in ES1 but 91% in ES2.

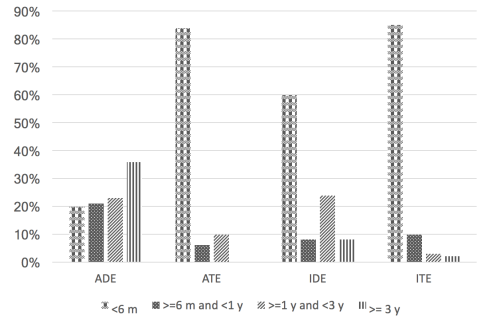

Figure 1: Experience of Participants - ES1

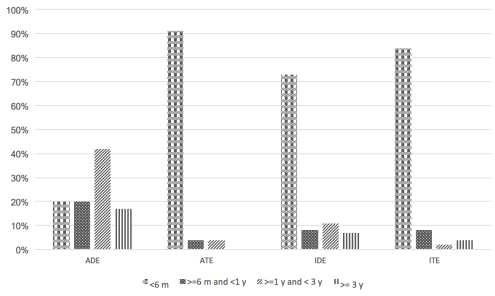

Figure 2: Experience of Participants - ES2

3.4.1 Training. For both sessions, *ES*1 and *ES*2, we trained the participants prior to the actual experiment executions. The participants were trained in multiple stages each year with the same instructors and also the same training content. They were taught and trained in functional (black-box) software testing techniques and the design of functional test-cases. The training in 2015 consisted of two lectures and one in-class exercise, whereas for 2016 the in-class exercise was replaced with home assignment merely due to logistic issues. The in-class exercise and home assignment utilised the same requirements specifications as a practice-object both years but it was different from the actual object of the experiment. Unlike the actual experimental sessions, we did not provide any supporting screenshot for in-class exercise and home assignment. However, we provided the same test-case design template for designing the test-cases for the practice-object to train participants for the experimental execution. Table 4 presents the sequence of the training phases and their contents.

Table 4: Training Sequence

Session	Lecture I	Class Exercise / Home Assignment	Lecture II
Duration	2 hours	2 hours / 1 week	2 hours
Content	Functional testing techniques (Equivalence partitioning, Boundary value analysis	Functional test-case designing exercise using the designed template; Solution Discussion	Test-case design techniques (Need and classification of techniques, Control flow techniques, Functional techniques); Test-case specification

3.5 Variables and Metrics

The variables and metrics of our study are explained in this section.

3.5.1 Independent Variable. Time-pressure is an independent variable of our study, with two levels, *time-pressure - TP* and *no-time-pressure - NTP*. We executed a pilot-run to determine the stretch of the two levels and as a result, 30 minutes were decided for *TP* group and 60 minutes for *NTP* group - further details on pilot are in Section 3.7. This was done to operationalise time-pressure.

Time-pressure was regulated psychologically in the two groups by announcing the duration in a different manner. Time reminders for *TP* group were announced thrice by the experimenter; first one after 15 minutes into the experiment and subsequent calls were made after every 5 minutes. Whereas, *NTP* group was announced of the total time available to them only in the beginning of the experiment, and experimenter made a termination call after 60 minutes.

A confounding variable of our study is *experimental session* with two levels, *ES1* and *ES2* per the design of our experiment in Section 2. We do not consider this variable as of primary interest, but incorporate it in the analysis as a blocking factor.

3.5.2 Dependent Variables. The dependent variables of our study are *perceived performance*, *actual performance* and *temporal demand*.

Perceived Performance: In order to measure perceived performance, we use the definition of the load attribute *performance* as provided in NASA-TLX documentation. Accordingly, it is a measure of how successful one perceives he was in achieving the objectives of the task and how satisfied he is with his performance [8]. NASA-TLX measures performance attribute (*PP*) on a scale from 0 to 100 - *failure to perfect.*

Actual Performance: We measure actual-performance in terms of coverage of specifications, i.e., ratio of the number of functional test-cases designed by a participant over the number of test-cases in a baseline test suite covering all specifications. We designed the test-case suite of MusicFone from researchers' perspective to serve as a baseline suite. This not only provided us with a possible total number of test-cases but also a test-suite to compare the extracted data with. Actual-performance therefore, is measured as:

$$AP = \frac{total\ TCs\ designed\ by\ the\ participant}{total\ TCs\ in\ baseline\ test\text{-}suite} * 100$$

$TCs = test\text{-}cases$

The original baseline suite consisted of 55 test-cases. In order to enhance the validity, we advanced our baseline suite with test-cases that were designed by the participants but were missing in our suite, during data extraction phase. This increased the total test-cases count to 68 in the baseline suite. This validity improvement step follows the recommendation of Mäntylä et al. [15].

Temporal Demand: Similar to *PP*, to assess the temporal demand experienced by the participants, we use the same definition as in NASA-TLX. The temporal demand (*TD*) attribute measures the time-pressure felt pertaining to the pace of the task elements' occurrences on a scale from 0 to 100, i.e., from low to high [8].

3.6 Hypothesis Formulation

We formulate the following hypotheses according to the goals of our study.

*H*1 postulates: *Actual performance under time-pressure differs from the actual performance under no time pressure.*

$$H1_A : \mu(AP)_{TP} \neq \mu(AP_{NTP})$$

$$H1_0 : \mu(AP)_{TP} = \mu(AP_{NTP})$$

Effect of time-pressure on perceived performance postulates as:

*H*2: *Perceived performance under time-pressure differs from the perceived performance under no time pressure.*

$$H2_A : \mu(PP)_{TP} \neq \mu(PP_{NTP})$$

$$H2_0 : \mu(PP)_{TP} = \mu(PP_{NTP})$$

On the relationship between actual performance and perceived performance, we postulate the following hypothesis:

*H*3: *Actual performance correlates with perceived performance.*

$$H3_A : r_{AP,PP} \neq 0$$

$$H3_0 : r_{AP,PP} = 0$$

3.7 Pilot Run

For the purpose of identifying the duration of the levels of time-pressure (*TP*, *NTP*), we performed a pilot with five post-graduate students. Another purpose of the pilot was to improve the instrumentation for our actual-experiment execution. The participants of the pilot-run completed the task of functional test-case designing with an average of 45 minutes. Considering thus the maximum possible duration for the actual-experiment execution, we decided 30 minutes for TP group and 60 minutes for NTP group. We also improved our requirements specification documentation based on the feedback from the pilot.

3.8 Analyses Methods

We employ two analyses methods, descriptive statistics and statistical significance tests of *F-test* family (*Two-way independent ANOVA*), and execute correlations' tests to test our hypotheses. Two-way independent ANOVA requires the following assumptions to be met [6]:

- Independence of observations, i.e., there should be no relationship between the observations of the groups.
- Normal distribution of the residuals.
- Homogeneity of variance for all the levels of the two independent variables.

The first assumption is related to the design of the study and our experimental design satisfies this assumption. Shapiro-Wilk test is used for assessing the normality assumption of the data. If data fails to meet the assumption of normality then we perform the non-parametric alternatives of the respective statistical tests. We report ω^2 as an effect-size value for 2-way ANOVA and correlation coefficient - r is itself an effect-size with 0.10 = small, 0.30 = medium and 0.50 = large effect [6]. We use significance level - $\alpha = 0.01$. All significance tests execute two-tailed tests except for the sanity hypothesis (Section 5.2.4), which is a one-tailed hypothesis. Analyses are performed in RStudio $ver.1.0.136$ using packages $stats$, car, $multcomp$ and $ggplot2$.

4 EXECUTION OF THE EXPERIMENT

4.1 Execution

The execution followed the same pattern for both the experimental sessions, *ES*1 and *ES*2. It comprised of three phases, pre- experimental data collection, experimental session execution, and post-experimental data collection.

Pre- experimental data collection involved filling in questionnaires and signing of consent forms reasonably prior to the actual experimental session each year. We conducted experimental session's execution of the two groups TP and NTP in parallel but in different lecture rooms. The allocation of the participants to the two groups followed randomisation. Every second or third participant arriving to the designated place was sent to other group by keeping the count of the participants balanced in each group. Additionally, participants were confirmed again of the filling of pre-questionnaire and consent form before their assignment to either of the groups. We kept the participants uninformed of the reason of random allocation. Randomised allocation resulted in 22 participants in TP group and 21 in NTP group of the session - *ES*1. *ES*2 contained 24 in TP and 24 in NTP group but 3 students in this group did not give consent to participate.

We developed two detailed scripts for the experimenters to guide through the experimental execution for both TP and NTP groups. The scripts contain time-stamped sequence of activities along with the instructions of conducting the activities and the verbal content to be communicated to the participants by the experimenters. The only difference between TP and NTP script is the administration of time-pressure in the two groups. Accordingly, 30 minutes were allocated to TP group and 60 minutes were allocated to NTP group for the actual task, per the operationalisation detailed in Section 3.5. After the actual task, post-experimental data collection was done

using NASA-TLX, for which the instructions were also part of the script. The actual duration of the experimental sessions, *ES*1 and *ES*2, remained consistent with the planned duration.

4.2 Data Collection

We extracted the data by counting the number of valid test-cases designed by the participants. Dropping the test-cases that did not qualify as a valid test-case was based on the following criteria:

- Repeated test-case: A duplicate test-case.
- Wrong test-case: Test-cases whose content of the columns of the test-case template are in contradiction to each other, i.e., a test-case depicting a wrong understanding of the requirements.
- Test-cases conflicting with the specifications: A test-case validating an unspecified behaviour of the application, which cannot occur simultaneously with the rest of the specified requirements.

The data extraction was done by one author but to lower the subjectivity in marking, we performed a pilot-marking (valid or not-valid test-case) on a randomly chosen subset of test-cases (108) from ES1 data only. This was proceeded by calculating an inter-rater agreement (kappa) between the two authors. We computed a Randolph's free marginal kappa= 0.963 because our case of assigning values to the categories qualifies for free-marginal kappa calculations [20]. The kappa value indicated adequate inter-rater agreement therefore, one author proceeded with the marking of test-cases. The extraction resulted in the dropping of 11 test-cases from ES1 data and 10 test-cases from ES2 data. This resulted in total of 371 test-cases in ES1 and 399 in ES2. During data extraction, we dropped one participant from ES1 (TP group) because all of his/her test-cases were dropped. The validity improvement step (Section 3.5.2) resulted in the addition of 13 test-case to the baseline test-suite, which increased the count to 68 from 55 test-cases. We finally, converted the extracted data into *actual-performance* data and extracted *perceived-performance* data from NASA-TLX.

5 RESULTS

5.1 Descriptive Statistics

We explore the data by presenting descriptive statistics and boxplots. Table 5 presents the descriptive statistics of actual performance-*AP* on designing test-cases by the participants in experimental sessions 1 and 2 - ES1 and ES2. In the table, *mdn* stands for median, *min* for minimum, *max* for maximum and *sd* for standard deviation. We can see in Table 5 that the values of TP and NTP of ES1 are quite close to each other and *sd* is almost equal for both the groups - 3.568 and 3.586. For ES2, a degree of difference can be observed for the values of mean and median between TP and NTP, whereas the rest of the values are similar. However, overall greater values of NTP compared to TP in ES1 and ES2 are indicative of more time availability in NTP sessions. Comparing values across ES1 and ES2, we can see that NTP values in ES2 are greater than NTP values in ES1, but it is opposite for TP. The standard deviations, *sd*, in ES2 are slightly greater than in ES1.

Descriptive statistics of perceived-performance are presented in Table 6. It is evident from the table for ES1 that the mean and median of perceived performance in NTP (47.381 and 50.000) are much greater than TP (39.048 and 35.000). However, in ES2, mean and median values of perceived performance in TP group (47.292 and 45.000) are greater than NTP group (43.333 and 40.000). Standard deviation of NTP, in both ES1 (21.944) and ES2 (20.391), is greater than TP (19.788 and 18.296). Comparison of values across the experimental sessions show that perception of performance in TP groups has increased from ES1 to ES2, whereas perception of performance in NTP group decreased from ES1 to ES2.

Table 5: Actual Performance - Descriptive Statistics

		Actual Performance				
ES	**Group**	**mean**	**mdn**	**min**	**max**	**sd**
1	**TP**	12.143	12.000	7.000	21.000	3.568
	NTP	13.810	13.000	9.000	22.000	3.586
2	**TP**	11.375	10.000	6.000	22.000	4.179
	NTP	14.857	15.000	6.000	24.000	5.092

Figures 3, 4 and Figures 5, 6 present the box plots for actual performance and perceived performance respectively. We can see from Figures 3 and 4 that the difference between the median values of the Groups in ES2 (NTP 15 and TP 10) is greater than the difference between the median values of the Groups in ES1 (NTP 13 and TP 12). Considering the range and spread of values of perceived performance, the difference in the median values within the experimental sessions are not considerably great. Although it is 15 units difference in ES1 (Figure 5) and 5 units difference in ES2 (Figure 6). For perceived performance, the difference of median values between the same groups, i.e., TP of ES1 vs. TP of ES2 and NTP of ES1 vs. NTP of ES2, has the same difference of 10 units.

5.2 Hypotheses Testing

5.2.1 H1 - Actual performance under time-pressure differs from the actual performance under no time pressure. The results of normality test of the residuals executed on actual-performance model satisfied the assumption with $p - value = 0.043$. The homogeneity of variance, assessed using Levene's test, for the two groups of experimental sessions is $F(1, 85) = 3.592, p - value = 0.061$ and for time-pressure groups is $F(1, 85) = 0.707, p - value = 0.403$, which indicates that the assumption is satisfied. Further, we also tested homogeneity of variance for the interaction of the two factors, i.e., among four groups, which is $F(3, 83) = 1.449, p - value = 0.235$, indicating that assumption still holds true. We are primarily interested in the effect of time-pressure therefore experimental-session is a nuisance factor for us. Considering this, we executed a 2-way ANOVA with blocking and computed type-III sum-of-squares for the model because our sample sizes are unequal [6], [29]. The results show a significant effect of time-pressure on the actual performance of the participants , $F(1, 84) = 8.523, p - value = 0.004, \omega^2 = 0.079$. We use *Tukey* for *post-hoc* analysis and the test also revealed that actual performance significantly differed between TP and NTP groups as an effect of time-pressure with $p - value = 0.004$. We could not observe a significant effect of experimental sessions on the actual performance of the participants, $F(1, 84) = 0.015, p - value = 0.901, \omega^2 = -0.011$. *Post-hoc test* for experimental sessions' effect

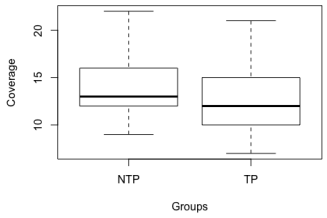

Figure 3: AP in ES1

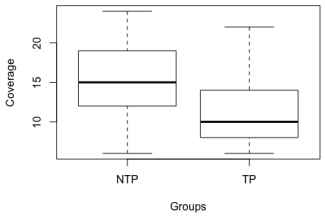

Figure 4: AP in ES2

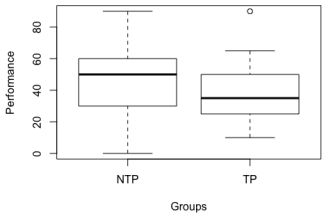

Figure 5: PP in ES1

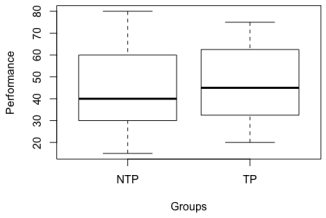

Figure 6: PP in ES2

Table 6: Perceived Performance - Descriptive Statistics

		Perceived Performance				
ES	**Group**	**mean**	**mdn**	**min**	**max**	**sd**
1	**TP**	39.048	35.000	10.000	90.000	19.788
	NTP	47.381	50.000	0.000	90.000	21.944
2	**TP**	47.292	45.000	20.000	75.000	18.296
	NTP	43.333	40.000	15.000	80.000	20.391

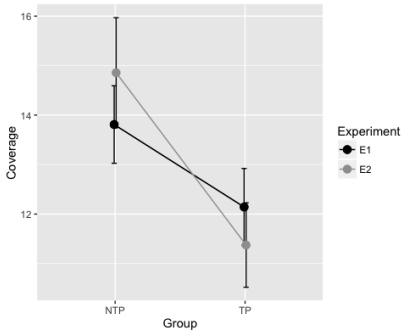

Figure 7: Line Error Bar-graph of Actual Performance

also reveal no significant effect on the actual performance with $p - value = 0.902$. Nonetheless, we reject the null hypothesis because actual performance significantly differ between TP and NTP groups as a result of the main effect of time-pressure.

Figure 7 presents a line error-bar graph (standard error of the mean) and shows a disordinal interaction. We are interested only in the main effect of time-pressure since we executed a blocked ANOVA for experimental-sessions. It is clear that the mean of actual-performance in NTP group is more than the mean in TP group for both experimental-sessions, $E1$ and $E2$. However, the mean of TP decreases in E2 compared to TP mean in E1, which results in a greater difference of the means of NTP and TP for E2.

5.2.2 H2 - Perceived performance under time-pressure differs from perceived performance under no time pressure. The results of normality test of the residuals of perceived-performance model satisfied the assumption with $p - value = 0.049$. Homogeneity of variance assumption is satisfied for both experimental sessions ($F(1, 85) = 0.137, p - value = 0.712$) and time-pressure groups

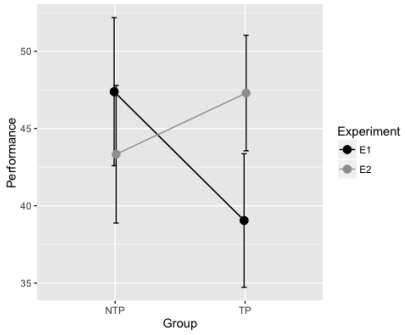

Figure 8: Line Error Bar-graph of Perceived Performance

($F(1, 85) = 0.923, p - value = 0.339$), and also for their interaction ($F(3, 83) = 0.203, p - value = 0.894$). Experimental-sessions is again a nuisance factor for us therefore, we execute the same type of 2-way ANOVA as for $H1$. The results do not reveal a significant effect of time-pressure on the perceived performance of the participants, $F(1, 84) = 0.210, p - value = 0.648, \omega^2 = -0.009$. *Post-hoc test (Tukey)* also confirmed the results of no significant difference due to time-pressure between TP and NTP groups with $p - value = 0.648$. We also could not observe a significant effect of experimental-sessions on the perceived performances, $F(1, 84) = 0.280, p - value = 0.598, \omega^2 = -0.008$. *Post-hoc test* also confirmed the result of the effect of experimental-sessions with $p - value = 0.598$. We therefore, fail to reject the null hypothesis that perceived performance is equal in TP and NTP groups.

Figure 8 presents a line error-bar graph (standard error of the mean) for the perceived performance. The graph shows a disordinal interaction. The mean of perceived performance of NTP group in $E1$ = ES1, is more than then perception of performance in $E2$ for NTP group. On the other hand, perceived performance of TP group in $E1$ is lesser than perceived performance in $E2$. Nonetheless, we could not observe a significant effect of time-pressure or experimental-sessions, as main effect, on the perceived performances.

5.2.3 H3 - Actual performance correlates with perceived performance. First we validate $H3_A$ for pooled data, i.e., finding correlation between actual and perceived performance without accounting for time-pressure groups and experimental-session groups. The normality test therefore, performed on the pooled data violated parametric assumptions thus, we compute *Kendall's tau* because there

was a tie in large numbers of ranks in the data. Based on Kendall's tau, $\tau = -0.12$ and $p - value = 0.1027$, we fail to reject the null hypothesis. According to this result actual-performance is not significantly related to perceived performance. Hence, we explore the relationship further by blocking the possible influence of time-pressure and experimental sessions. We therefore, execute a second-order partial correlation to assess a pure measure of the relationship between actual and perceived performances. The results showed a significant negative correlation between actual and perceived performances, $r = -0.284$ (small to medium effect), $p - value = 0.008$ and variance is $R^2 = 0.08$, when controlled for time-pressure and experimental sessions. This indicates that perceived performance accounts 8% of variance in actual performance. Pertaining to the observed result of the second-order partial correlation, we decided to explore the correlation further. We therefore, performed first-order partial correlation by controlling only for time-pressure's effect. The result again indicates a significant negative correlation of $r = -0.282$ (small to medium effect), $p - value = 0.008$ and the variance is $R^2 = 0.08$, which again indicates that perceived performance explains 8% of variance in actual performance.

When controlled for experimental-sessions and time-pressure together, a negative correlation was observed. Exploring the correlation further revealed that experimental-sessions have got no influence because the correlations and variance for first-order and second-order are approximately the same. Yet, we could not observe a correlation in the pooled data because time-pressure groups (TP and NTP) could have influenced in opposite directions, cancelling out each others' effect. Nonetheless, we reject the null hypothesis based on the partial correlation results.

5.2.4 A Sanity Hypothesis. We formulate a sanity hypothesis to validate operationalisation of time-pressure in our study. This hypothesis states that,

H: Testers in time-pressure group experience more temporal demand than testers in no-time-pressure group.

$$H_A : \mu(TD)_{TP} > \mu(TD)_{NTP}$$

$$H_0 : \mu(TD)_{TP} \leq \mu(TD)_{NTP}$$

We use the ratings of *temporal demand* attribute of NASA-TLX for testing our sanity hypothesis. Before executing a statistical test, we explore the data of this attribute by plotting a line error-bar graph. Figure 9 shows that NTP of ES1 endured more temporal demand compared to NTP of ES2. Whereas, participants of TP groups of ES1 and ES2 endured temporal demand with a minor difference from each other, with TP-ES2 more than TP-ES1. We test it for the pooled data, i.e., without accounting for experimental sessions. While there may be differences due to experimental sessions (characteristics of the participants), the instrumentation of time pressure is the same across both years. The assumption of normality got violated for TP groups' data, $p - value << 0.000$ therefore, we perform a non-parametric alternative of the independent *t-test*, i.e., *Mann-Whitney U test*. We observed a significant difference between TP and NTP groups with $p - value << 0.000$, effect-size $r = -0.517$ (large), $power = 0.518$ and $df = 85$. The null hypothesis is rejected that indicates that participants of TP group experienced more time pressure (temporal demand) than participants in NTP group.

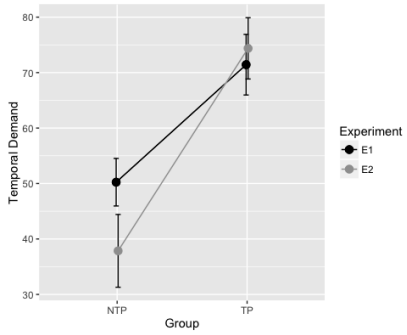

Figure 9: Line Error Bar-graph of Temporal Demand

6 DISCUSSION

We have observed time pressure to have a significant effect on the actual performance of the participants. Descriptive statistics inform us about the direction of the effect that participants in no-time-pressure groups designed more test-cases than the participants in time-pressure groups. This difference in the actual-performance increased more in the second experimental session (ES2), though we are not interested in the interaction. Figure 9 depicts the observed pattern of actual-performance (descriptive statistics), as a result of time pressure, in terms of temporal demand. We can see that less perceived temporal demand (time pressure) is related with the increased actual-performance for NTP-ES2 compared to NTP-ES1.

The observed patterns of actual-performance can be attributed to the characteristics of the participants employed in the two experimental sessions, ES1 and ES2. The collected background data of the participants of ES1 and ES2 show that participants of ES1 are relatively more experienced, except for industrial testing experience which is quite similar across experimental sessions. Therefore, for the actual-performance, the difference between TP and NTP groups of ES1 is relatively less compared to the groups of ES2. This indicates experience as a possible confounding factor with time pressure and actual-performance. Furthermore, the dropping of actual-performance in TP-ES2 compared to TP-ES1 could also be due to the experience of the participants, i.e., participants of TP-ES2 endured more time-pressure because of their lesser experience. Although, TP groups (ES1 and ES2) endured almost similar level of time pressure (temporal demand), per Figure 9. Another plausible factor contributing to the almost similar levels of perceived temporal demand for TP-ES1 and TP-ES2 is the experimenter. Since the experimenter in TP groups was the same for both experimental sessions, participants could have experienced similar levels of temporal demand, though time-pressure was regulated by following the guidance script. Apart from the effect of time-pressure, overall less percentage of actual performance by each participant could be pertained to the *complexity of test-case* aspect. Since MusicFone is a realistic object, designing test-cases from identified scenarios could have been difficult for the participants.

We could not observe time pressure to have a significant effect on the perception of performance of the participants. Despite not having a directional hypothesis for observing the probable effect of time-pressure on perceived performance, the results indicate an

inaccurate performance perception for the task because of the observed effect on actual performance. However, descriptive statistics inform that participants perceived to perform better under time pressure, compared to no-time-pressure group, in ES2, which is contrary to their actual performance in the same session. This can be a reason for the observed negative correlation between actual and perceived performance, when controlled for the effect of time-pressure and experimental-sessions. Moreover, participants in TP groups experienced more temporal demand than the participants in NTP groups, which indicates that we successfully regulated time-pressure and no-time- pressure conditions. The observed results of perceived performance are also not aligned with the perceived temporal demand, since the results suggest that regulated time-pressure affected the perceived temporal demand but still perceived performance remained unaffected.

6.1 Implications

The results imply that time-pressure affects actual performance of testers which may cause adverse effects on software quality. Because, giving less test coverage to the specifications may result into the ignorance of functionality that may appear trivial but are critical from integration or system level testing perspective. This may adversely affect external quality and increase development cycle costs. Self-assessment or the perceived performance by a tester is critical in this regard. This may lead to a misunderstanding of ones own actual performance, if self-assessments are not accurate. Especially in the case, when actual performance is lower than the perceived performance of the task undertaken. Hence, it is important for testers to improve in their self-assessments because this would lead them towards a better understanding of their real performance. It would also help in identifying their strengths and weaknesses when performing testing and to improve in their actual performance.

Practitioners can be improved in their self assessments of performance during their traineeship or probation period of their hiring. After performing the task, they should be asked to assess their own performance and then compare it with their actual performance, as a feedback by the supervisors. This would help them improve not only in their own self assessments but also the quality of the work. Thus, when they are engaged into scheduled assignments they are able to reflect well on their actual performance even under time pressure and manage a testing task at hand with a sound strategy. This practice may lead to a decrease of inaccurate risky self assessments under time pressure, and also for less challenged (time-wise) tasks, resulting in improved software quality, e.g., improved external quality and decrease in the number of feedback cycles between developers and testers. Furthermore, ability to assess performance well, may also help practitioners in providing better testing-task estimates during the planning phase of a project. Because, they know, how much time would they need to achieve a certain level of performance in terms of a quality testing.

7 THREATS TO VALIDITY

Following the list of validity threats provided by Wohlin et al . [28], we report only the relevant ones.

Conclusion Validity: We have addressed the threat of violated assumptions by carefully checking for the assumptions required by every statistical test and then choosing the right test for validating all the hypotheses. Reliability of measures threat is addressed by conducting a pilot-run, which improved our instrumentation and also established a common understanding for marking the test-cases during data extraction. Moreover, having an acceptable kappa-value for data extraction also reduced a subjectivity factor in the evaluation/marking of test-cases. We addressed the reliability of treatment implementation threat by developing detailed experimental execution scripts not only for treatment group but also for controlled group. The experiments were conducted in class-rooms therefore, it provided a control for possible external factors, e.g., noise and unexpected interruptions, introducing irrelevancies in experimental settings.

Internal Validity: Internal validity threats related to multiple group experiments are discussed in this section because our study comprised of multiple experimental groups. Our study is not prone to imitation of treatments threat because in both ES1 and ES2 sessions, treatment and controlled experimental sessions were executed in parallel. Additionally, all the participants were trained together on the same kind of training tasks in both academic years ($2015 - ES1$ and $2016 - ES2$) before the execution of the actual experiment. And were assigned in randomised order to the control and treatment groups in ES1 and ES2, this addressed interaction with selection threat to our study. None of the groups in both experimental sessions were given any kind of compensations which avoided the threat of compensatory equalisation of treatments.

Construct Validity: Our study is prone to mono operation bias threat because of our design, which involves only a single object (interaction of task and treatment). This confines the result of our study to one type of object. We calculated inter-rater agreement (kappa) to lower the subjectivity in marking of the test-cases to ensure the reliability of measurement, which addressed mono method bias threat. The metric that we used to assess actual-performance can be considered as *functional completeness* according to ISO/IEC25010; *"degree to which the set of functions covers all the specified tasks and user objectives"* [9]. If the current study is replicated then the validity improvement step (Section 3.5.2) may increase the number of total test-cases in the baseline test-suite; which will further lower the actual-performance. We do not consider it a validity threat because the actual and perceived performance are assessed separately for a potential impact of time-pressure, and perceptions can only be influenced if participants are aware of the possible total number of test-cases. This is not advised to be communicated as an experimenter. The participants were not aware of the treatment, therefore, it was not possible for them to guess the hypotheses. Additionally, the participants were informed that the activity (experimental task) would not be evaluated and also would not affect grading.

External Validity: Our study is prone to interaction of selection and treatment threat because we employed students for our study instead of professionals. Though considering the experience of the participants, they can be referred as representative of novice professionals. We conducted experiment in an academic setting, it is not a realistic environment but it provided a better control of the environment from other interfering factors. We limited the threat of setting and treatment by utilising a realistic task as an object of

our study. Although the task was a paper and pen activity, we do not consider it as a serious threat to generalizability because our scope was limited only to the design of test-cases.

8 CONCLUSION AND FUTURE WORK

Our study examined the effect of time-pressure on actual performance and perceived or self-assessed performance for the same executed task, along with the correlation between actual and perceived performance. We executed two controlled experiments, in two different academic sessions, with altogether 87 Master's degree students. Our research question, based on the results of the study is answered as:

Time-pressure significantly impacted the actual performance, i.e., the percentage of functional test-cases designed by the participants significantly differed between time-pressured and no-time-pressured groups. Although the perception of performance of the participants did not vary due to time-pressure for the implemented task, i.e., over-estimation. We also observed that actual performance is significantly correlated with perceived performance but negatively.

For testing practitioners, we recommend that they should improve in their perception or self-assessment of the task performed, especially in time pressured situations. This would enable them in becoming aware of their actual performance, reflect well on it and improve it, not only when challenged with time but also in normal paced situations.

Possible future work extensions of this work include conducting external replications and studying the confounding factors. For example, characteristics or experience of the participants, experimenter and the possible effect of using a different experimental object. Additionally, it would be beneficial for SE body of knowledge to examine the comparison of actual and perceived or self-assessed performance irrespective of other factors, e.g., time-pressure. Also, more research needs to be done on, how to raise sensibility of the impact of inaccurate perceived performance?

ACKNOWLEDGMENTS

This research is supported in part by the Academy of Finland Project #278354.

REFERENCES

[1] Jonlee Andrews and C Smith. 1996. In Search Factors of the Marketing the Imagination : of Products Mature Affecting Creativity for Marketing Programs. *Journal of marketing research* 33, 2 (1996), 174–187. https://doi.org/10.2307/3152145

[2] Nathan Baddoo and Tracy Hall. 2003. De-motivators for software process improvement: An analysis of practitioners' views. *Journal of Systems and Software* 66, 1 (2003), 23–33.

[3] Victor R. Basili, Gianluigi Caldiera, and Dieter H. Rombach. 1994. The Goal Question Metric Approach. *Encyclopedia of Software Engineering* 1 (1994). https://www.bibsonomy.org/bibtex/2ed38a7a0ec5148979dc72d0a65034eb4/stammel

[4] F Daniel Duffy and Eric S Holmboe. 2006. Self-assessment in Lifelong Learning and Improving Performance in Practice: Physician Know Thyself. *JAMA* 296, 9 (2006), 1137–1139. https://doi.org/10.1001/jama.296.9.1137

[5] D Dunning, C Heath, and J M Suls. 2004. Implications for Health, Education, and the Workplace. *Psychological Science in the Public Interest* 5, 3 (2004), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

[6] Andy Field, Jeremy Miles, and Zoe Field. 2012. *Discovering Statistics Using R*. SAGE Publications Ltd. 992 pages.

[7] Davide Fucci, Burak Turhan, Natalia Juristo, Oscar Dieste, Ayse Tosun-Misirli, and Markku Oivo. 2015. Towards an operationalization of test-driven development skills: An industrial empirical study. *Information and Software Technology* 68 (2015), 82–97.

[8] Human Performance Research Group and NASA. 1987. Nasa Task Load Index (TLX). (1987).

[9] ISo. 2018. ISO 25000 Software Product Quality. (2018). http://iso25000.com/index.php/en/iso-25000-standards/iso-25010

[10] M. Jørgensen. 2004. A review of studies on expert estimation of software development effort. *Journal of Systems and Software* 70, 1-2 (2004), 37–60. https://doi.org/10.1016/S0164-1212(02)00156-5

[11] Barbara Kitchenham, Shari Lawrence Pfleeger, Beth McColl, and Suzanne Eagan. 2002. An empirical study of maintenance and development estimation accuracy. *Journal of Systems and Software* 64, 1 (2002), 57–77. https://doi.org/10.1016/S0164-1212(02)00021-3

[12] Miikka Kuutila, Mika V. Mantyla, Maelick Claes, and Marko Elovainio. 2017. Reviewing literature on time pressure in software engineering and related professions: Computer assisted interdisciplinary literature review. In *Proceedings - 2017 IEEE/ACM 2nd International Workshop on Emotion Awareness in Software Engineering, SEmotion 2017*. 54–59. https://doi.org/10.1109/SEmotion.2017.11

[13] Leventhal L. M., Teasley B. E., and Rohlman D. S. 1994. Analyses of factors related to positive test bias in Software Testing. *International Journal of Human-Computer Studies* 41 (1994), 717–749.

[14] Mika V. Mäntylä and Juha Itkonen. 2013. More testers-The effect of crowd size and time restriction in software testing. *Information and Software Technology* 55, 6 (2013), 986–1003.

[15] Mika V. Mäntylä, Kai Petersen, Timo O. a. Lehtinen, and Casper Lassenius. 2014. Time pressure: a controlled experiment of test case development and requirements review. In *Proceedings of the 36th International Conference on Software Engineering - ICSE 2014*. 83–94. https://doi.org/10.1145/2568225.2568245

[16] Julie L. Montgomery and Wendy Baker. 2007. Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing* 16, 2 (2007), 82–99. https://doi.org/10.1016/j.jslw.2007.04.002

[17] I. Myrtveit and E. Stensrud. 1999. A controlled experiment to assess the benefits of estimating with analogy and regression models. *IEEE Transactions on Software Engineering* 25, 4 (1999), 510–525. https://doi.org/10.1109/32.799947

[18] Ning Nan and Donald E. Harter. 2009. Impact of budget and schedule pressure on software development cycle time and effort. *IEEE Transactions on Software Engineering* 35, 5 (2009), 624–637.

[19] Z Papamitsiou and A A Economides. 2014. Students' Perception Of Performance Vs. Actual Performance during Computer Based Testing: A Temporal Approach. In *Inted2014: 8th International Technology, Education And Development Conference*. Valencia, Spain, 401–411.

[20] J. J. Randolph. 2008. Online Kappa Calculator. (2008). http://justusrandolph.net/kappa/#dInfo

[21] Iflaah Salman, Ayse Tosun Misirli, and Natalia Juristo. 2015. Are students representatives of professionals in software engineering experiments? *Proceedings - International Conference on Software Engineering* 1 (2015), 666–676. https://doi.org/10.1109/ICSE.2015.82

[22] Iflaah Salman and Burak Turhan. 2018. Instrumentation for the experiment: Effects of Time-pressure on Perceived and Actual Performance. (2018). https://doi.org/10.5281/zenodo.1169185

[23] Hina Shah, Mary Jean Harrold, and Saurabh Sinha. 2014. Global software testing under deadline pressure: Vendor-side experiences. *Information and Software Technology* 56, 1 (2014), 6–19.

[24] D.I.K. Sjoberg, B. Anda, E. Arisholm, T. Dyba, M. Jorgensen, A. Karahasanovic, E.F. Koren, and M. Vokac. 2002. Conducting realistic experiments in software engineering. In *Proceedings International Symposium on Empirical Software Engineering*. 17–26. https://doi.org/10.1109/ISESE.2002.1166921

[25] Heikki Topi, Joseph S. Valacich, and Jeffrey A. Hoffer. 2005. The effects of task complexity and time availability limitations on human performance in database query tasks. *International Journal of Human Computer Studies* 62, 3 (2005), 349–379. https://doi.org/10.1016/j.ijhcs.2004.10.003

[26] Fiona Walkerden and Ross Jeffery. 1999. An Empirical Study of Analogy-based Software Effort Estimation. *Empirical Software Engineering* 158 (1999), 135–158. https://doi.org/10.1023/A:1009872202035

[27] David N. Wilson and Tracy Hall. 1998. Perceptions of software quality: a pilot study. *Software Quality Journal* 7, 1 (1998), 67–75. http://link.springer.com/article/10.1023/B:SQJO.0000042060.88173.fe

[28] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, BjÃűrn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering*. Springer. 236 pages. https://doi.org/10.1007/978-3-642-29044-2

[29] Jing Xu. [n. d.]. Randomized Blocks Designs and Two-Way ANOVA Randomized blocks designs and principles of experimental. ([n. d.]), 8 pages. http://www.bbk.ac.uk/ems/faculty/xu/xu-downloads/SML4.pdf