Conformance Factor in Test-driven Development: Initial Results from an Enhanced Replication

Davide Fucci
Dept. Information Processing Science
University of Oulu, Finland
davide.fucci@oulu.fi

Burak Turhan
Dept. Information Processing Science
University of Oulu, Finland
burak.turhan@oulu.fi

Markku Oivo
Dept. Information Processing Science
University of Oulu, Finland
markku.oivo@oulu.fi

ABSTRACT

Test-driven development (TDD) is an iterative software development technique where unit-tests are defined before production code. The proponents of TDD claim that it improves both external quality and developers' productivity. In particular, Erdogmus et al. (i.e., original study) proposed a two-stage model to investigate these claims regarding TDD's effects. Our aim is to enhance the model proposed in the original study by investigating an additional factor: TDD process conformance. We conducted a close, external replication of the original study accompanied by a correlation analysis to check whether process conformance is related to improvements for the subjects using TDD. We partially confirmed the results of the original study. Moreover, we observed a correlation between process conformance and quality, but not productivity. We found no evidence to support the claim that external quality and productivity are improved by the adoption of TDD compared to test-last development. Finally, conformance to TDD process improves the quality and does not affect productivity. We conclude that the role of process conformance is relevant in studying the quality and productivity-related effects of TDD.

1. INTRODUCTION

Test-driven development (TDD), despite its name, is also a design technique in which unit-tests — encompassing a small functionality—are written before the actual production code [1]. The developer using TDD then writes the minimal amount of code necessary to make the test pass to keep the code design as simple as possible. Subsequently, if necessary, the developer refactors the code (both production and test). The development continues following the three steps aforementioned: write a failing test, make it pass, refactor [1]. Although TDD has been studied using different approaches for over a decade in several studies [7, 9, 11], there is contrasting evidence about its claimed effects [1]. One of the reasons for such shallow evidence is the difficulty of handling the several variables that come into play when dealing with the factors involved in the software development process. One way to overcome this hurdle is the execution of several replications of the same study which progressively focus on different aspects in order to create a comprehensive vision of the phenomenon [6].

We have studied TDD in academic settings in our previous works [3, 4]. Our aim was to observe, through a controlled experiment, TDD's effects on the external quality of software artefacts and productivity of developers. In this paper we present a replication of the study by Erdogmus et al. [2]. We enhanced the original study by analysing an additional factor. In particular, we examined the relationships between process conformance to TDD and the independent variables (i.e., quality and productivity) used in the original experiment.

2. RELATED WORK

In this section we present two systematic literature reviews about the effects of TDD, along with primary studies addressing the role process conformance in TDD research.

Test-driven development has been studied in several primary studies through controlled experiments. These primary studies are summarised mainly in two secondary studies. The first is a systematic literature review [13] that includes 32 experiments spanning nine years (2000–2008) in both academic and industrial settings. The experiments, published in 22 peer-reviewed papers, compared TDD to a non-TDD technique. The results indicate that TDD has a positive effect on quality, both external and internal (e.g., code maintainability, re-usability, design), whereas no significant effect can be found regarding productivity. The second secondary study is a meta-analysis [10] that reported similar results. In particular, the authors reviewed 37 primary studies published until 2011 in 25 peer-reviewed papers. The studies were grouped according to different factors such as subjects' experience, use of other XP practices (e.g., pair-programming) or context factors such as industrial vs. academic studies. When comparing experiments in academia with experiments in industry, quality was improved by the employment of TDD in industrial settings, although with a drop in productivity when compared to academia. In both secondary studies, the authors identified process conformance as one of the factors threatening the validity of the primary studies.

Quantifying process conformance is critical in TDD research [12]. In their study, Wang and Erdogmus [14] try to measure process conformance with the goal of giving feedback to the developers using TDD and help them to improve their process. They present a proof-of-concept tool that mines

Hackystat[1] data looking for patterns typically associated with the TDD process. The same approach—exploiting low level data automatically acquired by an automatic tool—is used by Johnson and Kou [5]. They use a set of heuristics to identify 22 events that commonly take place during development, and categorise them into 8 classes. Each sequence of events are then categorised as TDD compliant or not TDD compliant according to the class they belong to. Their tool, Zorro, has been validated by comparing its performance with the manual analysis of the video recording of developers' activities. Zorro was able to correctly identify 89% of the development cycle in university settings and 85% in industrial settings. The tool used to measure TDD process conformance in this study is based on Zorro.

3. REPLICATION

Due to space limitations we report the original study's salient points in tabular form. Please refer to Erdogmus et al. paper [2] for further details. Note that in this study we tackled the same hypothesis of the original study for stage 1—examining differences between TDD and non-TDD in terms of testing effort (1T), productivity (1P) and external quality (1Q)—and stage 2 —focusing on the correlation between testing effort and the TDD productivity (2P) and external quality (2Q). Table 1 reports the original study, our previous close replication, and this replication's context variables. Table 2 summarizes the results. The motivations for this replication are twofold. First, we want to enhance the original study by limiting some of the threats to its validity. Second, we want to add data points to our previous replication [4] of the same study, to carry out a meta-analysis (not reported in this paper). We consider the study reported in this paper to be an enhanced, close, external replication [6], since our aim is to strengthen the design used in the original study by overcoming some of its limitations while keeping the core study intact.

As in our previous replication [3], the experiment took place in academic settings. The subjects were sampled by convenience among the students taking part to the *Software Quality and Testing* lab course during Fall 2013. The subjects were trained for six three-hour sessions during which they were introduced to unit-testing concepts in Java using the Eclipse IDE and JUnit, and test-driven development. During each session a hands-on tutorial was given by the instructors followed by programming exercises of increasing difficulty throughout the duration of the course. At the beginning of the course the subjects were given a pre-questionnaire to fill out in order to gauge their initial skills regarding object-oriented programming, unit-testing and the tools used during the course. The experiment took place during the last session; we asked the subjects to sign an informed participation form—without disclosing our research goal—so that the data could be retained only from the subscribers. We reached an unbalanced design after randomly dividing the subjects into TDD (experimental) or test-last development (TLD) (control) group. We asked the subjects to tackle a modified version of the Robert Martin's Bowling Scorekeeper kata, composed of 13 fine-grained user stories, using the development technique they were assigned to. We provided the subjects a stub project, containing the methods signatures necessary to run our acceptance tests (30 SLOC). Upon completion, we recorded the time and asked the subjects to fill out a post-questionnaire to gauge their opinion and feedback about the course and the use of TDD. At the end of the three hours, the remaining subjects returned their solutions and filled the post-questionnaire. We discarded 7 of the 41 solutions we yielded from the participants because these were either empty or failing to compile.

Besides the original study metrics, in order to test our additional hypothesis, we calculated one additional metrics; CONF that gauges the process conformance level of the TDD subjects. The value for CONF was calculated using a tool[2] that automatically captures and classifies sequences of development events into *development episodes*. Each development episode is then categorised as TDD or non-TDD compliant. CONF is the ratio between the number of TDD-compliant episodes and the total number of development episodes identified by the tool, normalised by 100. The range for CONF is [0, 100].

Since we introduce an additional factor, process conformance, that was not taken into account in the original study we formulated an additional hypothesis in which external quality and productivity are put in correlation with the level of conformance following assumption that the claimed effects of TDD should be more evident when the process is rigorously followed. These hypotheses are formalised as follows [3]:

3Q - Are process conformance and external quality linearly correlated?

H_0: $QLTY = \beta_0 + \beta_1 \times CONF, \beta_1 = 0$

H_1: $QLTY = \beta_0 + \beta_1 \times CONF, \beta_1 \neq 0$

3P - Are process conformance and productivity linearly correlated?

H_0: $PROD = \beta_0 + \beta_1 \times CONF, \beta_1 = 0$

H_1: $PROD = \beta_0 + \beta_1 \times CONF, \beta_1 \neq 0$

When planning the experiment we had to introduce changes in the experiment. In the original experiment the subjects were allowed to work on the task for several sessions and/or remotely. Hence, the net time varied from few hours up to 25 hours [2]. The need for this change was already known but, due to organisational limitation and the unavailability of the infrastructure for remote working, the experiment duration was fixed to a maximum of three hours. We acknowledge that this change might have an impact on the results of the experiment. The subjects might have felt discouraged in completing the task in such short time; particularly subjects in the TDD group might deviate from the prescribed development cycle due to the time pressure. Consequently we simplified the relevant parts of the tasks in the original experiment by removing the need for dealing with the handling of input and the formatting the output. The acceptance test suite used in this replication includes 56 tests (105 in the original study). We consider ours to be an external replication [6] since the original experimenters did not have an active role in the process. Nevertheless, we obtained from them a lab package containing the acceptance test suite and the spreadsheet to calculate the metrics. The original metric for QLTY was normalised by a difficulty factor. Such

[1] `http://www.hackystat.org`

[2] `https://github.com/brunopedroso/besouro`

[3] Please refer to the original study [2] for the definition of $QLTY$, $PROD$ and $TEST$

Context variable	Erdogmus et al., 2005	Fucci and Turhan, 2013	This replication
Subject type	46 undergraduate (35 after drop-outs)	33 graduate, 25 undergraduate	41 mixed (34 after drop-outs)
Subject unit	Individuals	Pairs and individuals	Individuals
Time to complete the task	Several lab sessions, remote work	Single lab session (3 hours)	Single lab session (3 hours)
Experiment design	One factor, two treatments	One factor, two treatments	One factor, two treatments
Control/treatment size	13/11	20/27	16/18
Variables	$TEST, PROD, QLTY$	$TEST, PROD, QLTY$	$TEST, PROD, QLTY, CONF$

Table 1: Original study and replications' contexts.

factor was not available to us and was not included when calculating QLTY. We foresee that this change might have flattened the distribution of the QLTY variable compared to the original study. The changes in the context and metrics do not allow a direct comparison between our study and the original one.

4. RESULTS

In this section we present the result of the statistical analysis. We proceeded by following the two stages of the original experiment and then checking the impact of the other factor; i.e., process conformance. The mean values for QLTY

Hypothesis	[2]	[3]	This replication
1T	✓	✗	✗
1P	✗	✗	✗
1Q	✗	✗	✗
2P	✓	✓	✓
2Q	✗	✗	✗
3Q	NA	NA	✓
3P	NA	NA	✗

Table 2: Results for the original study [2], Fucci and Turhan, 2013 [3], and this replication. ✗= Failed to reject H_0, ✓= Reject H_0

and PROD are higher in the TLD group, whereas the TDD group has a slightly better value of TEST. Note that the variable TEST shows a leverage point in the TLD group ($TEST = 46$). The descriptive statistics are reported in Table 3. We used a one-tailed Mann-Whitney U-test to com-

Variable	Median	Mean	Std.dev
	Cumulative — $n = 34(33)$		
QLTY	88.54	88.56	6.27
PROD	2.78	4.44	4.30
TEST	7.22 (7.22)	9.41 (8.30)	8.90 (6.22)
	TDD group — $n = 18$		
QLTY	87.46	87.41	6.12
PROD	3.33	5.18	3.18
TEST	7.77	8.36	5.73
	TLD group — $n = 16(15)$		
QLTY	89.38	89.85	6.37
PROD	4.40	6.68	4.79
TEST	6.39 (5.56)	11.21 (8.22)	11.60 (6.97)

Table 3: Descriptive statistics for the experiment variables. Value calculated when omitting the leverage point are reported in parentheses.

pare the two groups (TDD and TLD) in terms of TEST. We decided to use a non-parametric test since the variables, including QLTY and PROD, are not normally distributed according to the Shapiro-Wilk test. The significance level is set at 0.1 due to the small sample size, as recommended in the original study [2]. The test results did not show that TDD subjects' testing effort is greater than TLD subjects ($W = 140.5$, $p - value = 0.55$) also when the leverage point is removed ($W = 140.5$, $p - value = 0.43$). Hence, the null hypothesis in 1T failed to be rejected. The result of the test for variable QLTY show that TDD subjects did not achieve better quality than TLD subjects ($W = 112$, $p - value = 0.27$), the null hypothesis in 1Q failed to be rejected. The one-tailed Mann-Whitney U-test shows that TDD subjects were not more productive than TLD ones ($W = 117$, $p - value = 0.83$). The null hypotheses in 1P failed to be rejected. We were unable to show any significant difference between the two groups in terms of testing effort, and proceed to analyse the the relationship with the other response variables; i.e., external quality and productivity.

The second stage of the original study focuses on correlation analysis. In particular, hypothesis 2Q expresses a linear relationship between the testing effort and external quality. The distribution of the data points suggests that the relationship between QLTY and TEST is not linear. The regression line ($\beta = 0.05$, $p - value = 0.74$) remarks that such relationship does not exist. Hence, the null hypothesis in 2Q failed to be rejected. Remarkably, an increasing testing efforts leads TDD subjects to improved quality ($\sim +4\%$), whereas TLD subjects quality diminishes ($\sim -2\%$). We found a significant moderate relationship ($\beta = 0.48$, $p - value << 0.1$) between PROD and TEST as expressed in the alternative hypothesis in 2P. The regression line equation is $PROD = 1.60 + 0.48 \times TEST$. The equation's slope indicates that there is an expected increase of 0.48 in productivity of each testing effort unit; on the other hand, it is not possible to give an interpretation to the intercept since in our context we cannot have $TEST = 0$. The values of R-squared ($R^2 = 0.56$) and residuals standard error ($\epsilon = 2.66$) show that the model is moderately good, since testing effort accounts for more than half of the variability in productivity (56%), while the average error that might be encountered when predicting productivity from testing effort is 2.66 PROD units. Hence we reject the null hypothesis in 2P. When considering the two experimental group independently, the model is still significant (TDD: $\beta = 0.53$, $p - value << 0.1$. TLD: $\beta = 0.44$, $p - value < 0.1$). The equations expressing the models are $PROD_{TDD} = 0.75 + 0.53 \times TEST$ with $R^2 = 0.64$, $\epsilon = 2,37$; and $PROD_{TLD} = 2.45 + 0.44 \times TEST$ with $R^2 = 0.52$, $\epsilon = 3.04$. Both groups of subjects reach the same level of productivity (~ 11), although this happens faster for the TDD group (steeper slope) which compensates for the lower baseline (intercept). Both regression coefficient and R-squared values are better in the TDD group when compared to the overall dataset and TLD group, further reinforcing the idea that testing effort is pivotal for the manifestation of the claimed effects of TDD on productivity.

Furthermore, check whether the subjects in the experimental group (i.e., TDD) followed the prescribed development cycle. The CONF values range between 15 and 100, $median = 75.50$ and $mean = 67.20$. Moreover, the 3rd quantile value is 93, possibly indicating a positively skewed distribution. This data shows that the subjects actually used TDD to a good extent. The correlation between quality and process conformance is presented in the following model: $QLTY = 81.65 + 0.09 \times CONF$. The linear model underlying the relationship is statistically significant ($p-value = 0.09$, $R^2 = 0.13$, $\epsilon = 3.74$) yet weak, leaving a window of opportunity for further investigation. Nevertheless, the null hypothesis in 3Q is rejected. When assessing the correlation between PROD and CONF we obtain the following model: $PROD = 1.69 + 0.05 \times CONF$. The direction of the relationship is positive, but not significant ($p - value = 0.14$). Hence, the null hypothesis in 3P fails to be rejected.

5. THREATS TO VALIDITY

Being this work focused on theory testing rather than exploratory we gave priority to threats dealing with internal validity The main threat to the validity of our study is the low representativeness of the subjects when the population of reference is *software developers*. There are also some social threats that could have played a role. In fact, from the analysis of the post-questionnaire, TDD is perceived as difficult to apply by 70% of the respondents, although some subjects stated that the fine-granularity of the task (as in the experimental task) eases the application of TDD. For the internal validity, our work suffers from mono-operation and mono-method bias since we studied the constructs using only a single task and measuring each of them with a single metric. At the same time, there might be other constructs (i.e., internal code quality) that might have been affected by the treatment but were not part of the observed variables of the study. The conclusion validity of the study is threatened by the low statistical power we could achieve in the data analysis due to the limited sample size. Finally, the task and subjects are not representative of the real world. Nevertheless, other studies have showed that the more skilled student can perform at the same level of professional developers [8, 9], hence limiting the external validity threat to our study.

6. CONCLUSION

This replication partially confirmed the results of the original study, and fully matched the results of our previous replication. From both our replications it appears that there is no difference between TDD and TLD in terms of the testing effort, whereas such difference was found in the original study. We found a significant correlation between the testing effort and the productivity of the subjects, confirming the findings of our previous replication as well as the original study under different settings. We could not find a significant relationship between testing effort and external quality. Also this result holds throughout the three studies and under different replication settings. Finally, we explored the effects of process conformance—only for the subjects using TDD—on external quality and productivity. We conclude that there is no linear relationship between process conformance and productivity. Regarding the relationship between process conformance and quality we were able to show that a weak correlation exists. For high levels of conformance, the results show a high level of variance in terms of quality which might be explained by other covariate (e.g., pre-existing skill) that we did not consider in the conformance model. Finally, the material necessary for further replications is available at the first author's website[4].

7. REFERENCES

[1] K. Beck. *Test-driven Development: by Example*. The Addison-Wesley signature series. Addison-Wesley, 2003.

[2] H. Erdogmus, M. Morisio, and M. Torchiano. On the Effectiveness of the Test-First Approach to Programming. *IEEE Transactions on Software Engineering*, 31(3):226–237, 2005.

[3] D. Fucci and B. Turhan. A Replicated Experiment on the Effectiveness of Test-First Development. In *2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 103–112. IEEE, 2013.

[4] D. Fucci and B. Turhan. On the role of tests in test-driven development: a differentiated and partial replication. *Empirical Software Engineering*, pages 1–26, 2013.

[5] P. M. Johnson and H. Kou. Automated Recognition of Test-Driven Development with Zorro. In *AGILE 2007*, pages 15–25. IEEE, 2007.

[6] N. Juristo and S. Vegas. Using differences among replications of software engineering experiments to gain knowledge. In *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*, pages 356–366, Washington, DC, USA, 2009. IEEE Computer Society.

[7] L. Madeyski. *Test-driven development: An empirical evaluation of agile practice*. Springer, 2009.

[8] M. M. Müller and A. Höfer. The Effect of Experience on the Test-driven Development Process. *Empirical Software Engineering*, 12(6):593–615, 2007.

[9] M. Philipp. Comparison Of The Test-Driven Development Processes Of Novice And Expert Programmer Pairs. 2009.

[10] Y. Rafique and V. Misic. The effects of test-driven development on external quality and productivity: a meta-analysis. 2013.

[11] M. Siniaalto and P. Abrahamsson. Does test-driven development improve the program code? Alarming results from a comparative case study. *Balancing Agility and Formalism in Software Engineering*, 48(Chapter 24):143–156, 2008.

[12] S. Sørumgård. Verification of process conformance in empirical studies of software development. *Department of Computer and Information Science, The Norwegian University of Science and Technology*, 1997.

[13] B. Turhan, L. Layman, M. Diep, H. Erdogmus, and F. Shull. *How Effective Is Test Driven Development?* O'Reilly Media, 2010.

[14] Y. Wang and H. Erdogmus. The role of process measurement in test-driven development. In *4th Conference on Extreme Programming and Agile Methods*, 2004.

[4] `http://cc.oulu.fi/~dfucci/lab_package2012.zip`