Are Students Representatives of Professionals in Software Engineering Experiments?

Iflaah Salman, Ayse Tosun Misirli e Natalia Juristo

37th IEEE International Conference on Software Engineering - ICSE 2015

DCC890 - 16/2
Presented by: Igor Muzetti Pereira

Index

- Background.
- Related Work.
- Study Context.
- Research Method.
- Discussion.
- Threats to Validity.
- Conclusion.

Background

- Most of the experiments in software engineering (SE) employ students as subjects.
- Researchers use the students that they are teaching.
- **Conducting experiments with professionals in a real environment**, on the other hand, **is much more costly**, and the **research must be very well funded**.
- This raises concerns about the realism of the results acquired through students and adaptability of the results to software industry.

Related Work – Comparing Students and Professionals

- There are some **difficulties** of conducting controlled experiments **outside a laboratory environment**.
- Despite the valid reasons for **using students in experiments**, there are major concerns about the **generalizability of the results**.
- When the experiments lack realism, the findings are only valid in a specified experimental situation, making the results less significant for both applied and theoretical research.
- Moreover**, undergraduates are used more often than graduates**.

Related Work – Experimentation on Test-Driven-Development (TDD)

- **TDD** is an important development **practice** within **agile methodologies**.
- In this research, the authors focused on the effects of **TDD on code quality** in software projects implemented by **students and professionals**.
- **Internal code quality** as the design quality of a software system, and they list the set of metrics for quantifying internal code quality as **OO, cyclomatic complexity and Halstead metrics**.
- **External code quality** is usually measured in terms of the **number of pre-release and post-release system defects**.

Goal

- This paper presents an empirical study, whose goal is to **compare students and professionals to understand how well students represent professionals as experimental subjects in SE research**.

Study Context

- The comparison was made in the context of **two TDD experiments** conducted with **students in an academic** setting and with **professionals in a software organization**.
- The academic site chosen for the study was Universidad Polytechnic de Madrid (UPM), Spain.
  - 17 graduate students if different nationalities.
- The experiment with professionals was conducted at three different sites of the same software development organization.
  - 24 professionals participated in the Helsinki (Finland), Oulu (Finland) and Kuala Lumpur (Malaysia).

Study Context

- The research goal of both experiments was to observe the effects of TDD on quality and productivity.
  - Independent variables: **TDD** and the incremental Test-Last Development (**TLD**).
  - Dependent variables: **Quality** and **Productivity**.
- The experiment was **replicated twice**, **once in academia** and **once in industry**.
- However, the **replications differed** in terms of the **experimental design** regarding **assignment of tasks** to treatments and the **experiment protocol**.

Study Context

Comparison of Experimental Setup

	Industry Experiment	Academic Experiment
Experimental Design	ABB	ABBB
Objects/Tasks	MR, BSK, MP	MR, BSK, MP, S
Trainer	Same	
Time for the Tasks	Same	
Protocol	Five-day protocol	
Sampling	Convenience sampling	
Instrumentation	Eclipse IDE with JUnit testing framework; Java programming language; English as the training language	

One factor two-level, within-subjects

Sequence of the Experiments

Training on unit testing	ES	Training on TDD	ES	ES	ES
	TLD		**TDD1**	**TDD2**	**TDD3**
	MR in *industry*, MR, BSK, MP and S randomly assigned tasks in *academia*		BSK in *industry*, MR, BSK, MP and S randomly assigned tasks in *academia*	MP in *industry*, MR, BSK, MP and S randomly assigned tasks in *academia*	MR, BSK, MP and S randomly assigned tasks in *academia*

Study Context

Protocol					
	First day	**Second day**	**Third day**	**Fourth day**	**Fifth day**
Professionals	Training TDD and TLD impl.*	TDD training e TDD1 *	Practice session with TDD	Practice session with TDD	TDD2 *
Students	Training TDD	Training TLD and impl.*	TDD1 *	TDD2 *	TDD3 *

*Using the same technological setup in both settings, the data (source and test codes) for the specified tasks were collected at the end of each day.

Research Method

- The authors reports an empirical study in which data from two experiments are compared to analyze **how representative students are of professionals in SE experiments**.
- **Research goal (GQM)**:

Analyze students as experimental subjects

For the purpose of comparison with professionals

With respect to the code quality of software tasks implemented by both subject groups

From the point of view of the students and professionals

In the context of a TDD experiment conducted in academia with students and in industry with professionals.

Research Method

RQ 1.1: How much does the code quality of a task produced by students using TDD differ from the code quality of a task produced completed by professionals using TDD?

- $\boldsymbol{H_0}$**:** The code quality of a TDD task implemented by a professional and of a TDD task implemented by a student is the same.

$$\boldsymbol{H_0}\colon \boldsymbol{\mu(CQ)_{P-TDD} = \mu(CQ)_{S-TDD}}$$

Research Method

RQ 1.2: How much does the code quality of a task produced by students using TLD differ from the code quality of a task produced by professionals using TLD?

- $\boldsymbol{H_0}$**:** The code quality of a TLD task implemented by a professional and of a TLD task implemented by a student is the same.

$$\boldsymbol{H_0}\colon \boldsymbol{\mu(CQ)_{P-TLD} = \mu(CQ)_{S-TLD}}$$

Research Method

- **Variables**
  - Independent variable: Type of subjects (professionals and students).
  - Dependent variable: Code quality (it's measured in terms of static code attributes).

- **Objects**
  - Four objects (programming tasks) are used in the experiments.
    - Bowling Scorekeeper (BSK^{12}), Mars Rover (MR^{12}), Sudoku (S^{12}) and Music Phone (MP^{2}).

Research Method

- **Metrics**

Metrics representing code quality in this study

Cyclomatic Density (CD)	Halstead Programming Time (HPT)
Decision Density (DD)	Maintenance Severity (MS)
Essential Density (ED)	Branch Count (BC)
Cyclomatic Complexity (CC)	Condition Count (Cnd.C)
Essential Complexity (EC)	Decision Count (DC)
Halstead Difficulty (HD)	Lines of Code (LOC)
Halstead Length (H.Len)	Total Operands (T.Oprnds)
Halstead Volume (HV)	Total Operators (T.Oprtr)
Halstead Level (H.Ll)	Unique Operands Count (U.Oprnd.C)
Halstead Programming Effort (HPE)	Unique Operators Count (U.Oprtr.C)

Research Method

- **Data Collection**
  - Tool called “Prest” to extract the static code attributes.
- **Subjects**
  - 17 students.
  - 24 professionals.

Research Method

Experience Levels of Experimental Subjects

		Programming	Java Programming	Unit Testing	JUnit	TDD
Students	>10 years	0	0	0	0	0
	5-<=10 years	8	6	1	1	0
	2-<=5 years	5	7	2	2	0
	<2 years	4	4	14	14	17
	Total	17	17	17	17	17
Professionals	>10 years	10	3	2	1	0
	5-<=10 years	10	5	3	3	0
	2-<=5 years	4	9	14	8	5
	<2 years	0	7	5	12	19
	Total	24	24	24	24	24

Research Method

- **Analysis**

1. Their aims is to check the general differences between two subjects groups rather than differences in medians or variance.

2. Check if there are differences in terms of the minimum, maximum, median, mean and standard deviation of the metric data collected from the two subject groups.

Research Method

Treatment-Based Descriptive Statics (Median Values)

Metric	Students			Professionals		
	TLD	*TDD1*	*TDD2*	*TLD*	*TDD1*	*TDD2*
CD	1	1	1	1.5	1	1
DD	0	0	0	0	0	0
ED	0	0	0	0	0	0
BC	0	0	0	0.5	0	0
Cnd.C	0	0	0	0	0	0
CC	4	4	4	3	3	4
DC	0	0	0	0.5	0	0
EC	1	1	1	1	1	1
LOC	4	5	4	7	4	5
T.Oprnds	1	2	2	6	1	2
T.Oprtr	2	2	2	7	2	2
U.Oprnd.C	1	2	1	5	1	1
U.OprtrC	1	2	2	4	1	2
HD	0.5	1	1	2	0.5	1
H.Len	3	4	4	13.5	3	4
H.L1	1	0.714	0.666	0.348	0.666	1
HPE	0.477	0.602	0.477	0.928	0.477	0.477
HPT	0.026	0.033	0.026	0.051	0.026	0.026
HV	0.477	0.518	0.477	0.301	0.465	0.602
MS	1	1	1	1	1	1

Research Method

Comparison in Terms of Experiment Treatments

Metric	TLD	TDD1	TDD2
CD	0.018	0.029	0.091
DD	0.232	0.091	0.091
ED	1	1	1
BC	0.018	0.029	0.091
Cnd.C	0.232	0.091	0.091
CC	0.000	0.044	0.091
DC	0.018	0.029	0.091
EC	1	1	1
LOC	0.007	0.008	0.029
T.Oprnds	0.004	0.029	0.029
T.Oprtr	0.000	0.019	0.029
U.Oprnd.C	0.004	0.019	0.064
U.Oprtr.C	0.002	0.029	0.091
HD	0.000	0.008	0.029
H.Len	0.001	0.029	0.029
H.Ll	0.018	0.234	0.001
HPE	0.001	0.127	0.091
HPT	0.001	0.127	0.091
HV	0.026	0.611	0.001
MS	0.133	0.029	0.091

55% and 49% of code quality metric values differ between the two subjects groups for the TLD and for TDD1 treatments.

Research Method

Comparison in Terms of Treatment-Task Combinations

Metric	MR-TLD	BSK-TDD	MP-TDD
CD	0.954	1	1
DD	1	1	1
ED	1	1	1
BC	0.954	1	1
Cnd.C	0.999	1	1
CC	0.031	1	1
DC	0.999	1	1
EC	1	1	1
LOC	0.677	0.927	0.055
T.Oprnds	0.591	0.999	0.061
T.Oprtr	0.591	1	1
U.Oprnd.C	0.055	0.431	1
U.Oprtr.C	0.906	1	0.005
HD	0.677	0.927	0.010
H.Len	0.840	0.999	0.002
H.Ll	0.677	0.927	1
HPE	0.429	0.604	0.999
HPT	0.429	0.604	0.999
HV	0.906	0.604	0.319
MS	0.995	1	1

Discussion

- **TLD**: Professionals produced more lines of code per method, the cyclomatic complexity of their methods was small, produce big methods with more operators and operands.
  - Professionals are more experienced in programming skills than students.
  - However, professionals worked on only one task (MR) while students worked on four different tasks.

Discussion

- **TDD1:** Students produced more complex code in terms of median cyclomatic complexity, lines of code and unique operands and operators count metrics.
  - Both students and professionals were using TDD practice for the first time, neither had previous experience in the new approach.
  - They appear to act similarly during their first implementation.
  - Looking at all 100 hypotheses, however, we failed to observe major differences between the two subjects groups.

Discussion

- **TDD2:** Students and professionals performances are very similar in terms of median code quality metrics.
  - Considering all the hypotheses, however, we observe differences between students and professionals in 53% of the cases.
  - Differences between tasks could have affected the results.

Discussion

- For TLD and TDD applied on certain tasks, there are differences between the code quality of a task implemented by a student and a professional in terms of cyclomatic complexity, LOC, operator and operand counts.
- Inconsistencies between the comparisons based on experiment treatment and the treatment-task show that:
  - The selection of programming tasks for a particular development approach.
  - The experience of a subject in the development approach would significantly affect the findings.

Threats to Validity

- Internal Validity
  - Selection
  - Diffusion or imitation of treatments
- External Validity
- Construct Validity
  - Mono-operation bias
  - Mono-method bias
- Conclusion Validity
  - Violated assumptions of statistical tests
  - Repeated hypothesis testing

Conclusion

- Professionals and students differ in terms of code quality when they apply incremental TLD.
- Both subject groups perform similarly when they apply TDD for the first time.
- Neither of the subject groups performs better than the other when they apply a new technology during experimentation.
- Experience plays a significant role in subject performance.
- Results have the potential to incrementally build knowledge and contribute to the body of evidence.

Doubts?

Thanks!