The Journal of Systems and Software 137 (2018) 497–511

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

The impact of Software Testing education on code reliability: An empirical assessment

Otávio Augusto Lazzarini Lemos[a,*], Fábio Fagundes Silveira[a], Fabiano Cutigi Ferrari[b], Alessandro Garcia[c]

[a] Science and Technology Department, Federal University of São Paulo at S. J. dos Campos, Brazil
[b] Computing Department, Federal University of São Carlos, Brazil
[c] Informatics Department, Pontifical Catholic University of Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 9 April 2016
Revised 21 September 2016
Accepted 24 February 2017
Available online 10 March 2017

Keywords:
Software Testing
Computer science education
Student experiments

ABSTRACT

Software Testing (ST) is an indispensable part of software development. Proper testing education is thus of paramount importance. Indeed, the mere exposition to ST knowledge might have an impact on programming skills. In particular, it can encourage the production of more correct - and thus reliable - code. Although this is intuitive, to the best of our knowledge, there are no studies about such effects. Concerned with this, we have conducted two investigations related to ST education: (1) a large experiment with *students* to evaluate the possible impact of ST knowledge on the production of reliable code; and (2) a survey with *professors* that teach introductory programming courses to evaluate their level of ST knowledge. Our study involved 60 senior-level computer science students, 8 auxiliary functions with 92 test cases, a total of 248 implementations, and 53 professors of diverse subfields that completed our survey. The investigation with students shows that ST knowledge can improve code reliability in terms of correctness in as much as 20%, on average. On the other hand, the survey with professors reveals that, in general, university instructors tend to lack the same knowledge that would help students increase their programming skills toward more reliable code.

1. Introduction

Software Testing (ST) can be seen as one of the most important and least known aspects of software development. In fact, it is common that Computer Science (CS) students graduate into industry without knowing how to test a program (Clarke et al., 2014).[1] ST is thus one of the *dark arts* of software development (Myers et al., 2011). Yet, researchers and practitioners have always argued that testing demands a large share of the costs of a software project (Harrold, 2000). For instance, a survey conducted with 1560 senior IT executives and testing leaders from 32 countries revealed that the IT spend allocated to quality assurance and testing was 35% in 2015 (predicted to rise to 40% by 2018) (Capgemini Group et al., 2016).

Another aspect of ST that seems to be overlooked is the following: the mere exposition to its knowledge might help developers produce more correct – and thus, reliable – programs. In fact, there are several ideas in the ST body of knowledge that can produce positive effects in a programmers' skills. For example, consider the awareness that *virtually all programs contain faults* (Myers et al., 2011; Ammann and Offutt, 2008), a principle taught early in ST courses. Such an idea can instill a healthy skepticism in programmers towards their own code, making them more cautious. Moreover, the formal testing techniques themselves encourage designing programs with diligence. Take, for example, *boundary value analysis*, a functional testing criterion that requires writing tests for border inputs. Developers exposed to such a strategy can be more attentive about corner cases in their implementations, hence improving the correctness of their code.

Although the effect of testing knowledge on programmers seems to be intuitive, to the best of our knowledge, there is no empirical evidence to support it. In the literature, we can find substantial work on improving ST training in CS programs. For instance, Patterson et al. proposed the integration of testing

* Corresponding author.
E-mail addresses: otavio.lemos@unifesp.br (O.A. Lazzarini Lemos), fsilveira@unifesp.br (F. Fagundes Silveira), fabiano@dc.ufscar.br (F. Cutigi Ferrari), afgarcia@inf.puc-rio.br (A. Garcia).

[1] This article is an extension of a previous paper published at the *Proceedings of the 26th IEEE International Symposium on Software Reliability Engineering* (ISSRE 2015) (Lemos et al., 2015). It contains significant material added to the original contribution. In particular, this paper contains the following additional content: (1) a replication of the students' experiment, (2) a large control group, and (3) a survey with several professors to evaluate their level of ST knowledge.

http://dx.doi.org/10.1016/j.jss.2017.02.042
0164-1212/

tools into programming environments (Patterson et al., 2003); Jones has explored the integration of testing into introductory CS courses through testing labs and diverse forms of courseware (Jones, 2001); and Elbaum et al. presented a web-based tutorial to engage students in learning Software Testing strategies (Elbaum et al., 2007). However, to the best of our knowledge, there are no experimental studies into the effects of ST education on the developers' programming skills *per se*, in terms of reliability.[2]

Investigations into this topic are important because recent data shows that computing academic curricula tend to emphasize *development* at the expense of *testing* as a formal engineering discipline (Astigarraga et al., 2010; Clarke et al., 2014; Wong, 2012). In fact, as reported by Astigarraga et al. (2010), the overall academic CS curricula tend to place a heavy emphasis on design and implementation, rather than on quality assurance topics such as ST. On the other hand, even when ST courses are in fact present in curricula, it is unclear the extent to which the techniques that are taught are in fact adopted by the industry (*e.g.*, mutation (Madeyski and Radyk, 2002) and data-flow (Harrold, 2000) testing seem to be rarely put to practice). Empirical evidence showing that ST education might lead to more reliable programming is very important. In particular, it can motivate the creation or maintenance of these courses.

In this paper, we present two investigations related to ST education and software reliability: one involving *students* and another involving *professors*. The students investigation comprised a large controlled experiment that evaluated the impact of ST education on reliable programming, when compared to other types of knowledge; while the professors investigation involved a survey to evaluate the level of ST knowledge of instructors that teach introductory programming courses. The mains idea was to check: (1) whether ST education itself can have a significant impact on programming skills toward more correct – and thus reliable – code; and (2) whether CS instructors themselves possess a minimum level of ST knowledge to be able to instill these positive ideas early in their programs.

The students experiment involved a total of 60 senior CS undergraduate students, 8 auxiliary functions in 4 different domains (basic mathematics; array manipulation; string manipulation; and file input/output) with 92 test cases, and a total of 248 implementations. Subjects implemented two different functions before and after learning basic ST concepts and three techniques (functional – or *black-box* – testing, structural – or *white-box* – testing, and mutation testing), and the quality of the code produced before and after was compared. Our goal was not to verify how well the techniques were applied afterwards, but how ST knowledge could impact on the subjects' programming skills in terms of producing more reliable implementations (*i.e.*, we did not measure the quality of the testing code itself, neither which specific techniques were being applied). To evaluate the reliability of the implemented functions in terms of correctness, the produced implementations were executed against systematically developed test sets before and after the training took place. To improve the external validity of our experiment, we included two control groups in the experiment: one with 22 subjects taking a Software Engineering course, and another with eight subjects taking an object-oriented design course. These control groups help improve confidence that the outcomes observed with the ST course are not due to maturation, for instance. A small replication of the treatment group was also conducted to improve confidence in our outcomes.

[2] In this paper we use *correctness* as the main property to measure reliability. In our experiment the degree of reliability of a program is measured by its level of success against systematically developed test sets. We believe this is straightforward as a program can not be considered reliable if it is incorrect.

Our investigation with students provides evidence that ST knowledge can significantly impact on programming skills in terms of reliability. In fact, subjects in the main experiment were more than twice more likely to deliver correct implementations after learning the ST concepts and techniques (150%). Moreover, the correctness of the subjects' implementations was, on average, 20% higher after the exposition to ST knowledge took place. Interestingly, we noticed that the positive effect is present even when no specific testing technique is explicitly applied, possibly a consequence of the exposure to the testing theory itself. Results of the replication point to the same direction of the main experiment. On the other hand, subjects in the control groups did not perform as well as the ones in the treatment groups: results indicate that their performance in terms of reliability was not significantly affected, even after learning Software Engineering and object-oriented design concepts. Outcomes related to effort also indicate that subjects invest more time in their implementations after taking the ST classes. However, such effort did not result in the production of significantly more lines of application code, implying that the implementations produced afterwards were more reliable but not significantly larger than the ones produced in the first session.

Our investigation with professors, on the other hand, revealed that instructors that teach introductory programming courses lack proper ST knowledge: the mean score in the survey, which was composed of 10 basic ST questions, was only 4.24 out of 10.00. This is problematic because students might not be having the chance of learning concepts that can encourage more reliable programming (as shown in the students' experiment) early in their academic programs. In particular we noticed instructors lack the important notion of the *destructive* aspect of ST. For instance, most professors ignore the fact that the most successful test cases are the ones that tend to fail and thus reveal as yet to be discovered faults. As discussed by Myers et al. (2011), these ideas are important for an effective testing practice.

The remainder of this paper is structured as follows. Section 2 presents background knowledge required to understand our study, and Section 3 presents how our study was set up in terms of subjects, experimental design, metrics and statistical procedures. Section 4 presents the results and analysis of our experiment, while Section 5 discusses such results in more details. In the sequence, Section 6 presents our study limitations and Section 7 summarizes related research. Finally, Section 8 concludes the paper.

2. Software Testing knowledge

In this paper, functional, structural, and fault-based testing, together with basic Software Testing principles and concepts – such as the ones discussed in this section and in Section 1 – were taken as basic Software Testing knowledge (*ST knowledge*, from now on). Our goal is to evaluate whether such knowledge could impact on the programming skills of software developers, in terms of producing more correct, and thus reliable, implementations.

A *test case* (or simply, a *test*) consists of a set of inputs and expected output for a program (IEEE, 1990). The output is assessed via an *oracle*, which determines what is the correct result of the program under testing given an input (Binder, 1999). In our case, the oracle is a tester supported by an automated testing tool (in this paper, JUnit[3]) that implements *assertions*. Formally, a test case is an ordered tuple: $< (I_1, \ldots, I_n), O >$, where O is the expected output of the program when $I_1, \ldots, I_n$ are used as inputs.

Software Testing is defined as the execution of a program against test cases with the intent of revealing faults

[3] http://junit.org/ - 10/mar/2016.

(Myers et al., 2011). The varied testing techniques are defined based on the underlying artifacts used to derive test cases. Three basic testing techniques are *functional, structural*, and *fault-based testing*.

Functional – or *black box* – testing derives test cases from the specification of a program. Two of the most well-known functional-based testing selection criteria are *equivalence partitioning* and *boundary-value analysis*. Equivalence partitioning divides the input and output domains of a program into a finite number of valid and invalid *equivalence classes*. It is then assumed that a test case with a representative value within a given class is equivalent to testing any other value in the same class. This criterion requires a minimum number of test cases to cover the valid classes and an individual test case to cover each invalid class. Boundary-value analysis complements equivalence partitioning by requiring test cases to cover values at the boundaries of equivalence classes (Myers et al., 2011).

Structural – or *white-box* – testing is a technique that complements functional testing. It derives test cases from the internal representation of a program (Myers et al., 2011). Some of the well-known structural testing criteria are *statement, branch*, or *definition-use* (Rapps and Weyuker, 1985) *coverage*. These criteria require that all commands, decisions, or pairs of assignment and use locations of a variable be covered by test cases.

Fault-based testing has *mutation testing* as its most representative criterion. The main idea behind mutation testing is to define a set of mutation operators which, when applied to components of a program, introduce certain types of *faults* into the program. Typical mutations include changing variable names in expressions, or changing arithmetic operators (*e.g.*, a $>$ for a $<$). The goal then is to construct a set of tests cases T which will distinguish between a program P and any nonequivalent program P' which can be generated from the original P by the application of mutations to components of P (Howden, 1982). The quality of a test set is then measured by its capacity to distinguish the original program from all or most of the non-equivalent mutants (its *mutation score*).

3. Study setup

The goal of our study is to investigate the impact of ST knowledge on programming skills. Such impact is evaluated in terms of reliability. In this endeavor, we are interested in the following research question: **RQ_1** – Can ST knowledge help developers improve their programming skills in terms of delivering more reliable implementations?

As an additional investigation, we want to check whether developers tend to invest more (or less) effort in their implementations after learning ST, and whether there is a difference in the complexity of the code produced before and after the exposition to ST knowledge. Such additional investigation raises two other research questions: **RQ_2** – Does ST knowledge impact on the effort invested by developers on their implementations?; and **RQ_3** – Does ST knowledge impact on the complexity of the produced code?

Our investigation develops in terms of hypotheses H_1, H_2, H_3, and H_4, where the first is related to research question **RQ_1**, the second and the third are related to research question **RQ_2**, and the forth is related to research question **RQ_3**. The null (*0*) and alternative (*A*) definitions of each hypothesis are presented in Table 1.

The experimental setup adopted for our study shares several characteristics of a previous study conducted by the same authors (Lemos et al., 2012). The main difference is that in this paper we evaluate the impact of the testing knowledge on programming skills, while the previous experiment targeted agile practices. Moreover, in the experiment presented in this paper two functions and a control group were added to the experimental design. This was done to add more rigor to the study, since subjects implemented two functions in each session and we compared outcomes with groups of students that were not taking the ST course. The next paragraphs discuss our setup in detail.

Table 1
Hypotheses formulated for our experiment.

	Null hypothesis (0)	Alternative hypothesis (A)
H_1	$Correctness_{woTK} = Correctness_{wTK}$	$Correctness_{woTK} < Correctness_{wTK}$
H_2	$Time_{woTK} = Time_{wTK}$	$Time_{woTK} < Time_{wTK}$
H_3	$Size_{woTK} = Size_{wTK}$	$Size_{woTK} < Size_{wTK}$
H_4	$Complexity_{woTK} = Complexity_{wTK}$	$Complexity_{woTK} < Complexity_{wTK}$

Legend: H = Hypothesis; woTK = without testing knowledge; wTK = with testing knowledge; Size = code size; Complexity = code complexity.

Table 2
Characteristics of each experimental run involved in our study.

Type of group	Experiment	Sample size	Course taken
Treatment	**ST_1**	20	Software Testing
	ST_2	10	
Control	**SE**	22	Software Engineering
	OO	8	Object-oriented design

In total, our study involved four experimental runs, including two treatment and two control groups. Table 2 summarizes characteristics of each experimental instance included in our study. We also gave a name for each run to facilitate further reference. It is important to note that the main reason for adding control groups to our study is not to compare how the different types of knowledge would impact on reliability, but to improve confidence that the possible effects that can be observed in the treatment groups are really caused by the course content itself and not due to maturation (subjects might learn other things outside class over time that could have an impact on the outcomes).

3.1. Subjects, target functions, test sets, and tools

Subjects: Our main study involved 60 senior CS undergraduate students (20 in the main treatment group – **ST_1**, 10 in the small replication treatment group – **ST_2**, and 22 in the SE control group – **SE**, and 8 in the OO control group – **OO**). All students had basic Java programming knowledge learned along a one-semester, 72-h Object-Oriented Programming course (8 h per week). The students in the treatment groups were asked to perform the tasks of our experiment *before* and *after* learning ST knowledge while taking a 72-h Software Testing course. The first session took place on the first week of the semester, and the second took place around two and a half months later. The control groups did exactly the same thing, except that they were not taking the ST course. The first smaller control group was taking an Object-Oriented (OO) Design in Java course (mostly learning UML and OO patterns and their implementations in Java); and the second larger control group was taking a Software Engineering (SE) course. All three courses included significant amounts of programming exercises in Java. Tests were not mandatory but were recommended for all groups, based on the knowledge level of each subject (*e.g.*, using either main functions or JUnit tests).

All sessions were performed in a closed-lab environment. Students had two hours to complete their tasks, and were asked to use a Java IDE to implement their assigned functions. Since we were more interested in the effects of the testing theory itself on programming skills, not as much on the application of the specific testing techniques themselves, no functional, structural, or mutation testing tools were readily available to them. In this way, our results represent more the impact that *learning* the testing concepts and techniques have on programming skills. In fact, as we discuss in Section 5, by analyzing the code produced by the sub-

Table 3
Functions used in the experiment.

Domain	F	Description	Sample test case	# TCs
Array	$\mathbf{a_1}$	*Array equality*: given two arrays, the program should return *true* or *false* according to the contents of the arrays being equal or not.	< ([1, 2, 3], [1, 2, 3]), *true*>	20
	$\mathbf{a_2}$	*First index with different value*: given an array and a number, the program should return the first index of the array that contains a value different from the number.	< ([0, 0, 0, 0, 0, 1], 0), 5>	12
Math	$\mathbf{m_1}$	*Power of two*: given a number, the program should return *true* or *false* according to it being or not a power of two.	< (4), *true*>	6
	$\mathbf{m_2}$	*Factorial*: given a number, the program should return its factorial.	< (5), 120>	7
String	$\mathbf{s_1}$	*Capitalization of phrases*: given a string, the program should return the same string with the first letters of words capitalized.	< (“one two”), “One Two”>	7
	$\mathbf{s_2}$	*Maximum common prefix*: given two strings, the program should return the maximum common prefix between them.	< (“pref suf”, “pref fus”), “pref ”>	11
File I/O	$\mathbf{i_1}$	*Create text file*: given two strings, the program should create a text file whose content and location/name is indicated in the first and second string.	< (“abc”, “.\dir\text.txt”), *creates file. \dir\text.txt with “abc” as content*>	13
	$\mathbf{i_2}$	*File copy*: given two strings, the program should copy the file indicated in the first string to the location indicated in the second.	< (“.\test.file”, “.\tmp\”), *.\test.file is copied to. \tmp* >	16

Legend: F = Function; TCs = Test Cases.

Table 4
Equivalence classes and boundary values considered for testing *array equality* ($\mathbf{a_1}$).

Input cond.	Valid classes	Invalid classes	Boundary values
|ar1|	|ar1| > −1 (C1)		|ar1| = 0 (B1)
ar1 is *null*	No (C2)	Yes (C3)	
|ar2|	|ar2| > −1 (C4)		|ar2| = 0 (B2)
ar2 is *null*	No (C5)	Yes (C6)	
|ar1|, |ar2|	|ar1| > |ar2| (C7)		|a1| − |a2| = 1 (B3)
	|ar2| > |ar1| (C8)		|a2| − |a1| = 1 (B4)
ar1[i]	*Int.MIN* ≤ ar1[i] ≤ *Int.MAX* (C9)		ar1[i] = *Int.MIN* (B5)
			ar1[i] = *Int.MAX* (B6)
ar2[i]	*Int.MIN* ≤ ar2[i] ≤ *Int.MAX* (C10)		ar2[i] = *Int.MIN* (B7)
			ar2[i] = *Int.MAX* (B8)

jects of the treatment group, it appears that some of them have not explicitly applied the learned techniques, and still got better results.

Target functions. The features involved in our study were *auxiliary functions*, that is, supportive actions of software systems. It is important that these functions be developed with care because the history of software development shows that they can be the source of significant failures (Lemos et al., 2012). To select a representative and variable set of such features, we looked into the Apache Commons project,[4] which provides libraries of reusable Java components. We also selected functionalities that could easily be found through searches issued to code search engines; that is, we tried to identify commonly used auxiliary functions that were not readily available in the Java API. We categorized these functions into four domains: array manipulation (Array), basic mathematics (Math), string manipulation (String), and file input/output (File I/O). To obtain a richer set, we selected two functionalities within each domain. The auxiliary functions used in our study are listed in Table 3.

Another characteristic of the selected functions is that they are narrowly scoped. The idea is to perform a conservative evaluation: if particular knowledge can impact on the implementation of smaller features, we can expect them to further impact on larger ones. Another advantage is that this type of function enables the adoption of more systematic test case selection techniques to evaluate them in the experiment, such as functional testing. Such characteristic provides more control to the experiment.

Test sets. To evaluate the programs implemented by the subjects, we developed full functional test sets for each of the selected functions. The last column of Table 3 shows the number of test cases developed for each one. To construct the test sets, we applied the equivalence partitioning and boundary-value analysis criteria (see Section 2). These criteria were used to select representative test cases for each test set, trying to cover as many functional specificities of the functions as possible. To show an example of how test cases were developed, Table 4 presents the equivalence classes and boundary values (when applicable) for the $\mathbf{a_1}$ functionality. *ar1* and *ar2* are the $\mathbf{a_1}$ input arrays; |*ar*| represents the array *size*; and *arX*[i] represents an element of the array. *Int.MIN* and *Int.MAX* correspond to the minimum and maximum integer values. Since the study was conducted using the Java language, the highest and lowest possible integers were used as boundary values for that data type. A similar rule was applied to other types for other functions. Here, we do not use the specific values to represent the test cases independently of language.

Tools. Eclipse[5] was the IDE used to develop the functions, and JUnit was the framework used to develop the test cases. Students received instructions in order to make sure they would concentrate their effort only on implementing the intended auxiliary functionalities (and tests for them). For instance, the subjects were instructed to implement functions as static methods in a class with a predefined name. We did this because static methods are easier to implement since they do not require object instantiation. Moreover, auxiliary functions usually rely only on parameter values to fulfill their responsibility. This also enables the execution of our established test sets more easily.

3.2. Experimental design and procedure

For the conducted experiment, we adopted the *repeated measures* with cross-over and control group experimental design (or

[4] http://commons.apache.org/ - 10/mar/2016.

[5] http://eclipse.org/ - 10/mar/2016.

Table 5
Partial task assignments to subjects.

Subject	1^{st} Session Functions	2^{nd} Session Functions
01	a_1 & m_1	s_1 & i_1
02	a_1 & s_1	m_1 & i_1
03	a_1 & i_1	s_1 & m_1
04	s_1 & m_1	a_1 & i_1
05	s_1 & i_1	a_1 & m_1
06	m_1 & i_1	a_1 & s_1
07	a_2 & m_2	s_2 & i_2
08	a_2 & s_2	m_2 & i_2
09	a_2 & i_2	s_2 & m_2
10	s_2 & m_2	a_2 & i_2
⋮	⋮	⋮

pre-post test with control group), in which each subject implemented functions before and after acquiring the ST basic knowledge – in the case of the treatment groups – or OO design / SE basic knowledge – in the case of the control groups. Such type of design supports more control to the variability among subjects (Montgomery, 2006). To minimize the variability of the difference among functions, we randomized the assignments among students for both groups. Finally, to cancel function asymmetry, each function was assigned to be implemented before and after the treatment by different subjects.

Each experimental instance was conducted in two sessions. In the first session, prior to learning any ST, OO design, or SE concepts, students implemented two functions; and in the second session, after learning ST, OO design, or SE, they implemented other two functions. Since each subject implemented 4 functions, we collected a total of 248 implementations.

Students had to implement functions from different domains in the first and second sessions. For instance, a student implementing a_1 and i_2 in the first session would implement a Math and a String function in the second session. We did this to cancel the impact of function domains on each other while implementing the functionalities in the first and second sessions.

To help understanding the adopted experimental design, Table 5 presents part of the assignments used for the experiment for one of the groups.

3.3. Metrics

We adopted a straightforward metric to evaluate the *reliability* of the developed functions: their correctness in terms of their *Functional Test Set Success Rate* (FTSSR). For a given implementation, FTSSR is computed by dividing the number of successful test cases by the total number of test cases developed for a given function. The FTSSR is a continuous variable: it grades implementations from 0.0 to 1.0. For instance, an implementation of the $\mathbf{a}_1$ function that passed 10 test cases would receive an FTSSR score of 0.5, since there are 20 test cases developed for it in its test set (see Table 3).

For the effort evaluation, we measured the total development time in minutes subjects took to implement the two functions in each session, and the average number of produced lines of code (LoC). Some studies have found a positive correlation between the size of program modules in LoC and fault-proneness (Gyimothy et al., 2005; Subramanyam and Krishnan, 2003) (*i.e.*, the larger a module in LoC, the more faults it tends to present). Therefore, by measuring the difference in LoC from the first to the second session we are also secondarily evaluating an additional reliability metric. That is, if code produced in the second session is not larger than code produced in the first session, we have an additional evidence that reliability has not decreased afterwards from such a perspective.

To evaluate complexity, we computed the average McCabe cyclomatic complexity metric (McCabe, 1976) (M). We used the Eclipse Metrics to measure both LoC and M.[6] Subjects were responsible for registering the time taken to implement functions.

3.4. Statistical analysis

From a statistical standpoint, a simple observation of the means or medians from sample observations is not enough to infer about the actual populations. This happens because the reached differences might be a coincidence caused by random sampling. To check whether the observed differences are in fact significant, statistical hypothesis tests can be applied.

In our study, each subject developed functions before and after learning the ST, OO design, or SE concepts. In this case, the *paired* statistical hypothesis tests can compare measures within subjects rather than across them. Paired tests are considered to greatly improve precision when compared to unpaired tests (Montgomery, 2006). Since our results seemed to follow a normal distribution, according to a Shapiro-Wilk normality test, we decided to apply the paired Student *t*-test.

To have a more rigorous evaluation of our results, for the statistical tests ran in our experiment, we adopted a confidence level of 99%. Our analyses thus consider *p*-values below 0.01 significant. For the statistical tests we adopted the R language and environment.[7]

In this paper, we ran statistical hypotheses tests only for the experimental instances with a considerable number of subjects; namely, for $\mathbf{ST_1}$ and **SE**. Since $\mathbf{ST_2}$ and **OO** involved only about 10 students, for those groups we decided to perform only a descriptive statistical analysis, to look for general tendencies.

3.5. Survey with CS professors

An additional evaluation we want to make concerns the level of ST knowledge of CS professors that teach introductory programming courses. We believe it would be important that these instructors had an adequate notion of ST concepts, so that they could introduce such ideas to beginners (*e.g.*, the ideas discussed in Section 1). In order to check the level of ST knowledge of professors, we invited instructors who teach – or have taught – introductory programming courses to respond to a survey with 10 basic ST questions. Since Software Engineering professors would obviously have a higher level of ST knowledge, we focused our attention on professionals specialized in other CS fields. 53 professors from diverse Brazilian institutions completed our survey.

Table 6 presents the 10 questions we used in our survey with the answer templates we used to guide our grading. Answers were graded from 0.00–1.00 according to how close they were to the answer templates. For instance, in question (1) some professors answered that Software Testing intends to reveal errors, but did not mention that it entails the execution of a program. Such an answer would receive a 0.5 score, as it fails to mention an important aspect of ST. Participants were also instructed to state that they *did not know the answer* for a given question, if that was the case, which would result in a 0.0 score for that question.

Total scores were then in the range of 0.00 to 10.00. For the answer templates we used reference ST literature, that is, either famous textbooks – *e.g.*, Myers et al.'s "The Art of Software Testing" (Myers et al., 2011) – or other well-known references – *e.g.*, the IEEE 1012–2012 IEEE Standard for System and Software Verification and Validation (IEEE, 2012), Zhu et al.'s survey "Software unit test coverage and adequacy" (Zhu et al., 1997). We tried to

[6] http://metrics.sourceforge.net - 10/mar/2016.
[7] http://www.r-project.org/ - 10/mar/2016.

Table 6
The 10 questions used in our survey with their corresponding answer templates and references.

	Question	Template answer
1	What is Software Testing?	"[T]he process of executing a program with the intent of finding errors." (Myers et al., 2011)
2	What is a test case?	"[A] set of test inputs, execution conditions, and expected results developed for a particular objective, such as to exercise a particular program path or to verify compliance with a specific requirement" (IEEE, 2012)
3	In general, what constitutes a successful test case	"A successful test case is one that detects an as yet undiscovered error." (Myers et al., 2011)
4	What is the input domain of a program?	"Possible values that the input parameters [of the program] can have." (Ammann and Offutt, 2008)
5	What is functional - or black-box - testing?	"[A testing technique where] test data are derived solely from the specifications (i.e., without taking advantage of knowledge of the internal structure of the program)." (Myers et al., 2011)
6	What is structural - or white-box - testing?	"[A testing technique that] derives test data from an examination of the program's logic." (Myers et al., 2011)
7	What is boundary value analysis?	"[B]oundary-value analysis requires that one or more [input] elements be selected such that each edge of the equivalence class is the subject of a test." (Myers et al., 2011)
8	What is mutation testing?	A fault-based testing criterion that evaluates how good a set of test cases is at distinguishing the original program from a set of mutants (i.e., "alternative programs that differ from the original program in some fashion"). (Zhu et al., 1997)
9	What is a testing criterion?	"[A] criterion that defines what constitutes an adequate test." (Zhu et al., 1997) or "[A] rule or collection of rules that impose test requirements on a test set." (Ammann and Offutt, 2008)
10	Is it reasonable to argue that "virtually all program contain faults"?	Yes. (Myers et al., 2011)

Table 7
Outcomes for the treatment group.

S	1st Session						2nd Session					
	FTSSR			ΣT	μL	μM	FTSSR			ΣT	μL	μM
	F_1	F_2	μ				F_1	F_2	μ			
1	0.95	0.00	0.48	40	11.00	2.00	0.83	1.00	0.92	61	19.50	3.50
2	0.55	0.71	0.63	28	16.50	1.50	0.56	1.00	0.78	80	16.00	5.00
3	0.85	0.42	0.63	16	14.50	2.50	0.83	0.86	0.85	80	18.00	4.00
4	0.83	0.86	0.85	28	16.00	4.50	0.85	0.92	0.88	7	15.50	5.00
5	0.85	0.83	0.84	34	14.50	3.00	0.71	0.82	0.77	82	16.50	7.00
6	0.85	1.00	0.93	45	17.00	5.00	0.83	0.71	0.77	40	17.50	4.50
7	0.83	0.64	0.73	30	16.00	2.00	0.92	1.00	0.96	51	10.00	3.50
8	0.58	0.43	0.51	32	10.00	2.50	0.83	0.71	0.77	70	16.00	3.00
9	0.92	0.83	0.88	35	16.00	3.00	1.00	0.86	0.93	110	21.50	5.50
10	0.55	0.67	0.61	90	19.50	4.50	0.92	0.71	0.82	120	18.00	4.50
11	1.00	0.82	0.91	75	23.00	6.50	0.83	0.83	0.83	70	21.50	4.50
12	0.43	0.71	0.57	40	9.50	1.50	0.85	0.81	0.83	60	13.50	3.00
13	0.92	0.06	0.49	30	18.50	3.50	1.00	1.00	1.00	56	18.50	5.00
14	0.75	0.06	0.41	45	14.50	2.00	1.00	0.82	0.91	45	16.50	5.00
15	0.63	0.71	0.67	35	28.50	5.00	0.95	1.00	0.98	25	24.00	6.00
16	0.50	0.43	0.46	30	14.00	3.00	0.67	1.00	0.83	37	20.00	6.50
17	0.85	0.57	0.71	22	9.50	4.50	0.56	0.86	0.71	60	22.00	4.00
18	0.83	0.92	0.88	50	14.00	2.50	1.00	0.64	0.82	63	20.50	6.00
19	1.00	0.82	0.91	33	15.50	5.00	0.42	0.92	0.67	22	15.50	3.00
20	1.00	0.86	0.93	31	17.00	4.00	0.85	0.83	0.84	18	15.00	3.00
μ	0.78	0.62	0.70	38.45	15.75	3.40	0.82	0.87	0.84	57.85	17.78	4.58

Legend: S = Subject; FTSSR = Functional Test Set Success Rate; F_n = Function n; μ = Average; Σ = Total; T = Development Time (in minutes); L = Lines of Code; M = Cyclomatic Complexity.

keep most answer templates as verbatim from the texts as possible, with only minor modifications to fit an answer sentence.

4. Results and analysis

In this section we analyze results for each experimental instance involved in our study.[8] We group results by treatment and control groups.

4.1. Treatment groups

Table 7 presents the results of our experiment for the main treatment group **ST_1**. For FTSSR, the table shows the results for each function implemented and the average. For other metrics, it shows only the average (Cyclomatic Complexity - *M* - and Lines of Code - *L*) and total (Development Time - *T*). To allow a visual analysis of the main metric targeted in our study, Fig. 1 shows a boxplot of the FTSSR outcomes for that group. Note that some subjects were removed from our analysis either for producing outliers in terms of FTSSR, or for not completing all assigned tasks (for instance, some students did not implement one of the two functions required for each session). Outliers are discussed in Section 6.

We can notice that results related to reliability improved significantly and consistently after the ST course took place: the average FTSSR went from 0.70 to 0.84, a 20% improvement. Also, the number of implementations that passed all tests more than doubled in the second session: they went from 4 to 10. Moreover, the single subject that produced two implementations that passed all tests only did so in the second session (subject #13). Note that 13 out of the 20 subjects (65%) achieved better reliability results after taking the ST course. The boxplot also shows that the subjects'

[8] All data generated in our experiment can be found at http://www.ict.unifesp.br/fsilveira/data/SoftwareTestingExperiment_data.zip. We encourage readers to replicate our experiment – possibly in other settings – to obtain further evidence about this topic.

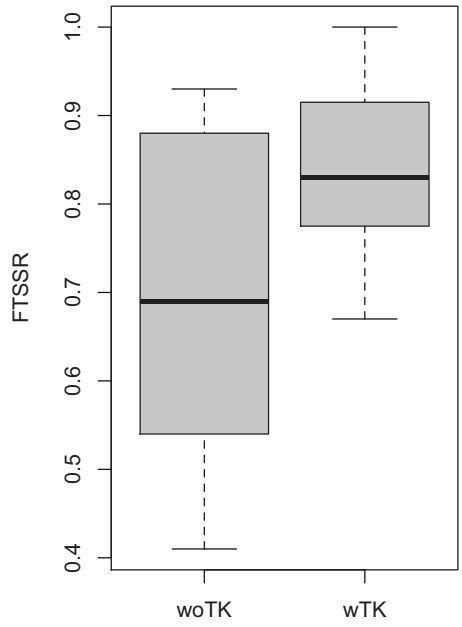

Fig. 1. Boxplot of the Functional Test Set Success Rate (FTSSR) outcomes of our experiment, for the $\mathbf{ST_1}$ treatment group. Legend: woTK = without Testing Knowledge; wTK = with Testing Knowledge.

performance was more varied in the first session, while in the second session they were consistently better. The second session box is smaller and higher than the first session one.

Other interesting outcomes are worth noting. For instance, the minimum FTSSR was improved by approximately 65% from the first to the second session (it went from 0.43 to 0.71). This indicates that even less proficient programmers can benefit from the ST knowledge, with respect to producing more reliable software. Also note that the subject that reached such a minimum in the first session, raised her FTSSR score to 0.91, on average, in the second session: a very significant improvement (one of the functions developed by the student in that session passed all tests). Another interesting outcome is that there is only a single function that did not pass any of the tests, and it was produced in the first session only. The same subject that developed such a function – subject #1 – developed another one that passed all tests in the second session.

To check whether the observed difference in terms of correctness was significant – the difference in average FTSSR –, we ran the *t*-test, which indicated a statistically significant difference at 99% confidence level (df = 19, *p*-value = 0.005168). Such result favors the *alternative* hypothesis (H_1-*A*) that subjects perform better after learning the ST concepts and techniques. We believe this is a key finding, since even for supportive functionality it appears that learning ST can bring benefits to developers. Moreover, the larger number of completely correct implementations for subjects in the second session shows that they tend to be more careful while implementing auxiliary functions after being exposed to ST knowledge, probably taking into account exceptional inputs that were included in the test sets.

Because we are considering reliability as the system's quality driver, a failing test case in our scenario is critical. This is particularly true for our experimental setting: since we applied functional testing, each test covers an important part of the functionality (*i.e.*, either an input or output equivalence class, or a boundary value), therefore a failing test case impacts significantly on the correctness of the system. In this sense, our results indicate that the exposition to ST knowledge could even help prevent serious problems caused by auxiliary functions, like the ones reported by major companies, as discussed in Section 3 and by Lemos et al. (2012).

In our additional investigation into the impact of learning ST theory on *effort and complexity*, we also reached interesting results. With respect to duration, subjects invested considerably more time in the second session: 19.4 min more, on average (more than 50%). To check whether the observed difference in terms of time was significant, we ran the *t*-test, which indicated a statistically significant difference at 99% confidence level (df = 19, *p*-value = 0.002406). Such result favors the *alternative* hypothesis (H_2-*A*) that subjects invest more time implementing functions after learning the ST concepts and techniques. This is an indication that they tend to be more cautious after the exposition to ST theory, either by investing more time in the application code itself, or in the tests.

On the other hand, when we look into the produced lines of application code, we did not find a significant difference: subjects produced, on average, 2.03 more LoC in the second session (15%). This is quite surprising as the implementations produced after learning the ST theory were more reliable than the ones produced before, but not significantly larger in terms of LoC. This indicates that ST theory might improve programming skills not only in terms of producing more correct code, but also in terms of producing leaner code. To confirm our intuition, we ran the *t*-test, which in fact did not indicate a statistically significant difference at 99% confidence level (df = 19, *p*-value = 0.03048). Such result favors the *null* hypothesis (H_2-*0*) that subjects produce similar number of lines of code after learning ST.

As commented in Section 3, some studies have found a positive correlation between the size of program modules in LoC and fault-proneness. Since implementations produced in the second session were not significantly larger in LoC, we believe this is an additional evidence that reliability did not decrease from the first to the second session.

With respect to complexity, code produced in the second session was sensibly more complex than code produced before, in terms of the McCabe's complexity metric: 1.18 points higher, on average (34.70%). Such difference indicates that the functions implemented after learning ST contained more conditional statements. This is expected, as those functions were more successful against the established test suites, which covered exceptional inputs. The *t*-test indicated a statistically significant difference at 99% confidence level (df = 19, *p*-value = 0.004649). Such result favors the alternative hypothesis (H_3-*A*) that subjects tend to produce more complex implementations after learning the ST concepts and techniques.

Small treatment group replication

Table 8 shows results of our small treatment group $\mathbf{ST_2}$. Since the sample size for this experimental instance was small – only 10 subjects participated in the replication –, we only conduct a descriptive analysis of the results.

Note that for this group the FTSSR also had a small increase from the first to the second session – that is, students performed slightly better with respect to reliability after learning ST, similarly to the main treatment group. Total Development Time also increased afterwards, by 14.4 min, on average. The other metrics, LoC and cyclomatic complexity also had a small increment.

The most interesting outcomes for this group, however, are related to limit values of FTSSR. Note that four functions developed by students in the first session did not pass any of the tests. On the second session, however, only a single function did not pass any of the tests (a 75% decrease). Also note that there were no implementations that passed all tests in the first session, while in the second, this number raised to three. While it is hard to affirm anything for certain based on these extreme cases, we believe they are consistent with our observation that ST knowledge can make

Table 8
Outcomes for the **ST_2** (small) treatment group.

S	1st Session						2nd Session					
	FTSSR			ΣT	μL	μM	FTSSR			ΣT	μL	μM
	F_1	F_2	μ				F_1	F_2	μ			
1	0.50	0.56	0.53	35	15.00	2.00	0.43	0.86	0.64	45	12.00	3.00
2	0.00	0.63	0.31	60	14.50	2.50	0.57	0.17	0.37	25	15.00	3.50
3	0.85	0.69	0.77	30	21.00	3.50	1.00	0.86	0.93	240	12.00	2.50
4	0.80	0.82	0.81	47	18.50	6.00	0.92	1.00	0.96	27	15.50	5.00
5	0.43	0.00	0.21	90	21.50	6.00	0.57	0.85	0.71	60	20.00	4.50
6	0.83	0.75	0.79	50	15.50	3.00	0.43	0.09	0.26	50	17.00	4.50
7	0.00	0.33	0.17	30	9.50	2.50	0.14	0.13	0.13	35	18.00	2.50
8	0.58	0.42	0.50	40	13.50	2.00	0.29	0.25	0.27	60	35.50	4.50
9	0.18	0.67	0.42	33	19.00	3.50	0.43	0.00	0.21	50	12.00	2.50
10	0.00	0.83	0.42	40	13.50	2.00	0.85	1.00	0.93	7	14.00	3.00
μ	0.42	0.57	0.49	45.50	16.15	3.3	0.56	0.52	0.54	59.90	17.10	3.55

Legend: S = Subject; FTSSR = Functional Test Set Success Rate; F_n = Function n; μ = Average; Σ = Total; T = Development Time (in minutes); L = Lines of Code; M = Cyclomatic Complexity.

Table 9
Outcomes for the **SE** main control group.

S	1st Session						2nd Session					
	FTSSR			ΣT	μL	μM	FTSSR			ΣT	μL	μM
	F_1	F_2	μ				F_1	F_2	μ			
1	0.43	0.18	0.31	73	13.50	2.50	0.85	0.67	0.76	32	12.00	3.00
2	0.58	1.00	0.79	31	9.50	2.50	0.00	0.71	0.36	41	23.00	2.50
3	1.00	0.57	0.79	22	13.50	5.00	0.18	0.67	0.42	45	17.50	3.50
4	0.95	0.50	0.73	57	19.50	2.50	0.27	0.19	0.23	74	17.50	2.50
5	0.43	0.82	0.62	32	12.50	3.50	0.17	0.13	0.15	50	20.00	3.50
6	0.92	0.18	0.55	28	20.00	5.00	0.13	1.00	0.56	29	20.50	3.50
7	0.67	0.58	0.63	70	13.00	2.00	0.85	0.45	0.65	70	19.00	4.50
8	0.43	0.13	0.28	9	16.50	2.50	0.82	0.58	0.70	6	15.50	4.00
9	1.00	0.25	0.63	15	14.00	2.50	0.85	0.55	0.70	11	11.50	3.50
10	0.36	0.67	0.52	17	14.00	2.00	0.14	0.55	0.35	18	15.00	3.00
11	0.69	0.00	0.34	100	9.00	1.50	0.33	0.00	0.17	34	15.50	2.50
12	0.14	0.00	0.07	23	18.00	3.00	0.75	0.75	0.75	22	19.50	3.00
13	0.92	0.71	0.82	32	16.00	4.00	0.50	0.63	0.56	32	24.00	4.00
14	0.85	1.00	0.93	10	10.00	3.50	0.27	0.69	0.48	9	15.50	3.00
15	0.71	0.67	0.69	89	20.50	3.00	0.92	1.00	0.96	31	16.00	5.00
16	0.85	0.69	0.77	23	15.00	3.00	0.71	0.45	0.58	85	28.00	6.50
17	0.29	0.83	0.56	100	15.50	2.00	0.50	0.85	0.68	25	18.50	4.50
18	0.92	0.82	0.87	90	15.50	3.50	0.13	0.83	0.48	22	14.00	4.50
19	0.92	0.00	0.46	17	13.00	3.00	0.67	0.43	0.55	25	12.00	2.00
20	0.14	0.58	0.36	4	11.50	1.50	0.75	0.71	0.73	60	20.00	6.50
21	0.85	0.86	0.85	75	20.00	4.00	1.00	0.42	0.71	30	12.50	2.00
22	0.33	0.29	0.31	33	14.50	2.50	0.13	0.58	0.35	22	16.00	2.50
μ	0.65	0.51	0.58	43.18	14.75	2.93	0.50	0.58	0.54	35.14	17.41	3.61

Legend: S = Subject; FTSSR = Functional Test Set Success Rate; F_n = Function n; μ = Average; Σ = Total; T = Development Time (in minutes); L = Lines of Code; M = Cyclomatic Complexity.

developers more cautious, and thus improve the reliability of their code.

4.2. Control groups

The control groups were added to our study to reduce threats to the internal validity of our experiment, in particular related to history (*e.g.*, students in the treatment groups could have been exposed to other types of knowledge that might also have had an impact on reliable programming). Next we analyze results for both control groups included in our study – **SE** and **OO**.

Table 9 presents the results for our main control group – **SE**, with a sample of 22 subjects. Again, to allow a visual analysis of the main metric targeted in our study, Fig. 2 shows a boxplot of the FTSSR outcomes for that group. Note that results related to FTSSR, our main metric, was very different from the treatment group. The difference becomes very clear when we look at the two boxplots presented, for the main treatment and control groups (Figs. 1 and 2). Note that while the second session box from the treatment group is located at a higher level than the first session box, the second session average from the control group is slightly lower than the first session average. In fact, at this time, students had a marginally worse performance in the second session when compared to the first (7% lower, 0.58 in the first session vs. 0.54 in the second session, on average). The *t*-test shows that the difference was not statistically significant at 99% confidence level (df = 23, *p*-value = 0.734). This is an important result, as it shows evidence that the ST knowledge can have a significant impact on reliable programming, while SE cannot.

When we look at the effort metrics, we can also see that they were not significantly affected by the SE knowledge either. Differently from the treatment group, subjects took a little less time on average to develop functions in the second session (8 min less on average than in the second session). The *t*-test confirms a non-significant difference at 99% confidence level (df = 23, *p*-value = 0.2157). LoC, on the other hand, was increased by 2.65, on average. However, the *t*-test also indicates a non-significant difference at 99% confidence level (df = 23, *p*-value = 0.0161). With respect to

Table 10
Outcomes for the **OO** (small) control group.

S	1st Session						2nd Session					
	FTSSR			ΣT	μL	μM	FTSSR			ΣT	μL	μM
	F_1	F_2	μ				F_1	F_2	μ			
1	1.00	0.75	0.88	67.00	13.50	2.50	0.60	1.00	0.80	23.00	19.50	6.00
2	0.58	0.29	0.43	35.00	12.00	3.00	0.00	0.75	0.38	11.00	17.00	3.00
3	0.71	0.75	0.73	27.00	19.50	3.50	0.82	0.56	0.69	39.00	24.00	5.00
4	0.95	0.06	0.51	49.00	12.50	1.00	0.92	0.75	0.83	30.00	14.00	2.50
5	0.92	0.43	0.67	41.00	18.00	5.00	1.00	0.85	0.93	11.00	10.50	3.50
6	0.43	0.58	0.51	18.00	17.50	5.00	0.56	0.83	0.70	53.00	15.50	3.00
7	0.50	0.29	0.39	45.00	16.00	3.00	0.67	0.29	0.48	28.00	13.50	2.00
8	0.92	0.29	0.60	8.00	11.00	3.00	1.00	0.57	0.79	28.00	17.00	5.50
μ	0.75	0.43	0.59	36.25	15.00	3.25	0.70	0.70	0.70	27.88	16.38	3.81

Legend: S = Subject; FTSSR = Functional Test Set Success Rate; F_n = Function n; μ = Average; Σ = Total; T = Development Time (in minutes); L = Lines of Code; M = Cyclomatic Complexity.

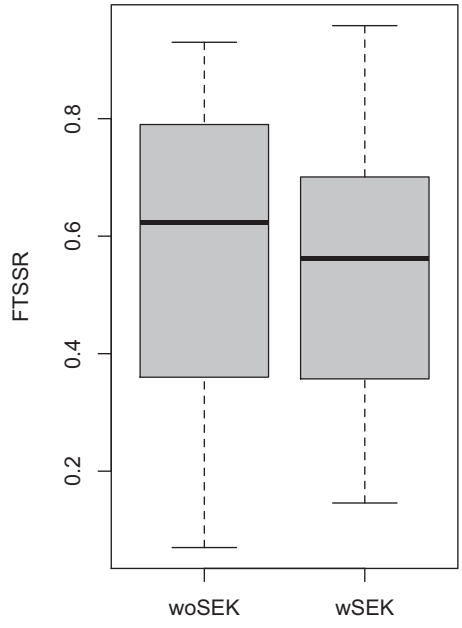

Fig. 2. Boxplot of the Functional Test Set Success Rate (FTSSR) outcomes of our experiment, for the **SE** control group. Legend: woSEK = without Software Engineering Knowledge; wSEK = with Software Engineering Knowledge.

complexity, the McCabe's complexity metric was improved only by 0.68 points. Such difference was also confirmed as non-significant by the *t*-test, at 99% confidence level (df = 23, *p*-value = 0.03593).

In summary, the performance of our main control group – **SE** – points out that the SE knowledge does not affect significantly the performance of developers, as the difference between outcomes for all metrics were found to be non-significant from one session to the other. This is a stronger evidence that the ST knowledge can in fact impact on the reliability of code, as when students take another course – and one that is also related to software development –, we do not observe the differences that were reached when subjects were taking the ST course.

Table 10 shows results for the **OO** control group. Although the sample size was much smaller for such a group – only 8 subjects – , some interesting outcomes are worth noting. In particular, results reached for this group reinforce the evidence that the ST theory can have more impact on reliable programming than other types of knowledge. Note that although FTSSR also improved in the second session for this group – differently than for the **SE** group –, the difference was not as significant and consistent as for the main treatment group. It is worthwhile noting that this group was taking a Java OO design course, so it was expected that their programming skills would be somehow improved afterwards, even though the OO concepts and patterns were not directly related to the algorithms implemented in the experiment rounds. On the other hand, the ST course, which does not focus directly on programming, had a more significant impact on reliable programming.

Other outcomes are worth noting. For instance, similarly to the main treatment group, there was a single implementation that did not pass any of the tests. However, for this control group, such implementation was produced in the second session (recall that for the treatment group no subject implemented a function that did not pass any tests in the second session). Also note that, differently from the main treatment group, cyclomatic complexity was not significantly higher in the second session (df = 7, *p*-value = 0.2251). This is an evidence that the students taking the ST course were probably more cautious about exceptional inputs, which made them add conditionals to their implementations, whereas the same did not happen here. On the other hand, similarly to the **SE** group, for the **OO** group the total development time reduced from the first to the second session. This is an interesting deviation from the treatment group, which took significantly more time to develop functions after learning ST (probably because they were more cautious and developed tests for their implementations). With respect to LoC, similarly to the treatment and **SE** control group, there was no significant difference from the first to the second session.

In summary, with respect to reliability, it can be noticed that the ST knowledge seems to have had more significant impact on the students than the Software Engineering and OO design knowledge. Also, it seems that the subjects that were exposed to ST theory became more cautious afterwards, by taking more time to implement functions and tests, and by adding conditionals to cover borderline cases.

4.3. Survey with computer science professors

In our first investigation with students, we have observed that teaching ST provides benefits. But are programming educators able to teach testing? In particular, what is the level of ST knowledge of instructors who teach introductory programming courses? Our second investigation tried to shed some light into this question.

We obtained 53 responses to our survey with the aim of assessing the level of ST knowledge of Computer Science professors that have taught introductory programming courses. The participants are from diverse Brazilian institutions. Table 11 presents statistics of the results from the survey. Outcomes reveal that instructors have a low level of ST knowledge, since the average score was 4.24. Note that more specific questions – such as Question 8 about mutation testing – received the lowest scores.

Table 11
Outcomes for our survey with CS professors.

Question	Average score
1	0.41
2	0.49
3	0.49
4	0.70
5	0.38
6	0.40
7	0.17
8	0.05
9	0.24
10	0.88
Overall score	
Avg.	4.24
Min.	0.00
Max.	7.75

To analyze the most common deficiencies, next we discuss the most frequent answers for each of the questions. With respect to Question 1: *What is Software Testing?*, we noticed that most subjects define ST as "a technique to verify whether the software corresponds to its requirements". Instructors fail to see the *destructive* aspect of ST, which is very important for an effective testing activity, as discussed a long time ago by Myers et al. (2011), since the first edition of "The Art of Software Testing". Also, most subjects fail to mention that Software Testing refers to a dynamic task that requires running software, as opposed to other static techniques, such as inspections (whose goals are also to check whether the "software corresponds to its requirements").

With respect to Question 2: *What is a test case?*, most subjects have only a high-level view of this term, commonly defining it as "a usage scenario of the program". This reveals a lack of understanding about the concreteness of a test case, which basically refers to a pair of input and expected output for a given program. In Question 3: *In general, what constitutes a successful test case?*, most professors responded that a test that passes is the most successful. Subjects again fail to see the that a failing test case is much more valuable to a developer, since it reveals an as yet undiscovered fault (given that the most realistic approach to software development is to expect that every program contains faults, as referred to in Question 10).

Most subjects answered Question 4: *What is the input domain of a program?* correctly. We believe that this might be because *domain* is a more general term that also appears in other fields (*e.g.*, in mathematical functions), so more subjects are familiar with such a concept. In Question 5: *What is functional - or black-box - testing?*, most professors recognize that it refers to a testing technique in which the tester does not have knowledge about the internal structure of the implementation. However, most fail to mention that the input data for functional testing comes from the inspection of specifications. With respect to Question 6: *What is structural - or white-box - testing?*, again most subjects correctly pointed out that it refers to a technique in which the tester has knowledge about the internal structure of the software. Nonetheless, most of them failed to mention that test data should be derived from such a knowledge.

With respect to the more specific ST questions, namely, Question 7: *What is boundary value analysis?*, Question 8: *What is mutation testing?*, and Question 9: *What is a testing criterion?*, most subjects responded that they did not know about such concepts. For Question 10: *Is it reasonable to argue that virtually all programs contain faults?*, most subjects responded correctly by saying 'Yes'. However, we found it surprising that there were still six respondents that answered "No" (that is, they failed to see that it is improbable for developers to produce even a simple program with total absence of faults, due to the complexity of software itself).

The results of our survey are important to show that instructors lack the adequate understanding of basic ST concepts, the same knowledge that, as indicated by our main study, can make developers produce more reliable software.

```
public static void main(String[] args) {
    //m1 - power of two
    boolean r;
    r=Util.isPowerOfTwo(5);
    System.out.print(r);
    System.out.println("");
    r=Util.isPowerOfTwo(1);
    System.out.print(r);
    System.out.println("");
    r=Util.isPowerOfTwo(4);
    System.out.print(r);
    System.out.println("");
    r=Util.isPowerOfTwo(8);
    System.out.print(r);
    System.out.println("");
    r=Util.isPowerOfTwo(16);
    System.out.print(r);
    ...
}
```

Fig. 3. A code snippet from tests developed by one of the subjects in the second session of our experiment.

5. Discussion

An interesting insight provided by our experiment is that subjects in the treatment group seemed to have performed better even when no specific testing technique was explicitly applied. As discussed in Section 3, no testing tools were readily available to the students when they performed the experimental tasks. This was done because we were more interested in the impact that the testing *knowledge* itself had on programming skills, not as much in the impact of the *application* of formal testing techniques.

Although students were also trained to write JUnit test cases, some wrote their tests on `main` programs. For instance, Fig. 3 shows a code snippet of tests implemented by one of the subjects on the second session in a `main` method. This student was one of the ones that obtained good results after learning ST theory. Note that nothing implies that a specific testing technique was being explicitly applied. Although it appears that a strategy similar to the one established by functional testing was followed (*i.e.*, some even and odd numbers were tried, the number one was tried), it does not seem that a formal partitioning of the input domain was performed, neither that all boundary values were considered. However, the student did perform better in the second session, indicating that the ST theory improved her/his programming skills. In fact, in one of the first session implementations by the same subject, none of the tests passed, indicating that significantly better coding was happening in the second session.

On the other hand, it must be noted that some of the students did write tests following a specific technique. For instance, Fig. 4 shows a partial JUnit class with tests developed by another subject that also had better outcomes on the second session. Note that, in this case, it appears that functional testing was being followed more thoroughly: there are additional tests for zero, a negative number, and a large number.

Another interesting result in our study refers to the increase in cyclomatic complexity from code produced after learning ST. We hypothesize that this increase is related to additional checks that subjects tend to add after being exposed to testing concepts.

```
public class UtilTest extends TestCase {
 ...
 public void testIsPowerOfTwo(){
  assertTrue(Util.isPowerOfTwo(4));
 }
 public void testIsPowerOfTwoWrong(){
  assertFalse(Util.isPowerOfTwo(9));
 }
 public void testIsPowerOfTwoZero(){
  assertFalse(Util.isPowerOfTwo(0));
 }
 public void testIsPowerOfTwoNegative(){
  assertFalse(Util.isPowerOfTwo(-2));
 }
 public void testIsPowerOfTwoLarge(){
  assertTrue(Util.isPowerOfTwo(4096));
 }
 ...
}
```

Fig. 4. Code snippet from tests developed by another subject in the second session of our experiment.

In particular, we noticed students were more careful while implementing their functions after taking the course. For instance, consider the code snippets presented in Fig. 5 which refer to functions implemented by the same subject before and after learning ST. We can see that in the first function there are no checks against variables (*e.g.*, to verify nullity of variables). On the other hand, in the second function the subject tested to see (1) whether strings did not have a particular format; (2) whether a particular object was not null; and (3) whether the file existed afterwards. While it is difficult to compare different types of functions that may require more or less checks, we can see that in the second implementation the subject was more careful and introduced more conditionals to the code. We believe that was the reason that cyclomatic complexity was significantly increased for such a sample, while the same was not observed for the control groups.

In their classic *The Art of Software Testing* (Myers et al., 2011), Myers et al. argue that "the software tester needs the proper attitude [..] to successfully test a software application". They go on to say that such *psychology of testing* establishes the most important considerations in ST. Our results indicate that the exposition to the principles behind such testing attitude, along with the techniques that derive from it, might produce positive effects on programmers themselves, with respect to reliable programming.

Agile development proponents seem to have identified the effect of such *testing attitude*, by encouraging testing activities along the development process (*e.g.*, the *test infected* condition Beck, 2000). In fact, in a recent survey with 326 agile developers, two of the top 5 most important agile principles identified by developers involved testing (*i.e.*, "Automated tests run with each build", and "Automated unit testing") (Williams, 2012). Although ST activities in the agile community are more on the lines of *practical* testing as opposed to more formal testing, both agile philosophy and classical testing education do share important principles with regards to ST. For instance, both traditional testing literature (Myers et al., 2011) and test-driven development proponents (Ambler, 2002) put forward the idea that *successful* test cases are the ones that find faults, and not the ones that simply "pass".

6. Threats to validity

It is common knowledge that all empirical studies have limitations (Briand and Labiche, 2004). On the other hand, we believe our study had several characteristics that made it more rigorous and thus with improved validity. In particular, the fact that it was conducted in an academic setting made us have more control over confounding effects. In any case, there are still limitations that are worth mentioning. In this section we discuss such limitations based on three types of threats to validity described by Wohlin et al. (2000). For each type, we list possible threats, measures taken to reduce each risk, and suggestions for improvements in future studies.

6.1. Internal validity

The lack of control of the subjects' skills on programming and testing (other than all being in the same year of the program) might have affected the internal validity of our experiment. However, the repeated measures design decreases the probability of this threat affecting our outcomes, because the same subjects implemented functions before and after learning ST and OO Design.

An aspect to be also considered is *mortality*. Since we had more students invited to participate in the experiment in the beginning (some of them did not complete the two sessions and were therefore excluded from the study), the actual tasks that took place did not follow the exact initial assignments. This could affect the balance of the assignments that was taken into consideration in the experiment design. However, we believe that since our sample was not too small for the treatment group, an adequate balance could still be maintained. Moreover, the initial assignment set contained some redundancies which helped circumvent such threat.

One may question the effect of the recent exposure to testing-related theory and practice on the performance of the treatment groups. A question that raises is: "*What should be the results of the experiment if the subjects were invited to perform similar assignments after some time (e.g., a few months)*?". To analyze this issue, we would need to run more rounds of the experiment with the same subjects and similar assignments. Nevertheless, we call the reader's attention to the fact that knowledge just acquired by the subjects of the control groups (namely, OO design and Software Engineering) did not significantly improve their programming skills in terms of reliability. That is, the exposure to other SE-related theory and practice did not lead CS students to achieve significant increase in code quality in terms of reliability.

Another threat to the internal validity of our experiment was the removal of some subjects that produced outliers with respect to the FTSSR metric in the treatment group.[9] In fact, when the same statistical test is run in the presence of the outliers, it indicates a non-significant difference at 99% confidence level. In order to justify the removal of such outliers, we investigated what could have affected these particular students in making them produce such extraneous results. In three cases, the FTSSR score was far below the average in the second session (0.21, 0.29, and 0.30), and in the remaining case, far below the average in the first session (0.31). The three subjects of the first group – results which could affect our conclusions – had strong reasons not to benefit well from the ST course: one of them was a foreigner with difficulty in understanding the language in which the course was taught; the second had a poor record of scores in many other courses taken; and the third was a part-time student. It is important to note that the same outlier identification process was conducted for all data produced in our study, and that the outliers only appeared for this group.

Finally, an additional threat to the internal validity of our experiment refers to the development time variable. Since subjects themselves were responsible for recording their own development time, some of them might have registered it wrong, which could affect our results. In any case we made sure to stress students to correctly register their starting and submission times as comments

[9] In all cases the outliers were identified by box-plotting our results in R.

```
...
String []split = s.split(" ");
StringBuffer []sbs =
  new StringBuffer[split.length];
for(int i = 0;  i < split.length; i++){
  sbs[i] = new StringBuffer(split[i]);
  Character c = sbs[i].charAt(0);
  c = c.toUpperCase(c);
  sbs[i].setCharAt(0, c);
}

s = "";
for(StringBuffer s1: sbs){
  s += s1.toString();
  s +=" ";
}
s = s.trim();
System.out.println(s);
return s;
...
```

(a) Code snippet of the s_1 function.

```
...
if(source.startsWith(" "))
  return;
if(target.startsWith(" "))
  return;
File f = new File(target+"/"+source);
if(f == null)
  return;
f.createNewFile();
if(!f.exists() )
  return;
...
```

(b) Code snippet of the i_2 function.

Fig. 5. Snippets of code implemented by a subject before and after learning ST.

to their code. Since tasks usually did not take long (2 h maximum), we believe most subjects registered their completion time correctly. On the other hand, since the same registering method was used for all subjects, imprecise outcomes that possibly occurred might have been equally distributed between sections.

6.2. External validity

The use of students as subjects for our experiment might have reduced its external validity. In fact, some experiments have shown opposite trends for students and professionals (*e.g.*, Arisholm and Sjoberg, 2003). However, according to other authors, students can play an important role in experimentation in the field of Software Engineering (Basili et al., 1999; Kitchenham et al., 2002). For instance, Canfora et al. (2007) have conducted a pair programming experiment in academia and replicated the same experiment in industry. According to the authors, the experiments produced similar results for both samples.

Another more recent study that involved Test-Driven Development (TDD) has also shown similar outcomes for both students and professionals. The same study points with more in-depth analysis that students can in fact sometimes be representatives of professionals in Software Engineering experiments, in particular when trying new approaches (Salman et al., 2015). Since in our case, students were being exposed to testing knowledge for the first time, it is likely that professionals with little testing background would perform similarly. Carver et al. (2010) also analyze characteristics in which empirical studies with students can make them more valid. According to the authors, when studying issues related to a technology's learning curve or the behavior of novices, students are exactly the right test population. In our case, since we wanted to evaluate how learning ST could impact on the behavior of programmers, we believe our experiment falls into such a category of study.

The students involved in our experiment – except for the single one mentioned before – were all Brazilians. It might be the case that students from other nationalities might perform differently. Moreover the experiment only involved students from a single school, the Federal University of São Paulo. Replications with students from other countries and schools would be required to be able to further generalize our results.

As commented in Section 4, *history* might also have affected our experiment. In particular, subjects of the treatment group could have gathered other knowledge that would also impact on the FTSSR metric. However, the fact that FTSSR outcomes were not significantly affected for the large control group – with students at the same level but who were not taking the ST course – helps increase confidence that the gathered ST knowledge explains the improvement in reliable programming for the treatment group.

6.3. Construct validity

A characteristic of our experiment that might have affected its construct validity is related to the metrics we have chosen to evaluate our results. For instance, we have used functional testing to develop the set of test cases to measure the reliability of the produced code. However, since we have taught such technique to the subjects, results might only indicate the extent to which they have applied it, but not the impact of the ST knowledge as a whole. However, as discussed in Section 5, some students did not appear to have explicitly used this technique, which indicates that at least for some subjects, other principles and techniques might have played a role in the final outcomes. For instance, mutation testing theory stresses common types of faults made by developers, modeled as mutation operators. Such theory might also have stimulated students to omit faults that were present in the first session implementations in their second session code. Moreover, as commented in Section 4, the non-significant increase in LoC in implementations from the first to the second session was also a secondary evidence that reliability has not decreased afterwards.

7. Related work

To the best of our knowledge, there are no studies that directly measure the impact of Software Testing education on programming skills, in particular with respect to the production of more reliable software. However, there are investigations into the testing proficiency of CS students, and the improvement of ST education in CS curricula. Also, there are studies that generally highlight the importance of ST education. Next, we sample some of these studies in comparison to ours.

In regard to the testing proficiency of CS students, Carver and Kraft (2011) conducted an empirical study to determine the testing ability of senior-level CS students. In particular, they wanted to evaluate whether students were able to create small, complete test suites for simple programs. Results show that a coverage tool can significantly help students produce better tests. While this investigation is related to ours, it is significantly different in that Carver

and Kraft's experiment evaluated the testing ability of students, while our study investigates the impact of ST education on programming skills.

With respect to improving Software Testing education, particularly in CS programs, some authors report on a positive "side-effect" in programming skills. Nevertheless, to the best of our knowledge, these studies do not report on controlled experiments to assess such "side-effect". For instance, Jones (2001) has explored the integration of testing into introductory CS courses through testing labs and diverse forms of courseware, including tool support for automated program grading. The author reports on the experience of an elective testing course compound by 80% of practice and 20% of theory. Beyond ordinary testing-related practices such as test case design, students have performed reverse engineering tasks to derive system specifications, have created test drivers and have written test scripts. According to the author, one major benefit obtained from this experience was the general improvement of students' software design and programming skills. Despite this, Jones (2001) did not report on objective measurement of the gains related to programming skills improvement and quality of produced code, that is, conclusions are more qualitative than quantitative. Our study, on the other hand, involved a quantitative controlled experiment, which provided more concrete evidence about such skill improvement.

Initiatives like Test Driven Learning (TDL) have also yielded gains in students' programming ability. For instance, Janzen and Saiedian (2006) introduced a combined, simultaneous testing-programming learning approach that can help both novice and experienced programmers improve their comprehension and ability, and hence produce high quality code design with reduced defect density. To verify their assumptions, the authors reported on an experiment that involved first-year CS students. The students have run four 50-min lectures whose topics were related to arrays and object manipulation. While one group of students was taught using a TDL approach, another group was exposed to non-TDL (*i.e.* traditional) teaching manner. A posterior evaluation revealed that the first group performed 10% better than the second on a quiz that covered the concepts and syntax from the experiment topics. Although our experiment did not consider a simultaneous testing-programming learning process, our results are well-aligned with Janzen and Saiedian's (2006).

More recently, (Offutt et al., 2011) also presented an approach to teach students to test better. The authors present an in-depth teaching experience report on how to successfully teach criteria-based test design using abstraction and publicly accessible web applications. Clarke et al. (2014), on the other hand, proposed a collaborative learning environment to integrate testing education into Software Engineering courses. Although these studies are important to improve the level of testing education, none of them investigate its impact on students' programming skills.

In the literature, we can also find several studies that highlighted the importance of testing education, and reported the lack of adequate ST training in CS curricula. For instance, Astigarraga et al. (2010) showed that most CS academic programs tend to emphasize *development* at the expense of *testing* as a formal engineering discipline. Wong (2012), on the other hand, argued that software testers are generally not adequately trained because most CS programs offer ST only as elective courses. Clarke et al. (2014) also argued that due to the large number of topics to be covered in SE courses, little or no attention is given to Software Testing, resulting in students entering industry with little to no testing experience. Since our experiment provides evidence that ST education can lead to the production of more reliable software, it presents a new argument to further motivate better testing education.

8. Conclusions

In a recent article, Vinton Cerf, one of the fathers of the Internet, called attention to the growing need for *responsible programming* in industry (Cerf, 2014). Such term is defined as a clear sense of responsibility programmers should have for their systems' reliable operation and resistance to compromise and error. We also believe programmers should be more responsible for the code they produce, specially today when so many depend heavily on software to work as advertised.

However, Cerf suggests that in order to achieve responsible programming, we need better tools and programming environments. While we agree with him, we believe that better *training* of professionals and instructors could also improve the quality of the produced software, in particular by the teaching of classical Software Testing theory. Other authors have argued that more exposure to Software Testing practices and tools is required for better training of software developers (Clarke et al., 2014; Wong, 2012).

In this paper we presented scientific evidence for such a claim. We conducted an experiment to verify the impact that testing knowledge has on programming skills, in terms of reliability. Our results suggest that after learning basic testing principles and techniques, developers are more than twice more likely to produce correct implementations. Moreover, our data suggests that although the code produced afterwards is more reliable, it does not tend to be significantly larger (in terms of lines of code). This indicates that the exposure to testing knowledge can make developers produce more reliable implementations with approximately the same amount of code. The performance of two control groups taking a Software Engineering and an OO design course was very different. In fact, for these groups the acquired knowledge did not seem to impact on the subjects' programming skills with respect to reliability.

On the other hand, we also wanted to evaluate the level of ST knowledge of instructors that teach programming courses themselves. In a survey involving 53 CS professors we show that these subjects lack the basic ST knowledge that was indicated in the students' evaluation to be effective for more reliable programming. This is an issue because programming apprentices might be failing to learn valuable programming concepts early in their programs.

Future work includes the replication of our experiment with professional developers and larger groups of students to shed more light on this subject. Moreover, it would be interesting to conduct experiments to further analyze how training in each separate testing technique can impact on programming skills (*e.g.*, by performing several programming tasks along the course).

Acknowledgments

The authors would like to thank for the following financial support: *Otávio Lemos*: FAPESP (grant 2013/25356-2, 2015/12787-0); *Fabiano Ferrari*: CNPq/Universal (grant 485235/2013-7); *Fábio Silveira*: CNPq/Universal (grant 455080/2014-3); *Alessandro Garcia*: FAPERJ (distinguished scientist grant E-26/102.211/2009), CNPq Productivity (grant 305526/2009-0) and Universal (grant 483882/2009-7), and PUC-Rio (productivity grant).

The authors would also like to thank the students who participated in the experiment.

References

Ambler, S., 2002. Introduction to test-driven development (TDD). Available from: http://www.agiledata.org/essays/tdd.html (accessed 10/mar/2016).

Ammann, P., Offutt, J., 2008. Introduction to Software Testing, 1 Cambridge University Press, New York, NY, USA.

Arisholm, E., Sjoberg, D.I.K., 2003. A Controlled Experiment with Professionals to Evaluate the Effect of a Delegated versus Centralized Control Style on the Main-

tainability of Object-Oriented Software. Technical Report 6. Simula Research Laboratory.
Astigarraga, T., Dow, E.M., Lara, C., Prewitt, R., Ward, M.R., 2010. The emerging role of software testing in curricula. In: Proceedings of the IEEE Transforming Engineering Education: Creating Interdisciplinary Skills for Complex Global Environments. IEEE, pp. 1–26.
Basili, V.R., Shull, F., Lanubile, F., 1999. Building knowledge through families of experiments. IEEE Trans. Softw. Eng. 25, 456–473.
Beck, K., 2000. JUnit test infected: Programmers love writing tests. Available from: http://junit.sourceforge.net/doc/testinfected/testing.htm (accessed 10/mar/2016).
Binder, R.V., 1999. Testing Object-Oriented Systems: Models, Patterns, and Tools. Addison-Wesley, Boston, MA, USA.
Briand, L., Labiche, Y., 2004. Empirical studies of software testing techniques: challenges, practical strategies, and future research. SIGSOFT Softw. Eng. Notes 29 (5), 1–3.
Canfora, G., Cimitile, A., Garcia, F., Piattini, M., Visaggio, C.A., 2007. Evaluating performances of pair designing in industry. J. Syst. Softw. 80, 1317–1327. doi:10.1016/j.jss.2006.11.004.
Capgemini Group, Sogeti, HP, 2016. World Quality Report 2015-16. Tech. Report. Cap Gemini S.A., Sogeti and Hewlett-Packard. https://www.capgemini.com/thought-leadership/world-quality-report-2015-16(accessed 10/mar/2016).
Carver, J., Kraft, N., 2011. Evaluating the testing ability of senior-level computer science students. In: Proceedings of the 24th IEEE-CS Conference on Software Engineering Education and Training, pp. 169–178.
Carver, J.C., Jaccheri, L., Morasca, S., Shull, F., 2010. A checklist for integrating student empirical studies with research and teaching goals. Empirical Softw. Eng. 15 (1), 35–59.
Cerf, V.G., 2014. Responsible programming. Commun. ACM 57 (7). 7–7.
Clarke, P.J., Davis, D., King, T.M., Pava, J., Jones, E.L., 2014. Integrating testing into software engineering courses supported by a collaborative learning environment. ACM Trans. Comput. Educ. 14 (3), 18:1–18:33.
Elbaum, S., Person, S., Dokulil, J., Jorde, M., 2007. Bug hunt: Making early software testing lessons engaging and affordable. In: Proceedings of the 29th International Conference on Software Engineering (ICSE 2007). IEEE, pp. 688–697.
Gyimothy, T., Ferenc, R., Siket, I., 2005. Empirical validation of object-oriented metrics on open source software for fault prediction. IEEE Trans. Softw. Eng. 31 (10), 897–910.
Harrold, M.J., 2000. Testing: a roadmap. In: Proceedings of the Conference on The Future of Software Engineering. ACM, pp. 61–72.
Howden, W.E., 1982. Weak mutation testing and completeness of test sets. IEEE Trans. Softw. Eng. 8 (4), 371–379.
IEEE, 1990. IEEE Standard Glossary of Softw. Eng. Terminology. IEEE Computer Society Press, New York.
IEEE, 2012. IEEE Std 1012–2012 (Revision of IEEE Std 1012–2004). IEEE Computer Society Press, New York, pp. 1–223.
Janzen, D.S., Saiedian, H., 2006. Test-driven learning: Intrinsic integration of testing into the CS/SE curriculum. In: Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education (SIGCSE 2006). ACM, Houston, Texas, USA, pp. 254–258.
Jones, E.L., 2001. Integrating testing into the curriculum – arsenic in small doses. In: Proceedings of the 32th SIGCSE Technical Symposium on Computer Science Education (SIGCSE 2001). ACM, pp. 337–341.
Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., Emam, K.E., Rosenberg, J., 2002. Preliminary guidelines for empirical research in software engineering. IEEE Trans. Softw. Eng. 28, 721–734.
Lemos, O.A.L., Ferrari, F.C., Silveira, F.F., Garcia, A., 2012. Development of auxiliary functions: should you be agile? an empirical assessment of pair programming and test-first programming. In: Proceedings of the 34th international Conference on Software Engineering (ICSE 2012). IEEE, pp. 529–539.
Lemos, O.A.L., Ferrari, F.C., Silveira, F.F., Garcia, A., 2015. Experience report: Can software testing education lead to more reliable code? In: Proc. of the 26th IEEE International Symposium on Software Reliability Engineering (ISSRE), pp. 359–369.
Madeyski, L., Radyk, N., 2002. Judy: a mutation testing tool for java. IET Softw. 4, 32–42.
McCabe, T.J., 1976. A complexity measure. IEEE Trans. Softw. Eng. 2 (4), 308–320.
Montgomery, D.C., 2006. Design and Analysis of Experiments. John Wiley & Sons.
Myers, G.J., Sandler, C., Badgett, T., 2011. The Art of Software Testing, 3rd ed Wiley Publishing.
Offutt, J., Li, N., Ammann, P., Xu, W., 2011. Using abstraction and web applications to teach criteria-based test design. In: Proceeding of the 24th IEEE-CS Conference on Software Engineering Education and Training. IEEE, pp. 227–236.
Patterson, A., Kölling, M., Rosenberg, J., 2003. Introducing unit testing with BlueJ. In: Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE 2003). ACM, pp. 11–15.
Rapps, S., Weyuker, E.J., 1985. Selecting software test data using data flow information. IEEE Trans. Softw. Eng. 11 (4), 367–375.
Salman, I., Misirli, A.T., Juristo, N., 2015. Are students representatives of professionals in software engineering experiments? In: Proceeding of the 37th International Conference on Software Engineering (ICSE 2015). IEEE, pp. 666–676.
Subramanyam, R., Krishnan, M.S., 2003. Empirical analysis of CK metrics for object-oriented design complexity: implications for software defects. IEEE Trans. Softw. Eng. 29 (4), 297–310.
Williams, L., 2012. What agile teams think of agile principles. Commun. ACM 55 (4), 71–76.
Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A., 2000. Experimentation in Software Engineering: An Introduction. Kluwer Academic Publishers, Norwell, MA, USA.
Wong, E., 2012. Improving the state of undergraduate software testing education. In: Proceedings of the 2012 Annual Conference of American Society for Engineering Education (ASEE 2012). IEEE, pp. 1–26.
Zhu, H., Hall, P.A.V., May, J.H.R., 1997. Software unit test coverage and adequacy. ACM Comput. Surv. 29 (4), 366–427.

Otavio Lemos received his MSc (2005) and DSc (2009) degrees in Computer Science from the University of São Paulo, Brazil. He currently holds an Assistant Professor position at the Federal University of São Paulo, Brazil, and is also a visiting researcher at the University of California, Irvine (2016). During his academic career, he has published a number of papers in important conference proceedings (such as ICSE 2012 and ISSRE 2015) and journals (such as JSS and IST). His research interests include software reuse, software testing, empirical Software Engineering, and agile development.

Fabio Silveira received his M.Sc. from the Federal University of Rio Grande do Sul in 2001 and his Ph.D. in Computer Science from the Technological Institute of Aeronautics (ITA) in 2007. He was a visiting researcher at the Technischen Universität Berlin from June to August 2009. He is currently an Associate Professor at the Science and Technology Institute of the Federal University of São Paulo (UNIFESP), doing research on object-oriented and aspect-oriented software testing, experimental Software Engineering, agile methods, and metadata.

Fabiano Ferrari received a Bachelor's degree in Informatics (2004) and a Ph.D. degree in Computer Science (2010), both from the University of So Paulo (ICMC/USP). He is now an Assistant Professor at the Computing Department of the Federal University of Sao Carlos (UFSCar). During his Ph.D., he has been a visiting student at Lancaster University/UK. His research interests include: software testing; Object/Aspect Oriented (OO/AO) Programming; Testing OO and AO Software; Adaptive Systems; Systematic Reviews; Experimental Software Engineering and Software Metrics.

Alessandro Garcia is an Associate Professor at the Informatics Department, PUC-Rio, since February 2009. He was a Lecturer at the Computing Department at Lancaster University, from February 2005 to January 2009. Garcia received his PhD in Informatics from the Pontifical Catholic University of Rio de Janeiro. His research mainly focuses on software modularity, software metrics, software architecture, error handling and empirical Software Engineering.