The Effect of Reasoning Strategies on Success in Early Learning of Programming: Lessons Learned from an External Experiment Replication

A. César C. França
Centro de Informática
Universidade Federal
de Pernambuco
(CIn – UFPE)
Recife, PE, Brasil
cesarfranca@gmail.com

Paulo R. M. da Cunha
Centro de Informática
Universidade Federal
de Pernambuco
(CIn – UFPE)
Recife, PE, Brasil
paulo.pi@gmail.com

Fabio Q. B. da Silva
Centro de Informática
Universidade Federal
de Pernambuco
(CIn – UFPE)
Recife, PE, Brasil
fabio@cin.ufpe.br

Background. Literal or theoretical replications are important to evaluate and assess empirical results. However, there are still few replications in software engineering, and fewer external replications, i.e., developed by researchers other than the original ones. **Aim.** This paper discusses the difficulties found and the lessons learned from performing two literal replications of an experiment involving human subjects. **Results.** Our results apparently contradict the conclusions of the original experiment. However, several differences in context made it difficult to achieve valid comparability. **Conclusion.** Experiments involving human subjects should collect and report as many qualitative context information as possible, so the results can be related to the conditions under which the hypothesis were found to be true. Besides, given the difficulties found in this study, literal replication does not seem to be the best strategy for experiments involving human subjects in software engineering.

Reasoning Strategies, Learning of Programming, Replication, Replicability, Lessons Learned

1. INTRODUCTION

The importance of replicating experiments and case studies in Software Engineering has significantly increased. According to Flyvbjerg (2006), a scientific discipline without a large number of thoroughly executed and replicated studies is an ineffective one. However, there are only few studies available in the literature. According to Lung et al. (2008), this dearth of replications is not a specific problem of the Software Engineering field, but common to many other disciplines.

Experiment replication is the repetition of an experiment to double-check its results (Juristo & Vegas, 2009). Multiple replications of an experiment increase the credibility of its results (Shull, et al., 2002). Nevertheless, replication is difficult. Even though reporting guidelines for some types of experiments in Software Engineering are available, the experience of Lung et al. (2008) has shown that many threats to validity of literal replications, especially when human participants are involved, are related not only to incomplete reports, but also to the experiment design and issues which cannot be easily controlled, such as

cost of the replication, variability of human behavior, or observer's tacit knowledge, inasmuch as they are likely to introduce biases to the results of replications (Jedlitschka & Pfahl, 2005). Lung et al. (2008) claim that unwanted (and often unavoidable) variations reduce the value of a replication, because they reduce the reliability of comparisons with the results of the original experiment.

This article describes a literal replication of the experiment studied in Lung et al. (2008). This experiment was replicated at the Universidade Federal de Pernambuco (UFPE) and the difficulties of performing this replication were compared to the ones faced and reported in Lung et al. (2008).

The experiment was first designed by Dehnadi and Bornat in 2006 (Dehnadi & Bornat, 2006). The researchers claimed to have developed a test able to predict which students would succeed in an introductory programming course, by assessing the student's reasoning strategies. Even though that paper was unpublished and strongly criticized, researchers started to replicate the experiment in different countries (Dehnadi, et al., 2009; Lung, et al., 2008). Later, Dehnadi et al. (2009) performed a

meta-analysis combining the results of six replications, and found some correlation to the original experiment results. However, in 2008 Lung et al. (2008) had published the results of a replication of the same experiment. They discussed a set of problems that made the comparison of their results to the original experiment impossible. This raised the contention that other comparisons of replications of this same experiment, included that above cited meta-analysis, might not make sense as well.

This paper presents the results of two literal replications of Dehnadi and Bornat's experiment, and discusses issues related to literal replication with human subjects. We chose to perform a literal, instead of theoretical, replication because we were mainly interested in replication issues in experiments that involve people, and not so much on the theory (although we believe it is an interesting and useful work)

The results of our replications do not corroborate the conclusions of the original experiment, although variations in context threat the validity of the comparison of the results. On the other hand, almost all of the difficulties reported by Lung et al. (2008) were faced in our replications as well.

Our conclusions go in two complementary directions. First, given the difficulties mainly related to isolate and control context variables, experiments involving human subjects should collect and report as many qualitative context information as possible, so the results can be related to the conditions under which the hypothesis were found to be true. Second, following the conclusions of Lung et al. (2008), literal replication does not seem to be the best strategy for experiments involving human subjects.

2. BACKGROUND

2.1 The Original Experiment

Dehnadi's experience of teaching led him to believe that some novices appear to use prior rational mechanisms, distinct from those taught in the initial programming course, to explain program behaviour. Therefore, he designed a test to predict success or failure before students have had any contact with any programming language. In Dehnadi and Bornat (2006), "success" means that a participant passes the course in which he or she is enrolled (Lung, et al., 2008).

The test consists on a set of 11 different anticipated mental models, by which students are likely to interpret assignment instructions (Table 1) and short sequences of assignments (Table 2).

The result of the test is based on the hypothesis that students who answer the test questions using mostly the same anticipated mental model are able to build and use a mental model in a consistent way. Thus, these "consistent" students tend to succeed in initial programming courses, whereas "inconsistent" students tend to fail. Figure 1 shows a sample assignment question.

Table 1: Anticipated mental models of $a=b$

Model	Description	Effect
M1	right to left move	$a \leftarrow b$; $b \leftarrow 0$
M2	right to left copy	$a \leftarrow b$
M3	left to right move	$a \rightarrow b$; $0 \rightarrow a$
M4	left to right copy	$a \rightarrow b$
M5	right to left move and add	$a \leftarrow a+b$; $b \leftarrow 0$
M6	right to left copy and add	$a \leftarrow a+b$
M7	left to right move and add	$a+b \rightarrow b$; $0 \rightarrow a$
M8	left to right copy and add	$a+b \rightarrow b$
M9	no change	
M10	equality	$a=b$
M11	swap	$a \leftrightarrow b$

Table 2: Anticipated mental models of $a=b$; $b=a$

Model	Description
S1	$a=b$; $b=a$ Conventional sequential execution
S2	$a=b \parallel b=a$ Independent assignments, independently reported
S3	$a, b=b, a$ Simultaneous multiple assignment, ignoring effect upon source

It is important to notice that, as shown in Figure 1, only one mental model is bound to the correct answer (CM2). However, the study focuses on consistency instead correctness, and uses the notion of rational misconceptions to support the fact that even though the chosen mental model is wrong, the student can be considered consistent.

<p>1. Read the following statements and tick the correct answer in the front column.</p> <pre>int a = 10; int b = 20; a = b;</pre>	<p>The new values of a and b are:</p> <div style="display: flex; flex-direction: column;"> <div>☐ a = 30 b = 0</div> <div>☐ a = 30 b = 20</div> <div>☐ a = 20 b = 0</div> <div>☐ a = 20 b = 20</div> <div>☐ a = 10 b = 10</div> <div>☐ a = 10 b = 20</div> <div>☐ a = 20 b = 10</div> <div>☐ a = 0 b = 10</div> <div>☐ If none, give the correct values: a = b =</div> </div>
---	---

Figure 1: A Sample Test Question

In order to verify the effectiveness of their test, Dehnadi and Bornat have administered it to 30 students on a further-education programming course at Barnet College and 31 students in the first-year programming course at Middlesex University, at the very beginning of their course and

a second time to the same subjects after the topic had been taught. Dehnadi and Bornat, then, correlated the test results of the first and second administrations and the course pass/fail results of the end-of-course exams. The χ^2 test showed that consistent students (C) tend to succeed, whereas inconsistent (I) students tend to fail (Table 3). Blank (B) refers to those students who did not answer enough questions to be considered consistent or inconsistent.

Table 3: Test and exams results

Test	Pass	Fail	Total
C	22	4	26
I	8	15	23
B	4	6	10
Total	34	25	59
$\chi^2 = 13.944, df = 2, p < 0.001$ Highly significant			

At first, the experiment was strongly criticized because of the presence of several threats to its validity. For instance, no information was recorded about earlier education, programming experience, age or sex. Also, the judgement of consistency was not explicit and repeatable. Later, Dehnadi et al. (2009) improved the research protocol by adding background questions, adding two more anticipated mental models, making the judgement of consistency explicit and repeatable, and identifying 4 different levels of consistency (C₀, C₁, C₂, C₃).

The improved experiment was repeated twice at Middlesex University and also at six more universities, by different experimenters. The results of six of these experiments have supported the hypothesis of the relation between consistency and success on early learning of programming. Moreover, a meta-analysis of these six experiments showed that programming background has only a weak and insignificant relation to the novices' success. It also showed that the success rate of "wrong-model" consistent students is almost as good as the consistent group. The overall result confirms the result of the original experiment, although the details of each sample or casual adaptations in the original procedure were not reported.

Dehnadi et al. (2009) described also two experiments which are, as they call them, apparent refutations. Both experiments found no correlation between consistency and success. Dehnadi et al. (2009) say that some confounding factors, for example lenient examinations, might have affected the results of these replications. On the other hand, Wray (2007) argue that since the objective of an introductory programming course is to teach students and train them to pass the exams, a high pass rate should not be considered as leniency,

and trying to predict which students are likely to succeed or fail may not make much sense.

2.2 University of Toronto's replication

In an attempt to better understand replications, Lung et al. (2008) performed a replication of Dehnadi and Bornat's experiment at the University of Toronto. According to the authors, that experiment was chosen for several reasons: it had generated a considerable buzz on the internet in 2006; several groups around the world had set out to replicate the experiment; the results were interesting; the designed procedure appeared to be sound, but had a number of potential threats to validity; and the experimental materials were readily available from the original experimenters.

Therefore, they opted to perform a literal replication, that is, to perform identical measurements on similar experimental units. Ideally, the only difference between the original experiment and a literal replication would be the set of participants involved. However, even though they did not intentionally modify the procedure to improve it, they faced several inevitable changes, mainly for local circumstances. Table 4 shows the differences and changes to the experimental phase in the University of Toronto's replication.

Table 4: Differences and changes to the experimental phase

Original	Univ. of Toronto's Replication
Participants drawn from courses at Barnet College and Middlesex University	Participants drawn from a course at the University of Toronto.
Prior mathematics experience not explicitly required	All participants had grade 12 mathematics or equivalent.
Some participants were contacted, interviewed, and tested prior to taking their first programming course	Participants were not tested until the second week of classes and were not interviewed at all.
The predictive test was administered twice to each participant	The predictive test was administered once to participants.
The course from which participants were drawn were taught or tutored by an experimenter	None of the participants were taught or tutored by anyone affiliated with the experiment.
Recruitment method unknown.	Participants were given a chance to win \$50 DVD gift certificates
Participants were asked what A-Level courses or equivalents were taken.	Participants were asked what courses were taken to the highest high school level.
Responses by participants were coded using a subjective system.	Responses by participants were coded using an automated tool.
Time used by each participant to complete the test was recorded.	No timing.

In the data analysis phase, some modifications were intentionally made in order to improve the

reliability of the comparison. For example, they developed a Python™ script that calculates the degree of consistency, avoiding any subjective mistake; they excluded students who did not complete the course; and they also considered a different threshold for pass/fail than the actual course results, better matching the Dehnadi and Bornat's pass/fail meaning. Table 5 shows the differences and changes to the analysis phase.

Table 5: Differences and changes to the analysis phase

Original	Univ. of Toronto's Replication
<i>Tests for self-selection not performed/shown.</i>	<i>Tests for self-selection performed</i>
<i>Transformations of data not performed/shown.</i>	<i>Recoding responses of HTML-only experience as having no experience. Exclusion of data of participants who claimed prior programming experience. Removal of data of participants who had no final mark available</i>
<i>No comparisons of data.</i>	<i>Data compared to Dehnadi and Bornat's.</i>
<i>Correlation examined between consistent mental models and passing/failing a course</i>	<i>Correlation examined between consistency and being above/below the median</i>
<i>Blank and inconsistent participants were combined during analysis</i>	<i>Blank participants were not included in analysis.</i>

Table 6: Confounding factors

<i>Participants who were able to attend were taking fewer courses; a lighter workload might mean more time to devote to the course and do better.</i>
<i>It is possible that timing participants affected the outcome, e.g., by inducing time pressure.</i>
<i>The population's mathematical experience may have a material impact on the result</i>
<i>Students may have reacted differently in the experiment and/or exam because their instructor or tutor was running an experiment in which they were involved</i>
<i>Participants may have down-played their experience to avoid (imagined) negative consequences.</i>
<i>The criteria and guidelines for awarding passing and failing marks may differ between academic institutions</i>
<i>Since so few students fell into the failing group at the University of Toronto, it is hard to draw any conclusions about those who pass and those who fail.</i>
<i>Dehnadi may have inadvertently taught the material differently</i>

In University of Toronto's replication, being consistent has no significant correlation with being successful at the end of the course. Moreover, they found no significant difference in average marks between the consistent and inconsistent groups.

That result is expected to weaken the theory. However, Lung et al. (2008) identified a set of relevant confounding factors which could not be treated in the replication, and which can reduce the validity of this comparison. These factors are listed in Table 6.

3. METHODOLOGY

3.1 Goals and Hypothesis

The main objective of this research was to execute two literal replications of the Dehnadi and Bornat's experiment, which means to perform the same experimental design and analysis procedure in a different population. However, some adaptations had to be done to make the experiment feasible at the Universidade Federal de Pernambuco (Federal University of Pernambuco), but these adaptations were carefully planned in order to preserve the validity of the comparison of its results to the original experiment. Additionally, this research aimed to compare the difficulties faced to perform the two replications to the difficulties reported in Lung et al. (2008), in order to better understand some experimental design issues which can risk the replicability of an experiment. The set of hypotheses and variables investigated in this research are specified in Table 7 and Table 8.

3.2 Experiment Design

Two replications were performed, one in the Computer Science Introduction to Programming (CS-IP) course and other in Computer Engineering Introduction to Programming (CE-IP). Moreover, the tables of contents and the examination questions are similar to Dehnadi and Bornat's original experiment. However, while in the CS-IP course the students learn, and also are supposed to not have any previous contact with, Java™ programming language, in the CE-IP course they learn C, which uses a very similar syntax of assignment and sequential instructions.

The authors of the original paper were contacted first to explain some questions, related to experimental design details that were not completely clearly reported in the past publications. Then, Dehnadi sent by e-mail the material and information necessary to run our replication using the same improved procedure of the six replications carried out in Dehnadi et al. (2009). Therefore, the instruments used were exactly the same of Dehnadi and Bornat's improved experiment, with no adaptations. Since the students are required to know English to take CS-IP and CE-IP courses, the test was not translated to Portuguese.

Some of the difficulties reported in Lung et al. (2008) were also faced in our replication, for instance, in order to comply with the University Ethics Committee, our students were asked to sign an *Informed Consent Form* (ICF), explaining that attending the tests would be facultative and confidential. All variations of our adaptation are summarized in Table 9.

Table 7: Variables

Variable	Scale
Consistency	C_0, C_1, C_2, C_3 , not C , Blank
Consistency Mental Model	CM1, CM2, ... CM11
Success	Pass, Fail
Prior Programming Experience	Yes, No

Table 8: Hypotheses

Hypothesis	Scale
H_1	C_0 level of consistency correlates to success
H_2	C_0 - C_3 levels of consistency correlate to success.
H_3	Prior programming experience is not correlated to success.
H_4	CM2 consistency is not correlated to success
H_5	C_0 level of consistency correlates to success even with no CM2 students included

Table 9: Differences and changes to the experimental phase

Original	UFPE's Replication
Participants drawn from courses at Barnet College and Middlesex University	Participants drawn from two courses at the Universidade Federal de Pernambuco (UFPE), Recife, Brasil.
Some participants were contacted, interviewed, and tested prior to taking their first programming course	The participants were neither contacted nor interviewed before the tests.
All participants had their course results known	The participants who did not finished the course were removed from data analysis
Recruitment method unknown	The test covered all enrolled students from both courses
Written tests were administered in the classroom	Written tests were administered in the laboratory
The tests were not administered during classes time	The tests were administered during classes time
The course from which participants were drawn were taught or tutored by an experimenter	None of the participants were taught or tutored by anyone affiliated with the experiment.

Given the Dehnadi and Bornat's test results and the final courses results, the anomalous data (students who took the test, but did not finished the course) were identified and removed from the analysis, and the individual hypotheses were tested using the Fisher exact test, inasmuch as the χ^2 (chi-squared) test is not accurate for a small sample ($N < 40$).

4. RESULTS AND ANALYSIS

4.1 Sample

Even though it was clearly stated that attending the research was facultative, the level of participations could be considered high, covering 38 valid students in the CS-IP course and 35 valid students in the CE-IP course. Tables 10-13 show respectively the distribution of the final course results, consistency levels, prior programming experience, and the CM2 consistency level.

Table 10: Distribution of Final Course Results

	CS-IP	CE-IP
Pass	47%	66%
Fail	53%	34%

Table 11: Distribution of Consistency Levels

	CS-IP	CE-IP
C_0	74%	97%
C_1	5%	0%
C_2	3%	0%
C_3	0%	0%
Not C	16%	3%
Blank	2%	0%

Table 12: Distribution of Prior Programming Experience

	CS-IP	CE-IP
Yes	61%	60%
No	39%	40%

Table 13: Distribution of CM2 Consistency Model

	CS-IP	CE-IP
CM2	87%	91%
Not CM2	13%	9%

The proportion of people belonging to the blank, consistent, and inconsistent groups was significantly different in the two instances of the experiment, so the results are not comparable to the original experiment at all. On the other hand, it also shows that the test is not psychometric, such as in Wray (2007). Another important comparison shows that in almost all past replications, CM2 people are about 90 per cent successful, whereas in our data set they were between 50 per cent (CS-

IP) and 65 per cent (CE-IP), which is also significantly different.

4.2 Hypothesis Testing

In our two set of data, we found no significant correlation between consistency and success. Almost all results were the opposite of Dehnadi and Bornat original experiment, but H_3 , which states that there is no correlation between relevant prior programming experience and success in passing the programming course. Moreover, none of the variables tested showed any correlation.

Table 14 shows the H_1 hypothesis test, from which C_0 consistency level and success are not correlated. Table 15 shows the H_2 hypothesis test, from which $C_0 - C_3$ consistency levels and success are not correlated either. Table 16 shows the H_3 hypothesis test, from which prior programming experience and success also are not correlated. Table 17 shows the H_4 hypothesis test, from which CM2 mental model of consistency and success are not correlated. Finally, Table 18 shows the H_4 hypothesis test, from which “wrong” mental models of consistency (CM2 excluded) and success are not correlated.

Table 14: H_1 hypothesis testing

	CS-IP		CE-IP	
	Pass	Fail	Pass	Fail
C_0	15	13	23	11
Not C_0	3	6	0	1
		$P = 17.80\%$ NOT CORRELATED	$P = 34.29\%$ NOT CORRELATED	

Table 15: H_2 hypothesis testing

	CS-IP		CE-IP	
	Pass	Fail	Pass	Fail
$C_0 - C_3$	17	14	23	11
Not C	1	5	0	1
		$P = 9.00\%$ NOT CORRELATED	$P = 34.29\%$ NOT CORRELATED	

Table 16: H_3 hypothesis testing

		CS-IP		CE-IP	
		Pass	Fail	Pass	Fail
Prior Prog. Exp.	Yes	13	10	13	8
	No	5	10	10	4
		$P = 10.23\%$ NOT CORRELATED		$P = 24.41\%$ NOT CORRELATED	

Table 17: H_4 hypothesis testing

	CS-IP		CE-IP	
	Pass	Fail	Pass	Fail
CM2	14	13	21	11
Not CM2	3	1	2	1
		$P = 30.26\%$ NOT CORRELATED	$P = 46.39\%$ NOT CORRELATED	

Table 18: H_5 hypothesis testing

	CS-IP		CE-IP	
	Pass	Fail	Pass	Fail
C_0 (CM2 excluded)	1	0	2	0
Not C_0	2	1	0	1
		$P = 75.00\%$ NOT CORRELATED	$P = 33.33\%$ NOT CORRELATED	

4.3 Comparison to the University of Toronto's replication

As in Lung et al. (2008), many unintended an unavoidable variations were faced during the planning and executing phases of the replication. Although the adaptations were planned to maintain the comparability of our replications to the original experiment, some variations could not be anticipated. Table 19 shows the differences between the University of Toronto's replication and ours.

Table 19: Differences and changes to the experimental phase

Univ. of Toronto	UFPE's Replication
Participants drawn from a course at the University of Toronto.	Participants drawn from two courses at the Universidade Federal de Pernambuco (UFPE).
All participants had grade 12 mathematics or equivalent.	There is no math pre-requisites to the programming courses in the experiment
Participants were given a chance to win \$50 DVD gift certificates	There was no prize for attending the experiment
Written tests were administered in a classroom	Written tests were administered in the laboratory
The tests were not administered during classes time	The tests were administered during classes time
Responses by participants were coded using an automated tool.	Responses by participants were coded using a subjective system.

Our research data have shown no correlation between consistency and success. However, our experience has had almost all weaknesses described by Lung et al. (2008) in Table 6. Therefore, even though the procedure was planned to be as similar as possible to the original, we agree with Lung et al. (2008) when they say that their results could not strictly be compared to those in Dehnadi and Bornat, so neither does ours.

However, instead of testing several distinct operationalizations in order to find some justification for our results, as in Lung et al. (2008), we have concentrated in strictly following the plan, to avoid both adding comparison problems and fishing results.

4.4 Discussion

Our results clearly show no correlation between consistency and success. However, it does not mean that the theory is wrong or not useful. Our results may be indicating that the theory is not covering all possible variables, though. One possible interpretation is that this test is not psychometric, as discussed in Wray (2007).

Another possible interpretation for our results is that the slight adaptations were sufficient to completely invalidate comparisons of our results to the original experiment. However, according to Juristo and Vegas (2009), when we find that the results of two non-identical replications are inconsistent, we should not always consider the replications to have failed and give up trying to aggregate the results. We should examine, one by one, which experimental conditions have changed and then try to discover new influential variables.

Additionally, Lung et al. (2008) have reported many flaws in the original experiment design which could make it impossible to be literally replicated. Several of these flaws are related to the understanding of the theory. For instance, "success" as defined by Dehnadi and Bornat, means passing in the course examinations, but the course examination is nothing but a set of subjective marking system, completely dependent on the tutor or the institution passing policy.

Three other replications have achieved results similar to ours. In Dehnadi et al. (2009), the theory authors blame the leniency on examinations, which could bias the research results. However, they do not consider that good tutors should be able to prepare their students to pass even in the most difficult examination, so how is leniency on examination distinct from efficacy on teaching?

The Dehnadi and Bornat's results are relevant inasmuch as it shows an existing correlation

between consistency and success. It is important to notice, though, that this existent correlation might not be interpreted as an "effect" of consistency on success. Conversely, it could be interpreted as evidence that something beyond this test may be related to success. Therefore, they may be looking for cognitive models – which are defined as the way people build and use mental models (Lycan, 1999). There are some cognitive theories whose relation to success in learning programming should be indeed broadly investigated.

Moreover, their interpretation of mental models may not be completely right, since mental models should be created during the learning process, instead before learning some subject. As Lung et al. (2008) also say, an alternative interpretation of this theory is that inconsistent people should be more flexible when it comes to learning and thus should have a higher success rate.

At this point, we should argue why we did not relate our results with the other replications available in the literature using meta-analysis, a validity and important tool to increase statistical power of empirical results (Glass, 1977). The main reason is that a known weakness of the method is important in this context: meta-analysis does not control sources of biases or design flaws. In fact, a meta-analysis of a badly designed experiment will not produce good statistics. In this context, we are particularly concerned about the design problems of the original experiment, thus using meta-analysis does not seem to be a good strategy.

4.5 Lessons Learned

In the planning phase of our replications, it was clear that it is difficult to retrieve all details of the experiment procedure from reported papers. According to Shull et al. (2002), the main difficulty is to understand the concepts underlying the techniques under study and to master the knowledge involved in running the experiment. Additionally, transferring the tacit knowledge may be impossible in some cases. The tacit knowledge refers to information that is important to the experiment but is not coded into the lab package.

In our replication, the authors of the original experiment were completely available both to supply tools and information, and also to clarify any doubts eventually found about the overall procedure. This was indeed extremely important to planning these two replications.

The replication design was simple, since we could access all the instruments designed in the original experiment. It was necessary to make some changes to adapt the experiment to the Universidade Federal de Pernambuco's local

characteristics and environment, but without risking too much the comparability of the results (Juristo & Vegas, 2009). The experience reported in Lung et al. (2008) also helped to predict difficulties and variations in the execution.

During all the replication process, we have noticed that several flaws in the original experiment design challenged the replication. The problems could be categorized as follows:

A) Theory understanding: the theory has to be extensively covered to generate replicable experiments. In this specific case, the experiment has problems in the definition of “success” and also in use of “mental models” prior to learning:

- we cannot say whether or not someone has learned to program, but only that some individual has earned a certain grade as assessed by the instructor;
- theory should be adjusted to limit its scope to programming aptitude on tests;
- as a humble suggestion, Dehnadi and Bornat could be looking at cognitive models instead mental models.

B) Hypotheses generation: the hypothesis generation can also be a problematic phase in the research design. In the original paper, the hypotheses are not explicitly declared. The process of generating hypotheses should be extensively covered in the reports. This experiment also had problems related to hypotheses generation:

- The psychometric characteristic of the test should be addressed before raising hypotheses about the relation between consistency and success;
- An alternative interpretation of this theory is that inconsistent people should be more flexible when it comes to learning and thus should have a higher success rate;
- There are several facts that lead students to pass, struggle, or fail in a programming course. Therefore, a qualitative research should have captured these characteristics before, so the quantitative study could have adequately isolated them.

C) Interference: for any type of research, it is fundamental to take good decisions on how to verify the variables and test the pre-defined hypotheses. In their experiment design, Dehnadi and Bornat may not have taken the best decisions towards the experiment replicability:

- Dehnadi may have inadvertently taught the material differently;
- It is possible that timing participants affected the outcome, e.g., by inducing time pressure;

- Students may have reacted differently in the experiment and/or exam because their instructor or tutor was running an experiment in which they were involved.

D) Interpretation of the Results: the experimenter’s expectancy effect may also turn to be a significant problem to make experiments replicable. As discussed before, the initial results of Dehnadi and Bornat’s experiments may have been over-startling, but after some replication attempts and critics, they organized and improved the original design. However, some problems still can be seen in the interpretation of their experiment:

- External validity of the results was not addressed in the original experiment;
- Correlation tested with χ^2 does not clarify the cause-effect relation between consistency and success as described by the authors;
- How is leniency on examination distinct from efficacy on teaching?

One of the most important lessons learned from this replication is that even though an experiment seems to be easy to be replicated, it does not necessarily mean that it is suitable to be replicated. The replication’s experimenter should consider checking the theory understanding, the hypotheses generation process, the interference reliability, the interpretation of the results and the comparability of the experiments results.

According to Juristo and Vegas (2009), it is possible to use differences among replications of Software Engineering experiments to generate knowledge, either by improving the procedure, revealing new confounding factors, changing hypotheses or improving the analyses. However, since improving the original experiment has to be listed as a goal of a replication, it should be clearly stated in what ways the experiment can be improved.

5. FINAL CONSIDERATIONS

This paper presented a literal external replication of an experiment on the effect of reasoning strategies on success in early learning of programming. The original experiment proposed a test which could predict success or failure in introductory programming courses, before students have had any contact with any programming language. The results presented by Dehnadi and Bornat (2006) have motivated several researchers to replicate that experiment. However, while some replications do achieve the same results of the original experiment, others do not. We performed two replications, none of which have confirmed the hypothesis of correlation between consistency

according to the Dehnadi and Bornat's test and the success of students in passing introductory programming courses.

However, even though we decided to perform a literal replication, by repeating the same procedure executed in the original experiment, a set of planned adaptations, unintended variations, and unpredicted environmental variables may have impacted the validity of the comparison between our experiment and the original.

According to Juristo and Vegas (2009), one of the most effective ways to generate knowledge from non-exact replications is by giving suggestions to improve the overall experiment. Therefore, the main contributions of this paper are not only the results of the replicated experiment, but a set of suggestion of improvements to the experiment.

Our conclusions go in two complementary directions. First, given the difficulties mainly related to isolate and control context variables, experiments involving human subjects should collect and report as many qualitative context information as possible, so the results can be related to the conditions under which the hypothesis were found to be true. Second, following the conclusions of Lung et al. (2008), literal replication does not seem to be the best strategy for experiments involving human subjects.

As further steps, we plan (1) to redesign the Dehnadi and Bornat experiment, focusing the study of cognitive models and personality profiles to investigate the relation of these factors to the success in learning and developing programming skills; and (2) perform a systematic literature review on replication of software engineering experiments in order to gain more knowledge on the pros and cons of performing replication of experiments involving human subjects in software engineering.

6. ACKNOWLEDGEMENTS

The authors would like to thank Dehnadi Saeed, one of the authors of the original experiment, for his attention in answering our questions about the original study. Additionally, we would like to thank professor Ricardo Massa, Ph.D and professor Adriano Sarmento, Doctor, for access to their students. Samsung Informatics Development Institute (SIDI) partially funded the participation of the third author in the EASE'2010 Conference. Prof. Fabio Q. B. da Silva holds a research grant from the Brazilian National Research Council (CNPq).

7. REFERENCES

- Caspersen, M. E.; Larsen, K. D.; Bennedsen, J. (2007). Mental models and programming aptitude. In *ITiCSE '07: Proceedings of the 12th annual SIGCSE conference on Innovation and technology in computer science education*, pages 206–210, New York, NY, USA, 2007. ACM.
- Dehnadi, S.; Bornat R. (2006), "The camel has two humps (working title)," In *Little Psychology of Programming Interest Group*, Coventry, UK: 2006, pp. 1-21.
- Dehnadi, S.; Bornat, R.; Adams, R. (2009) "Meta-analysis of the effect of consistency on success in early learning of programming," In *Proceedings of 21st Annual Psychology of Programming Interest Group Conference*, Limerick, Ireland: 2009, p. 10pp.
- Fix, V.; Wiedenbeck, S.; Scholtz, J.(1993) "Mental representations of programs by novices and experts. In *CHI '93: Proceedings of the INTERACT '93 and CHI '93 conference on Human factors*, Amsterdam: 1993, pp. 74-79.
- Flyvbjerg, B. (2006). Five Misunderstandings About Case-Study Research. *Qualitative Inquiry*, 12(2), 219-245. doi: 10.1177/1077800405284363.
- Glass, G. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, 5: 351-379.
- Jedlitschka, A., Pfahl, D. (2005). Reporting guidelines for controlled experiments in software engineering. In *2005 International Symposium on Empirical Software Engineering*, 2005. (pp. 92-101). Ieee. doi: 10.1109/ISESE.2005.1541818.
- Juristo, N.; Vegas, S. (2009) "Using Differences among Replications of Software Engineering Experiments to Gain Knowledge," In *Proceedings of Third ESEM: International Symposium on Empirical Software Engineering and Measurement*, Lake Buena Vista, FL: IEEE Comput. Soc, 2009, pp. 356-366.
- Lung, J.; Aranda, J.; Easterbrook S.M.; Wilson, G.V. (2008) "On the difficulty of replicating human subjects studies in software engineering," In *Proceedings of the 13th international conference on Software engineering - ICSE '08*, New York, New York, USA: ACM Press, 2008, p. 191.
- Lycan, W. G. (1999). Mind and cognition: an anthology. In *Volume 8 of Blackwell philosophy anthologies*. Wiley-Blackwell, 1999, Second Edition. 540pp. ISBN 0631205454, 9780631205456.

Shull, F.; Basili, V.; Carver, J.; Maldonado, J.; Travassos, G.; Mendonca, M.; Fabbri, S.(2002) "Replicating software engineering experiments: addressing the tacit knowledge problem," In *Proceedings International Symposium on Empirical Software Engineering*, IEEE Comput. Soc, 2002, pp. 7-16.

Wray, S. (2007). SQ minus EQ can predict programming aptitude. In *Proceedings of the PPIG 19th Annual Workshop*, Finland, 2007.