Investigating Replication Challenges through Multiple Replications of an Experiment

Coding Synthesis

Notes regarding the participants excerpts quoted:

- The focus group sessions were conducted in Portuguese. All the excerpts are translated to English, but some sentences might feel unnatural for an English speaker. We made our best effort to be as faithful as possible to the original statements.
- When participants say “he” or “him”, they usually refer to the original paper author(s).
- The expression “(UNIN)” refers to unintelligible audio.
- A question mark in parenthesis “(?)” refers to non-clear audio, which we made our best effort to understand.

Notes about codes:

- A list of the codes synthesized here can be seen in the file “code_list.pdf”.
- In the synthesis, a parenthesis with red text (“Red Text”) indicates intersection between the current code described and another code.

Subjects Allocation:

Instance	Group	Participant
1st	A	Odin
		Thor
		Ymir
	B	Saturn
		Mercury
		Helios
	C	Neptune
		Pluto
		Minerva
	D	Vulcan
		Diana
		Bacchus
2nd	E	Athena
		Artemis
		Dionysus
	F	Cronus
		Apollo
		Hestia
	G	Zeus
		Poseidon
		Persephone

1 SETTING UP THE ENVIRONMENT

1.1 C1

1.1.1 Extraction of code.zip file
(Addressed by 3 teams → C, D, E)

Some teams had difficulties extracting code.zip file, which was the file containing the source codes of the projects. That happened because when setting the configuration up of the virtual machine, Virtual Box allocates by default 10 GB as a limit to the virtual Hard Drive Size. However, after installing Ubuntu and extracting the folder, this hard drive's size was not sufficient to store the extracted files. (Pluto: Because when unzipped, it was bigger than the virtual machine's HD.) (Vulcan: Because we tried to run with "code.zip", we tried to download internally, then we tried to uninstall things inside the system to make room and we couldn't.) (VM's Size) So, the extraction operation made the Virtual Machine to freeze for one team, (Minerva: It was the size of the virtual machine. Then when I was going to extract the "code.zip", it crashsed.) and for another made a participant to lose a Virtual Machine (in his words "broke the image"). (Dionysus: I got some size limitations and everything, so it ended up breaking the virtual machine. I missed that one, the image. I had to import the virtual machine from scratch because... I don't know if it was the code [code.zip], something like that, in the set of projects that ended up crashing.

A participant even suggested that the issues with the extractions might have impacted the replication quantitative results, however he wasn't able to explain clearly the reason and he wasn't sure of that either. (Vulcan: Small difference. We don't know exactly what caused it. But it was a small difference. I suppose it could be related to using "code.zip" or not. I suppose. Because the size of the folder that was inside the experiment was less, so it might have a relationship. But I'm not sure.)

One of the teams recurred to Daniel to see if the participants were executing it appropriately. This shows insecurity from the participants, so we can infer that for them the extraction of the projects was not as smooth as one would expect. (Because I got my hands on everything already. I had already tested it. But when it was time to unzip... I had already shown it to you, remember? To see if it was right. On the day I closed everything. Then when it came time to unzip the file showed a problem).(Author's intervention)

This difficulty was regarded by one of the participants as being the most time consuming of the whole replication. (Vulcan: That was, I think, the greater problem that we lost more time in the whole experiment, that was it.)(I4)

1.1.2 Projects Extraction
(Addressed by 6 teams → A, B, C, D, E, F, G)

(I've inspected the source code to see where the algorithms expect the codes to be. According the file BugsChecker.Java lines 132 and 142, the codes should be at the root of eclipse project, in a folder called 'code'. The instructions on the website says 'Then, download the source code of the projects and add the folder code to your project.' It says you must extract the folder to the project but do not specify a specific location.)

After the code.zip was extracted, there was an additional step: placing the projects folder in a location the algorithms could detect and run appropriatelly. One team, composed by seasoned

developers, judged this task to be straightfoward (Thor: And there was the detail that I had to put the code folder ("code.zip") with the project, and it wasn't complicated ...).

Another teams missed the presence of more descriptive instructions; where the folder should be placed or imported to the eclipse project containing the algorithms (Minerva: It also doesn't talk about how to import. The folders are there, it doesn't say how the project mattered. So, where should I put? It doesn't say how it performs. So it was made by…). One team even mentioned a trial-and-error approach until they could find the appropriate location. (Saturn: There's this code ("code.zip") folder, which he sends to download and puts in the project folder... I think he could detail more specifically which folder to play.(…) So, sometimes, I threw it in a location that there was an error... I think Helios also had some error... Then later, as Mercury had managed to do it, he said: "No, it's in this other folder here. ") (Clarity of the Instructions)

*Another issue that made a particular team confused about the projects extraction: on the eclipse project there was a folder called 'bugs', which contained C source code related to the corpus of faults (section 4.1 original paper) needed by the algorithms to perform the samplings operations. The corpus of faults source code was structured by project in a structure resembling the structure found in code.zip content. All the projects on the compressed file had a matching entry in bugs folder. It was possible to run the algorithms without extracting the code.zip, but the results returned by them would obviously not correct. This single team thought that the 'bugs' folder were the actual C projects. After extracting code.zip and seeing the same folder structure made they think that they had do replace the files in bugs folder, one by one. They expressed this confusion in several moments on their Focus Group session, showing negative feelings towards the absence of clarity of instructions related to that part. (And there was a path to this folder where the examples were. And that's it, these examples were running. You wouldn't need to get the "code.zip" to be able to do anything. And then okay, we wanted to run the way he had proposed there, we got the "code.zip". But at no point in his study did he say that he needed to change something. Well then, we analyzed the code, saw where we needed to change it, and then changed it. In some, it worked, and some didn't.) (The question of code.zip too. Whether I would use it or not. Because there he says to use it. But he doesn't say where to use it or how to use it.)(Clarity of the Instructions) Although they adopted a way of proceeding that was not the expected, they managed to find convergent results for most of the algorithms, on the first study, first part. *(Spreadsheet.) (Put in other section or in this one, I dunno: Maybe a better explanation about the bugs folder could led them to understand the distinction between the code.zip projects and the bugs folder. Another solution would be the algorithms return an error message informing that the projects folder was missing.)

Very important → One other team observed an interesting phenomenon: when running the algorithms without extracting code.zip, it came with converging results for number of bugs (while configurations came diverging.) Participants were perplexed since it is expected that the bugs were depending on the analysis of the projects. **They suggested the authors performed a pre-computation of the algorithms so the replication artifacts retrieved the bugs value independently from the C projects source code sampling, which is quite undesirable for a replication since the results of the pre-computation can't be double checked by independent researchers.** (Artemis: Because we looked, even without the codes, and it gave the right amount of errors. But when you looked at the amount of settings, it had different values. Because it kind of already came a pre... pre-filled there an XML ...)

For the second part of the study, another participant related an additional point of confusion: the two projects needed for only that part were already on the set of 24 projects analyzed on the first part. They assumed it was necessary to delete all the other projects in order to make the second study work correctly, which was not necessary. However, this was a unique case and the instructions did not mention anything regarding a deletion. (Minerva: when we downloaded in the first analysis, in the first study, we download the package that comes with all the projects. Including this one. We

erased the others and left only these two.) (Clarity of the Instructions) Anyways, for that team in particular, the instructions weren't clear enough for the second part regarding the analyzed projects. (Clarity of the Instructions)

Another difficulty mentioned regarding the extraction of the projects was the following: one of the teams felt that the instructions were negligent concerning setting up the projects for that second part. The instructions simply do not address this subject. The team members came to different conclusions when reading the instructions (Apollo: Everything is fine. But then for others, for another study, for the second study, he doesn't say anything. So much so that Cronus turned around and said: "hey pal, my data is very different from yours. Why?" And we went to see that that was it: he wasn't playing the code folder inside the other projects to be able to run. And I did it. So I did it because I thought it was intuitive, but not for him.) In the end, as a team, they were able to find out that no setting up was needed since the projects used on the second part are already extracted on the first part. However, they felt the instructions should explicitly state that.(C2 [**C2 is a superset of Clarity of the Instructions]**).

The site says that it was necessary to extract TypeChef but one of the project had the dependency already. That led to some confusion by the participants. (Dionysus: Here is the project. Then we unzipped it. OK. Let's generate the [.]jar. Then, there was a lot of trouble there... Then, at the time, man, we were even having... Like, but trying to run TypeChef to run the projects, passing it as a parameter, you know? Then, "no, it is not necessary to import..." And then we saw that one of the projects was already with TypeChef there).

1.1.3 Package Installation
(Addressed by 5 teams → A, C, E, F, G)

- One team complain that regarding libraries versions and dependencies was not clear (Zeus: And the other, in relation to another part of 'thing', was compatibility of the libraries, in relation to version, dependency between them... That's where I think it wasn't clear.)(Clarity of the Instructions)
- Missing dependency on the tutorial: flex (Dionysus: For example, flex, which we had no way of guessing where to install that dependency.) (Apollo: There was the problem of flex too. Dependencies that didn't work. We had to ask you to be able to solve there.) One of the teams reported that if it wasn't for Daniel's intervention, they would be stuck for longer. (There was the problem with the flex that if you hadn't answered, we'd be stuck... to be able to find out what it is and such...) (Author's intervention)
- Missing dependency on the tutorial: PUMA (although I do not recall it being needed.) (Zeus: The one with the PUMA […] Poseidon: It was, that we had to look for how to install it...)
- One team complained that installing dependencies on MAC caused conflicts with other applications installed on the OS. (Persephone: Other than that, trying to install on my computer, disconfigured(?) my entire environment.) (Other Operating Systems)

There were several complaints about installing the Linux packages needed for run the algorithms. Participants reported three libraries in which they got problems: flex, git and PUMA. Two teams complained that for this part as well the instructions weren't clear and precise enough. One of the teams even mentioned that Daniel's intervention was helpful for allowing them to not get stuck at package installation.

One team mentioned difficulties installing the packages on MAC OS. Although they complained that somethings in the OS became 'disconfigured', they were able to install the dependencies on MAC. (See sections "Other Operating Systems").

1.1.3.1 Undertaker (C4)
(Addressed by 7 teams → A, B, C, D, E, F, G)

- According to the paper, it is used for calculating statement-coverage ("We used statement-coverage as implemented in the Undertaker [52] tool suite.").

- For the constraints parts, they had to customize the tool to run the BusyBox besides the kernel. (Conceptually we can use the original implementation of statement-coverage, as part of Undertaker [52], but the tool is not flexible to handle other projects than Linux. Thus, we use an alternative im)That means that the if you install the version included on the repositories you shouldn't be able to get results for statement-coverage in constraints.")

- Teams that installed from apt-get:

- Odin (Thor: In the source code, there was the package he wanted, so he just made an apt and installed it.)
- They said installing from the sources didn't work (Thor: However, the main program he used to be able to do the tests didn't work at first. So, I had to do some research to understand what the problem was. Then I found out that in... in the source there where the Ubuntu packages are...).
- Helios (Saturn: There was a library that he had installed that he didn't install. By the link he sent, it didn't install. So, I had to search out… Mercury: I ran apt-get. There's that too.)
- They said installing from the sources didn't work (see quote above)
- Minerva (Neptune: This one, oh. Undertaker. Yeah, that I installed normal. I put "sudo apt-get install"... Then it installed. Then the second time I tried to follow as I was there. But then I had difficulty. Then after I installed it normal, I put "test ti" there and it worked normal.)
- The participant said he had difficulties on installing (see quote above) (different from other teams that said that it didn't work. Maybe his intention was saying the same thing than the other teams, but with more polite words.)
- Diana (Vulcan: There was one you had to download. You downloaded a zip, and then you had to run that zip there, and you didn't need any of that. You would apt-get and install it. Did you understand? The instructions didn't add up. We apt-get and it was ok.)
- Perceive how this participant states that the instructions are wrong and installing from the repository was sort of a workaround (which is not true at all) (Vulcan: that its installation instructions on the website did not match the correct installation. But then we managed to get around it.)
- Hestia: (Apollo: Cronus took it and already tried to install the undertaker in another way, and not the form that was being detailed there on the website)

- Did not mention (at least on this part):

- Persephone

- One team mentioned the lack of information about suported OS versions for Undertaker. (Dionysus: In this project it has been saying there that it works on Ubuntu and Undertaker [he meant Debian]. Even so, on Undertaker he didn't make the Ubuntu and Debian versions that run.) (Clarity of the Instructions)

1.1.4 Operating Systems

1.1.4.1 Other Operating Systems

(Addressed by 6 teams → A, B, C, E, F, G)

Not a difficulty, but an interesting fact: Some teams attempted to run on MAC. One of them used a feature called MAC ports, that enables Linux libraries to be run on MAC. (Ymir: "And I installed a feature on the MAC OS OS called MAC Ports, which it ports these Linux libraries to run there. So, after I managed to configure MAC Ports with the dependencies, it ran smoothly") Another team wasn't able. Maybe it's related to how familiar the participant was with MAC or the specific tools such as MAC ports. (Neptune: I tried to install on MAC too; I couldn't. (UNIN). Another team tried to install on MAC with no success either. The participant was unable to install two dependencies on MAC (Which ones?) (Dionysus: I'm trying to get it to work on MAC and it has two dependencies that don't work at all) Yet another one tried (Hestia: he tried to run it on MAC and it was also one of the problems (…) Then I said: "no, it won't run. There are the following dependencies...) Yet another team tried to install on MAC without success, and ended up screwing up their system. (Persephone: And not being able to run on MAC (?). Then I ended up deconfiguring everything, I don't know how. It took a lot of work to be able to configure the entire environment of my project again .) (Lack of Familiarity)(Package Installation)

- One participant said that running the most updated version was a good manner to avoid dependency/versions problems (Ymir: And I ran unlike most computer students, I only work with the most up-to-date versions.)

- One team tried to make it work in Debian, with no success (Why?) (Artemis: Dionysus tested on Debian, it didn't work.). (ER3)(With Debian's/Ubuntu's version explicitly declared, the replication could be facilitated) (Package Installation) This teams had tried to install n on MAC previously.

1.1.5 Problems with Versions
(Addressed by 3 teams → A, E, G)

- One of the participants has experience with installing/configuring Operationg Systems. So, he already has expertise in how to handle dependency issues, like the ones that happened on the study. (Ymir: What I think maybe this is related to the experience that the person has. So, like this: natively, I'm from the support area. So, these problems to configure, little problem of configuration... You already have a pre-defined scheme in your head.) (Lack of Familiarity)

The instructions weren't clear regarding libraries and dependencies between them. (Zeus: And the other, in relation to another part of 'thing', was compatibility of the libraries, in relation to version, dependency between them... That's where I think it wasn't clear.) (Clarity of the Instructions) (Package Installation)

1.1.5.1 Problems with Eclipse
(Addressed by 5 teams → A, B, D, E, F, G)

- One team stated that installing eclipse wasn't a problem at all. (The one with the more experient developers). (Ymir: No, no. With eclipse, all right.)

- The replication site does not specify which eclipse version should be used neither gives a link (Dionysus: It doesn't make a very clear version, if that influences...)(Clarity of the Instructions) (Comments About Original Paper's Artifacts Website)

- Three teams complained that they installed a more recent version of eclipse, and it wasn't compatible with Java 7, which is the one present on Ubuntu 14.04 repositories. They solved that problem by installing Java 8 (Diana: We had a problem with the eclipse too, I think. We downloaded… Vulcan: The eclipse version. We installed a current version of recent… Diana: Because of Java .8. Then, after that, I only used the 7, if I'm not mistaken. We had to uninstall and install another version that took Java 7. | Dionysus: So in version 14, we had a problem with the eclipse that, like, we downloaded the eclipse. Got the latest. "No, this one only works with Java 1.9, 1.8." What version of Java in version 1.7? "Damn, let's get a version of the eclipse now that works on version 1.7.") Once again, there was a team complaining about the absence of eclipse version description on the replication website (Problem with Java)(C2)

- One team reported difficulties in the process of installing eclipse. The issue here was the process itself on Linux, which required downloading a tar zip file and then extracting. It's important to remember that is more common in Linux installing through programs using the distro package manager, but in many occasions you can't get a recent program version using an old version of the distro using only, for instance, apt-get. (Dionysus: "No, click on inst here" Then you click on inst, nothing comes of it. "Damn, let's go there, let's go to the command line there..." Then, unzip the tar, copy it to that folder. Then, ok, you made everything to work. Then when it's going to open... Crash! Well, then, how do you solve it? "Not. You have to open a clean eclipse. Then, when it works, the right button doesn't work, then you: "oh my God". [laughter]. Man, man! "I'm going to kill Daniel.") (Package Installation) (Ubuntu's Version)

1.1.5.2 Problem with Java
(Addressed by 3 teams → A, D, E)

- One participant reported that he installed an eclipse package which contained a wide range of dependencies included, including Java JDK. (Ymir: For eclipse I downloaded full Java. And then some students said (sic): "No. The core version was ok". I said: "Boy... I don't want to be on this issue." I went there and downloaded it complete...) This is the approach he usually takes to minimize dependency problems. (Ymir: So, sometimes you end up installing too many dependencies already to make sure you don't have a problem. As I told you: I installed the entire developer package. I didn't even know, got it? Java CC…)

- Once again, there was the complaint of absence of Java's version required to be installed. (Artemis: If you had the right "install the version of Java", we would install the right version. We wouldn't need to test the 8, go back, test the 7 to see which one would run.)(Clarity of the Instructions)

1.1.5.3 Ubuntu's Version
(Addressed by 2 teams → E, G)

- One team reported the difficulty of finding old working versions of dependencies required to run the experiment. Sometimes it's required to have old versions since they used Ubuntu 14.04 and the newer dependencies are not compatible with older versions of Ubuntu (unless you update a huge range of dependencies which is not the most elegant solution). (Artemis: So, if you have to install an old system, updates always ask for now, current versions... You get a version, [there] when you go to get a plug-in, the plug-in is up to date. Then you go and take the previous one, the previous one is not working.) (Package Installation)

- The same team tried to used Ubuntu 16.04, but they weren't able to set up the environment since some dependencies did not work (which ones?) (Dionysus: We tried to run on Ubutnu version 16, [...], the dependencies no longer work.)

- Daniel intervened indicating them a working OS, which is Ubuntu 14.04. On their point of view, they would lose some time until finding the Ubuntu version capable of running the experiment and their dependencies taking into account their level of experience with OSes. (Artemis: Time is a very crucial thing. I think… Athena: I was going to try it on Windows from now on. I was going to try it on the MAC now; wouldn't work. Then we would talk more here(?). Then I would arrive and find out: "ah, the article is from 2014. Let's test it with (UNIN.)" Then I would test it on Windows, it would take forever, I would test it on Ubuntu, until...) Another team which Daniel assisted judged the intervention to be essential. (Cronus: So it was essential. Because, like, if you said it was Ubuntu 14.04, I don't think we had this problem, but I think Dionysus actually ran with Ubuntu up to date. We... I don't know how it was going to be, but we were going to try to run with Ubuntu up to date and it wouldn't work.) (ER2)

- One team complained that one the replication site there was a description of the OS on which the experiment was originally ran. This team felt that if, if those information were present, it would be easier to recreate the conditions for replications similar to the original. **The team also proposed that it would be to actually have a complete replication environment already set.**(So, if there were versions... "Look, I ran on operating system X, with library X..." So, we would try to rebuild that environment. Or even, he could provide: "look, it's here.")(Clarity of the Instructions) (Replication Artifacts).

2 ALGORITHMS EXECUTION

2.1 E1
(Addressed by 3 teams → E, F, G)

- According 2 teams, passed the difficulties of setting up the environment, the algorithms execution was straightforward, in general. (Dionysus: The execution itself is smooth, it's fast. The big problem we had was setting up the environment. | Apollo: The execution was normal.)

- One team mentioned that there could be a list of files to be executed (in my observations there is the mismatching list for the first part and no list for the third part). The participants complained that there were some files which were not needed to be executed among the ones which were required to be executed. (Zeus: It would also be good for him to show at execution time which files should be executed in the experiment. Because there's a file in the middle that didn't even need to be executed. Persephone: It worked, but we didn't use it for anything.) (Clarity of the Instructions)

- Participants stated that on their perspective the execution was straightforward. It's interesting to notice that on their statements they emphasize that this part in specific was not difficult, in contrast with the phase of Setting up the enviroment (Dionysus:The execution itself is smooth, it's fast. The big problem we had was to define the environment. | Apollo: The execution, no.)

2.1.1 Inconsistencies

2.1.1.1 Name Inconsistencies
(Addressed by 6 teams → A, B, C, E, F, G)

- One team reported that it was easy to identify the algorithms by the output, matching the results described on the paper in terms of algorithms name. (Thor: Because it's also well discriminated when you run the algorithms, there are the results: the right name and value there. Then, looking at the result, looking at the article and looking at the table...).

- There were instances of differences of the algorithms names on the paper and the execution outputs/names on the source code (Saturn: But not all the names there of what they were like (UNIN) were what appeared there in the results | Minerva: There were a number of things that didn't add up. The name of the algorithm in the article was one thing. It was a certain name. On the console there, in the execution, it was another...):

- most-enable-disabled. Five teams reported they implied that all-enabled-disabled were referring to the same algorithm. Some teams admit that they just assumed because the name was similar, and didn't try to reason about the names meaning (Saturn: Then, you have to go and infer. I think that the most-enabled-disabled, there wasn't this most-enabled-disabled there in the result. So, it's the most similar to the one we ended up using. Mercury: All-enabled-disabled | Minerva: Yeah. One example was that we assumed that "all" is the same as "most-enabled-disabled." So, like, we don't even know if it's right. put it like this in the table. It started with the assumption. | Pluto: In the console it said "all". In the spreadsheet it had "most-enabled-disabled" | Artemis: It's the same as the question of the algorithm: There was an "all" and there was a " most". We inferred that it was the same algorithm. So, we did it. But... it's a lot of guesswork, right? I think that's it. | Persephone: Oh, yeah. Yeah. The all-enabled-disabled e the most-enabled-most-disable (she meant most enabled-disabled. We assumed it said the same thing.)

- Another team complained that on the source code, the classes names that needed to be ran were not much intuitive, for instance, a classe called “Combine Two Algorithms”. (Apollo: The class names are a bit intuitive, right? They talk about... There's a class there that is 'combine two algorithms', something like that... So, you can understand more or less there what it is intuitively. But I think there must be one more thing (UNIN.))

- One team composed by experienced researchers recommended consistency between the names on the artifacts an on the paper (Zeus: In short: the nomenclature he has to use in the experiment is the same as the website, the same as the article. So the recommendation is this: use the same terms.)

(Note: Codes intersection verified. There are no codes intercepting ‘**Name Inconsistencies**’)

2.2.2.1.1 E3 (See note below)

(In this code, there are only 5 labeled snippets that are outside of E3, but that are more linked to Confusion Factor. I'll leave the analysis on that part there, because the snippets don't answer the E3 question directly.)

2.1.2 Abnormalities in Execution
(Addressed by 1 team → E)

- Random file commented → See (**Correspondency Between Tutorial and Artifacts)**, last line.

2.1.2.1 Execution Failures
(Addressed by 5 teams → A, B, E, F, G)

- Two teams stated that they attempt to make changes in the source code in order to make non-functioning algorithms to run (each team dealt with a different situation), but they were not able to make their respective algorithms to run correctly. (Helios: We already called(?), he said: I changed, I tried to change, it worked, it didn't work... This happened in general in the activity. | Artemis: The error was in an import there that we couldn't modify there, because it was a bib. Then, we ended up not being able to) (E4)

- Another team, facing a failure in execution, though about contacting the original study authors. (Hestia: Because in some algorithms, they get an error, and I just wanted to ask Apollo: “Apollo, did you get an error? How did it happen?" And, for example, if this was my work, and Apollo said he would also do that, it was to pull my work, “what was the mistake, let's get in touch with the authors.) (Communication with Authors).

- See Error on path
- See E4

- One team complained facing a Null Pointer Exception on algorithm combination classes in Headers. However, I was able to execute the algorithms and all the teams filled that spreadsheets regarding this part. Some teams put NA (because some algorithms weren’t implemented), but no other team reported that. (Apollo: Execution was normal. There, I think there are two that break: there is one that gives a null pointer; the other I don't remember what the error was.)

2.1.2.1.1 Error on path
(Addressed by 4 teams → A, B, E, G)

- On the constraints part, on the algorithms one-enabled one-disabled and statement coverage (Athena: the algorithm I think is the one-disable, one-enabled and statement-coverage), the algorithm contained a path pointing to the paper's first author's computer, instead of inserting a relative path (as one participant suggested). That caused a Null Pointer exception. (Dionysus:but then we had another problem, which was the path. It was "/home/flavio"… | Zeus: Afterwards, he could, to avoid this thing on the way there, he could say: "oh, copy and paste like that," and instead of taking the whole path, he would just take the project, right? This he can do there in Java. Take only the relative path). However, two teams considered this flaw as something normal that can be overlooked during the rush when developing a software (Zeus: One thing: when we're going to implement it: it's funny, right? We implement it starting, putting our first path. Then, later when (UNIN.), we change to the relative path. But it's something that actually goes unnoticed sometimes.) Some teams reported they were not able to correct this problem (Ymir, Helios, Dionysus) and one reported they were successful when fixing that. (Persephone) (Dionysus: And then, when it came to this part of the constraints, these paths, I saw that it had been made available in the zip there, I tried to redirect it, but then it turned out that this specific one didn't work. | DANIEL: It was the path, right? And then it worked out, that you changed the path? Zeus: Yeah, yeah.)

2.1.2.2 Long Wait Time for Some Algorithms
(Addressed by 3 teams → A, D, E)

- Some teams complained about algorithms that took a long time to be executed, namely, statement-coverage on the first experiment and Random, on first and second experiments. For Random algorithm on Global Settings, another team complained this algorithm took too long to execute as well, with this team giving up on the algorithm likewise. (Artemis: There was a row where we left the random of the second study, we left it and went to lunch, came back and the result... But it took so long that we said: "this is not working")

A common problem on those situations is that there was not feedback of the time estimated to run the algorithm or a progress indicator, which make the participants to wonder if the execution was running ok or not. (Diana: I remember I was like: "Wow, I don't know when it's going there." Then Vulcan looked and said: "No, this one goes up to...") (Execution Output) In Random algorithm on the first study there is actual output shown on the console, but it wasn't meaningful enough for the participants. (Diana: We look. I remember that we looked at code... the "random" file too. That was giving an error. Like, we couldn't execute. It never left 1 in the console output…) (Execution Output)

SYNTHESIS:
- Diana: Random study 1 (Spreadsheet: - When running Random it displays the output 1..2..3...) (We looked. I remember that we looked at code... the "random" file too. It was giving an error. Like, a we couldn't execute it. It wouldn't leave the 1 there…)

- Artemis: Random study 2 in global (There was a row where we left the second study random, we left it and went to lunch, came back and the result... But it took so long that we said: "this is not working here") → Team gave up.

2.2 E2
(Addressed by 3 teams → A, D, G)

One team related the experience of a person to the capability of running the experiment, with the imperfect level of description the participants judged this experiment to have. Both experienced and inexperienced researchers will face a problem, but only the experienced one will overcome it. (Vulcan: The experienced person may have the same problem, but he will be able to solve it. The difference is this. That we had problems too, but we turned around and managed to solve) (Lack of Familiarity) We can imply that this teams judges a better to be fundamental for a researcher without familiarity of the subjects and tools related to this paper.

2.3 E4
(Addressed by 4 teams → C, D, E, F)

Some team inspected the source code structure when trying to make the algorithms to work. For them, they needed to match the folder code projects name with bugs folder projects name. So then, they did that. (Vulcan: The code.zip file size is bigger than it is there. But the folders are the same. So, we don't know exactly what's there and what's not there. And what should we do to replace it. There's clearly no way you're going to replace it, right? In this case, we replaced it manually, right? We went there in the code, moved and replaced. A few ran. There were others that required us to manually rename files... folder names) (Projects Extraction) (Clarity of the Instructions). This team didn't try to understand the source code though. (Vulcan: "Oh, trying to understand the algorithm." No. We didn't do any of that. But, for example, those from the folders, we went to analyze which folder he was looking for.) Somehow, they managed to find correct results for experiment 1, although they were not able to run statement-coverage and random for some results for experiment 1 and 2. (SPREADSHEET)

- - Random file commented → See (**Correspondency Between Tutorial and Artifacts)**, last line.

2.4 Execution Output
(Addressed by 4 teams → C, E, F G)

- One team complained that, when running the 'random' algorithm on the first experiment, at first the algorithm took a lot of time to execute. They were unsure about the meaning of the output information, since it showed only '1' on the console. [Later on the algorithm continued to the point where it showed the algorithms results. However there was no information indicating that the algorithms execution was in progress, making the participants to wonder if the algorithm was indeed being run or not] (Diana: We look. I remember that we looked at code... the "random" file too. That was giving an error. Like, we couldn't execute. It never left 1 in the console output…) (Long Wait Time for Some Algorithms)

- Another team complained in a more generic way, that overall there were cases where the algorithms output had a lot of verbose information, where in other cases in showed only the plain result. This was a consistency issue that made they wonder why the algorithms output were different among them (Dionysus: One algorithm (UNIN.) in one way. The other (UNIN) in a completely different way, you know? And then you don't know. Did... It generated a lot of output information there. Is this going to be useful for anything? In the other one, it just generated the specific result there, you know?) For three teams, those cases of high verbosity were an obstacle when extracting the data, since the participants had a hard time understanding what was relevant on the output and what was not (Dionysus: the way he organized the data in the article is completely

different, for example, from the way the system there, his environment displays the information. He only brings there, let's say, the "pretty" information, treated little and such. And in the execution, it's not like that. Things don't talk very well there. | Diana: The second study, it throws a lot of litter. I think for... He should be doing some testing within his code, to see if it was running, to see where it was working... So much so that there are a lot of 'yes' showing up there. But then, from the point of view, for me who just wanted to collect the result, that was horrible. | Zeus: Because in these last executions there, for example, we saw that there were some paths with some configurations that, for sure, are an intermediate thing, which he used to ensure that his final configuration there was the sum of the configurations of the paths. who's running, it’s just disturbing)

3 PAPER INTERPRETATION

3.1 I1
(Addressed by 3 teams → A, E, F)

- One team: overall, the paper was well written, although it was not trivial to understand (Saturn: The paper for me is well written. It explain the concepts, especially for those who have already mastered it, I think it's quiet there. It's not trivial, but it's understandable.)

- One team reported that although the paper was comprehensive and well reported, they felt the level of information for a replication on the paper insufficient (Athena: I think for its purpose, which was to compare the algorithms, it was on point. But I don't think he thought about replication. These things were missing.)

3.2 I2

3.3 I3 (See note below)

(Nothing beyond questions' answers. I wasn't able to understand Helios's comments so I skipped them.)

3.4 Lack of Familiarity
(Addressed by 7 teams → A, B, C, D, E, F, G)

In many opportunities, participants stated that experience played a role on experiment replication.

- One participant (Ymir) reported having a solid experience with MAC. He said that for him using the VM was more difficult than setting up the environment on a native MAC installation.(Ymir:I'm usually a beta tester and such, and I was with the last one I had available there. (UNIN) normally generates some kind of incompatibility, but I already have experience with operating systems; I had no problem configuring it. I had more problem with the virtual machine than with MAC itself.) (Other Operating Systems)(Virtual Machine – Complaints)

- Two participants have experience in software testing: Ymir (statements before) and Thor (Thor: Because I already have some experience in software testing. And I already knew what I was doing (…)).

- The same participant was aware of how to solve resolution problems caused by the absence of VM-Additions. Many other teams complained of that problem and they weren't that familiar with VMs and solving problems with them. (Odin: There were a lot of people complaining about performance too. How to increase the memory to be able to run the… Ymir: I've always worked with a virtual machine. There's always a little extra tool package that installs the drivers.) (VM's Resolution) (Interaction Between Teams)

- One participant suggested: some participants might have had difficulties because some others might have only academic experience. So in practical tasks they might lack skills (Odin: It's because you have different profiles. Sometimes, some students, they are from a totally academic line, got it? (…) Sometimes the guy isn't... he's not familiar with development, he's never actually

needed to install a VM.). However, this is something I mitigated on the study's design when assembling teams.

- Similarly, two teams mentioned some participants might not have familiarity with software programming in practice or not have affinity with that. (Mercury: Because if you have other people... other profiles that the guy doesn't like to see even a line of code. (…) This kind of thing that not everyone works with on a day-to-day basis. | Apollo: There are people who have never worked with development or anything...) Sometimes, they might feel that the instructions are written for a person which had a certain notion of how performing the tasks (Hestia: In fact, at this point in the second study, it's not even making it clear: it's really indicating. Apollo: you take, for example, our team itself, I have a little more knowledge of programming and such. Hestia doesn't have much.)(Clarity of the Instructions)

- Three teams were familiar with the themes addressed on the paper: Ymir (Odin :So, as I had already seen the topic, already(?) I was more familiarized, I found it interesting the approach of you comparing the algorithms to see the feasibility of executing some and not others.), Persephone (Persephone: I work with it. I work with pair-wise. Zeus: The person who showed me this article for the first time was Eduardo. It hadn't been published yet, practically.), and Dionysus (Dionysus: Yeah... Actually, I was already interested in this work. I was even... I had already done the reading before.)

- Another team mentioned that familiarity with themes of the paper play a role: the lack of it (Saturn: The article itself, I wouldn't say is an easy read. There are concepts there that are implied by those in the area.). (Difficulties in Interpreting the Paper)(I2)

- In several occasions, participants mentioned that being able to work with Operating Systems setting up is a must.

- 3.4.1 C3 (See note below)

This question has a speculative nature. It doesn't add much to phenomena explanation.

3.4.2 Understanding about Replications
(Addressed by 4 teams → A, D, E, G)

- One team reported that they were unsure about how much intervention they were allowed to do for their replication still be considered as it. This happened when they had the idea to perform modifications on the source code. They wondered if they could modify or not in order to not 'spoil' the replication. (Mercury: If I got there, looked at the code, and saw how it was doing the calculation in the other case, you could do a similar thing. But then, we were left with the discussion: "Is this within the scope of the replication of the experiment?" And we decided not to. That replicating the experiment would not be taking and changing the code that the guy made and generating it the way we think it is. And then, we ended up not moving on.) One of the participants said they could not modify the source code, because they had to reproduce exact like the original, so no modifications allowed. However, this statement is wrong and shows that he is not aware of replications definitions. There is no such thing like exact replication in SW (CITE), that's why the name is "close replication". This same participant said that due to the fact that he felt the experiment somehow was replicated different, the results obtained are invalid. (Helios: It's not possible to make the comparison... For me, it's not possible. There were also many doubts about how this experiment happened. And I think the experiment (UNIN) was different in some way.) A

similar opinion was given by a participant from another team (Bacchus: I believe it interferes. Because from the moment you cheat the way, you are biased, perhaps, there. Because you do not know...)

- Another good example, by another team.: (Persephone: The focus of the article is not to get the person to replicate from it. It is necessary that the person understands why it is important and what its results are and how it is (UNIN) that can help with the data. The replication has to be, in fact, in the extra material.) (I1)

- Persephone: experience only with internal replications: (Persephone: I had never participated in a replication of anyone, actually. (UNIN.) I had just replicated the study itself. I had never replicated someone else's study. So, it was interesting to see the difficulties, and such... I think that's why...) (ER4)
- Poseidon: seasoned researcher: (Poseidon: I would say, not that I saw much different than what I've seen, so it didn't summed much.)

3.4.3 Skills with Linux
(Addressed by 5 teams → A, C, D, E, G)

- A participant says that except from him, the other people on their team, without expertise with Linux, struggled a bit to set up the environment. He suggested that the paper's author might have had written the instructions with a person which was familiar using Linux (Neptune:Even because, I believe they thought that person would already have mastery of Linux. (UNIN). Everybody started trying. Then I set up the virtual machine environment and shared it with the team. So for me it wasn't so much trouble. But people already had. [They] don't have all that ability with Linux.) (Clarity of the Instructions)

- Another team mentioned that considered that, without having knowledge with Linux they wouldn't be able to reproduce the experiment (Athena: So, if we didn't know anything about Linux, we wouldn't go anywhere.)

3.4.4 Insufficient Academic Background
(Addressed by 1 team → A)

- One team had a perception of some participants from other teams that some participants didn't have a solid academic background. They noticed some participants were not familiar with fundamental programming language concepts (Ymir: But I keep thinking, after the conversation I had that day, there's (?) a bunch of people there who didn't have much knowledge of programming.) One participant from this team had strong opinion about that, saying that everybody with a degree in computing should be aware of fundamental concepts.

3.4.5 ER8
(Addressed by 2 teams → A, G)

- A team, which participated on the discipline of Empirical Software Engineering, who had just seen Controlled Experiments reported that they didn't see the experiment as structured as they seen on the course. However, I assume that the professor might have given a broad overview of the subject, and since those subjects were novel on experiments, they had difficulties to perceive the nuances of a subject they had just learned on a real world experiment or to some information it is there, but not

explicitly highlighted. (Minerva: Not! No. No. No. As the Professor explained, I didn't think so. I didn't see that there...). Furthermore, a team of seasoned developers commented that for conferences, many times the authors have eliminate empirical information that might be considered not that relevant depending on the venue (to save page space); a situation they experienced themselves. (Poseidon: In our case, we know the weight of the conference, we know the limitation and we have experience of works already published in which we had to cut this part, so if [for] the focus of the conference is not so important.)

3.5 Difficulties in Interpreting the Paper
(Addressed by 7 teams → A, B, C, D, E, F, G)

- ***Team Ymir:*** This team inserted '@'as a generic sign on the spreadsheet. To them it means "no result". But they don't specify what happened: for example, the coverage declaration, which the other teams claim took too long, here the participants insert the @ char. For Null Pointer Exception, they do the same thing. They also put this signal on those that should have no results, such as the 5-wise in the study, for example. (Thor: Now, what happened was that some of the results, as you saw in the spreadsheet that I even put a caption on, some results didn't come. Then, I don't know if it's because something was missing, or if it didn't apply to that test case with that algorithm. Then I put an at sign there.)

3.5.1 Confusion Factor
(Addressed by 5 teams → B, C, D, E, F)

** Conceptually related to (E3)

* To me, it's actually desirable that they realize that on the second study samples/file is not present. So they can reason twice and try to dig deeper on the terms and meanings of the concepts related to the results.

- Teams mentioning (five): Helios (Mercury: But some of the algorithms it ran, it just showed the number of settings. Then we were in doubt whether we had to infer...), Minerva (Minerva:But on the other hand, there was also the question, I think in the third study, that this sample size did not come.), Diana (Diana: If it's by file, I think so. The problem is that it is by configuration only. There is no sample.), Dionysus (Dionysus: Then, we saw that not all systems could run those combinations he was proposing.), and Hestia (Apollo: In study 1. Study 2 doesn't even have that.)

- Incited discussions in 1 team (Mercury: Now, maybe interpreting the results and filling out the form was our biggest problem. Because there was a session where it asked for the number of configurations per file, right? But some of the algorithms it ran, it just showed the number of settings. Then we were in doubt whether we had to infer (…) This prompted a lot of discussion on Whatsapp itself. "Do we have to infer, do we have to know, is there another part that has the number of files, then we have to match one thing with the other?") (ER7)

MERGED: Admission (Addressed by 1 team → D, F)
- One team mentioned deduction on the second part, related to sample/file (Diana: In the beginning, it was file/sample [Diana was confused] still. He was right. Vulcan: Then he changed. Then he had some that was another name. It didn't match what we had to put in the spreadsheet. We deduced: "I think this must be it.") However, according to this team spredsheet, they just put configurations on this part. However, they admit they did not care about the accuracy/relevance of the results on this part. (Vulcan: We looked at the table of the original article, sometimes the fields were not the same

ones you were using... So we were always in doubt. So, we put the output he gave there. Not us. We weren't really concerned about whether it was statistically relevant or not.) Another team mentioned that they used the value 'configuration' for 'sample/file' even knowing that it wasn't the same thing, but it was the only number they had at hand (Mercury: In fact, I found the following: I was aware that the information I was putting there was not a sample/file. I was just putting in number of settings. (...) But, as Saturn said, as we only had this result, we decided to put this result on the form, because it was what we had been able to see.) (Interpretation Effort)

- Teams mentioned non-existing algorithms for the second part and for certain combinations (Dionysus: Then, we saw that not all systems, it was possible for you to run those combinations that he was proposing.)

3.5.2 Attempt to Calculate the Confusion Factor
(Addressed by 3 teams → B, E, F)

- One team started to calculate the values, but later they gave up because they thought this activity was out of the scope of replication. (Mercury: Right. In fact, I even talked to Saturn and Helios at the time it was possible to extract this information from the algorithm. If I got there, entered the code, and saw how it was doing the calculation in the other case, you could do a similar thing. But then, we were left with the discussion: "Is this within the scope of the replication of the experiment?" And we decided not to.) (Understanding about Replications) (Hypot. Design Activity) (E4)

- One team attempted to reverse engineer to find out how the original authors performed the calculus. They found out the algorithms use the number of 50078 files. (Athena: So, for the first part, we took his calculation, reverse-engineered it and found out. There were 50078 files.) They complained, that for the second study this number wasn't applicable. (However, this makes all sense, since on the second study there are the header files, configuration files, so it's natural the number will be different.) They didn't tell how they came to that conclusion [that the total number of files was different on the second study]. In their opinion, the total number of should have been listed on the paper (Athena: I think the problem is not that he does it differently or prints differently. The problem was that he didn't say the total file quantity. In the article, somewhere, he would have to tell us to do this calculation).

- Another team tried to reverse-engineer to find a way to calculate the possible results, but they were n't able to get close to paper's results (I imagine they tried to 'calibrate' their calculi on the first study which had the results for sample/file. But they do not mention anything about it). (Apollo: I tried to do some crazy calculations there with some information in the paper, with file numbers. (...) ...number of settings it gives there, I added it up, tried to divide, inverted, did a lot of things, but I couldn't get close to the result of the paper.)

3.6 Convergence of Results
(Addressed by 2teams → C, F)

- One team: In a broader sense, yes (Pluto: Most of them, yes. Neptune: As we said, there are some points (UNIN), few points. Minerva: Yes, but most of them do.)

- Another team considered that since they were not able to execute everything and the time for analyzing the results was insufficient, the replication was not performed. (For me, the result is binary: "man, I couldn't replicate it." For me, replicating is being able to find the result of the table

and be able to answer your question. As I can't, for me I didn't replicate the results.) (Difficulties in Interpreting the Paper) (Subjects Underestimating Tasks)

3.6.1 I4
(Addressed by 1 team → C)

One team was confused seeing 0 on statement-coverage, and they wondered if 0 had the same meaning as the dash (which the paper puts for the algorithms that do not scale.) In my opinion, since the paper states clearly that some algorithms do not scale, they shouldn't be that impressed of seeing a 0 or an invalid result on the algorithms which clearly do not scale (Minerva: Right. There was also a thing about the dash. And there he puts the dash. The dash we didn't understand if it was 0 or why it didn't appear. In our case, there was one that was 0; in the article was dash. There was one in our case that nothing appeared; was there the dash. We don't know if the dash…) (Difficulties in Interpreting the Paper)

3.6.2 Divergence on the Extraction of code.zip
(Addressed by 1 team → E)

- One team noticed a diversion on the results regarding algorithms combinations results, that in most of the cases, for the first study, it showed the number of bugs with one more bug than what is on the paper results. (Dionysus: When we unzipped, she unzipped and created another code folder inside code. Then he was showing one more number. Artemis: One more number in all.) They impute that to an additional parent 'code' folder being created after extracting code.zip, and this additional folder make all algorithms number to increase by 1. They said that after deleting this additional folder, they were able to get the exact result that was on the paper (Dionysus: Then, after we resolved, we removed this inconsistency, then we started to give the right number). However, their spreadsheet do not contain the same number as the paper; and for algorithms combination C2 and C4 they still reported the bugs number to be 1 more than the paper, which make me wonder if the explanation they gave was true. I ran the algorithms as well, got the different results, but their solution didn't solve my problem as well. (Projects Extraction)

3.7 Execution Variations due Lack of Details
(Addressed by 1 team → B)

- One team addressed that the lack of enough details (like versions, missing steps) might lead to unintended variations on the execution. (Mercury: If the intention was for everyone to replicate it, do everything the same, then I think the script is really flawed. But I think it doesn't stop they from doing it. You can do it. Now, there will be variances. Exactly. You'll have people who can do it, installing a different tool, with a different version, another one here...) (Clarity of the Instructions)

4 REPLICATION EXPERIENCE

4.1 ER1
(Addressed by 3 teams → A, B, D)

- How to properly fill out the form. (Ymir: No, I had that initial difficulty, which I even asked for... so I would know exactly how to fill it out, right? I didn't have... In the entire experiment, you didn't have that information. After Thor filled in there I understood. But, like this, my question was more about the intention of not making mistakes in filling it out exactly, so as not to have rework. The idea, basically, is this one.) However, I explained them the activity. Maybe they could have been more clear in which part wasn't clear for them or if they just forgot.

- I could have explained better that the VM was a clear Ubuntu 14.04 installation, with no extra setting or nothing prepared by default. It was just like using a local machine. (Mercury: Because, at first, I had thought that that machine had something that we were going to need from that machine. After the information came that it was a clean machine, "oh, so I'll do it on mine right here". Then, when I went to do it in mine, I didn't have any more difficulties.) (Migration to the Native OS Installation)
(Vulcan: Right. Because you delivered a clean virtual machine, right? Only with Ubuntu installed. If we knew that we could use it this way, we would have created a virtual machine with Ubuntu, of the necessary size, we would have done everything)

- Maybe I should have been even more clear about the objective (it's on the slide 16, but somehow it was not clear to them. Or I could have said to refer to the slide.) (Helios: Unless that was the goal, right? Because I still wonder if it was the algorithms or something else. (UNIN)) (Study's Objective was not Clear) (Difficulties in Interpreting the Paper)

- One of the participant perceives that if we had more time for the whole experimental study, it could have been explored applied in a better manner

4.2 Author's intervention (ER2)
(Addressed by 6 teams → A, C, D, E, F, G)

- Interventions: extracting code.zip, installing dependencies Undertaker (Neptune: Then we stopped there and after receiving your help...), and 'flex' (two teams), which I didn't explained previously and it's not described on the website (Dionysus: Dependency problems, which were not described in the environment, that we had to resort to you. For example, flex, which we had no way of guessing where to install that dependency. | Apollo: There was the problem with the flex too. Dependencies that didn't work. We had to ask you to be able to progress there.)

- One team was trying to replicate Study 2 on the second part (despite the fact I told them to not do it on the training session.) So I had to explicitly tell them to skit part 2 after they were already performing the replication (I honestly don't remember in which circumstance that happened.) (Ymir:I think maybe the fact that we, before you said "part 2, no", we were trying to make everything straight.)

- One member from another team has reached me many times in order to make sure his team was performing the replication correctly. However, I tried to make interventions as few as possible, in order to minimize a possible bias coming from me. Another team joked about this attitude of

somehow hiding the information (but I reiterate that I was avoiding creating bias on the participants, that's why I had this attitude.) (Poseidon: (laughs) You looked like the Dungeon Master.)

4.3 ER3
(Addressed by 6 teams → A, B, C, D, E, F)

*I'm not inserting here suggestions that come motivated by sheer inexperience. (e.g. "Minerva: Not! No. No. No. As the Professor explained, I didn't think so. I didn't see that there...") This participant was having contact with designing experiments by the first time. You can't expect papers in the real world to be scored and distributed as an example from a textbook. That is way too naive. But maybe for replications, having a structure like that could be interesting. (Pluto: It would make understanding more clear.)

- In general, there was the feeling that some minor details missing, both from the paper and the replication site. According to the teams, those kind of information is essential on a replication (Mercury: If he wanted it to be a more scripted business like that, it could be more detailed.)
One team mentioned that this may help researchers with different backgrounds to perform the replication more uniformly (Apollo: So, if you are providing a study replication package, you have to detail as much as possible, precisely to be able to get these guys with both expertise, with expert; how much with the rookie guy.) → **VERY COOL.**

- 2 teams: A description of the machines which the algorithms were run originally (Thor: I think there was a lack of detail about which environment he necessarily ran.), including the OS (Athena: But I don't even remember him saying that he Ubuntu… | Apollo: I think the big problem was not having a minimum configuration requirement to run, like "oh, you guys need a 20 machine GB of HD with at least as much RAM"...). One seasoned participant from another team mentioned that, although it is not common to have this kind of information in SE papers, it would be very handy for a replication (so they could configure the VM with the correct setting), at least on the site with the artifacts (a.k.a replication package) Dionysus: But for replication, I think it would make all the difference. Because then you: "Wow, I'm going to create a virtual machine with these minimum specifications here that he defined, for hardware and operating system." Which would make life easier for the replicators there.)

- The process of extracting code.zip, end to end (Saturn: It has this folder ("code.zip"), which its sends to download and place in the project folder... I think he could detail more specifically which folder to place.)

- Version of Java, eclipse, libraries, OS(as stated above) (Vulcan: I think not. Because I think everyone knows how to install Java, everyone knows how to install eclipse. The question is which version you are going to use. Get it? | Apollo: I could also have defined this: which versions of the libraries they needed, the programs they needed, the system...) (many other statements)(Problem with Java) (Problems with Eclipse)

- They forgot to mention the installation of library flex, which was required (Athena: it should say that we had to install flex.)

- Although the calculus is explained on the paper, the total number of files is never mentioned for study 1 and 3. (Athena: From the table too, it mentions configuration by file. But it doesn't mention the amount of files at any time.)

- Mention the estimated time for each task on the experiment. (Hestia: One thing I think would help in this regard: adding the task description of the environment assembly requirements was just if I had the time: “look, you will need these requirements, on average, together(?) X time to be able to assemble such an environment.")

4.3.1 Clarity of the Instructions
(Addressed by 6 teams → B, C, D, E, F, G)

(Artemis: It's common for him. For him it is something trivial. So, it passes in a way that something is missing, a few bits [that] for us to understand is [be] easier. But why for him, this is so obvious!)

In general, there was the feeling that the instructions were not complete enough for a replication. The participants praised the fact there is a supplementary site, but many information is missing. (see sections intersecting with this and ER3)

- Instructions are not in a step-by-step format (Minerva: Despite not being step by step on the site...) (Comments About Original Paper’s Artifacts Website)

- Contrasting with the most statements regarding clarity of the instructions, one participant said that overall the instructions were clear on the website (see his background later) (Vulcan: In terms of installing the environment itself, it was clear what the instructions said there on the site.) (Comments About Original Paper’s Artifacts Website)

- Two teams stated that it’s important for you to be precise when giving instructions about the replications, regardless of the background you assume the your audience you have. You have to give step-by-step description no matter the person that will replicate is skilled or not for the sake of uniformity (Diana: And as much as people know, he has to explain the step by step correctly. Independent, I think. (…) That something(?) will be replicated in the same way | Apollo: So, if you're providing a study replication package, you have to detail as much as possible, just to be able to get these guys with both expertise: the expert and the novice guy.) (Lack of Familiarity)

- One participant suggests that sometimes the lack of clarity might happen due to space limitations on the paper, since it’s only 10-page long (Dionysus: Another thing that I think it impacts there is the issue of space limitations. And then, I didn't get to check it, but it already has an IEEE Transactions journal, there are other articles by it, from the group, which bring this analysis with more information). However, another participant arguments that the information you can’t put on the paper you can place it on the supplementary website, so space wouldn’t be an issue (Apollo: But I think the space issue is less. Because the space you can play in the complementary material, make it available there... Great.)

- In one occasion, one team couldn’t figure out the problems in setting up the enviroment due to the lack of descriptiveness, which required Daniel’s intervention to solve (Hestia: But it's alright. This isn't even setting up the environment. There was no description there. We spent three afternoons trying to identify the problems in the environment…)

4.3.1.1 Correspondency Between Tutorial and Artifacts
(Addressed by 3 teams → C, D, E)

- Three teams complained about the following phenomenon: on the replication website, there are instruction indicating to run the files BugChecker.java, Bug2CombinationsChecker.java,and

Bug3CombinationsChecker,java. However, the algortihms random and statement-coverage were not in those files. There were additional java files to be ran in order to execute those algorithms, but they were not listed on the tutorial. (Minerva: And there on the website, they didn't say that it was to execute this file that was there separately. We didn't execute it. And there in the article, we didn't have the result. So we assumed that that algorithm that we didn't execute could be executed (?), but in the tutorial that followed there wasn't. (Here they refer to step 3 of study 1, which actually instructs the execution of only BugCheckers.java) | Diana: It was. So we ran it, it ran five algorithms. Only one it didn't run; we had to go there in a separate file and run that file itself. That wasn't in the experiment. We went there and ran that algorithm, got it? I mean, that file. | Athena: It could. Because, like, it created... All was in a single file. The statement and the random were in a single file. Which we discovered because we read the name of the class) (Clarity of the Instructions) (ER3)

- One team pointed a discrepancy between the zip file name and the extracted folder name (build system model) (Minerva: There was also a question that it asks a file to be downloaded that we unzip; he gets another name. Remember? We kept looking. I know that in the first one he was with build system model.) [SamplingBuildSystem → SamplingConfigurationModel] This caused some confusion to that team, which later they were able to figure out what was the problem.

*****IMPORTANT**
- One team reported that there was a code responsible for running 'random' algorithm in Sampling Constraint was commented. (I later investigated it was in SamplingConstraints/ BugCheckerStmtCoverageRandomSampling). The team wondered if they should execute the algorithm as it was or if it should uncomment the commented code and obtain the results from it. (Athena: And inside random was commented, and the other one was... I think it's okay to run both; I don't know why he commented on random. [the two it refers to are statement-coverage and random]) (Abnormalities in Execution)(E4)
Inspecting the spreadsheets, only two teams (Diana/Dionysus) were able to execute this one and get results. That means in the end Dionysus team actually uncommented random. Probably the other teams did not have the initiative to uncomment or didn't see the random in ConstraintSampling. I really consider it was a huge flaw from the researchers to leave such important code commented. Their bad!
- Remember that for this part, first the file BugCheckerStmtCoverageRandomSampling cannot be executed because of the 'Error on path'. After you solve that, you have a file with Statement Coverage and a Random algorithm commented. If you wanna see them both, you have to uncomment Random.

4.4 ER4
(Addressed by 1 team → F)

- One participant states that it was the first time he wasn't participating in a study where participants performed a replication, where one observe other researchers replicating. He even refers to the exploratory study as a case study, due the novelty of the study methodology. (Cronus: Also because it was a different study: a case study, what are you doing. So, you have to follow people's work, but it's not necessarily work that we do... It's not that exploratory case study: you're not watching the guy working there. So, you take it, put us to work... Yeah. It's a little confusing. First time I'm participating in a study like this.)

4.5 ER5
(Addressed by 1 team → E)

- One participant, which work in research with HCS, would like to have the output of the products generated. Thus, this study could be more meaningful to him and maybe could have given insights or his research (Yeah... Actually, I was already interested in this work. I was even... I had already done the reading, and everything else, to see if it had any use for my research. (...) One of the things that I think was missing, at least for me, is because it has the output of the generated products. (...) Which for me is what would be useful.) (Motivation)

4.6 ER6 (See note below)

Conceptually linked to Time Availability.

4.7 ER7
(Addressed by 6 teams → A, B, C, D, F, G)

Separated by teams:

TEAM Ymir: Thor executed the algorithms and filled the spreadsheet. (Ymir: Because I didn't execute. Who performed it all was Thor. When I put the light(?), I was still organizing the thing, which I went to look at, it had already filled out.)

TEAM Helios: The work was well-distributed and collaborative (Helios: And one thing that happened, I don't know if I have to talk about it now or later, is that it was very interesting to work with my colleagues. Because one person did one thing, another person did another...) They made use of Whatsapp for task coordination and communication among the members for clarifying questions about the paper. (Saturn: We even sent it via Whatsapp: "Hey, did the guy mean this right here?" This at the time we were running the algorithms, to see if it was that sequence, or how to interpret the data that was coming out to fill in the form later.)

TEAM Minerva: They made use of Whatsapp for communication among the members for clarifying questions about the paper.

TEAM Diana: Team worked together, in meetings. (Diana: And then it was also complicated because we had to meet and do it only in his, when the three of us were there.)

TEAM Hestia: Team split efforts, trying to do the same tasks in parallel. (Cronus: Then I got it one way, Hestia was trying to run it another way.), although they were allowed to exchange artifacts among them (Apollo: And as I saw that they were both having a lot of trouble to be able to set up the environment, then we found this solution: to make a clone of the machine to share with them.) Teams recognize one member had more expertise, in this case, Apollo (Hestia: But the best answer would be that he maybe, or in addition, good would be Apollo's vision. Because he has greater expertise regarding…)

TEAM Persephone: This team performed the replication remotely: Zeus was in Brazil, Persephone was on the US and Poseidon on GB.
Only Zeus ran on Linux. (Persephone: Only Zeus did it. He has to speak.)

They tried to do togehter, sharing the screen through SKYPE. (Persephone: We spent three hours on Skype, without success.) It seems they interacted through SKYPE in several occasions, even when only two members were available. (Persephone: The first research question I made with Poseidon)

4.8 Time Availability
(Addressed by 4 teams → C, D, E, F)

- One team mentioned they were at the end of the semester, so they were overwhelmed with a lot of activities from many courses. (Vulcan: The end of the semester came and we are overloaded with a lot of things...) (ER6)

- Another team wished the activity was not comprised within the course, so maybe they would have more time two do, e. g. two months instead of two weeks (Pluto: Or even for the time of the discipline. We had a tighter time. So, maybe if it wasn't within the discipline, we could, instead of two weeks, do it in two months, for example.)

- Two teams mentioned they did not have time to dig deep on the source code (Vulcan: Then, it wasn't a deep source code reading. It wasn't trying to understand the algorithm. We didn't even have time for that.) (Hestia: No. I felt like it when I saw the error, but then I didn't have time left.) (E4).

- One team said they started the task as soon as I launched the activity. In their opinion, they believe that because they have started so soon it was essential for them to be able to finish the task. (Athena: We started right away. As soon as the task was announced (…) I think that's why there was time.) They spend long hours and even skipped lunch in an occasion. (Dionysus: And then, every day we tried to take at least a period of time, in the morning, to stretch out, to... There was a day when the girls even went without lunch...)

- INTERESTING ---> This same team mentioned they at the beginning they were too optimistic, estimating that the replication would take less time. (Dionysus: I think we were optimistic. [Thinking] that we would do it and finish it. But I'm glad we started right away. Because if we left it to the last minute...) Although other teams do not mention it, that might have happened to other teams. That would explain a lot of misunderstands and participants attitudes towards the study (even rushing the replication when seeing that the time would not be enough since they started late).

- There was also the complaint that the time was not enough to make a deep analysis/interpretation on the results. According to them, setting up the environment took too much time ffrom them, so little time was left for analyzing the results. (Hestia: Regarding the results analysis, I wanted to deepen the understanding, which was the last table (UNIN.) And I didn't feel that, so we had time available just to be able to see this.) (Difficulties in Interpreting the Paper)

5 CODES NOT MATCHING OTHER CATEGORIES

5.1 Comments About Original Paper's Artifacts Website
(Addressed by 5 teams → B, C, D, E, G)

- The overall feeling is that the website was not detailed enough (Mercury: It's not even from the article. It's the website…), except for one participant (Vulcan: In terms of installing the environment itself, it was clear what the instructions said there on the site.) (Clarity of the Instructions)
Some stated that this is inexcusable since you have 'unlimited' space on the supplementary website. (I already addressed that in ER3) (Zeus: But look: if he didn't have space to put in the article, let him do it on the website.)

- 3 teams: The website contains the dependencies for running the algorithms, but do not indicate versions nor include links for the necessary programs/libraries (Saturn: The information that is on that page there. It doesn't make a very clear version, if that influences. (…) I don't even remember there if it has a link for you to download the eclipse. | Vulcan: He, on the site itself, he didn't put what the versions were. He said that we had to run in the eclipse, he didn't say exactly which version of Java. | Athena: And on the website he could also have put this information later, and he didn't do it either.) (Problems with Eclipse) (Problem with Java) (Ubuntu's Version)

- Although the website it's not clear enough in terms of details (from the participant viewpoint), with some effort it's possible to perform the tasks (Saturn: So, the information on the site is not very clear that way. You can do it, but you can manage a lot. I think it leaves a lot in the hands of those who are here…)

- Not step-by-step (Minerva: Despite not being step by step on the site...) (I've already addressed that in (Clarity of the Instructions))

- The fact that it was not that well detailed was a demotivating issue (Minerva: when we reached in the part of replicating the study, the website was poorly explained, which also discouraged us a lot more.) (Motivation)

- The site is stored at a university's server. This makes it easier to go offline than storing in a source code repository like SourceForge or GitHub. (Zeus: That's one thing, and the other is that I think he should put the artifacts and things he used in SourceForge or a Github repository. Not to be dependent on the university. Because his website is at the university.) (Suggestions) (Replication Artifacts)

5.2 Original Papers' Link Tilde Character
(Addressed by 2 teams → A, B)

- In the paper, there was a link to the experiment website containing the artifacts for replication. However, the tilde operator character(~) was incorrectly replaced by the small tilde operator (˜). Therefore, whenever the participants tried to access the link, they obviously were not able to reach the website. (Ymir: Just the tilde. Just the tilde, to access the site... That character that (UNIN)... | Saturn: I think that on the auxiliary site that is linked in the article, by the way, the link doesn't work...)

5.3 Replication Artifacts
(Addressed by 3 teams → E, F, G)

- One participant mentioned that the author didn't think about artifact deprecation. (Athena: So maybe it worked on Debian at that time and the machine must have been from another version that didn't work and had some version that did. I think what was most lacking was for him to think about the future: if someone uses this in a while, other than now, what are the problems they will have?)

- One of the participants said that the execution of the algorithms is not transparent enough regarding what happens between the input and the output. In his view, this is quite important for a replication (Dionysus: With regard to learning, I think, so it was clear that the way I have to make information available to the community, I think I need to take that a little more into account to allow for replication. (...) And then, people don't describe: how do I get this input and how do I generate this output.) (Criticism to Research Field)

- Three participants suggested that the authors could have created a VM (and upload in Docker as two participants mentioned) with all the dependencies set. Thus one interested researcher could simply access this VM and run the experiment (Dionysus: This question of the Undertaker's dependencies and everything else, if it took it to the Web, let's say: "ah, there's nowhere to put a virtual machine." But that question with Docker: it could create an image of his project on his machine that is working, he created an image for Docker, uploaded it to a server, and it would be more agile for people to replicate. | Hestia: It's also because, to provide a virtual machine with the environment, there are many GB, but if he had made it available in a ready-made virtual machine: "here is the environment for you to just 'play'", then it minimizes immensely. Because virtual machine is not a new subject(?).)

5.4 Communication with Authors
(Addressed by 1 team → A, C, E, F, G)

- One team expressed the opinion that the lack of access to the authors might cause problems (Ymir: Just because you don't have access to the original researcher. Because you could talk, exchange some ideas, I don't know... This non-availability may cause a problem.)

- Three team spontaneously mentioned the desire to mail the authors in order to clarify some questions or things they do not found clear (Neptune: Then I would send him an email and all, to clear up any doubts… | Hestia: But I thought about sending an email to Márcio [original paper's Author]: "how did you run?") . (although one participant from this teams was not comfortable with that idea and would hesitate)(Minerva: But there are things that it's even annoying to send an email. You will send an email to the author asking how to set up the environment there… How to unzip…). One participant knows the first author in person (Dionysus: I know Flávio [original paper's Author], . And I even considered the possibility of contacting Flávio, got it?) Another worked under Kästner's [original paper's Author] supervision in the past (Persephone).

- One participant gave the suggestion that the supplementary site could have a 'comment' section to facilitate the communication between the replicating researchers and the paper's authors. (Poseidon: I, in addition to Docker, as Zeus said, I don't think I've seen this on any replication page, data availability, it's you having a part of comments. Because maybe it will be very difficult for us now to pick up and go back there and send the email to Flávio: "look; had this and this point." And there, we could comment.) (Suggestions) (Comments About Original Paper's Artifacts Website)

5.5 Comments About Other Replication
(Addressed by 1 teams → D, E, F)

- One of the participants said that replicating this experiment was useful experience since she would perform other replications, so facing the problems would give her transferable knowledge to use on her further replications. (Diana: I really needed to know how it worked. And then, in a way, I knew that he, the guy couldn't have explained how this guy worked(?); I had never replicated an experiment. So, just for the experience of replicating a study, it was at least interesting.) (Motivation)

- One other replicated a simmilar study, which the participant had to ran algorithms. He praised that the other study gave the VM and guidelines, with dependencies and version to run the study. Naturally, this gave background to give comments to the replication of "A comparison of 10 sampling algorithms..." (Dionysus: There was an article I reviewed that people made the virtual machine available, got it? "Look, here is the virtual machine, you have to open the eclipse, and you have to run the classes such". In addition to giving the virtual machine, he gave a guideline of the steps you had to follow in order to replicate that study, got it?) He mentioned that after this other study replication he is aware of the need to be precise and comprehensive when elaborating the replication package, giving the artifacts in an ease-to-use fashion and being clear about what happens in the execution as well (Dionysus: So, my goal now is in the next studies, which is what I'm working on now, is to make the information available to the staff in a way that is possible, to make the replication easier.)
