Setting up the Environment

- C1
  In the environment setting phase, which were the major difficulties?
- - Extraction of code.zip file
  The 24 C projects to be run under the sampling algorithms were compressed in a single .zip file, available in the experiment website. After extracted, the projects occupy plus than 2 GB on the hard disk. This code(speaking in terms of research analysis) is labeled on the statements that participants said about unzipping of code.zip.
- - Projects Extraction
  This code tags occurrencies of when participants had problems or were unsure about the process of extracting the C/C++ projects to the appropriate location and make the algorithms to analyze them correctly.
- - Package Installation
  The Java project with the sampling algorithms depends on a number of dependencies installed on the underlying Operating System. The code 'package instalation' represents the statements regarding those dependencies packages.
- -- Undertaker (C4)
  Among several dependencies needed for the replication to run, Undertaker is a noticeable one, often troublesome. This codes maps the affirmations that participants said regarding this specific component.
- - Operating Systems
  This code aggregates categories related with phenomena that involve directly Operating Systems.
- -- Other Operating Systems
  In the instructions about how to run the replication, based on the experiment's website, we can assume that you should run the Java project under a Debian-based Linux distribution (since it describes the apt-get commands that one should run to install the dependencies). However, some participants might attempt to run the replication in a Operrating System other than Linux. This codes matches the statements describing this phenomenon.
- - Problems with Versions
  Category that aggregates codes reporting problems with version conflicts.
- -- Problems with Eclipse
  This code is used for tagging any issues related to the eclipse IDE, from instalation to problems concerning execution including versions conflicts.
- -- Problem with Java
  This code tags problems with Java Runtime Environmnent and Java Developent Kit versions
- -- Ubuntu's Version
  Ubuntu Software Repositories work in a way that, after a distribuition version is released, any software on the repositories is not updated anymore (except for a limited number of applications and libraries in the backports repositories). Therefore, if one desires to have an updated version of a certain software, it must update the whole Ubuntu system. This tag marks instances of mismatches between the software needed for run the experiment and software in a different version of Ubuntu.

Comments About the Virtual Machine

- - Migration to the Native OS Installation
  Some participants did not want to run the experiment in a virtual machine, but rather on working OS running directly on the hardware, without virtualization. This tags marks cases when participants told situations like that.

  PS: When I mean native installation, I mean on Ubuntu 14.04. Whenever they use other OS, we put the statements under the tag "Outros SOs"
- - VM's Performance
  Performance issues and slowdowns perceived by the participants due the fact they were using a virtual machine to run the experiment (in the case they did).
- - RAM da VM
  Instances participants complaint about the given amount of RAM available in the virtual machine (althought this configuration can be easily changed).
- - VM's Resolution
  Instances where the participants mentioned negatively the Virtual Machine's screen resolution, or had problems to configure it properly.
- - VM's Size
  It was suggested the participants to run the replication in a Virtual Machine. However, sometimes participants did not put attention to the Hard Disk size the guest machine would possess. This code is labeled on the participants statements in which the VM size is said to be a replication difficulty.
- Virtual Machine - Complaints
  Complaints and reports about problems on the virtual machine.
- Virtual Machine - Comments
  Statements about the virtual machine, not necessarily complaints.

Replication Experience

- ER1
  Was there any point on the training session you think could be better explained? How exactly? What do you think I could have done to make the concepts and the activity more clear?
- Author's intervention (ER2)
  Due to time restrictions, and somehow, playing the role of the original researcher, whenever the subjects felt some step on the study was impassible, the first author of this study made interventions to prevent them getting stuck for too long. This codes is used to tag mentions to this phenomenom.
- ER3
  How the experiment could be better detailed in order to facilitate the replication?
- - Clarity of the Instructions
  Although there are instructions to perform the replication, they might not be clear enough. Or, depending on the participant perception, they might be perfectly described. Whenever this topic is addressed, this code will be used.
- -- Correspondency Between Tutorial and Artifacts
  The tutorial in the website intends to describe a sequence of steps that will lead to a valid execution of the sutdy. In some situations, the steps described cannot be fully reproduced, due to inconsistencies on the artifacts existing in the Java project and the description on the Replication Package. This code marks instances of that situation.
- ER4
  Do you think this activity was useful for your academic background formation? (Please be honest; you do not have to please me at all).
- ER5
  Ranging from 1 to 5, what score would you give to describe how interesting the experiment was to you?
- ER6
  Do you think if the activity would have been applied in the beginning of the course would it change something on the experiment execution in terms of subject's motivation and time availability?
- ER7
  How was the task division among you (being in mind that no one will be humiliated or pusnished by the content of any answer given). Was it straightfoward to synchronize tasks and meetings?
- Time Availability
  Whenever the subjects said that time availability was an obstacle for performing the replication in they way they wished they could, we mark this tag. Notice that we do not try to infer whether the participants are telling the truth or not (at least not

at this moment); only if they state so.

Algorithms Execution

- E1
  Was it difficult to run the algorithms? Why? List some hardships.
- - Inconsistencies
  Inconsistencies perceived in the replication artifacts.
- -- Name Inconsistencies
  Throughout the paper, website and artifacts, there are references to the terms and artifacts addressed on the study. However, the names on them might be insconsistent. This codes tags comments about instances of that phenomenon.
- --- E3
  During the activity I have noticed that for some teams was strange to perceive that samples/file and configurations per file were the same thing. Why do you think that, even though these concepts are defined on the paper, many people had this difficulty?
- - Abnormalities in Execution
  Whenever paticipants reported the execution did not happen seamlessly as expected, the categories under this one are tagged.
- -- Execution Failures
  This code is used to highlight testimonials describing faults on the execution of the algorithms.
- --- Error on path
  In the source code, there is a reference to a path that point to Flavio Medeiro's home folder. When you try to run that, unless you change the code to point to your local machine, you will get an error. This code mark instances where participants mentioned that (the error and/or the correction)
- -- Long Wait Time for Some Algorithms
  An algorithm that took too long to be executed was often mentioned by the participants. This tags marks that.

  Edit: statement-coverage
- E2
  By the forms answers, most of the people agreed that the algorithms execution tutorial (from the study's website) could be better. But do you think that a more descriptive tutorial would be only something that would be better or would it be fundamental for the replication success?
- E4
  Did you manage to check the algorithms' source code? Did that helped the task somehow? In what exactly?
- Execution Output
  This tag marks problems regarding lack clarity on the outputs, including the level of verbosity on the outputs and lack of feedback concerning algorithms progress execution.

Paper Interpretation

- I1
  What was your overall impression about the paper (in terms of structure, writing, etc)? Do you think that the way the study was reported was appropriate for a replication? Why?
- I2
  In general was the paper reasonable to understand (even you not having a wide domain of its subject?)
- I3
  Can you recall some specific part that was not so clear?
- Lack of Familiarity
  This code marks instances where familiarity with the subjects addressed on the paper are cited. The subjects are the following, not restricted to: Software Product Lines/Configurable Systems, and sampling algorithms.
- - C3
  In your opinion, is it possible that the authors could have thought that people with a degree in computing would have at least a notion of how to install in IDE and that is why they did not took took up much space explaining that on the web site? Do you think that this applies to things related to Linux and command line usage? Why?
- - Understanding about Replications
  Obviously, one of the main subjects for a researcher to understand in order to replicate an experiment is replication itself. If someone is not aware of the definitions of replication in Empirical Software Engineering, this might lead to inaccuracies when performing a repetition of a prior study. On the other hand, being familiar with replications is desirable for this study. This code was used to tag that phenomenon.
- - Skills with Linux
  In order to run the algorithms, the experiment's website gives instructions for Debian-based Linux distributions. In the case of Undertaker, a person attempting to install it must compile this software from the sources. Thus, if one is not familiar with Linux, probably will face difficulties to set the environment. This code is used for quotes when the participants relate skills on Linux to the activity of running the replication.
- - Insufficient Academic Background
  Sometimes, participants can make comments about the how subjects' education background might have influence on how the experiment was replicated by them. This tag marks such comments.
- - ER8
  Do you think the lack of experience with scientific research (not restricted to Software Engineering) was something that played againsted you when executing the experiment's task? Why?
- Difficulties in Interpreting the Paper
  The participants might have different degrees of understanding the baseline study. Sometimes, this understanding might influence on the replication execution and their motivation to perform the experiment. This code tags quotes about that.
- - Confusion Factor
  Although after setting the environment up, is relatively straightfoward to execute the algorithms. Subjects could easily simply run the algorithms, fill the spreadsheet without a basic understanding of what the results mean. This tag marks whenever users face the situation where they have to reason twice in order to understand the relation between the output and the additional columns in the spreadsheet.

  ** EDIT: Now this code refers to all changes I've done on the second study to make them reason about the results. Not only related to sample/configuration stuff anymore.*
- ---- Admission
  Whenever participants affirmed they have used 'configuration' in the cell for 'sample per file' in the extraction spreadsheet concerning the numbers for RQ4, RQ5 and RQ6, because was the number they had at hand (and therefore did not try to interpret the numbers they gor), this code is used.
- -- Attempt to Calculate the Confusion Factor
  Following the previous code, some participants actually tried to calculate the number of configurations per files from the 2nd study algorithms. This tag marks sentences where participants mention testimonials related to that.
- Convergence of Results
  This code mark statements about convergence of results between the paper and the replication results. More than simply numeric results, we intend to capture participants impressions and the reasoning behind converging or diverging results.
- - I4
  When I asked you in the form "do the results diverge", what was your comprehension about the divergence or not divergence of the results?
- - Divergence on the Extraction of code.zip
  Difference in the result of bugs due to the creation of a subfolder in the code.zip extraction -> It happens in C1, C2 and C4.
- Execution Variations due the Lack of Details
  This tags marks whenever the participants suggested that the

lack of details might lead to unintended variations.

Subjects' Attitude

- - Hypot. Paper
  This tags marks whenever subjects attempt to guess the intentions and design decisions behind the baseline study paper.
- -- Communication Failure Between Original Study's Authors
  This marks a participant hypothesis that a failure of communication among original researchers themselves led to inaccuracies when reporting the paper.
- - Hypot. Design Activity
  In a similar manner than the previous code, in some occasions, the participants try to guess aspects about the design of the observatory study. They, sometimes, spend more or less effort based of what they guess about the goals or the variables of this study (which at the Focus Group moment was hidden to them).
- - Hypot. Objective
  The observational study's main goal was not disclosed to the participants. However, some statements said by them give hints that they might have guessed correctly the actual objective of this study. This tag marks instances of those statements.
- - Hypot. Other Teams Impacting
  Sometimes the participants change their attitute towards the study because they hyipothesize about how other teams might be performing the study. Therefore, they might be encourage or discouraged to do a certain action based on how they imagine how other teams are acting towards the replication.
- - Hypot. Other Teams Not Impacting
  Those are the instances where participants of a specific team raise hypothesis concerning the other teams, but does not appear to interfere on how that first team performed the replication.
- Subjects Effort
  This code marks instances in which the engagement/dedication to the study can be a factor of influence on performing the replication, positively or negatively.
- Posture of Subjects in the Training Sessions
  The subjects received a training session where the study's basic concepts weere explained. However, the participants attitude in the training session might influence on their understanding about the study. Some off these attitudes are: stress, mistruss, lack of motivation, distraction, and so on. This tags marks testimonials about this topic.
- Interpretation Effort
  Statements that mention the effort or the absence of effort spent to the participants to understand the replication they were performing.
- Hypothesizing
  Participants often make assumptions about many aspects involving the observational study. This is what we call here "hypothesizing". This act of creating hyphotheses sometimes make them to have a different attitute towards the study based on what they guess about factors hidden to them.
- Patience to Follow Rules
  In some cases, the participants might imagine that the process of setting the environment is so trivial, as well as running the algorithms, that they might tend to ignore some steps in the tutorial, spoiling the process. This tags holds the participants' discussions on that theme
- Subjects Underestimating Tasks
  In many cases, participants underestimated the effort required to perform the replication. That made them to leave the replication to the last moment. When they perveived that setting the environment is not that trivial, they started to stress out and be frustrated. This tag marks instances of that.

Codes not Matching Other Categories

- - C2
  Some participants said the instructions were clear. Others said they here not clear. After all, in your opinion, the instructions were clear or not? Why do you think that there was this discrepancy? Can you give an example of some situation that the lack of clarity was an obstable on the development of the activity?
- - Comments About Original Paper's Artifacts Website
  This mark aggregates comments about the supplementary website.
- - Subjects Feeling 'Cheated'
  The expectation of how much effort the participants would spend on the experiment by them was much lower than the actual effort required to perform the replication. This tag mark instances where subjects report they felt "betrayed" to not be informed of the difficulties of the study beforehand.
- - Original Papers' Link Tilde Character
  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.
- Replication Artifacts
  Comments about replication artifacts, all of them.
- Communication with Authors
  Sometimes, when facing hardships running the replications, the participants expressed the desire to e-mail the baseline study's author. We highlighted, then, when subjects mentioned that with this code.
- I'm not sure
  Self explanatory. It is different from "[no code]".
- Motivation
  We assume that being motivated to participate in the study is beneficial to performing the activity, while being unmotivated can make the participant to not perform the replication in a diligent manner. Therefore, whenever we detect some statement that shows motivation, with a positive or negative attitute towards the study we use this tag.
- Comments About Other Replication
  Subjects familiar with other replications might make comparison between this study and the other they have read or performed. This tags mark testimonials about that situation.
- Suggestions

Complaints

- - Environmnent was not Ready
  Complaints about the replication environment has not been given to the participants already set up.
- Criticism to Research Field
  Subjects familiar with Software Engineering or Sampling Algoritmhs research might criticize some aspects these research fields can improve to get more maturity.
- Criticism Towards the Original Study
  This tag marks whenever participants make comments about points they missed or could be better on the baseline study.
- Criticism Towards Study Design
  Even though the observatory study's design was not disclosed to the participants, in some instances they criticize the way the study was conceived, following their perception.
- Study's Objective was not Clear
  Mainly in the first application, the participants complained that the study's objective was not clear enough as soon as they started the replication task. In the second application, this complaint was also present, although there was a bigger effort to explain the study's objective (although not fully disclosed). This code marks instances of subjects reporting said they were not fully aware of the objective.
- Study's Design Disclosure Request
  Because they were playing the role of researchers, the subjects sometimes wished they knew deeply about the observatory study. They even expressed that not knowing the design of this upper level study prevented them to realize the activity in a

better way. This code is used for sign the times when statements of this kind were said.

- Implementation
  Complaints about the algorithms implementaion.
- University's Internet Quality
  This tag marks instances where the participants make complaints about the Internet connection quality on the University (since many of them perform academic activities there).
- Extraction Spreadsheet
  Complaints about phenomena surrounding the extraction spreadsheet.
- Other Complaint
  I can't tag with the higher hierarchy (Complaints) on QDA Miner, so I created this tag.
- Manual Labor
  This tag is use to highlight affirmations where partcipants complain about the manual effort they had to spend on the study (for instance installing the virtual machine, setting up the IDE and so on)

Threats

- Forgetting Effect
  In this study's first application there was a time gap of approximately one month between the replication and the focus group. Although we executed some actions to mitigate the forgetting effect, it still might appear. Therefore, this code marks instance of this effect being reported by the participants.
- Interaction Between Teams
  It was explicitly told the participants from a team could not not ask participants from other teams for help. Nevertheless, this might have happened. This code emerged from participants reports about that phenomenon.