# – Evaluation

## Introduction

The main motivation of this chapter is to answer the following research questions defined at Chapter 1:

* Is it possible to use provenance analysis in order to understand events that transpired during the game?
* Is it faster to analyze the game by using provenance instead of watching a replay of the game session?
* Is it better to do provenance analysis or watch a replay of the game to know how the events occurred during the game?

To assess the possibility of using provenance analysis for improving understanding, it was necessary to generate a replay of a game session and compare the replay analysis with a provenance analysis using a provenance graph. This comparison was conducted through a questionnaire containing specific questions about the events occurred during the recorded game session. Volunteers were required to watch the video and then answer the questionnaire by consulting it again (denominated Replay group) or by analyzing the provenance graph (denominated Provenance group).

Lastly, to compare the results obtained by both groups, we utilized two metrics: accuracy and time. The first metric, accuracy, has the intention to verify the answers provided by both groups. The second metric, time, is used to measure the time each volunteer took to answer all questions in the questionnaire, thus allowing to know which method (replay or provenance) is faster.

This chapter is organized as follows: Section 6.2 describes details about the experiment planning. Section 6.3 explains the experiment execution, while section 6.4 provides a statistical analysis by detailing tests, their results and conclusions on the data obtained. Section 6.5 describes threats to validity of the experiment. Lastly, Section 6.6 presents the final considerations of this chapter.

## Experiment Planning

In order to verify the possibility of using provenance for aiding in understanding the events in a game, it would be required to analyze a game session. Initially, was planned to allow the volunteers to play the game and to answer a questionnaire at the end of the game session. Half the volunteers would answer the questionnaire while having access to the provenance of the game session. The other half would answer the questionnaire by only using his/her memory. However, the formulation of the questionnaire would be impractical due to the random nature of a game session, even when restrictions are placed in order to control the outcome. Thus, events that occurred to one player might not be true to another.

To deal with this issue, we opted for a more controlled environment. Instead of playing the game, volunteers will watch a game session played by a third person, which was previously recorded. Thus, the questionnaire can be customized to the game session, allowing asking specific questions about events that occurred. Also, the questionnaire is designed to measure the accuracy of the answers and the time volunteers took to finish it.

Lastly, volunteers are required to read and watch tutorials due to the unfamiliarity with the game and the *Prov Viewer*. Thus, the experiment is divided in three stages, as illustrated by Figure 1: Generating the questionnaire, running the experiment with volunteers, and analyzing the results. The analysis of the results is done by a statistical analysis using the hypothesis test in order to compare the obtained results from both methods (provenance and replay).

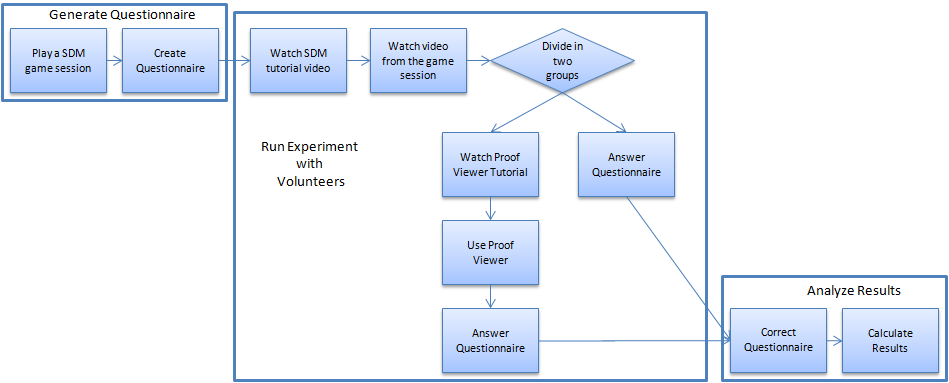


Figure 1: Experiment Execution Flowchart

Another important factor for the design of the experiment concerns the definition of the significance level to be used for the statistical analysis. For the experiments performed in this work we used a confidence interval of 95%, which translates to α = 0.05 where α is the maximum probability of rejecting a true null hypothesis.

## Experiment Execution

Following the plan from Figure 1, the first stage (Generate Questionnaire) is executed before running the experiment with volunteers. We created a video of a recorded game session from SDM that at the same time narrates the player’s decisions throughout the game. Then the questionnaire was designed based on the video, which consists of ten questions. The first and last questions are related to time measure, the times when the volunteer started and ended the questionnaire. The second question is designed to identify which group the volunteer belongs to: Provenance, which uses *Prov Viewer* while answering the questionnaire, or Replay, which answers the questionnaire without using the tool. The other seven questions are related to events that transpired during the game, with four questions designed for provenance analysis (questions 3, 5, 6, and 9) and three to be easily answered by both methods (provenance and replay).

The next stage is running the experiment with volunteers. Before participating in the experiment, volunteers were required to read and sign a consent form. Then, volunteers watched a tutorial video from SDM, explaining details about the game’s interface, and read a written document explaining key features. Then they watched the gameplay video and were divided in two groups: those that would use *Prov Viewer*, and those that would not. After the gameplay video, the volunteers were handed the questionnaire. However, the provenance group had to watch another tutorial video for the tool before getting the questionnaire. This stage also had a time limit to avoid fatigue. All documents used at this stage are available at Appendix A.

Before running the experiment, we previously ran test session in order to test the proposed plan. The test was applied to an undergrad class composed of 21 volunteers. It followed the plan proposed at the previous section with a different order for videos: SDM tutorial, *Prov Viewer* tutorial, then gameplay video. This order was changed in the experiment due to volunteers were reviewing the *Prov Viewer* tutorial while answering the questionnaire. This was happening because they were forgetting how to operate the tool after watching the 7 minutes gameplay video. Another change made for the experiment was related to the questions in the questionnaire. Some questions were leaving room for different interpretations, which caused too many mistakes on both groups. Thus we decided to create a new scenario (and video) with a different set of questions. Lastly, during the test we allowed each volunteer to watch the videos at their own pace, causing chaos during the test because of undisciplined behavior from the volunteers. They were also deceiving the time they took to answer the questionnaire. Thus, during the experiment we followed a stricter timetable. Thus the results obtained during the test were archived and not used.

After the test and making the appropriate changes, we applied the experiment in two different undergrad classes, composed of 18 and 19 volunteers each. From those 37 volunteers, only 32 were able to finish the experiment in the allocated time, thus discarding 5 partially answered questionnaires. After running the experiment on both classes, the questionnaires were analyzed yielding the described results in Table 1 and Table 2. All questions have the same weight and values varying from 0 to 1, depending on the answer provided, and duration values expressed in minutes.

Table : Provenance Group Results

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Results** | | | | | | | | | | | | | | | | |
| Q3 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| Q4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| Q5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| Q6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q7 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Q8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.5 | 0 | 0.5 | 0 | 0.5 |
| Q9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| Duration | 25 | 18 | 19 | 21 | 18 | 19 | 21 | 21 | 28 | 21 | 28 | 29 | 26 | 27 | 20 | 30 |

Table : Replay Group Results

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Results** | | | | | | | | | | | | | | | | |
| Q3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q4 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| Q5 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q7 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0.5 | 0.5 | 0.5 |
| Q9 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| Duration | 20 | 32 | 33 | 32 | 30 | 30 | 48 | 42 | 38 | 38 | 31 | 14 | 19 | 8 | 29 | 19 |

## Statistical Analysis

The statistical analysis was performed with the intention of checking the obtained results and to verify if they have any significant difference. The main idea is to compare the results obtained from the questionnaire and the elapsed time from both methods of analysis of a game flow. All tests were done in the free software R[[1]](#footnote-1), which is commonly used for statistical analysis and graph construction with the IDE *RStudio*[[2]](#footnote-2).

### Normality Test

The fundamental part of a statistical analysis of an experiment is the hypothesis test (WOHLIN *et al.*, 2000). In the hypothesis test, two hypotheses are proposed and used to validate the collected data. On a normality test, the first hypothesis is the null hypothesis H0, which states that the data collected have a normal distribution. The second hypothesis, H1, states that the data collected does not have a normal distribution. However, testing hypothesis involve two types of risks: Type-I and Type-II. The Type-I error refers to the rejection of the null hypothesis H0 even when it is true, while the Type-II error accepts the null hypothesis H0 when it is false. These errors are dependable on the power of the test C, which is the probability of 1 - α that the test is true if H0 is false, where α is the probability of committing the error Type-II. Given this, a normality analysis from the obtained data is required to decide between parametric and non-parametric tests. Thus, we used the Shapiro-Wilk test (SHAPIRO; WILK, 1965) with the following hypotheses:

This test is executed in R by the command *shapiro.test(x)*, where *x* is the vector containing the data to be analyzed. It is provided as output the statistical value *W*[[3]](#footnote-3) from the Shapiro-Wilk test and its *p-value*[[4]](#footnote-4), as can be seen by FIGURE. The null hypothesis is rejected if *p-value* is lower than the significance level α, thus concluding that the data did not have a normal distribution.



Figure 2: Example of R’s output for Shapiro-Wilk test

The normality assumption was violated for all obtained results from the experiment because *p-value* < 0.01. It is possible to verify that *p-value* < α since α = 0.05 and *p-value* < 0.01, thus rejecting the null hypothesis. The results can be seen at Table 1. Note that *Replay’s* duration’s *p-value* is greater than 0.05. However, *Prov’s* *p-value* is 0.04337, which is lesser than α = 0.05. Question 6 result is null because all volunteers zeroed in both methods.

Table 3: Normality Test Results

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Duration |
| Prov | 4.34e-05 | 4.553e-08 | 1.575e-08 | Null | 2.566e-05 | 1.213e-05 | 1.575e-06 | 0.04337 |
| Replay | 4.553e-08 | 3.408e-07 | 1.575e-06 | Null | 5.272e-06 | 1.33e-05 | 4.34e-05 | 0.7363 |

Therefore, non-parametric tests were used for statistical analysis of data. The test used to compare the means was Mann-Whitney, which is also known as Wilcoxon rank-sum[[5]](#footnote-5) test. There are other non-parametric tests (Chi-2, Kruskal-Wallis), however Mann-Whitney was chosen because it compares two means from two different samples against the same alternative hypothesis. The next section presents the results obtained from Wilcoxon test to verify if both methods (using provenance and watching replay) results are equals.

### Comparison of Means

The hypothesis used for Wilcoxon test to verify the results are:

The mean is calculated for each question from the questionnaire and the duration that each volunteer took to finish it. Table 2 illustrates the resulting values for the mean and standard deviation for both methods:

Table 4: Mean and Standard Desviation for the results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Duration |
| Prov | **Mean** | 0.5 | 0.9375 | 0.1875 | 0 | 0.375 | 0.1562 | 0.8125 | 23.1875 |
| **Standard Deviation** | 0.5164 | 0.25 | 0.4031 | 0 | 0.5 | 0.3010 | 0.4031 | 4.2461 |
| Replay | **Mean** | 0.0625 | 0.875 | 0.1875 | 0 | 0.25 | 0.1562 | 0.5 | 28.9375 |
| **Standard Deviation** | 0.25 | 0.3416 | 0.4031 | 0 | 0.4472 | 0.2394 | 0.5162 | 10.5797 |

It is possible to assert that exist a difference in means if the null hypothesis H0 is rejected. The test is performed by the command *wilcox.test(x, y, conf.int = T)*, where x and y are vectors to be tested and *conf.int* is used to display the confidence interval. Figure 3 illustrates an example of the output from this command in R with α = 0.05, while Table 3 illustrates all results obtained.



Figure 3: R's output for Wilcoxon test

Table 5: Results obtained from the Wilcoxon test

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| α = 0.05 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Duration |
| p-value | 0.007259 | 0.5757 | 1 | Null | 0.467 | 0.8099 | 0.07049 | 0.03595 |
| CI | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 |

The null hypothesis is not rejected if *p-value* is greater than significance level α. In other words, there is not enough evidence to assert a difference between results. When the null hypothesis is rejected (*p-value* < α), it is necessary to identify which method is superior by analyzing the confidence interval *CI*. If *CI* – α < 0, then . Otherwise . By analyzing the *p-values* from Table 3, the usage of provenance analysis provided better results on question 3 and in the time required to finish the questionnaire (duration), in comparison with replay analysis, while the other occasions there is no evidence to assert difference between results (*p-value* > α).

The *boxplot* graphs shown at Figure 4 aims to summarize the distributions for both Prov and Replay methods, providing another view of the tests described above. In these graphs, the boxes represent part of the central distribution, which contains 50% of data. Thus, the data scattering is proportional with the box’s height. The median is represented by a black line inside the box. 25% of data is between the box’s edges and the median. The median location indicates that if the distributions are symmetrical in the experiments. Lastly, circles indicate outliers, which are data with more than 1.5 interquartile range (Q3 – Q1) from other data.

By analyzing Figure 4, it is possible to infer that question 3 yielded better results by using provenance while questions 4 and 5 had equal results. Meanwhile, questions 7 and 8 results were similar but with varying scattering. Even though results are matching with wilcoxon test data, question 9 has a different behavior due to the small difference from *p-value* to α (*p-value* = 0.07 against α = 0.05). By analyzing the *boxplot* for question 9, the results for using provenance is greater than the replay method. While replay’s data is scattered around the margins with the median at the middle, the provenance’s median is located at the maximum value, yielding greater results. Lastly, as proven by the wilcoxon test, using provenance for analysis provides faster answers than analyzing the game session’s replay. This is clearly seen by comparing the medians between methods and the box’s scattering (height) and position. Even though question 5 had the same results, question 6 boxplot was discarded because both methods had equal values and were zero (without any outliers).

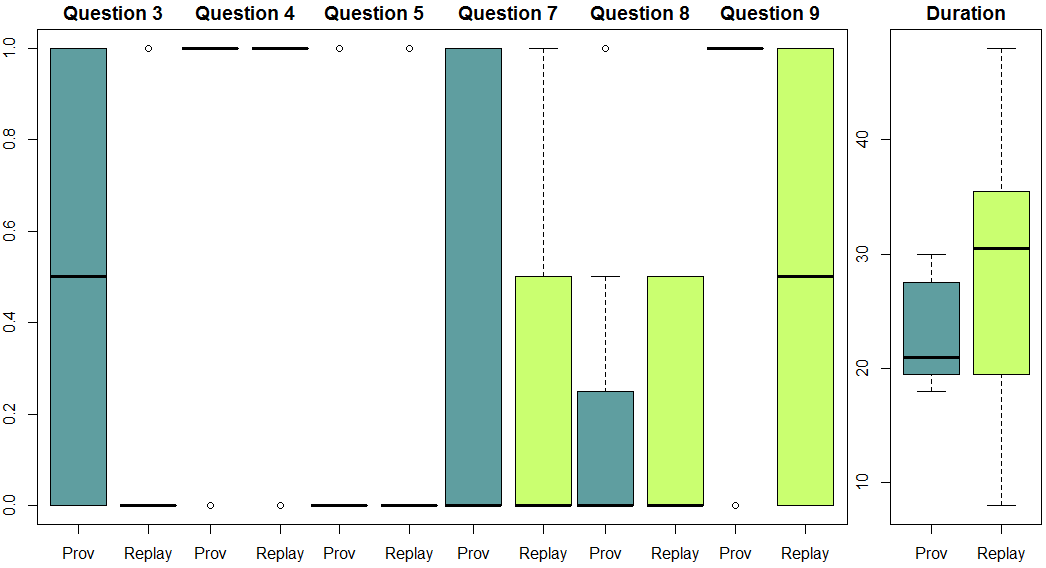


Figure 4: Boxplots from the experiment

## Threats to Validity

Despite the care in reducing the threats to the validity of the experiment, there are factors that can influence the results. In relation to internal validity, the selection for both groups (Provenance and Replay) can affect the results because of the natural variation in human performance. Furthermore, the experiment was executed with volunteers, which generally are more motivated for executing tasks. Anyone from the class could choose to be dismissed from the experiment and be released earlier. Lastly, the experiment was the first time volunteers had contact with both the game and the tool. The lack of experience with either can affect the results, even when minimized by the usage of tutorials. For external validity, volunteers were from the same course (“Prog1”), which is from the first period for the undergrad courses.

For construct validity, the questionnaires were composed of multiple questions to reduce threats related to failing to answer the questions by not identifying the correct reason. Another risk is related to people being afraid of being evaluated, thus try to look better by changing the answers. This is specially the case of the duration they took to finish answering the questionnaire. To minimize this, we had a strict timetable for each activity, stating the exact time they began answering the questionnaire and verifying the time they finished it.

In relation to conclusion validity, there is the reliability of measures. This is dependent on factors like question wording, which may allow for different interpretations, and the graph layout. To minimize this, we answered any doubts voiced by volunteers related to the questions in the questionnaire or the tool (*Prov Viewer*). Another threat is related to the size of the population evaluated, which is composed of 37 volunteers. Thus to minimize this threat, we used non-parametric tests to evaluate the obtained results.

## Final Considerations

## References

SHAPIRO, S. S.; WILK, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, v. 52, n. 3/4, p. 591, dez. 1965. Acesso em: 5 jun. 2013.

WOHLIN, Claes *et al.* *Experimentation in software engineering: an introduction*. Norwell, MA, USA: Kluwer Academic Publishers, 2000.

1. http://www.r-project.org/ [↑](#footnote-ref-1)
2. http://www.rstudio.com/ [↑](#footnote-ref-2)
3. The W statistic checkes if the sample is from a normal distribution. Data normalization is shown by low values. [↑](#footnote-ref-3)
4. *p-value* is the lowest level of significance at which the null hypothesis could be rejected for the given observations. [↑](#footnote-ref-4)
5. <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html> [↑](#footnote-ref-5)