# – 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 the events transpired during the game?
* Is it faster to analyze the game by using provenance instead of watching a replay of the session?
* Does provenance analysis help to find answers with more accuracy than watching a replay of the game session?

Experiments were elaborated in order to answer these questions.

## Experiment Planning

In order to verify the possibility of using provenance to aid understanding of events in a game, it would be required to analyze a game session. Initially, it 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 nature that each game session is different from another, even when restrictions are placed in order to control the outcome. Events that occurred to one player might not be true to another. Thus, the questionnaire would be required to cover almost all possibilities or have generic questions.

To deal with this issue, we opted for a more controlled environment. Instead of playing the game, all volunteers would watch the same game session being played by a third person. Thus, solving the problem of having multiple, and different, game scenarios. However, the problem of formulating the questionnaire would still persist. To solve this, we decided to previously record a game session and the provenance information. Then, the questionnaire would be elaborated. Thus volunteers would watch the recorded game session instead of watching it in real time. The questionnaire can be customized to the game session because it was previously recorded. This allows for specific questions to be asked about specific events and their respective outcomes in the questionnaire, instead of making generic questions trying to guess the possible outcomes.

Thus, volunteers were required watch a gameplay video and answer a questionnaire. To answer the question about using provenance analysis for aiding understanding, some volunteers answered the questionnaire by using the *Proof Viewer*, while the others only by re-watching the gameplay video.

## Experiment Execution

Due to the unfamiliarity with the game and the *Proof Viewer*, the experiment execution followed the plan detailed at the previous section with a few additions: tutorials. The execution process is divided in three stages, as illustrated by Figure 1: Generating the questionnaire, running the experiment with volunteers, and analyzing the results.

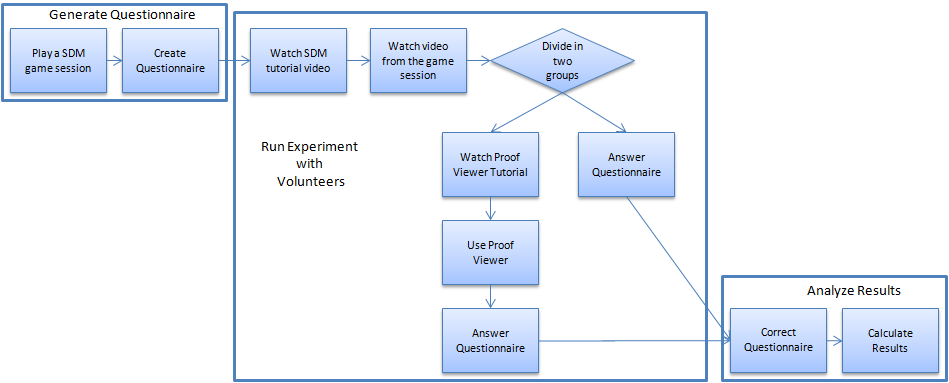


Figure 1: Experiment Execution Flowchart

The first stage is generating the questionnaire.

The next stage is running the experiment with volunteers. Before participating in the experiment, volunteers were required to read and sign a consent form. First, volunteers were required to watch a tutorial video about the SDM, which explained details about the game’s interface and also had a written document explaining key features. Then they were allowed to watch the gameplay video and were divided in two groups: those that would use *Proof Viewer*, and those that would not. After the gameplay video, the volunteers answered the questionnaire. However, the group for *Proof Viewer* saw a tutorial video and got a written document before being allowed to answer the questionnaire, both detailing the *Proof Viewer’s* interface and explaining all features in it. The other group was allowed only to re-watch the gameplay video. All documents used at this stage are available at Appendix A.

Lastly, analyze results.

The results obtained are: …

## 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 1: 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 | 4.778e-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 2: 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 | 0.1719 | 0.5 | 28.9375 |
| **Standard Deviation** | 0.25 | 0.3416 | 0.4031 | 0 | 0.4472 | 0.2366 | 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 3: Results obtained from the Wilcoxon test

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Duration |
| p-value | 0.007259 | 0.5757 | 1 | Null | 0.467 | 0.6098 | 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…

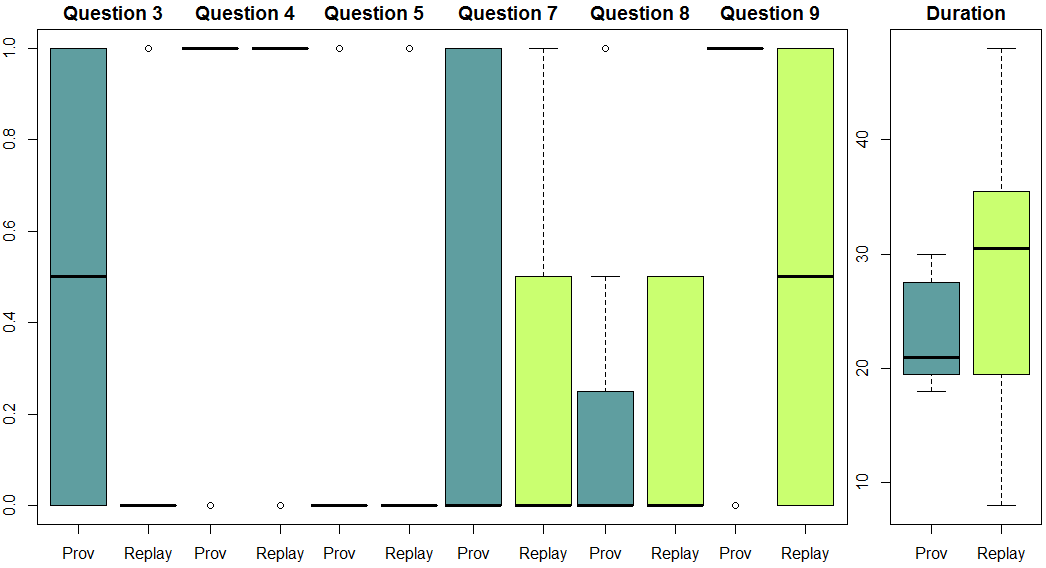


Figure : Boxplots from the experiment

## Threats to Validity

First time with SDM.

First time with Proof Viewer.

Few volunteers.

Questionnaire with multiple questions.

Questionnaire’s correct answers.

Time restriction.

First period students (prog1).

## 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. A estatística W verifica se a amostra provém de distribuição normal. Menores valores evidenciam normalização dos dados [↑](#footnote-ref-3)
4. Menor nível de significância no qual a hipótese nula poderia ser rejeitada para as observações dadas. [↑](#footnote-ref-4)
5. <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html> [↑](#footnote-ref-5)