# – Evaluation

## Introduction

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

* Does provenance analysis help to understand events that emerged during the game?
* Is provenance analysis faster than only watching a replay of the game session?
* Is provenance analysis more accurate than only watching a replay of the game session?

To assess the possibility of using provenance analysis for improving understanding, we generated a replay of a game session and compared it with provenance analysis using a provenance graph. This comparison was conducted through a questionnaire containing specific questions about events that occurred during the game session. Volunteers were divided into two groups: replay group and provenance group. Both groups watched the replay of the game session. The provenance group also had access to the provenance graph. At the end, both groups answered the questionnaire.

Lastly, we used two metrics to compare the results obtained by both groups: precision and time. The first metric, precision, has the intention to verify the correctness of 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 obtained data. Section 6.5 discusses some threats to validity of the experiment. Lastly, Section 6.6 presents the final considerations of this chapter.

## Experiment Planning

We opted for a controlled environment in order to reduce independent variables that were beyond our control. Instead of playing the game, volunteers watch a recorded game session played by a third person. Thus, the questionnaire can be customized to the game session, allowing asking specific questions about events that occurred in that particular session. Also, the questionnaire is designed to measure the precision of the answers provided by both groups (replay and provenance) and the time volunteers took to finish it. Precision (BAEZA-YATES; RIBEIRO-NETO, 1999) is a traditional metric for information retrieval and can be seen as a measure of correctness, which is the percentage results that are relevant. Time is the measure of the minutes taken to complete the questionnaire, which is the difference of the time when the volunteer finished the questionnaire and the hour he started it.

Before filling the questionnaire, volunteers are required to read and watch tutorials due to the unfamiliarity with the game and the *Prov Viewer*. Furthermore, we ran a pilot of the experiment in order to determine the experiment structure, which was initially structured as follows: volunteers were divided into two groups and start the experiment by watching the SDM tutorial, then the *Prov Viewer* tutorial (only for the provenance group) and the replay of the game session video. Lastly, they receive the questionnaire.

This order was later changed for the experiment due to the fact that volunteers were reviewing the *Prov Viewer* tutorial while answering the questionnaire. This happened because they were forgetting how to operate the tool after watching the replay of the game session video, which takes around seven minutes. 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 pilot we allowed each volunteer to watch the videos at their own pace, causing chaos because of undisciplined behavior from the volunteers. They were also deceiving the time they took to answer the questionnaire. Thus, we decided to impose a stricter timetable, providing the questionnaire only after all volunteers of the same group finish watching the videos.

With the changes made after the pilot, the experiment plan is illustrated by Figure 1 and is divided in three stages: Generating the questionnaire, running the experiment with volunteers, and analyzing the results. According to the plan shown in Figure 1, the first stage (Generate Questionnaire) is executed before running the experiment with volunteers. We created the replay of a recorded game session from SDM that narrates the player’s decisions throughout the game. Then, the questionnaire was designed based on the video[[1]](#footnote-1), consisting of ten questions. The first and the last questions are related to time measurements: the times when the volunteer started and finished the questionnaire. The second question is designed to identify the group of the volunteer: 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 emerged during the game and have the same weight with values varying from 0 (wrong) to 1 (correct), depending on the answer provided. A value of 0.5 means the answer was partially correct, meaning that only one item was correctly identified. These questions explore different aspects from the game, and some questions require a deeper knowledge of the game.

The next stage is to run the experiment with volunteers. Before participating in the experiment, volunteers are required to read and sign a consent form. Then, volunteers watch a tutorial video from SDM, which explains details about the game interface, and read a written document summarizing key features. Subsequently, they watch the replay video and are divided in two groups: those that will use provenance and those that will not. After watching the replay video, the volunteers are handed the questionnaire. However, the provenance group watches another tutorial video for the tool before receiving the questionnaire. This stage also has a time limit to avoid fatigue. All documents used at this stage are available at Appendix A. Lastly, we performed a statistical analysis over the results by means of hypothesis test in order to compare the obtained results from both methods (provenance and replay).

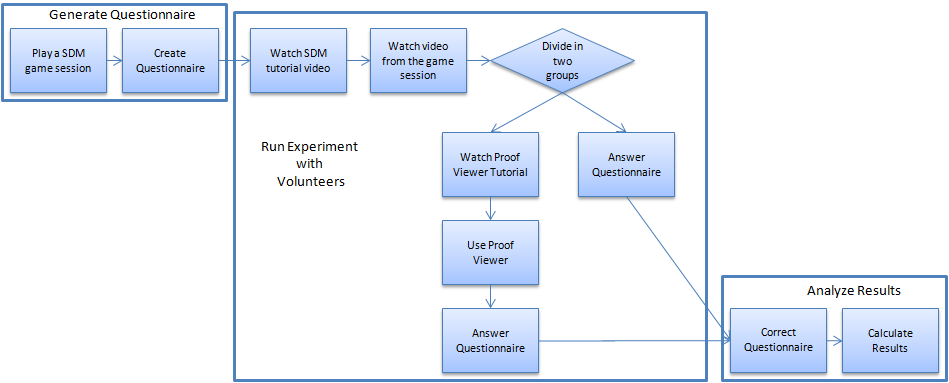


Figure 1: Experiment Execution Flowchart

Another important factor for the design of the experiment concerns the definition of the significance level to be used during 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 incorrect rejecting the null hypothesis (Type I error).

## Experiment Execution

The pilot was applied to an undergraduate class composed of 28 volunteers. The obtained results are described in Table 1 and Table 2, where μ represents the mean and σ the standard deviation. The duration values are expressed in minutes. However, this data was not used for the experiment or the statistical analysis due to the changes made after the pilot.

Table 1: Pilot Provenance Group Results

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Results | | | | | | | | | | | | | | μ | σ |
| Q3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0.3 | 0.48 |
| Q4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| Q5 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0.54 | 0.52 |
| Q6 | 0.5 | 1 | 1 | 0.5 | 0.5 | 0.5 | 0 | 0.5 | 1 | 0.5 | 1 | 0 | 0 | 0.54 | 0.37 |
| Q7 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1 | 1 | 0.5 | 0.5 | 0.5 | 0 | 0.54 | 0.25 |
| Q8 | 0 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0 | 0 | 0.5 | 0.5 | 1 | 0.5 | 0.46 | 0.32 |
| Duration | 15 | 20 | 22 | 23 | 25 | 25 | 25 | 33 | 34 | 40 | 36 | 24 | 41 | 27.9 | 8.03 |

Table 2: Pilot Replay Group Results

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Results | | | | | | | | | | | | | | | | μ | σ |
| Q3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0.07 | 0.26 |
| Q4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| Q5 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0.53 | 0.51 |
| Q6 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0.33 | 0.49 |
| Q7 | 0 | 0.5 | 0.5 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.43 | 0.17 |
| Q8 | 0 | 0.5 | 0.5 | 1 | 1 | 0 | 0 | 0.5 | 0.5 | 0 | 0.5 | 0 | 0.5 | 0.5 | 0 | 0.37 | 0.35 |
| Duration | 10 | 24 | 10 | 15 | 16 | 25 | 17 | 11 | 15 | 30 | 15 | 22 | 8 | 16 | 34 | 17.9 | 7.62 |

After the pilot and making the appropriate changes in the plan, 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 5 partially answered questionnaires were discarded. After running the experiment on both classes, the questionnaires were analyzed and yielded the results described by and Table 4.

Table 3: Provenance Group Results

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Results | | | | | | | | | | | | | | | | | μ | σ |
| Q3 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0.5 | 0.5164 |
| Q4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.9375 | 0.25 |
| Q5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0.1875 | 0.4031 |
| Q6 | 0 | 0 | 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 | 0.375 | 0.5 |
| Q8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.5 | 0 | 0.5 | 0 | 0.5 | 0.1562 | 0.3010 |
| Q9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0.8125 | 0.4031 |
| Duration | 25 | 18 | 19 | 21 | 18 | 19 | 21 | 21 | 28 | 21 | 28 | 29 | 26 | 27 | 20 | 30 | 23.1875 | 4.2461 |

Table 4: Replay Group Results

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Results | | | | | | | | | | | | | | | | | μ | σ |
| Q3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0625 | 0.25 |
| Q4 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.875 | 0.3416 |
| Q5 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1875 | 0.4031 |
| Q6 | 0 | 0 | 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 | 0.25 | 0.4472 |
| Q8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0.5 | 0.5 | 0 | 0.0938 | 0.2015 |
| Q9 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0.5 | 0.5162 |
| Duration | 20 | 32 | 33 | 32 | 30 | 30 | 48 | 42 | 38 | 38 | 31 | 14 | 19 | 8 | 29 | 19 | 28.9375 | 10.5797 |

## Statistical Analysis

A fundamental part of the 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. However, hypothesis testing involves two types of error: Type-I and Type-II. The Type-I error refers to the rejection of the null hypothesis even when it is true, while the Type-II error refers to the acceptance of the null hypothesis when it is false. These errors depend on the power of the test C, which is the probability of 1 - β that the test is true if H0 is false and β is the probability of committing the error Type-II. Moreover, the hypothesis test can be parametric or non-parametric. Parametric tests have a greater power C, thus produces more accurate and precise estimates. However, parametric tests can only be used if samples follow a normal distribution. Nevertheless, non-parametric tests does not require normality and are recommended when samples are small (WOHLIN *et al.*, 2000).

The statistical analysis was performed with the intention of checking the obtained results and verifying if they have any significant difference. The main idea is to compare the results obtained from the questionnaire and the elapsed time of both groups. All tests were done in the open source software R[[2]](#footnote-2), which is commonly used for statistical analysis and graph construction, within the IDE *RStudio*[[3]](#footnote-3).

### Normality Test

On a normality test the null hypothesis H0 states that the collected data follows a normal distribution. The alternative hypothesis, H1, states that the collected data does not follow a normal distribution. Given this, a normality analysis from the obtained data decides between using parametric or 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*[[4]](#footnote-4) from the Shapiro-Wilk test and its *p-value*[[5]](#footnote-5), as can be seen by Figure 2. The null hypothesis is rejected if *p-value* is lower than the significance level α, thus concluding that the data do 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 in Table 5. Note that replay’s duration *p-value* is greater than 0.05 (0.73). However, provenance’s *p-value* is 0.04337, which is lesser than α = 0.05. Null values in the tables come from constant observations, thus not being possible to test normality with Shapiro-Wilk, but indicating that the data does not follow a normal distribution.

An important fact not observed until now was the presence of outliers. Outliers are data far from the norm for the population, typically with more than 1.5 interquartile range (Q3 – Q1) from other data. They can have detrimental effects on statistical analyses, increasing the variance error and reducing the power of statistical tests. However, not all outliers are illegitimate contaminants (BARNETT; LEWIS, 1994). The outliers detected in the sample are legitimate cases, since they are directly related to the correctness of the answers provided by volunteers. In any case, Table 6 illustrates the results for the normality test without outliers, which were removed following the 1.5 interquartile range definition.

Table 5: Normality Test Results with Outliers

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

Table 6: Normality Test Results without Outliers

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Duration |
| Prov | 4.34e-05 | Null | Null | Null | 2.566e-05 | 3.481e-06 | Null | 0.04337 |
| Replay | Null | Null | Null | Null | 5.272e-06 | Null | 4.34e-05 | 0.7363 |

Therefore, non-parametric tests were used for statistical analysis. The test used to compare the means was Mann-Whitney, which is also known as Wilcoxon rank-sum[[6]](#footnote-6) test. There are other non-parametric tests, such as Chi-2 and Kruskal-Wallis, however Mann-Whitney was chosen because it compares two means from two different samples against the same alternative hypothesis, which fits to our experiment design. The next section presents the results obtained from Mann-Whitney test to verify if the group results, with and without provenance, are equals.

### Comparison of Means

We adopted the following hypothesis in our tests:

The mean is calculated for each question from the questionnaire and for the duration that each volunteer took to finish it. Table 7 illustrates the mean the standard deviation of each question for both methods, based on the data presented in Table 3 and Table 4.

Table 7: Mean and Standard Deviation for each question

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 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 | 0.0938 | 0.5 | 28.9375 |
| **Standard Deviation** | 0.25 | 0.3416 | 0.4031 | 0 | 0.4472 | 0.2015 | 0.5162 | 10.5797 |

It is possible to assert that there is a difference in means if the null hypothesis is rejected. The Mann-Whitney test is performed in R 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. As default, the *wilcox.test* paired attribute is set to false, representing the Mann-Whitney test. illustrates an example of the output from this command in R with the default α value, which is 0.05, while Table 8 illustrates all obtained results.



Figure 3: R's output for Mann-Whitney test

Table 8: Results obtained from the Mann-Whitney test

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| α = 0.05 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Duration |
| p-value | 0.007259 | 0.5757 | 1 | Null | 0.467 | 0.6371 | 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 8, the usage of provenance analysis provided better results in question 3 and in the time required to finish the questionnaire (duration), while there is not enough evidence to assert difference between results for the other questions (*p-value* > α).

The *boxplots* shown in Figure 4 aim at summarizing the distributions of both provenance and replay methods. 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. This way, 25% of data is between the box’s edges and the median. The median location indicates if the distributions are symmetrical in the experiments. Lastly, circles indicate outliers.

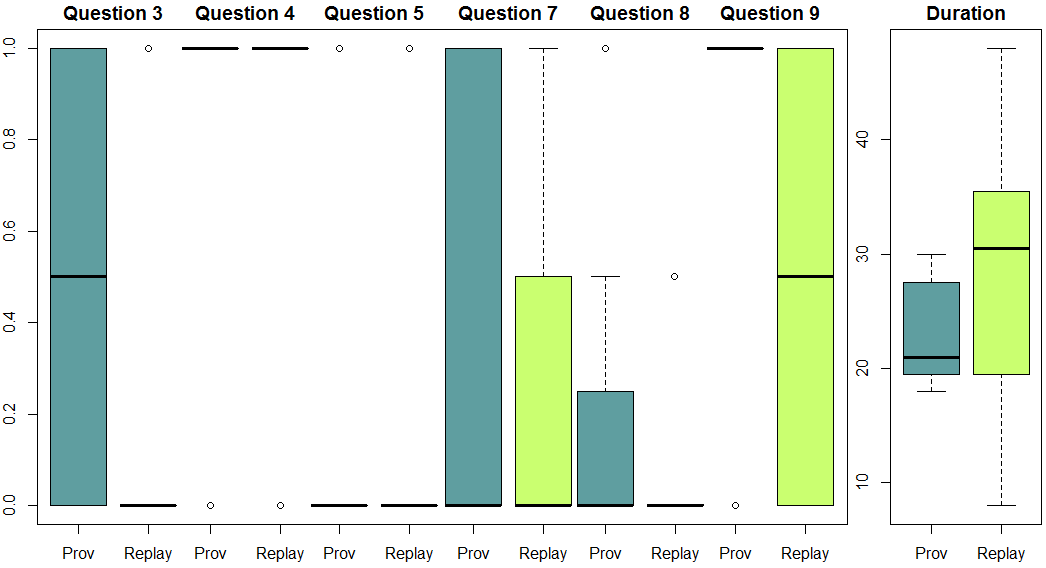


Figure 4: Boxplots from the experiment

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 Mann-Whitney 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 are greater than the replay method. While replay’s data is scattered around the maximum and minimum values with the median at the middle, the provenance’s median is located at the maximum value. Lastly, as shown by the Mann-Whitney test, using provenance for analysis provides faster answers than analyzing the game session’s replay. This is clearly seen by comparing the medians between both methods and the box’s scattering (height) position. The *boxplot* for question 6 was discarded because both methods had equal values and were all zero (without outliers).

## 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 contact of the volunteers with both the game and the tool. Thus, the lack of experience can affect the results, even when minimized by the usage of tutorials. For external validity, to level the experience of volunteers, they were from two different classes of the same discipline (Introduction to computer programming), which occurs in the first period of undergraduate course in Computer Science at *Universidade Federal Fluminense*.

Regarding construct validity, the questionnaires were composed of several questions to reduce threats related to a lack of knowledge from the game, thus exploring different aspects from the game. Another risk is related to people being afraid of being evaluated, thus trying to “look better” by lying. This is the case of how long 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 and delivered the questionnaire.

A threat related to conclusion validity 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 the threat, we answered any doubts voiced by volunteers related to the questions in the questionnaire or regarding the tool (*Prov Viewer*). Another threat is related to the size of the population, which is composed of 37 volunteers.

## Final Considerations

This chapter presented the evaluation of the adoption of provenance to analyze a game flow. This evaluation was performed using statistical analysis on the values obtained from the experiments. The results demonstrate the obtained results for analyzing with provenance were equal or greater than watching a replay. Furthermore, analyzing the game flow with provenance is faster than only watching a replay of the game session.

In relation to correctly identify the causes of the events in the game, using provenance provided better statistical results in at least one case (question 3), and slightly better results in another (question 9). The other cases were not statistically different with the current sample size.

## References

BAEZA-YATES, Ricardo A.; RIBEIRO-NETO, Berthier. *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

BARNETT, Vic; LEWIS, Toby. *Outliers in Statistical Data*. Chichester; New York: Wiley, 1994.

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. The replay video used for the experiment is available at LINK [↑](#footnote-ref-1)
2. http://www.r-project.org/ [↑](#footnote-ref-2)
3. http://www.rstudio.com/ [↑](#footnote-ref-3)
4. The W statistic checks if the sample is from a normal distribution. Data normalization is shown by low values. [↑](#footnote-ref-4)
5. *p-value* is the lowest level of significance at which the null hypothesis could be rejected for the given observations. [↑](#footnote-ref-5)
6. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html [↑](#footnote-ref-6)