Exam

Maath Musleh (19860127-8859)
4ME312 Advanced Topics in Media Technology—FT2020

September 16, 2020

Abstract

1 Overview

We were presented with results of a usability testing of two systems. The data set provided included 82 observations and 13 variables. These variables could be divided into three different sets:

1. Self-reporting of user demographics (field and experience)
2. usability and performance evaluation
3. session data (order, task score and time)

The first set of variables provides description of the field and experience of the user with similar systems and System A in particular. The *job* vector is categorical with three different possibilities: industry, academia, or other. The *experience* vector is a Likert scale data that spans between 1 (no experience) to 7 (substantial). information on the field of the user. The variable *used_A* is a Boolean that describes whether the user has experience with System A, in particular.

The second set of variables automatically logged by the system. The order by which the systems in comparison are presented to the participants is logged in characters AB or BA. The system also logs the time in seconds it took users to finish the task *A_time* and *B_time*, for System A and System B respectively. Additionally, the system calculates an automatic score as an integer for user's performance in System A and System B, as *A_score* and *B_score* respectively. However, the score scale was not clarified. We assume that the score is out of 100.

The third set of variables are the user's self-reported evaluation during the session. the usability and performance of the system is reported on a fiver-point Likert scale for both A and B systems, in *A_use*, *B_use*, *A_perf* and *B_perf.* The B system Usability score, however, was inverted, 1 good to 5 bad. We processed the vector to match the other scores order, 1 bad to 5 good.

Finally, each observation in the data set is uniquely identified by the *timestamp* variable.

2 Methodology

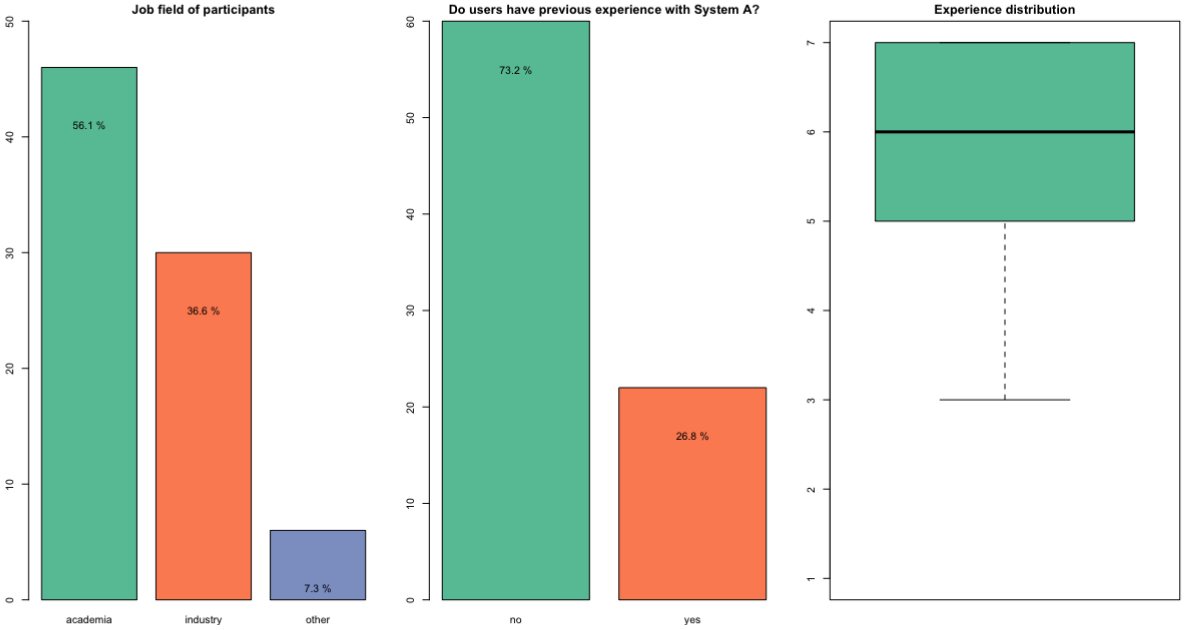
We used R[1] to conduct the data analysis and plot the graphs. We started by reading the CSV file and manually observing the data set to note any possible issues. We then, conducted a basic descriptive analysis of our data set. We used bar plots and box plots to understand our data. We also tried to understand the data in relation to demographics.

We analyzed the date by dividing them into three sets as it was described in section 1. Thus, we tried to connect the system logs and evaluation scores to demographics.

Thereafter, we conducted an inferential analysis on the scores that was logged automatically by the system. We conducted a paired t-test on the scores of systems A and B.

3 Descriptive Analysis

We conducted an analysis of the participants demographics. As we see in Figure 1, More than half the users participating in the evaluation works in Academia. Only a small number, 7.3%, have a job outside Academia or Industry. Thus, we could define two distinct categories throughout our analysis, academics and non-academics. The users self-reported a high-level of experience with similar systems with a median of 6, as seen in the distribution in the box plot. 75% of the users claimed an experience of 5 in a 7-point scale. This is explained by a self-reporting of experience in System A by 73.2% of the users.


Job field of participants

Do users have previous experience with System A?

Experience distribution

Figure 1: Description of the participants background

A general view of the evaluation results, seen in Figure 2, shows us a larger distribution of scores for System A. Despite, a similar median, System B's mean score (M = 3.541, SD =

[1] Code is attached with the exam submission.

.52) was larger than System A (M = 2.78, SD = 0.63). To get a clearer picture, we plotted the difference between usability score given by participants to System A and B (refer to Figure 4).

We notice that 25% of the participants did not think System B was better in terms of Usability. This could be related to the demographics of the participants (see Appendix B). We notice that half the academics did not think that System B was better than System A, in comparison to 75% of the workers in the industry who though System B was better. Similarly, 75% of participants who evaluated System B second, gave it a better score for usability. None of the participants, who had previous experience with System A, thought that System A is better than B in terms of Usability.

It is a different picture for the performance score that received poorer score for System B. We did not notice any significant difference in terms of demographics (see Appendix C). and Figure 3

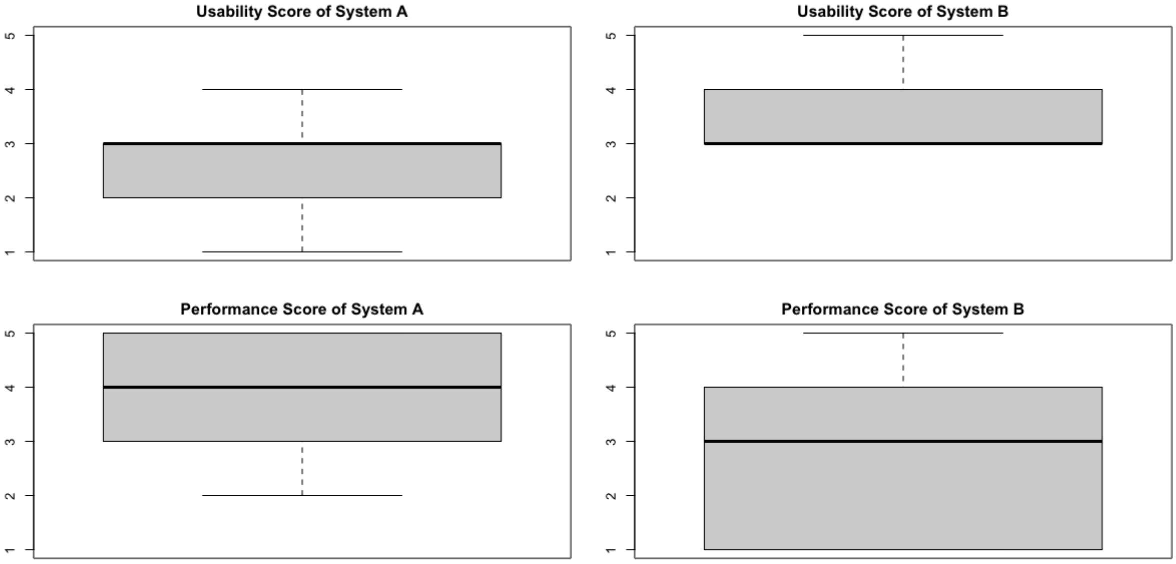


Figure 2: Distribution of evaluation scores reported by participants.

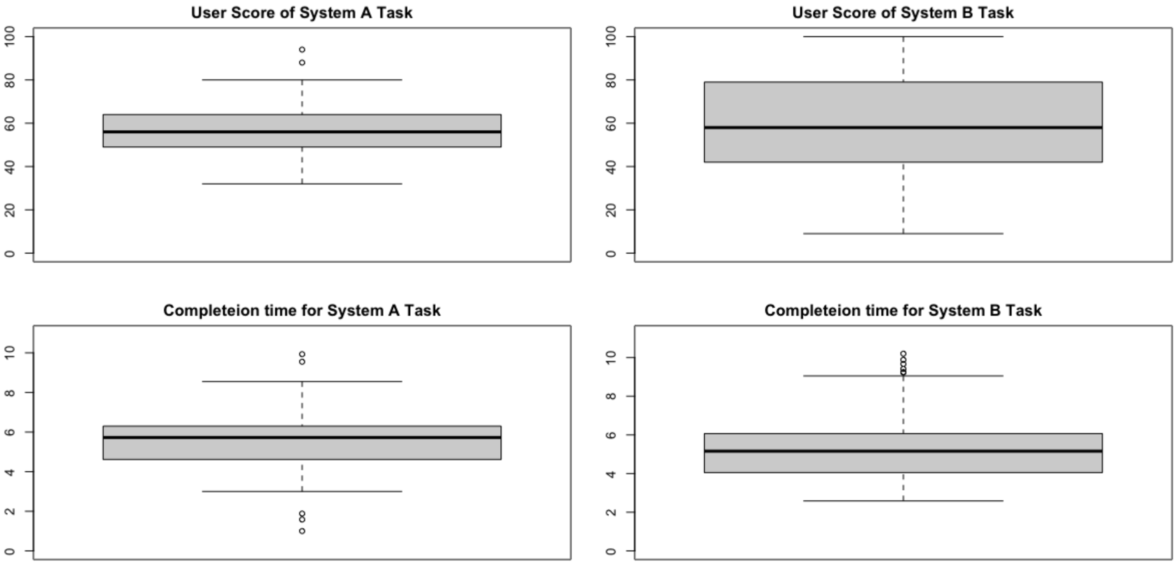


Figure 3: Distribution of automatic system logs of sessions.

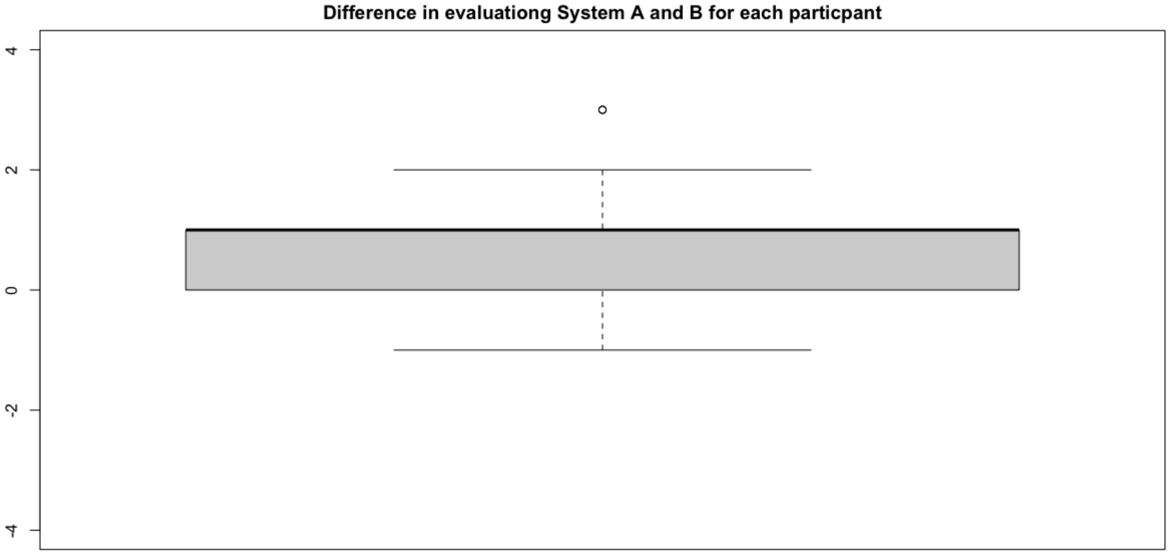


Figure 4: Difference in usability score given by participants between System A and System B.

As seen in Figure 5, we can see that users scored higher in System A task when they had previous experience with the system. However, we do not know how experienced were the users with System A. In the second row, we can notice that previous experience with similar systems did not affect participants scores. This is similar to the pattern we see in the scores of Task B based particularly on previous experience with System A. Furthermore, despite the difference in the score variance based on jobs, we cannot detect substantial differences in other aspects.

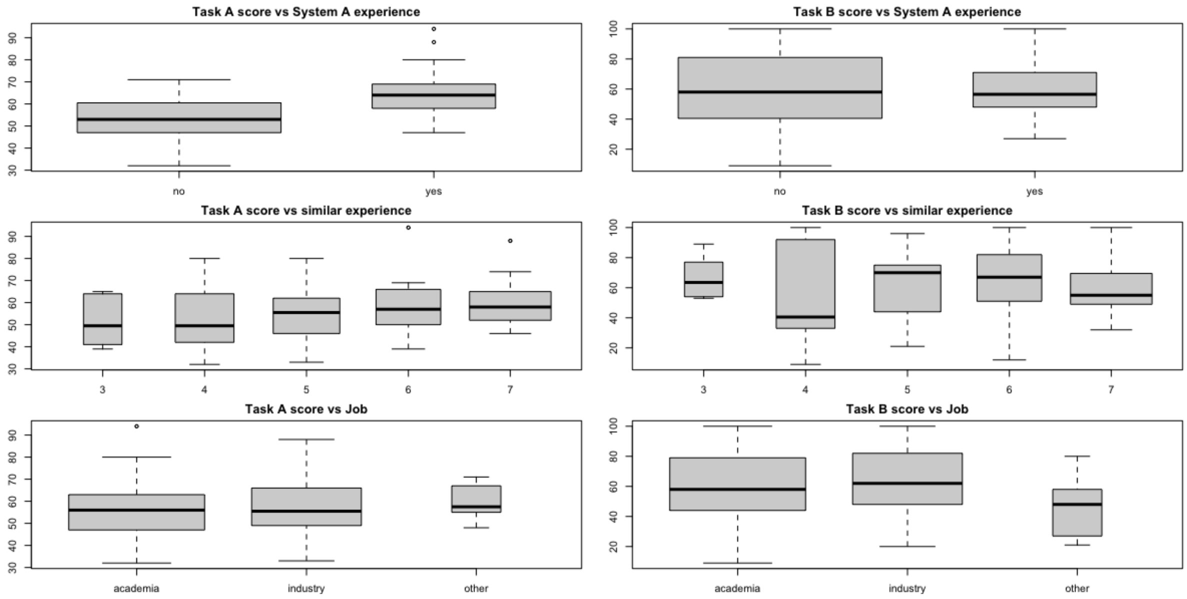


Figure 5: Description of how users scored on tasks A and B based on demographics.

In Figure 6, we notice a high variance in the time it took academics to finish both task A and task B. We notice couple of outliers who display more proficiency. In completing task A, the gap in proficiency between users who work in the industry is much less. The variance increases in task B.

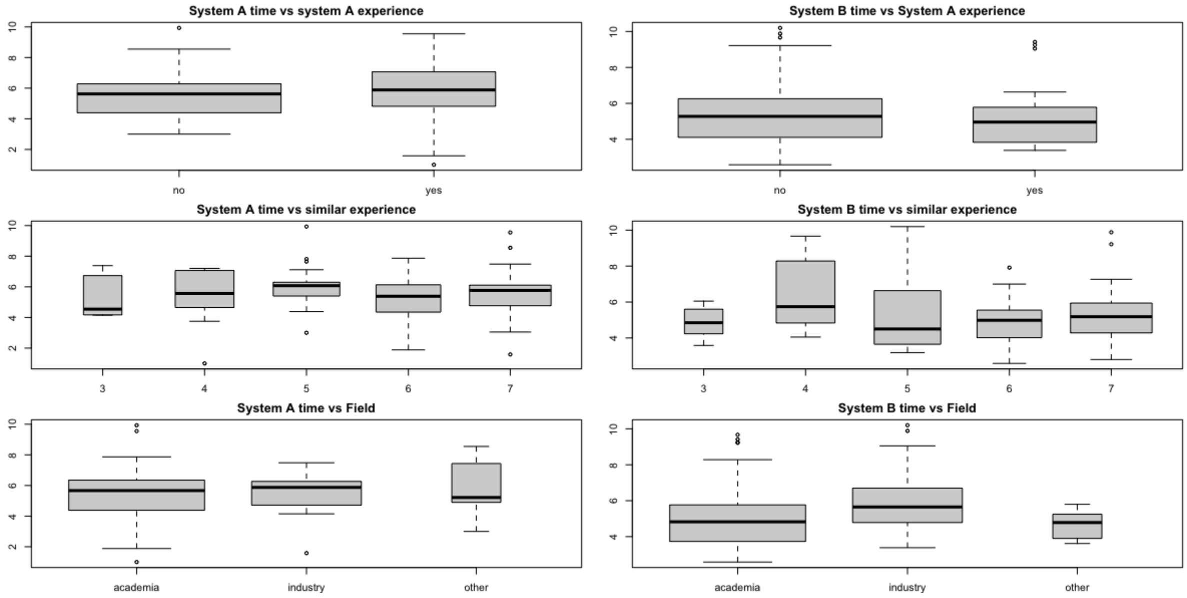


Figure 6: Description of how long (in minutes) it took users to finish tasks A and B based on demographics.

Figure 7 shows a high variance in how users evaluated the performance in the system. We notice a higher performance score for System A, especially amongst users who have previous experience. The performance score of system A amongst users in academia and the industry is very similar in it distribution. However, evaluation of the performance of System B plummets more amongst users who work in the industry.

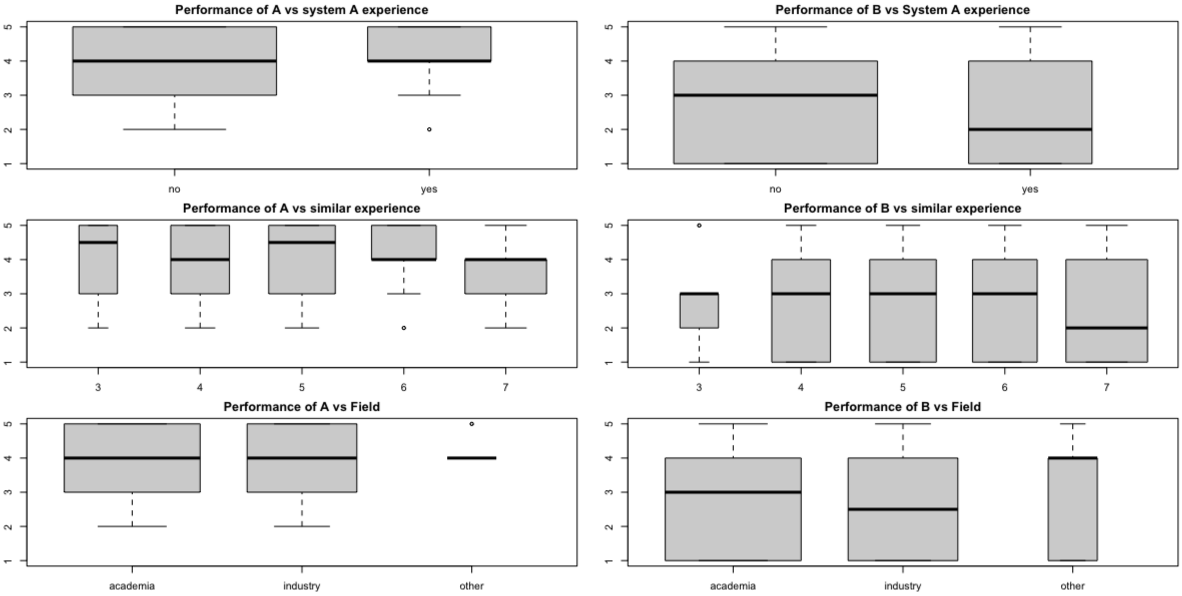


Figure 7: Description of how users evaluated performance of Systems A and B based on demographics.

In Figure 8, we see a substantial difference in the distribution of usability scores. Users with previous experience using System A are homogeneous in their evaluation of the usability of both systems. However, experiences with similar systems does not have a big impact. Similarly, the field of users jobs did not impact their evaluation of the systems. Generally, both systems were evaluated equally in terms of usability.

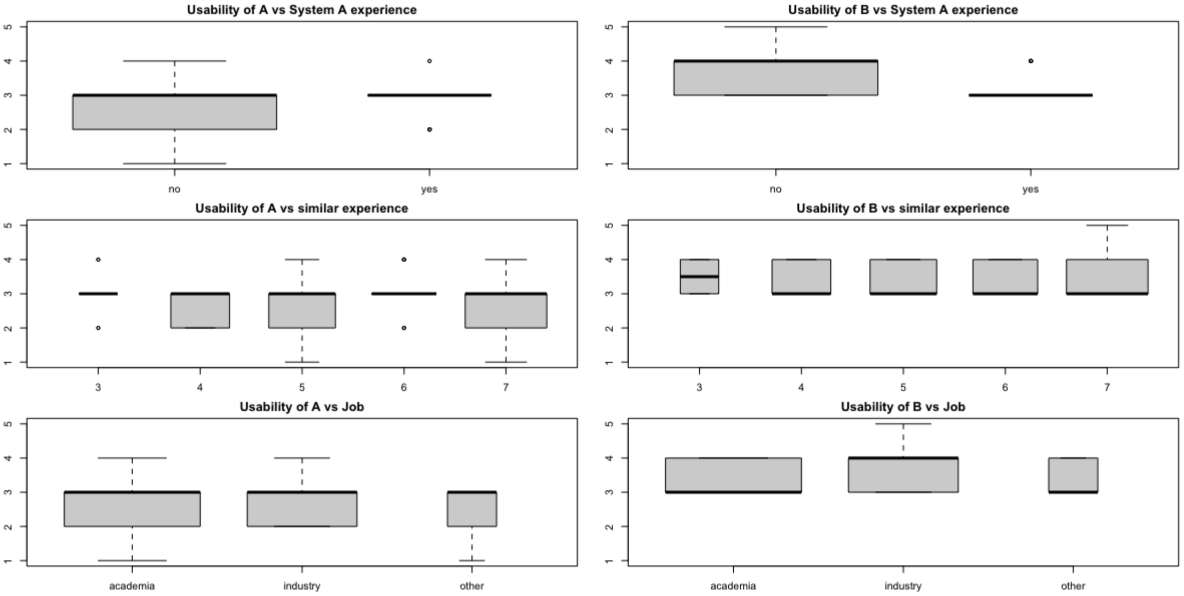


Figure 8: Description of how users evaluated performance of Systems A and B based on demographics.

4 Inferential Analysis

We checked for the normality of the scores (see Appendix A, given by the system for both tasks A and B. We conducted a Shapiro-Wilks Test and plotted a histogram. We set the significance rate at 10%. Although the data seems to be normally distributed for scores of A, it is not significant enough. Data of scores of Task B does not seem to be normally distributed.

Thus, we decided to conduct a paired t-test. Our alternative hypothesis H_a is that the means of both score sets are different. We set the significance value at .05. We rejected the H_a as the p-value reported is .13 with df = 81. Thus, we do not have any evidence that there is a significant difference int he mean of the two sets.

We conducted a Chi-square test to see if there is any relation between users who have previous experience with System A and the difference in their usability evaluation between system A and B. This aims to understand if experience plays a role worth investigating. We set the null hypothesis as there is no significant difference and we set the significance value at .05. We rejected the null hypothesis and accepted the alternative hypothesis. There's a possible relation between experience with System A and difference in usability evaluation X^2 (N = 83) = 9.70, p = .045.

5 Discussion

Although participant numbers could not give us indication on how users job influencing their mastering of the systems,this is a matter to be investigated further. Nonetheless, based on the data we acquire, we notice that System B might have been modified to meet the skills of users who work in the industry. This could be supported by the results in Figures 10.

The increase in the Usability evaluation is connected to the job field and previous experience with System A. A possible explanation is that users in the industry have more experience with the technology. However, in terms of performance, users believed that System A is better. It is worth investigating more whether this is related to user expectations.

We notice a gap between developers evaluation of the system and users expectation. We recommend that developers conduct deeper studies to understand the expectations of their target users, especially in academia. Meanwhile, we recommend that continuous deployment of System A and the disregard of System B.

A Appendix 1: Histogram of system scores of Task A and Task B

* We used an R code from the example codes provided by course co-ordinator Aris Alissandrakis.

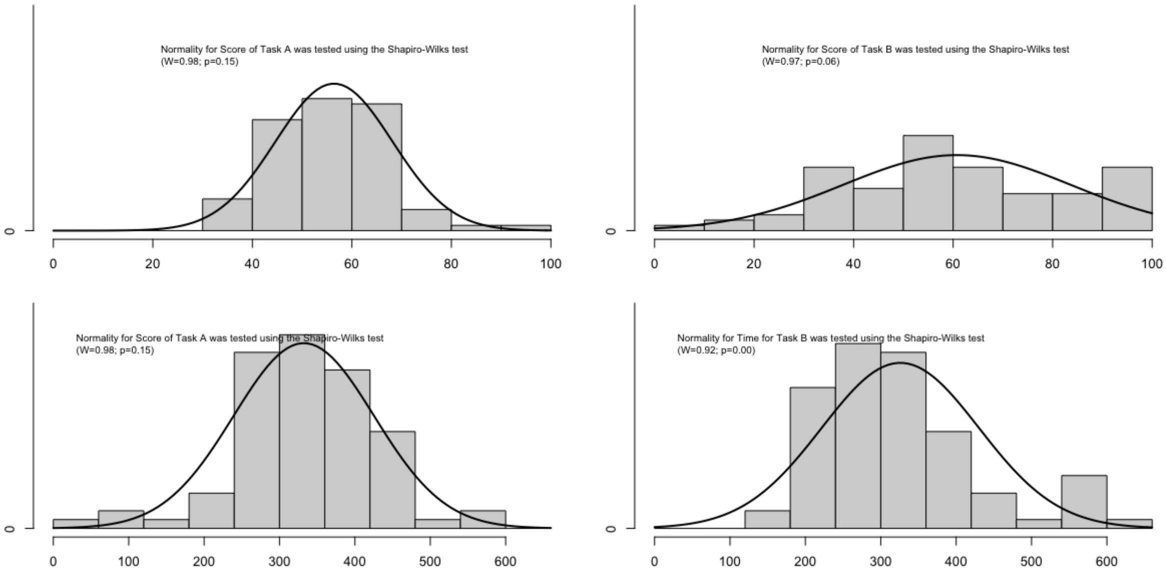


Figure 9: Normality test for scores of task A and task B.

B Appendix 2: Difference in Usability Rating between System A and System B (B_use - A_use) by demographics and order of evaluation.

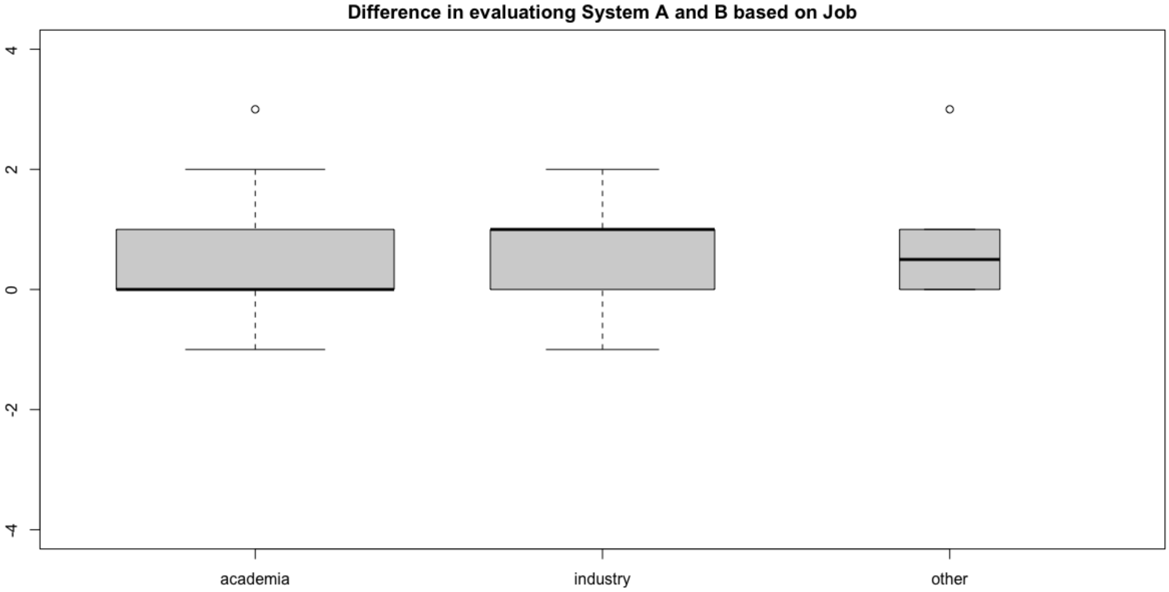


Figure 10: Difference in Usability Score by Job.

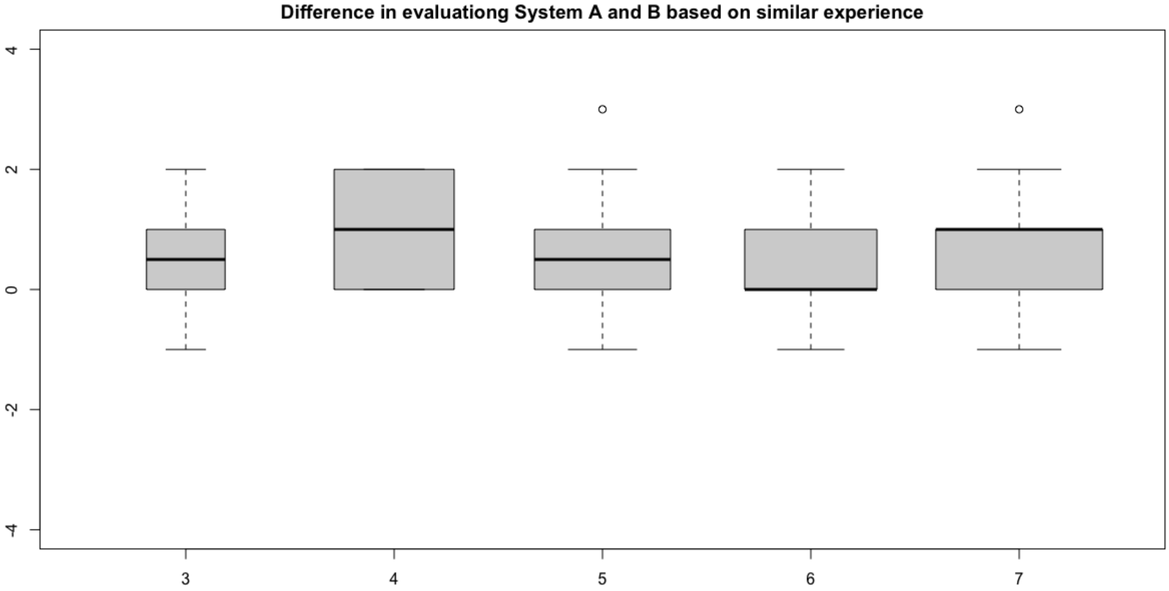


Figure 11: Difference in Usability Score by Similar Experience.

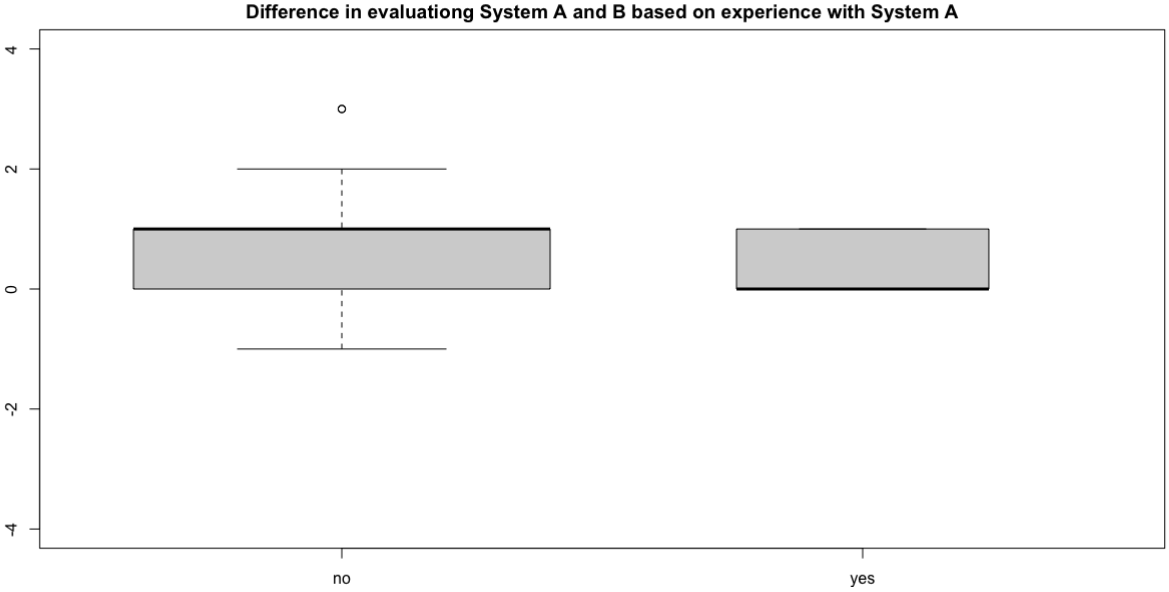


Figure 12: Difference in Usability Score by Experience with System A.

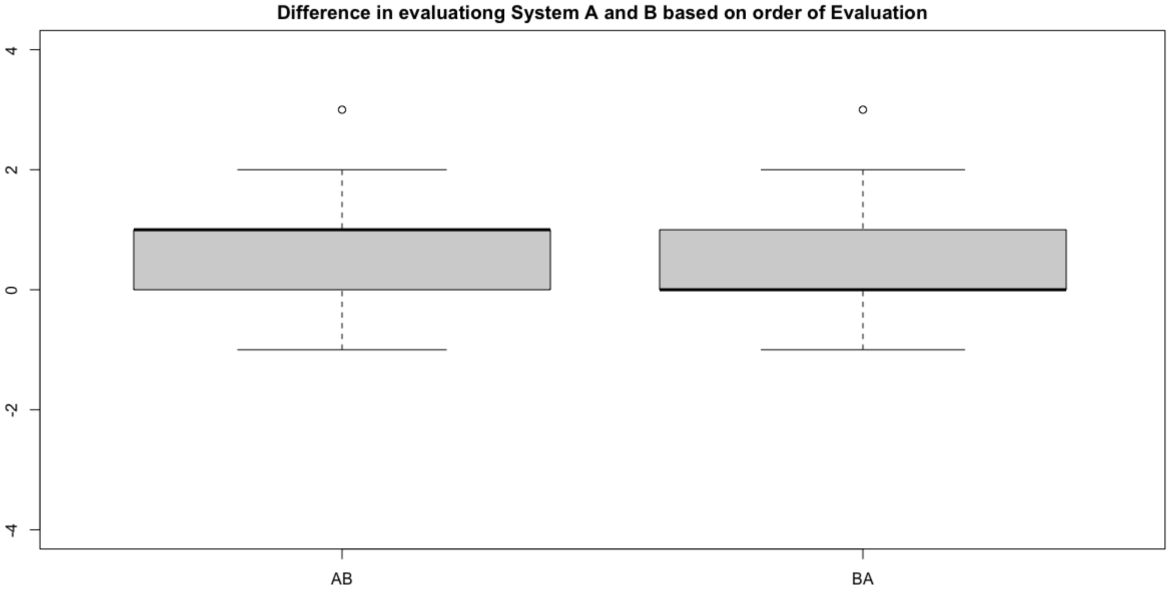


Figure 13: Difference in Usability Score by Experience by order of Evaluation.

C Appendix 3: Difference in Performance Rating between System A and System B (B_perf - A_perf) by demographics and order of evaluation.

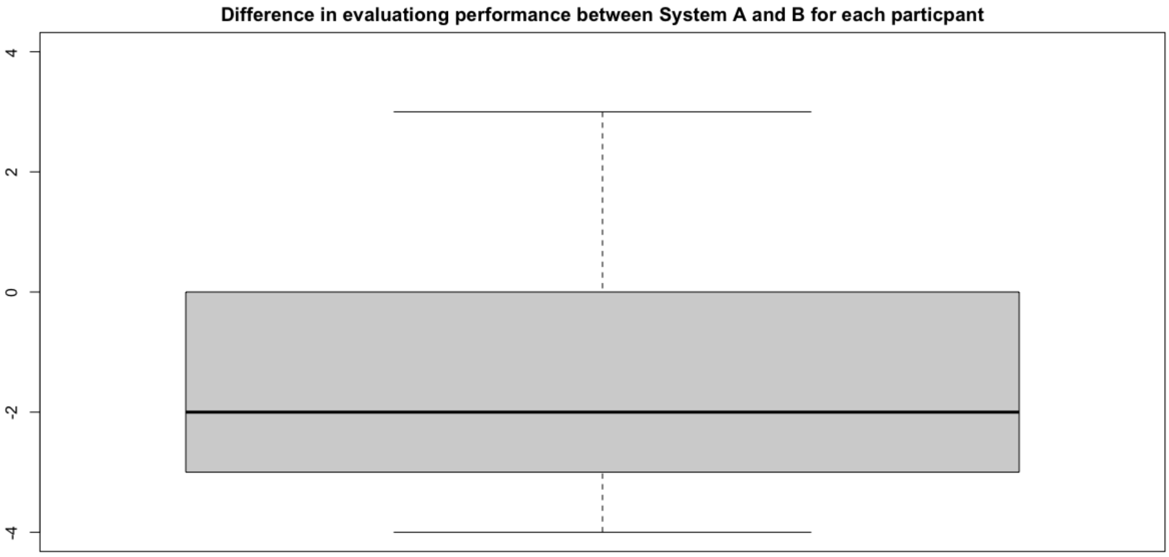


Figure 14: Difference in Performance Score.

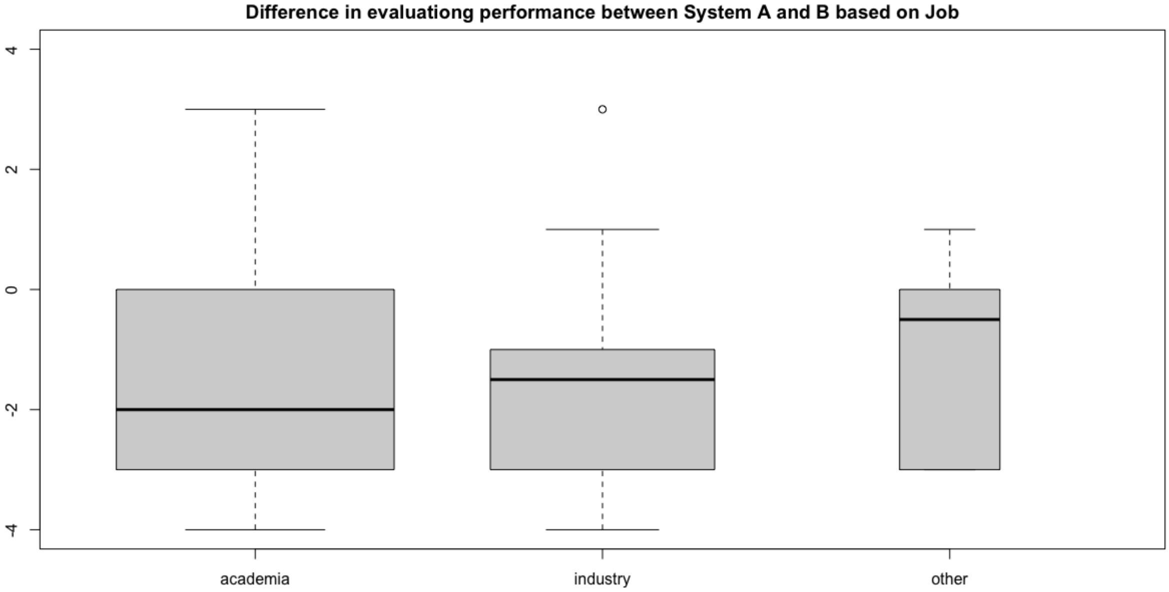


Figure 15: Difference in Performance Score by Job.

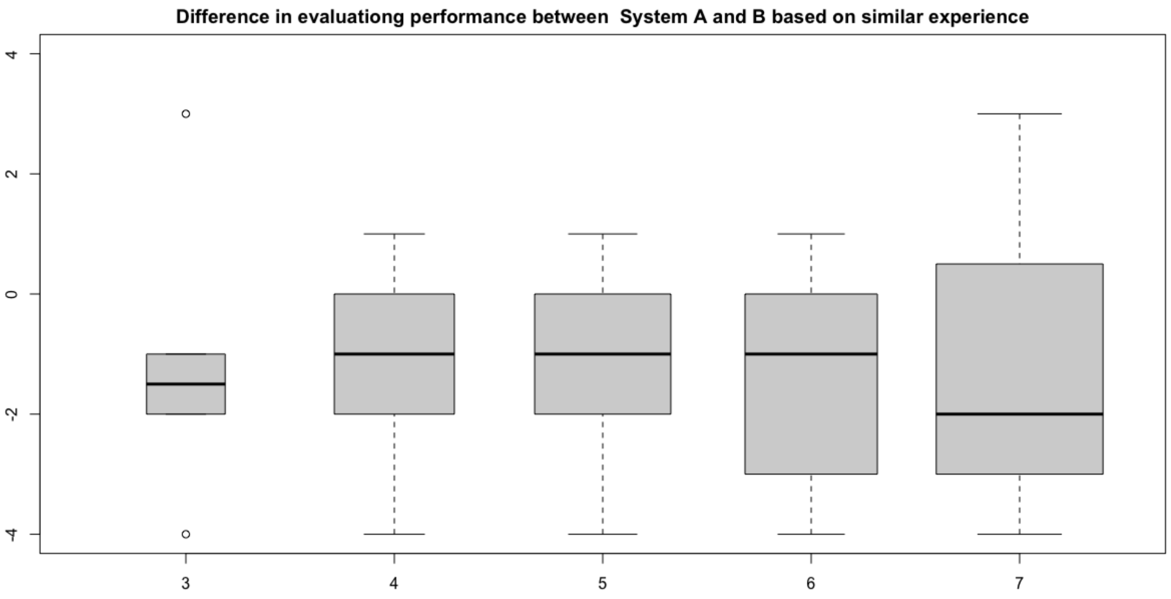


Figure 16: Difference in Performance Score by Similar Experience.

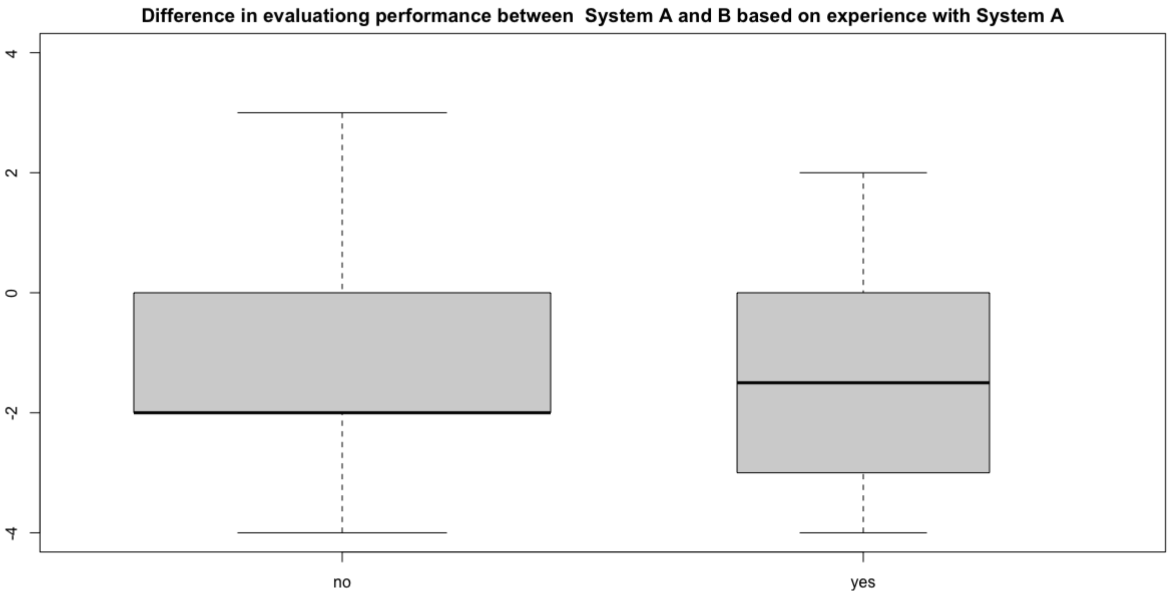


Figure 17: Difference in Performance Score by Experience with System A.

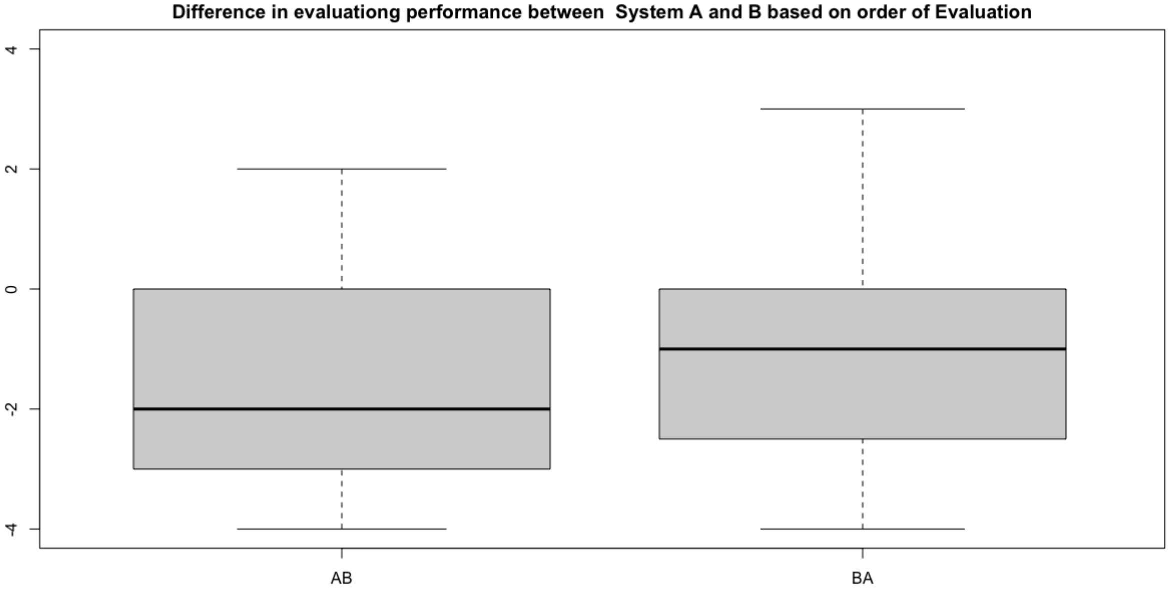


Figure 18: Difference in Performance Score by Experience by order of Evaluation.