1. SUS Questionnaire Outcomes

Figure 1 presents the outcomes of the System Usability Scale (SUS) questionnaire collected from 10 participants after they interacted with both versions of the system: the version that adheres to Nielsen's heuristics and the other that deliberately does not. The figure indicates that participants with the identification number 01, 02, 03, 04, and 05, representing Group 01, tested the more advanced application, while participants with the identification number 06, 07, 08, 09, and 10, representing Group 02, interacted with the basic application. Each participant answered the questionnaire by indicating their level of agreement or disagreement with each of the 10 statements using a 5-point Likert scale, where 1 means "strongly disagree" and 5 means "strongly agree".

The questionnaire statements address various usability aspects including the level of the application's complexity and the training and assistance required for users, which makes it a reliable and valid usability assessment tool. However, as the score for each statement is not valuable on its own, a SUS score for every participant must be calculated. This value is out of 100 and it represents the overall usability of the system.

To determine the SUS score, a total score needs to be calculated first. This is done by adding up the contribution of each statement. The contribution of the odd-numbered statements is obtained by subtracting 1 from the score, whereas the even-numbered statements is by subtracting the score from 5. The SUS score is then obtained by multiplying the total score by 2.5. Afterwards, this score indicates whether users are satisfied or dissatisfied with the system's usability: a score of 80.3 or more suggests a very good usability, a score of 51 or less indicates potential usability issues that need to be addressed, and a score around 68 which is the average, reflects that the system's usability is considered acceptable but could be enhanced for a better user experience.

Application Tested	Participant Identification Number	SUS Statement Scores											
		St1	St2	St3	St4	St5	St6	St7	St8	St9	St10	Total Score	SUS Score
The application that follows Nielsen's heuristics	01	4	1	5	2	4	2	3	2	5	1	33	82.5
	02	4	3	4	3	5	1	5	1	5	1	34	85
	03	5	1	5	1	4	1	4	1	5	2	37	92.5
	04	5	1	5	1	5	2	4	1	5	3	36	90
	05	5	1	5	1	5	1	5	1	5	1	40	100
The application that doesn't follow Nielsen's heuristics	06	3	5	1	4	2	4	3	4	1	4	9	22.5
	07	2	5	1	3	1	4	2	5	1	4	6	15
	08	1	5	2	5	1	5	1	4	2	4	4	10
	09	1	5	3	5	1	4	3	5	1	5	5	12.5
	10	2	4	3	5	3	4	3	4	1	4	11	27.5

Figure 1. SUS Questionnaire Outcomes

Based on the feedback gathered from the 10 participants presented in Figure 1, it can be seen that there is a contrast in the SUS scores obtained from the 5 users who tested the application that follows Nielsen's heuristics and the 5 users who tested the poor usability application. The SUS scores received from the first 5 participants were 82.5, 85, 92.5, 90, and 100, all of which exceed 80.3. This indicates that users were satisfied with the application and found it intuitive and easy-to-use. In contrast, the SUS scores received from the second 5 participants were 22.5, 15, 10, 12.5, and 27.5, all of which fall below 51. This reflects that users were dissatisfied and less likely to use the application again. Overall, these results indicate that Nielsen's 10 usability heuristics play an important role in designing usable applications.

Furthermore, to have an overview of the data collected from both groups and to better understand its distribution, the descriptive statistics illustrated in Figure 2 include the mean, minimum, maximum and

standard deviation for group 01 and group 02. Afterwords, the box plot of Figure 3 was created to visually represent this distribution.

SUS_Score

Group	Mean	N	Std. Deviation	Minimum	Maximum
Group01	90.0000	5	6.84653	82.50	100.00
Group02	17.5000	5	7.28869	10.00	27.50

Figure 2. Descriptive statistics of Group 01 and Group 02

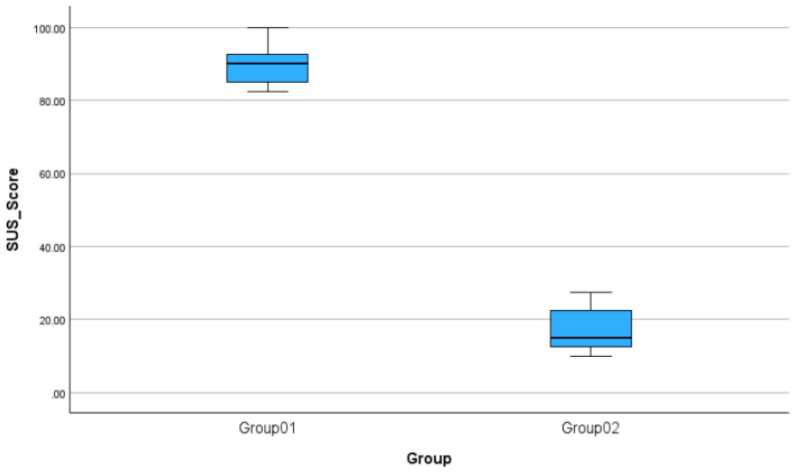

Figure 3. Box plot explaining the difference of SUS scores received from both groups

The box plot clearly indicates that the SUS scores received from both groups are different, with the scores from the first group being higher than those from the second group.

For the first group, which was composed of 5 samples (N =5), the mean score was 90. The SUS scores received from this group range from 82.5 to 100, with a standard deviation of 6.847. This deviation indicates that while participants were generally satisfied with the application, there were variations in their opinions about its usability.

The second group, on the other hand, which was also composed of 5 participants, recorded a mean of 17.5 with 10 as the minimum SUS score and 27.5 as the maximum SUS score. The standard deviation of this group was 7.289, which is slightly higher than the first group. This indicates that participants were generally confused about the application, with some of them found it more difficult to use than others.

2. Mann-Whitney U Test for SUS scores

While the SUS scores reflected the participants' approval or disapproval of the applications' usability, they didn't confirm the difference between the two applications to see whether the one that follows Nielsen's heuristics is significantly more usable than the basic version, thus, the Mann-Whitney statistical test was used for this project to test this difference.

In this test, the Z-value indicates the difference between both samples in terms of standard deviations, whereas the p-value shows how likely the difference can happen by chance. The two samples for this study were the group of participants who tested the application that follows Nielsen's heuristics and the group of those interacted with the poorly designed application. The generated Mann-Whitney test is illustrated in Figure 4.

Since all participants representing the second group recorded the lowest SUS scores, the rank sum was 15, while for the first group was 40. This was obtained by sorting all scores from the smallest to the largest value then assigning ranks to them and adding up the ranks for each group. Additionally, the mean rank values indicate that the scores of group 02 are much lower compared to those of group 01. The data presented in the test statistics table of Figure 4 show that the U-value was 0, the asymptotic p-value (2-tailed) was 0.009. the exact p-value (2-tailed) was 0.008 and the Z value was -2.611. Based on the Z value which is more extreme than the critical value (1.96), and the corresponding p-values, which are both less than the significance level (0.05), the null hypothesis was rejected and so there is a significant difference between both groups. Furthermore, as group 01 had a mean rank of 8 which is higher than that of group 02, therefore the application that follows Nielsen's heuristics is significantly more usable than the other application that does not.

Mann-Whitney Test

Ranks

	Group	N	Mean Rank	Sum of Ranks
SUS_Score	Group01	5	8.00	40.00
	Group02	5	3.00	15.00
	Total	10		

Test Statistics[a]

	SUS_Score
Mann-Whitney U	.000
Wilcoxon W	15.000
Z	-2.611
Asymp. Sig. (2-tailed)	.009
Exact Sig. [2*(1-tailed Sig.)]	.008[b]

a. Grouping Variable: Group

b. Not corrected for ties.

Figure 4. The Mann-Whitney test generated to test the difference between Group 01 and Group 02

3. Mann-Whitney U Test for each statement of the questionnaire

The SUS questionnaire is an assessment tool that provides an overview of a system's usability. However, it does not specify the areas of the user interface that require improvement. Analysing the questionnaire statements, on the other hand, can offer insights about how participants found different aspects of the application tested. As a result, the Mann-Whitney U test was generated for each statement of the questionnaire to identify the factors that contributed to the overall difference between the group of participants who tested the application adhering to Nielsen's heuristics and those who tested the application with poor usability.

The results of the Mann-Whitney test displayed in Figure 5 show that the SUS questionnaire statements 01, 03, 05, 07, and 09 which reflect positive usability aspects and an overall good user experience, had higher mean rank values in group 01 (8, 8, 8, 7.70, and 8 respectively) than group 02 (3, 3, 3, 3.30, and 3 respectively). This indicates that the first group who tested the advanced version of the application rated those statements more positively to reflect their satisfaction with the usability rather than the group who tested the basic version.

The questionnaire statements 02, 04, 06, 08, and 10, which suggest problems with a system's usability such us the difficulty of use, had higher mean rank values in group 2 (8, 7.90, 8, 8, and 8 respectively) than group 01 (3, 3.10, 3, 3, and 3 respectively). This indicates that participants of the second group

agreed with those statement more than participants of the first group to reflect their dissatisfaction with the usability of the application they tested.

Regarding the test statistics table presented in Figure 5, each statement had a Z-value below -2. Additionally, the asymptotic p-value (2-tailed) for statement 01, 02, 03, 04, 05, 06, 07, 08, 09, and 10 were 0.008, 0.005, 0.007, 0.010, 0.007, 0.006, 0.017, 0.006, 0.004, and 0.006 respectively. All statements had an exact p-value of 0.008, except for statement 07, which had 0.016. Both the asymptotic and the exact p-values for each statement are lower than the significance level of 0.05, thus, there is a significant difference between the first and the second group.

The nineth statement "I felt very confident using the system" had the smallest p-value and so the most significant difference between the groups. This highlights that users felt much more confident using the application that follows Nielsen's heuristics compared to those who tested the other application, who did not feel confident at all. Hence, these guidelines have a big impact on user confidence and the experience with the system.

However, while the p-value of the seventh statement indicates that there is a statistically significant difference, it is the highest p-value amongst all statements. This means that both groups had similar responses for the statement. Furthermore, based on the SUS questionnaire outcomes illustrated in Figure 1, it is shown that 4 out of 10 participants selected the midpoint of the scale (3) for statement 07, which means a neutral response. This may be because the aspects of applications related to the ease of learning did not differ too much between the good and bad applications, therefore all participants were a bit confused whether other users will learn to use the applications quickly or not.

Ranks

	Group	N	Mean Rank	Sum of Ranks
Statement01	Group 01	5	8.00	40.00
	Group 02	5	3.00	15.00
	Total	10		
Statement02	Group 01	5	3.00	15.00
	Group 02	5	8.00	40.00
	Total	10		
Statement03	Group 01	5	8.00	40.00
	Group 02	5	3.00	15.00
	Total	10		
Statement04	Group 01	5	3.10	15.50
	Group 02	5	7.90	39.50
	Total	10		
Statement05	Group 01	5	8.00	40.00
	Group 02	5	3.00	15.00
	Total	10		
Statement06	Group 01	5	3.00	15.00
	Group 02	5	8.00	40.00
	Total	10		
Statement07	Group 01	5	7.70	38.50
	Group 02	5	3.30	16.50
	Total	10		
Statement08	Group 01	5	3.00	15.00
	Group 02	5	8.00	40.00
	Total	10		
Statement09	Group 01	5	8.00	40.00
	Group 02	5	3.00	15.00
	Total	10		
Statement10	Group 01	5	3.00	15.00
	Group 02	5	8.00	40.00
	Total	10		

Test Statistics[a]

	Statement01	Statement02	Statement03	Statement04	Statement05	Statement06	Statement07	Statement08	Statement09	Statement10
Mann-Whitney U	.000	.000	.000	.500	.000	.000	1.500	.000	.000	.000
Wilcoxon W	15.000	15.000	15.000	15.500	15.000	15.000	16.500	15.000	15.000	15.000
Z	-2.668	-2.785	-2.712	-2.578	-2.685	-2.739	-2.386	-2.739	-2.887	-2.730
Asymp. Sig. (2-tailed)	.008	.005	.007	.010	.007	.006	.017	.006	.004	.006
Exact Sig. [2*(1-tailed Sig.)]	.008[b]	.008[b]	.008[b]	.008[b]	.008[b]	.008[b]	.016[b]	.008[b]	.008[b]	.008[b]

a. Grouping Variable: Group
b. Not corrected for ties.

Figure 5. The Mann-Whitney test generated for each statement of the SUS questionnaire

4. Experiment Evaluation

To summarise, the results of the p-value and Z-value obtained from the Mann-Whitney U test generated for the SUS scores, which were 0.009 and -2.611 respectively, along with a higher mean rank (8) for the group of participants who tested the application that follows Nielsen's heuristics compared to those who tested the application that does not (3), strongly indicate that Nielsen's 10 usability heuristics are highly effective to design user friendly applications and play an important role in contributing to an overall good user experience. Moreover, the statistically significant differences between the first and the second group, when each of the questionnaire statements was analysed separately, underscore the importance of these heuristics for user satisfaction.