Chapter 7

Evaluation: Results and Numerical Analysis

7.1 Introduction

In previous chapter, we have described ~~in depth~~ user study design including questionnaire presentation, data collection procedure, data structure, data storing mechanism, ~~etc~~. In this chapter, we discuss and analyze the study generated data with the help ~~different~~ statistical ~~algorithms and~~ principles which are commonly used for user studies such t-test, and ANNOVA. The goal of the study was to evaluate user performance and user experience of our newly designed approach of uncertainty visualisation and generate quantitative data. We will use that ~~quantitative~~ data in analysis, prepare results, eventually discuss the findings, and finally come to conclusion.

7.2.1 Sample Population Demographics

Age and Gender

The sample population of 32 participants had a distribution of 78.12% male (25/32), 21.88% female (7/32). Given that we ~~didn't~~ have ~~any~~ plan to control gender within the recruitment policy, we have recruited on ~~the~~ first come first join basis. All participants were in the age range of 22-35 years old.

Education

There were 25% CS grad students (8/32), 28.12% CS undergrad students (9/32), 34.37% ICT grad students (11/32), 3% Statistics undergrad students (1/32) and 9.37 telecom professionals (3/32).

Prior experience in visualisation

- All CS and ICT students had taken at least one course of visualisation/graphics design in their undergraduate/graduate level and 12 of them had conducted their undergraduate thesis related to visualization or graphics or image processing.
- Telecom professionals also came from CS background, so they had taken Computer graphics course in their undergraduate level.
- All participants had played computer games ~~so~~ many times.

- 15 participants have knowledge about animated movies.
- Every participant came from science background and that's why they have very good geometric knowledge such as identifying bubble, square, grid, and knowledge about measurement of thickness of objects.

7.2 Study results

We have obtained several kinds of data from user study such as:

i. Quantitative Questionnaire Results.
ii. Time utilization data for each component.
iii. SUS data for CA and VSUP.
iv. NASA-TLX for CA and VSUP.

We analyse all these data in various ways in the following sections which helps to reach ~~out~~ in conclusion from the study.

7.2.1 Quantitative Questionnaire Results

As we have four core components, we designed study content for each component individually and collected the log data for each component separately. We already stated, there were 8 questions for each component and every question carried 1 point. For answering correctly, the participant gains one point and do no lose/~~carry~~ any point for wrong answers. So, a participant can gain minimum 0 point and maximum 8 points for a component. That point achievement is considered as user performance of the study and we are going to analyse the user performance on the basis of ANOVA for four components and t-test for two grouped (CA and VSUP) components.

7.2.1.1 One-way repeated measures ANOVA

The user performance results that we received from the study can be summarized as Table 7.1 and the complete raw data is attached in APPENDIX-L.

Groups	N	Mean	Std. Dev.	Variance	Std. Error.
CA + Bubble	32	6.2813	1.301	1.692	0.23
CA + Grid	32	5.5938	1.2916	1.668	0.2283
VSUP + Bubble	32	5.6563	1.4053	1.975	0.2127
VSUP + Grid	32	5.1875	1.2032	1.456	0.2127

Table 7.1: Data summary

The results of a one-way ANOVA can be considered reliable if the following assumptions are met:

1. the response variable (the dependent variable) is normally distributed.
2. the samples are independent.
3. the variances of populations are equal.

Since the sample are taken from independent interfaces of questionnaire, the requirement (2) fulfilled. Again, from Table 7.1, we see that variances are equal which conforms condition (3). Now Shapiro-Wilk Test of Normality in Table 7.2 shows that the distributions of each component significantly departed from the normality which dissatisfies the requirement (1). Again, from Figure 7.1 box plots, we see that the distributions are not normal since there sizes are significantly varying. Next, we check the Kolmogorov-Smirnov Test of Normality in Table 7.3, and it shows the P and D values are small, hence distributions do not differ significantly from that which is normally distributed. That means it met the requirement (1).

Component	W	P	Status
Ca + Bubble	.915	.015	Significant departure from the normality.
Ca + Grid	.932	.045	Significant departure from the normality.
VSUP + Bubble	.911	.012	Significant departure from the normality.
VSUP + Grid	.913	.013	Significant departure from the normality.

Table 7.2: Shapiro-Wilk Test of Normality

Component	Skewness	Kurtosis	P-Value	D-Value	Status
Ca + Bubble	-0.562275	-0.055645	0.2151	0.18146	Normal
Ca + Grid	0.066708	-0.785852	0.22398	0.17976	Normal
VSUP + Bubble	-0.155834	-0.985813	0.20936	0.10434	Normal
VSUP + Grid	-0.147245	-0.733659	0.34901	0.15993	Normal

Table 7.3: Kolmogorov-Smirnov Test of Normality

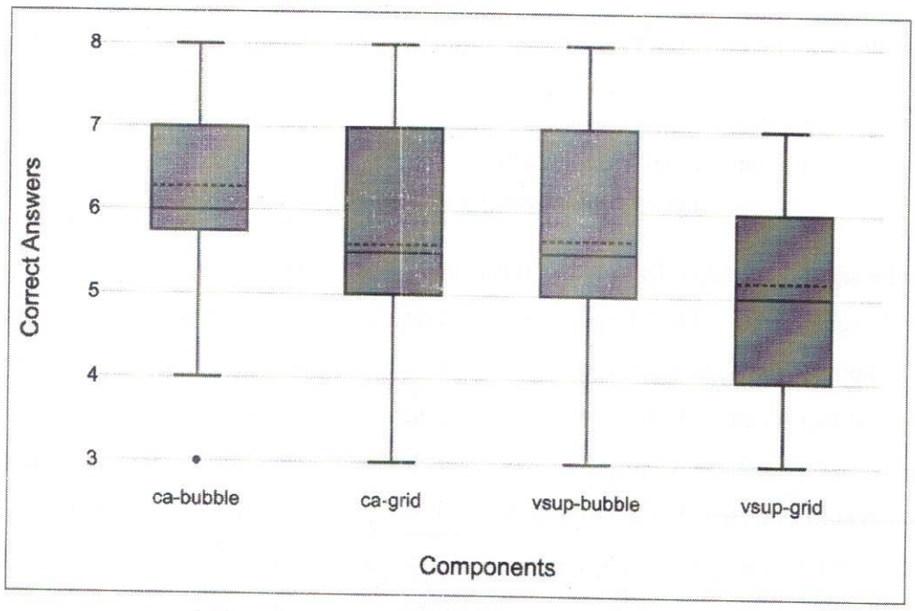


Figure 7.1: Box plot of user performance

Non-parametric Test:

To evaluate the difference of user experience across four components, we used the non-parametric Kruskal-Wallis test (α =0.05). The H statistic is found 10.2365 (3, N = 128). The p-value is .01666. So, the result is statistically significant at $p < .05$.

Parametric Test:

We get the ANOVA summary as Table 7.4.

Source	Degrees of Freedom DF	Sum of Squares SS	Mean Square MS	F-Stat	P-Value
Between Groups	3	19.5875	6.5292	3.8499	0.0113
Within Groups	124	210.2851	1.6958		
Total	127	229.8726			

Table 7.4: ANOVA Summary

So, we briefly point out the findings from the ANOVA test as follows:

(1) Null and Alternative Hypotheses

The following null and alternative hypotheses need to be tested:

Ho: $\mu 1 = \mu 2 = \mu 3 = \mu 4$ (Performances were equal for all components)

Ha: Not all means are equal (Performances were not equal for all components)

The above hypotheses will be tested using an F-ratio for a One-Way ANOVA.

(2) Rejection Region

Based on the information provided, the significance level is α=0.05, and the degrees of freedom are $df1$=3 and $df2$=3, therefore, the rejection region for this F-test is R = R = {F: F > 2.678}.

(3) Test Statistics

The computed test statistic F equals 3.8499, which is not in the 95% region of acceptance: [-∞: 2.678].

(4) Decision about the null hypothesis

p-value equals 0.0113, [p (x ≤ F) = 0.988735]. It means that the chance of type1 error (rejecting a correct H0) is small: 0.01126 (1.13%). The smaller the p-value the stronger it supports H1. Again, from the sample information we get that F = 3.85 > Fc=2.678, it is then concluded that *the null hypothesis is rejected.*

(5) Conclusion

It is concluded that the null hypothesis Ho *is rejected.* Therefore, there is not enough evidence to claim that not all 4-population means are equal, at the α=*0.05* significance level. In other words, the difference between the averages of some groups is big enough to be statistically significant.

The following Figure 7.2 summarizes the results of the One-Way ANOVA:

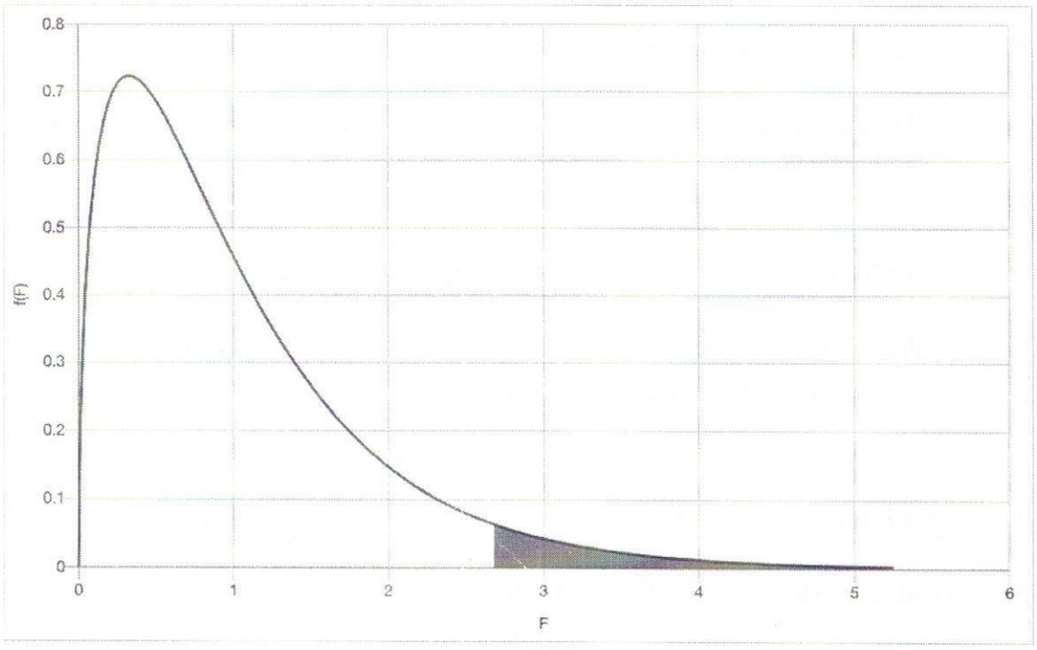


Figure 7.2: ANOVA Results: F=3.85, p-value=0.0113, Ho rejected.

< <font size>

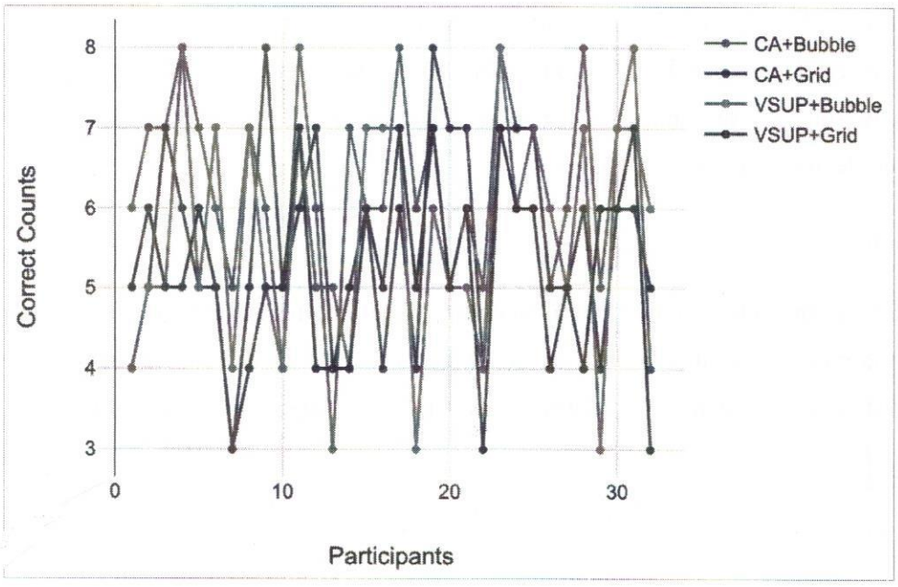
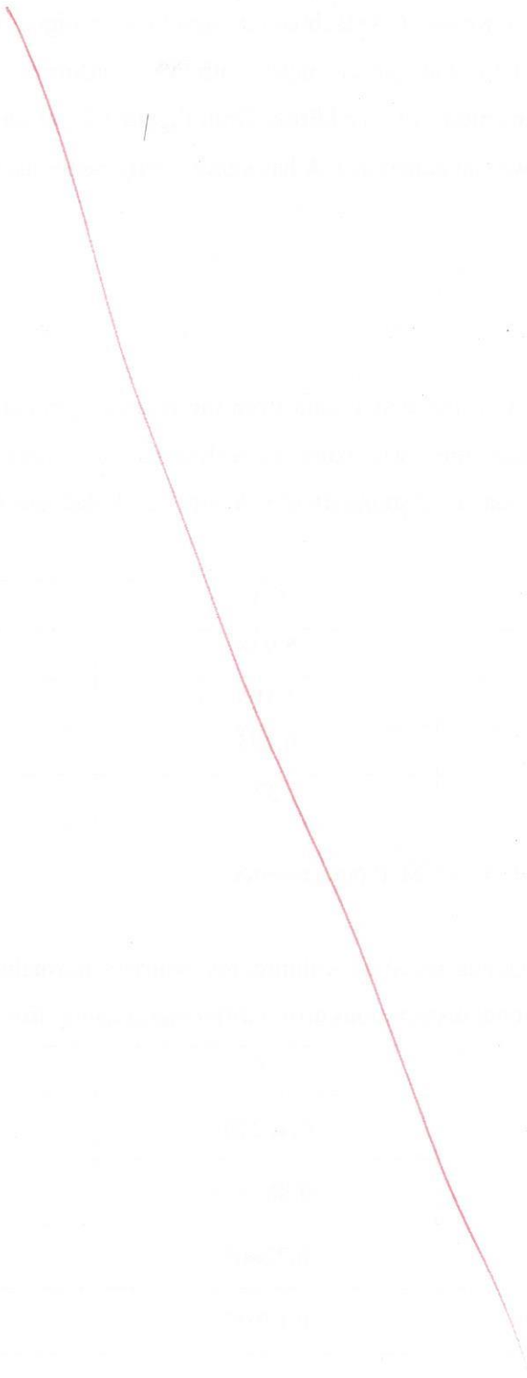


Figure 7.3: Line chart presenting user performance

< this is hard to understand, maybe by two graphs.
1) CA+Bubble and VSUP+Bubble.
2) CA+Grid and VSUP+Grid. ??

↳ or just remove it?



However, from Table 7.1 we see, CA+Bubble has significantly higher means compared other distributions and CA+Grid has closer mean with VSUP+Bubble, and VSUP+Grid has significantly lower mean among all. In addition, from Figure 7.3, we see CA groups dominates the VSUP groups. So, we can conclude CA has significantly better user experience compared to VSUP.

7.2.1.2 Paired t-test

We have generated the CA and VSUP data from the four components performance data by grouping and averaging the two pairs (CA+Bubble, CA+Grid and VSUP+Bubble, VSUP+Grid). Now the statistical summary of CA and VSUP data are shown in the following Table 7.5.

Group	CA	VSUP
Mean	5.938	5.422
SD	1.105	1.078
SEM	0.195	0.191
N	32	32

Table 7.5: Summary of CA vs VSUP performance

why not Shapiro W.Th?

We present test statistics and result of Kolmogorov-Smirnov normality test in the following table 7.6 where we find both distributions do not differ significantly from normally distribution.

Group	CA	VSUP
Skewness	-0.462203	0.071066
Kurtosis	-0.865819	-0.873663
p-value	0.22608	.64088
Test statistic (D)	0.17937	.12636.

Table 7.6: K-S Test of Normality

The following steps show the paired t-test results for the given data and draws conclusion from the test:

(1) Null and Alternative Hypotheses

The following null and alternative hypotheses need to be tested using paired t-test:

Ho: $\mu_D = (\mu 1 - \mu 2) >= 0$ (interpretation?)

Ha: $\mu_D = (\mu 1 - \mu 2) < 0$ (interpretation?)

This corresponds to a left-tailed test, for which a t-test for two paired samples be used.

(2) Rejection Region

Based on the information provided, the significance level is α=0.05, and the critical value for a left-tailed test is $t_c = -1.696$.

The rejection region for this left-tailed test is $R=\{t : t < -1.696\}$

(3) Test Statistics

The computed t-statistic = 3.61

(4) Decision about the null hypothesis

Since it is observed that $t = 3.61 \geq t_c = -1.696$, it is then concluded that *the null hypothesis is not rejected.*

Using the P-value approach: The p-value is $p = 0.9995$, and since $p = 0.9995 \geq 0.05$, it is concluded that the null hypothesis is not rejected.

(5) Conclusion

It is concluded that the null hypothesis Ho *is not rejected.* Therefore, there is not enough evidence to claim that the performance mean difference $\mu_D = \mu 1 - \mu 2$ is less than 0, at the α = 0.05 significance level.

Confidence Interval: The 95% confidence interval is $0.224 < \mu_D < 0.807$.

We can visualize the paired T-test scenario graphically as follows:

(what does it mean?)

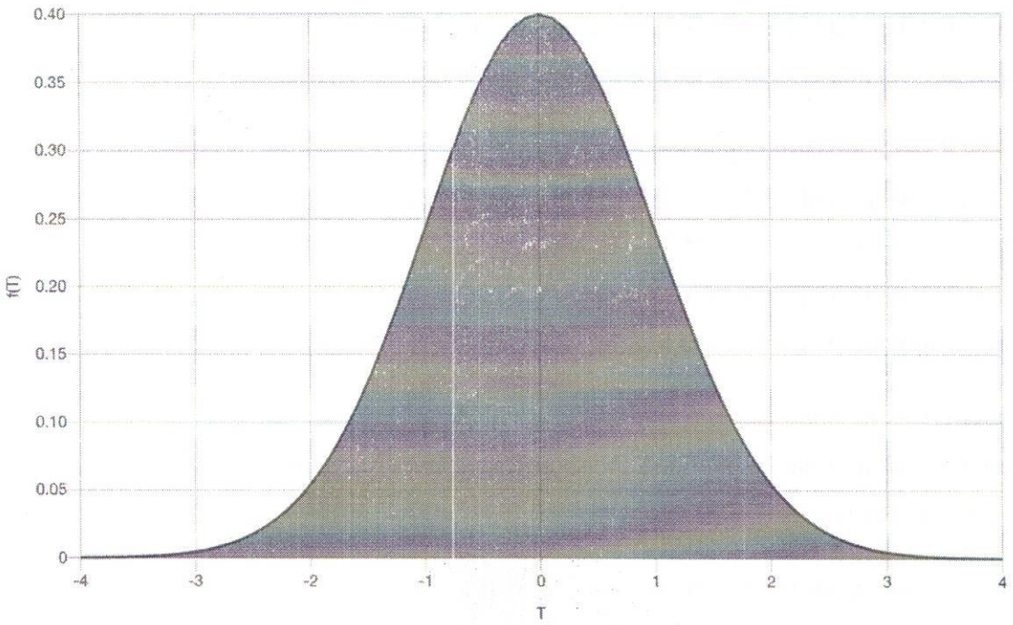


Figure 7.4: Paired t-test sample with p-value=0.9995 for CA vs VSUP performance.

↳ all grey?

7.2.2 Time Utilization Results

Our automated system tracked effective response time for every component separately. The statistical summary of the timing data is represented in the following table 7.7

Group	CA	VSUP
Mean	8.675	9.647
SD	2.320	3.123
SEM	0.410	0.552
N	32	32

Table 7.7: Summary of CA vs VSUP timing

The Shapiro-Wilk tests on both distributions showed that they met normality test with the following results:

For CA = W(32) = .959, p = .254

For VSUP = W(32) = .977, p = .716

The following steps show the paired t-test results for the given time data and draws conclusion from the test:

(1) Null and Alternative Hypotheses

The following null and alternative hypotheses need to be tested:

Ho: $\mu_D = (\mu 1 - \mu 2) <= 0$ (CA response was equal or faster than VSUP response)
Ha: $\mu_D = (\mu 1 - \mu 2) > 0$ (CA response was slower than VSUP response)

This corresponds to a right-tailed test, for which a t-test for two paired samples be used.

(2) Rejection Region

Based on the information provided, the significance level is $\alpha = 0.05$, and the critical value for a right-tailed test is $t_c = 1.696$.
The rejection region for this right-tailed test is $R=\{t : t > 1.696\}$

(3) Test Statistics

The computed t-statistic is equal to -2.656

(4) Decision about the null hypothesis

Since it is observed that $t = -2.656 \leq t_c = 1.696$, it is then concluded that *the null hypothesis is not rejected.*
Using the P-value approach: The p-value is $p = 0.9938$, and since $p = 0.9938 \geq 0.05$, it is concluded that the null hypothesis is not rejected.

(5) Conclusion

It is concluded that the null hypothesis Ho *is not rejected.* Therefore, there is not enough evidence to claim that the timing mean difference $\mu_D = \mu 1 - \mu 2$ is greater than 0, at the $\alpha = 0.05$ significance level.

(also clearly state what the interpretation of the statistical results mean here)

Confidence Interval

The 95% confidence interval is $-1.718 < \mu_D < -0.226$.

We can visualize the paired T-test scenario graphically as follows:

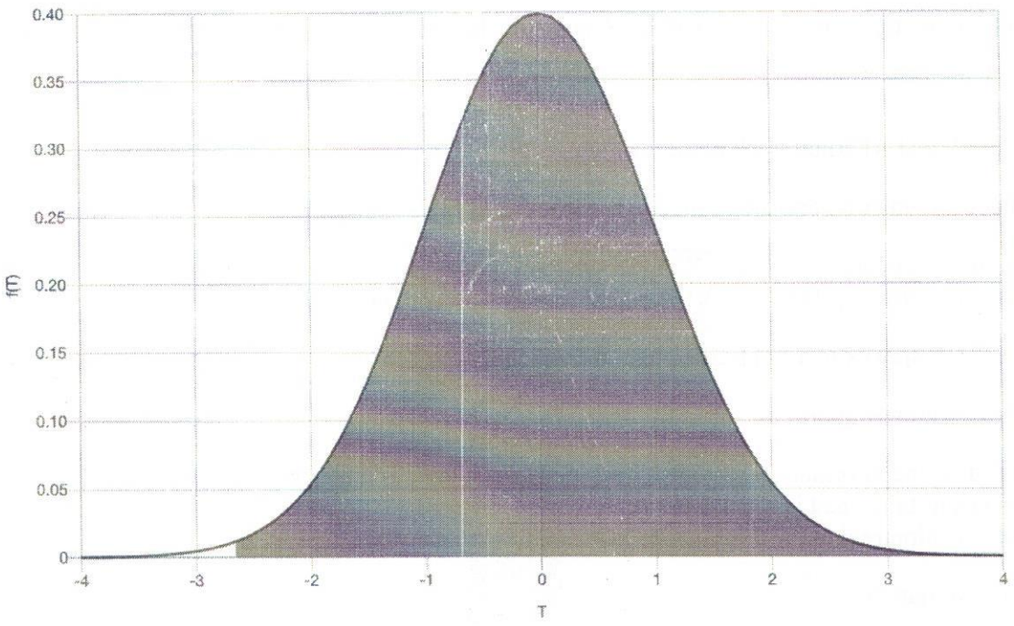


Figure 7.5: Paired t-test sample with p-value=0.9938 for CA vs VSUP timing.

7.2.3 SUS Results

The SUS provides a quick tool for measuring the usability of various kinds of systems based on user experience. It consists of a 10 items questionnaire with five scale response from participants starting from Strongly agree to Strongly disagree. It doesn't have any right or wrong evaluation of any question and hence collectively its use is in classifying the ease of use of the system being tested. The best way to interpret the results is to normalize the scores to produce a percentile ranking. By convention, we converted SUS results to SUS scores by the following rules:

↳ (do you have a reference for this?

- For odd items: subtract one from the user response.
- For even-numbered items: subtract the user responses from 5
- This scales all values from 0 to 4 (with four being the most positive response).
- Add up the converted responses for each user and multiply that total by 2.5. This converts the range of possible values ~~from~~ 0 to 100 instead of ~~from~~ 0 to 40.

to a range from.

The statistical overview of the scores is given below:

Group	CA	VSUP
Mean	60.078	61.094
SD	16.307	14.227
SEM	2.883	2.515
N	32	32

Table 7.7: SUS scores summary of CA vs VSUP

The Shapiro-Wilk tests on both distributions showed that they do not meet normality test with the following results:

For CA = $W(32) = .913, p = .013$
For VSUP = $W(32) = .889, p = .003$

The following steps show the Kruskal-Wallis Test results, which is non-parametric alternative to the One-Way ANOVA test since the distributions are not normal. The purpose of the test is to assess whether or not the samples come from populations with the same population median.

(1) Null and Alternative Hypotheses

The following null and alternative hypotheses need to be tested:

Ho: The samples come from populations with equal medians

Ha: The samples come from populations with medians that are not all equal

The above hypotheses will be tested using the Kruskal-Wallis test.

(2) Rejection Region

Based on the information provided, the significance level is α=0.05, and the number of degrees of freedom is $df = 2 - 1 = 1$. Therefore, the rejection region for this Chi-Square test is $R = \{\chi 2: \chi 2 > 3.841\}$.

(3) Test Statistics

The computed H statistic is = 0.146

(4) Decision about the null hypothesis

Since it is observed that $\chi^2 = 0.146 \leq \chi_c 2 = 3.841$, it is then concluded that the null hypothesis is not rejected.

Using the P-value approach: The p-value is $p = 0.702$, and since $p = 0.702 \geq 0.05$, it is concluded that the null hypothesis is not rejected.

(5) Conclusion

It is concluded that the null hypothesis *Ho* is not rejected. Therefore, there is not enough evidence to claim that not all population medians are equal, at the $\alpha=0.05$ significance level.

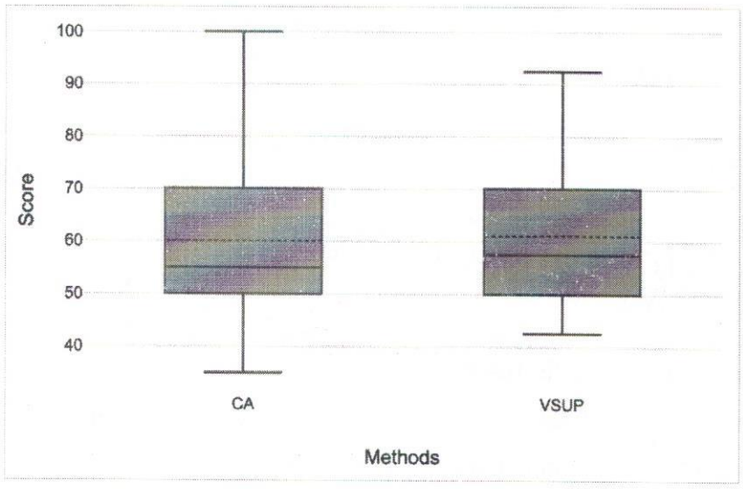


Figure 7.6: SUS rating plots for visualization methods

Although the scores of the methods are slightly varying according to Figure 7.5, the differences ($\chi2 = 0.146$, $p = 0.702$, $df = 1$) were not statistically significant as per Kruskal-Wallis test at $\alpha = 0.05$.

7.2.4 NASA-TLX Results

The TLX stands for Task Load Index and is a measure of perceived workload. Just like SUS data, we have collected Nasa-TLX load test data from our online system. A TLX method increments of high, medium, and low estimates for each point result in 21 gradations on the scales. To score, we subtract 1 from the given rating in the range of 1-21, and multiply by 5. For example, if user gives a rating 5, the score would be 20: (5-1) x 5.

Methods	NASA-TLX	Shapiro-Wilk Normality Test ($\alpha = 0.05$)		
		Test Statistic (W)	p-value	Status

<put this table all together on the same page>

CA	Mental Demand	.906	.009	Not normal
	Physical Demand	.914	.014	Not normal
	Temporal Demand	.948	.128	Normal
	Performance	.932	.044	Not normal
	Effort	.942	.085	Not normal
	Mental Frustration	.916	.017	Not normal
VSUP	Mental Demand	.863	.001	Not normal
	Physical Demand	.903	.007	Not normal
	Temporal Demand	.938	.067	Not normal
	Performance	.887	.003	Not normal
	Effort	.901	.006	Not normal
	Mental Frustration	.877	.002	Not normal

Table 7.8: Normality test results of NASA-TLX score

Since almost all datasets didn't follow the normal distribution, we used the Kruskal-Wallis non-parametric test to evaluate the differences across the two methods of uncertainty representations (CA and VSUP) on NASA-TLX ratings. The following null and alternative hypotheses need to be tested with Kruskal-Wallis test.

Ho: The samples come from populations with equal medians

Ha: The samples come from populations with medians that are not all equal

The following table shows the summary of such test results of Kruskal-Wallis test at the $\alpha = 0.05$ significance level:

NASA-TLX	X2	P	df	H	Conclusion
Mental Demand	0.19	0.6626	1	0.19	Not Rejected
Physical Demand	0.062	0.8038	1	0.062	Not Rejected
Temporal Demand	0.018	0.8932	1	0.018	Not Rejected
Performance	3.61	0.0574	1	3.61	Not Rejected

Effort	0.062	0.8038	1	0.062	Not Rejected
Mental Frustration	0.173	0.6772	1	0.173	Not Rejected

Table 7.9: Kruskal-Wallis test results of NASA-TLX

No statistically significant differences were found between the ~~learning~~g conditions ~~on~~: mental demand ($\chi_2 = 0.19$, $p = 0.6626$, $df = 1$), physical demand ($\chi_2 = 0.62$, $p = 0.8038$, $df = 1$), temporal demand ($\chi_2 = 0.018$, $p = 0.8932$, $df = 1$), performance ($\chi_2 = 3.61$, $p = 0.0574$, $df = 1$), effort ($\chi_2 = 0.62$, $p = 0.8038$, $df = 1$), and mental frustration ($\chi_2 = 0.61$, $p = 0.6772$, $df = 1$) for the significance level $\alpha = 0.05$.

7.3 Summary of Results..

<Briefly summarise results here>

<Did the users express any comments?>

Chapter 8

Conclusions and Future Work

In this thesis, we propose a novel approach of uncertainty visualisation in terms of Chromatic Aberration in web platform. There is an existing uncertainty visualisation system namely VSUP that presents a different approach of uncertainty visualisation. We conduct a within subject comparative user study with VSUP vs our system to find user performance accuracy/error rate, task completion time, subjective assessment with NASA-TLX and SUS. From numerical analysis and evaluation of the results, we see user performance and perception is statistically significant and faster compared to VSUP whereas in the subjective assessment they do not vary statistically significantly.

Nevertheless, we admit that in real aberration the picture blurring happens very slowly from inner edge to outer edge but in our case, it just gives us a range of uncertainty for the prediction, so the whole edges are with the same bright color. However, our simplified implementation allows us to reduce the aberration to both double and/or single parameter, which facilitates chromatic aberration tuning with regards to the amount of represented uncertainty. To mitigate the blurring effect additional research can be conducted such add adding additional color effect and imposing 3D effect.