CS3106 HCI Practical 2 Report

170011474

November 2019

1 Overview

This practical requires students to use one-way ANOVA to analyse data, to implement a application to show pie menus and linear menus, and to conduct a between-subject experiment for two levels of menu types: a pie menu and a linear menu.

2 Question 1

The table for Question 1 is appended in the Appendix A.

3 Question 2 and Question 3: A Between-Subjects Experiment

3.1 4 Factors to consider

There are 4 factors to consider when deciding on whether to implement a linear menu or a pie menu under specific context.

1. **Error rate**, referring to the frequency of errors occurred. A satisfactory menu is expected to reduce the error rate for users and so accurate results are generated. If a menu style brings misleading and error-prone operations for users, then this menu style could not be the best for them.

2. **Task time**, referring to the time users need to accomplish a specific task. A desirable menu style is supposed to minimize the task time, particularly seeking time for users to locate the correct menu item. If the menu requires longer task time for tasks, then it could affect its functionality.

3. **User experience**. Users prefer the menu style which brings them more comfortable and positive perceptions rather than a complicated menu even with more functions. Easy and simple manipulation probably benefits the error rate and the task time further.

4. **Item sets**. If there are diametrically opposed item sets (for example, open / close) or direction items (up /down), then a pie menu could be the best choice. If items are mainly "first", "second", "third", and etc., then a linear menu could be more powerful than a pie menu.

Overall, when we are deciding which menu style to use, we still need to focus on the current context and circumstance we have. Different situations can result in different optimal choices.

3.2 Experiment

3.2.1 Set Up

A between-subjects design was adopted to execute the experiment to avoid asymmetrical skill-transfer effects. Hence, each participant was exposed to only one condition: either a pie menu or a linear menu.

3.2.2 Hypothesis

H_1: The pie menu will have lower error rate than the linear menu.
H_0: There will be no difference in error rates between the pie menu and the linear menu.
H_1: The pie menu will take less task time than the linear menu.
H_0: There will be no difference in task time between the pie menu and the linear menu.

3.2.3 Participants

Participants were 6 university students who were experienced with computers, for avoiding effects on experimental conditions caused by the lack of computer experience.

3.2.4 Variables

The independent variable are two menu types and the type of required tasks.
The dependent variables are the number of errors (when the user selected an unexpected menu item) and the task time (the time to complete a task).

3.2.5 Apparatus and Material

Equipment are Thinkpad T480s PC with Linux Ubuntu 18.04 and an application implemented by qml based on Qt Creator.
A practical task for tutorial was used: Find out the "go back" button in a pie menu.
A paper with 4 tasks written were designed respectively: Find out 8^{th}, 5^{th}, 1^{st}, 7^{th} section.

3.2.6 Procedure

6 participants was recruited in University of St Andrews. They were randomly assigned to a linear menu or a pie menu, 3 for each. Before starting, a tutorial was provided for pie menu testers to show how to use a pie menu by a task with different content. The task list was written on a paper which was given to them. The typical sequence of events is as below:

- An application window which draws in Overleaf browser window is displayed to subjects.
- Participants press the left mouse button to invoke a menu. The menu appears in the cursor's position. If the cursor locates in the edge of the window then the menu will change its position to fit.
- Participants select a menu item and then the menu will be closed.
- The task time (starting from the invoking time to selecting time) appears on the right top of the window and the error code will be written into a file in the back-end.

3.2.7 Possible Threats and Solutions

There are 3 major possible threats to the experimental validity.

1. (External validity) The conclusion is limited within this certain context of "jumping to a specific section". It is merely a linear item set from the first section to the last one, but cannot be applied to other context. Other contexts, including tasks in geographical or directional sense or other semantics, might give a different result. Therefore, the experiment is not externally general enough.
2. (Internal validity) Nonequivalent baseline of experience in pie menus. Some participants probably used pie menus before but the others did not, causing a skill-transfer effect to the experiment.
3. (Internal validity) The area or size of the pie menu. If the menu is too large or too small, then it could bring effects to the experiment. But there was no way to design equivalent sizes for pie menus and linear menus.

To minimise effects brought by those threats, 3 solutions are possible respectively.

1. For the first threat, further studies can take other contexts into count to generalise the result. However, requirements for this assignment only base on Overleaf sections, so no other context was adopted.

2. For the second threat, in order to reduce the risk, in the recruitment process, candidates who had no experience in pie menus but were experienced in computer were selected to guarantee a baseline of knowledge of pie menus. Then, a tutorial was provided before starting the experiment, to present what a pie menu looked like and how to manipulate it. To exclude experimental interference to experimental tasks, a totally different task in different context was shown, so no participant would predict real tasks.

3. For the third threat, since it is very difficult to measure the effect of the area of a pie menu, a relatively proper size of a pie menu was set compared to the size of the linear menu.

3.2.8 Results

All collected data was recorded in an Excel spreadsheet. Since those distributions were normal, a one-way ANOVA was adopted to analyse the data with the embed tool in Excel (see Appendix B).
As for error occurrences, the pie menu resulted in fewer average errors than the linear menu (0 vs. 0.5 respectively). We assumed these errors were normally distributed. Analysis of variance at a significance level of $\alpha = 0.05$ showed that this difference was not statistically significant ($F = 3, p > 0.05$), **so the H_0 was not rejected.**
For task 1 and task 3, the linear menu resulted in fewer average task time than the pie menu (942 ms vs. 994.6667 ms; 682.3333 ms vs. 747.6667 ms respectively). Analysis of variance at a significance level of $\alpha = 0.05$ showed that this difference was not statistically significant($F = 0.0847, p > 0.05$; $F = 2.4509, p > 0.05$).
In task 2 and task 4, the pie menu resulted in fewer average task time (846.6666 ms vs. 1017 ms; 767 ms vs. 1019.6667 ms). Analysis of variance at a significance level of $\alpha = 0.05$ showed that this difference was not statistically significant ($F = 1.2943, p > 0.05$; $F = 2.3732, p > 0.05$). **Therefore, the H_0 was not rejected.**

3.2.9 Discussion and Conclusion

An interesting fact for task 3 is that the possible reason for better performance of the linear menu is that when participants invoked the linear menu, the menu popped up under the mouse where the nearest menu item was "Section 1". According to Fitts's Law, participants could take advantage of the current distance and took less time to complete the task (selected the "Section 1").
From another study conducted by Callahan in 1998, the pie menu performed better than the linear menu in task time, while this experiment did not give a significant result. There were 4 potential issues and reasons to be illustrated.
The first issues is the size of the menu. Since it is difficult to measure how size will affect the result, there is no method to fairly balance different sizes of a pie menu and a linear menu.
The second issue could be the small number of participants. Since there were only 6 participants in total, some random noises would significantly affect results.
The third issue is that the current item set is linear. The pie menu might not perform best in this context (Callahan et al., 1998).
The fourth issue is that different people tend to have different response capability, so individual difference can be significant. Therefore, there was no uniform standard to measure the ability of different menu types. A within-subjects design may be a better choice.
Since the context is limited to the current linear item set, other item sets including diametrically opposed or directional item sets were not applied. Further studies can be conducted in such area.
Additionally, measuring effects brought by different sizes of different menu types are also areas of further research.
Deeper levels of menu can be an area as well (Callahan et al., 1998). A linear menu can invoke a new sub-menu and a pie menu can generate a sub-menu for the clicked item.

3.3 Extension: A within-subjects experiment

3.3.1 Set up

A within-subjects design was adopted to execute the experiment, and each participant was exposed to both levels of menu types: a pie menu and a linear menu. *Counterbalancing method was used to reduce asymmetrical skill-transfer effects.*

3.3.2 Hypothesis, Participants, Variables, Apparatus and Material

The same as the between-subjects experiment above. Participants were different from those who were in the between-subjects experiment.

3.3.3 Procedure

The same as the between-subjects experiment above. *One difference is that 3 participants started with pie menus, while the other 3 participants started with linear menus.* In this way, counterbalance was applied. The typical sequence of events was the same as the Procedure of the between-subjects experiment.

3.3.4 Possible Threats to Internal Validity and Solutions

There are 2 possible **threats** to internal validity:

1. The baselines of experience in linear menus and pie menus are not equivalent. In other words, most people encountered linear menus most in daily life, while they were not familiar with pie menus as much as with linear menus. Therefore, they might perform better in the linear menu rather than in the pie menu.

2. The skill-transfer effects would be brought into the experiment. If all participants finished the experiment in a same order (for example, started with linear menus and ended with pies), then they could learn from those tasks.

There are 2 possible **solutions** to minimise threats respectively:

1. Try to give a tutorial to participants who had no experience in pie menus before, so they could all have a relatively identical baseline of knowledge of pie menus.

2. Counterbalancing method was applied. Half participants were given linear menus first and another half participants were given pie menus first. In this way, there would have been no confounding of order of tasks.

3.3.5 Results

T-Test paired two sample for means method was adopted to analyse data, since there were paired observations of the same participant.
As for **error occurrences**, the pie menu resulted in fewer average errors than the linear menu (0 vs. 0.5 respectively). We assumed these errors were normally distributed. Analysis of variance at a significance level of $\alpha = 0.05$ showed that this difference was not statistically significant ($p = 0.1817 > 0.05$), **so the** H_0 **was not rejected** (see Appendix C).
As for task time, in **task 1**, the pie menu resulted in less task time than the linear menu (955.6667ms vs. 1045.5ms). Analysis of variance at a significance level of $\alpha = 0.05$ showed that this difference was statistically significant ($p = 0.0163 < 0.05$).
As for task time, in **task 2**, the pie menu resulted in less task time than the linear menu (927.6667ms vs. 991.8333ms). Analysis of variance at a significance level of $\alpha = 0.05$ showed that this difference was not statistically significant ($p = 0.4808 > 0.05$).
As for task time, in **task 3**, the linear menu resulted in less task time than the pie menu (723ms vs. 872.6667ms). Analysis of variance at a significance level of $\alpha = 0.05$ showed that this difference was statistically significant ($p = 0.0228 < 0.05$).

As for task time, in **task 4**, the pie menu resulted in less task time than the linear menu (962.83333ms vs. 1032ms). Analysis of variance at a significance level of $\alpha = 0.1$ showed that this difference was statistically significant ($p = 0.0899 < 0.1$).
By using the average task time table, the pie menu resulted in less average task time than the linear menu. Analysis of variance at a significance level of $\alpha = 0.05$ showed that this difference was not statistically significant ($p = 0.7559 > 0.05$).

3.3.6 Discussion, Conclusion

Compared to the between-subjects design above, the within-subjects design has an advantage: it does not depend on the random assignment of participants. But the between-subjects experiment has more external validity, as the users in daily life only face one choice.
Results were more significant than results of the between-subjects experiment. It could be explained by the fourth issue. However, the null hypothesis was still not rejected since the evidence was not strong enough. The issues were still similar to issues of the between-subjects experiment.
Further possible studies are illustrated above in Conclusion of the between-subjects experiment.

References

[1] Callahan, J., Hopkins, D., Weisert, M. and Shneiderman, B. (1998). *An empirical comparison of pie vs. linear menus.* [online] https://dl.acm.org/. Available at: https://dl.acm.org/citation.cfm?id=57182 [Accessed 21 Nov. 2019].

A Appendix A: Table for Question 1

Results	Table 1	Table 2	Table 3
SS_{error}	80	72	596
SS_{total}	142.5	112	1440.6667
SS_{effect}	62.5	40	844.6667
n Participants	10	10	21
m Groups	2	2	3
df_{error}	8	8	18
df_{effect}	1	1	2
MS_{error}	10	9	33.1111
MS_{effect}	62.5	40	422.3333
α Confidence Level	0.05	0.05	0.001
F-ratio	6.25	4.4444	12.7550
Critical Value	5.3177	5.3177	10.3899
Significant?	Yes, 6.25 > 5.3177	No, 4.4444 < 5.3177	Yes, 12.7550 > 10.3899
Table 1 Reporting	The Keyboard QWERTY resulted in fewer average errors than the Keyboard QuickPath (12 vs. 17 respectively). We assumed these errors were normally distributed. Analysis of variance at a significance level of α = 0.05 showed that this difference was statistically significant ($F_{1,8} = 6.2500$, $p < 0.05$)		
Table 2 Reporting	The Mouse SurePen resulted in fewer average errors than the Mouse QuickPen (7 vs. 11 respectively). We assumed these errors were normally distributed. Analysis of variance at a significance level of α = 0.05 showed that this difference was not statistically significant ($F_{1,8} = 4.4444$, $p > 0.05$)		
Table 3 Reporting	The SR Siri resulted in fewer average errors than the SR Alexa, and the SR Alexa resulted in fewer average errors than the SR Google Home (34 vs. 38 vs. 39 respectively). We assumed these errors were normally distributed. Analysis of variance at a significance level of α = 0.05 showed that this difference was statistically significant ($F_{2,18} = 12.7550$, $p < 0.001$)		

Figure 1: ANOVA Table Results

B Appendix B: Results for Between-Subjects Experiment

Error Table		
Task / Menu	Linear	Pie
Task 1	1	0
Task 2	0	0
Task 3	0	0
Task 4	1	0

Task 1 Task Time		Task 2 Task Time		Task 3 Task Time		Task 4 Task Time	
Linear	Pie	Linear	Pie	Linear	Pie	Linear	Pie
878	695	1154	956	752	790	1098	945
852	1258	861	976	662	721	772	650
1096	1031	1033	708	633	732	1189	756

Summary Table					
Results	Error	Task 1	Task 2	Task 3	Task 4
SS_{error}	1	87010	10449.3334	140809.333	
SS_{total}	1.5	200537.333	115163.5	16852	224353.333
SS_{effect}	0.5	4160.6667	28153.5	6402.6667	83544
n Participants	8	6	6	6	6
m Groups	2	2	2	2	2
df_{error}	6	4	4	4	4
df_{effect}	1	1	1	1	1
MS_{error}	0.16667	49094.1667	21752.5	2612.3333	35202.3333
MS_{effect}	0.5	4160.6667	28153.5	6402.6667	83544
α Confidence Level	0.05	0.05	0.05	0.05	0.05
F-ratio	3	0.0847	1.2943	2.4509	2.3733
Critical Value	5.9874	7.7086	7.7086	7.7086	7.7086
Significant?	No	No	No	No	No

C Appendix C: Results for Within-Subjects Experiment

Error Table		
Task / Menu	Linear	Pie
Task 1	0	0
Task 2	1	0
Task 3	0	0
Task 4	1	0

Average Task Time Table		
Task / Menu	Linear	Pie
Task 1	1049.5	955.6667
Task 2	991.8333	927.6667
Task 3	723	872.6667
Task 4	1032	962.8333

Task 1 Task Time		Task 2 Task Time		Task 3 Task Time		Task 4 Task Time	
Linear	Pie	Linear	Pie	Linear	Pie	Linear	Pie
1238	1161	1362	1049	815	1065	1131	1176
737	652	849	642	687	700	718	687
975	858	1113	1175	799	807	1088	936
1117	908	799	578	501	657	1096	925
1229	1178	964	1134	919	1143	1120	1054
1001	977	864	988	617	864	1039	999

t-Test: Paired Two Sample for Mean: Error		
Results	Linear	Pie
Mean	0.5	0
Variance	0.3333	0
Observations	4	4
Pearson Correlation		
df	3	0
t Stat	1.7321	
$P(T <= t)$ one-tail	0.0908	
t Critical one-tail	2.3534	
$P(T <= t)$ two-tail	0.1817	
t Critical two-tail	3.1824	

t-Test: Paired Two Sample for Mean: Task 1		
Results	Linear	Pie
Mean	1049.5	955.6667
Variance	35573.5	39214.6667
Observations	6	6
Pearson Correlation	0.9453	
df	5	0
t Stat	3.5574	
$P(T <= t)$ one-tail	0.0081	
t Critical one-tail	2.0150	
$P(T <= t)$ two-tail	0.0162	

t Critical two-tail	2.5706	

t-Test: Paired Two Sample for Mean: Task 2		
Results	Linear	Pie
Mean	991.8333	927.6667
Variance	45281.3667	65196.2667
Observations	6	6
Pearson Correlation	0.6245	
df	5	
t Stat	0.7614	
$P(T <= t)$ one-tail	0.2404	
t Critical one-tail	2.0150	
$P(T <= t)$ two-tail	0.4808	
t Critical two-tail	2.5706	

t-Test: Paired Two Sample for Mean: Task 3		
Results	Linear	Pie
Mean	723	872.6667
Variance	22894.4	38157.0667
Observations	6	6
Pearson Correlation	0.8168	
df	5	
t Stat	-3.2441	
$P(T <= t)$ one-tail	0.0114	
t Critical one-tail	2.0150	
$P(T <= t)$ two-tail	0.0229	
t Critical two-tail	2.5706	

t-Test: Paired Two Sample for Mean: Task 4		
Results	Linear	Pie
Mean	1032	962.8333
Variance	24684.4	26658.9667
Observations	6	6
Pearson Correlation	0.8737	
df	5	
t Stat	2.0988	

$P(T <= t)$ one-tail	0.0449	
t Critical one-tail	1.4759	
$P(T <= t)$ two-tail	0.0899	
t Critical two-tail	2.0150	

t-Test: Paired Two Sample for Mean: Average Task Time		
Results	Linear	Pie
Mean	949.0833	929.7083
Variance	23299.9722	1676.34028
Observations	4	4
Pearson Correlation	0.9687	
df	3	
t Stat	0.3416	
$P(T <= t)$ one-tail	0.3776	
t Critical one-tail	2.3534	
$P(T <= t)$ two-tail	0.7552	
t Critical two-tail	3.1824	