Improving Command Selection with Command Maps

Dianna Hu and Joy Ming

CS279 Assignment 2

1. Design and Analysis

1.1. Experiment design.

In this study, participants completed tasks for each of the two interfaces with 18 trials for the practice phase and 36 trials for the testing phase, totaling 108 trials ((18 practice + 36 testing) x 2 interfaces). We counterbalanced the order of interfaces that participants encountered as well as the command set that was associated with each interface.

Furthermore, we replicated the study with three specific participant types: in-person volunteers (N = 10), online volunteers (N = 13), and paid online mechanical turk (MTurk) participants (N = 14). The table below shows the spread of demographics:

Table 1. Demographic data of participants for each study.

	Age (avg)	Gender
In-person	21.3	7 M, 3 F
Online	22.9	6 M, 7 F
MTurk	35.7	9 M, 5 F

1.2. Analysis Process.

We analyzed the data first within each type of experiment (see sections 2.1 to 2.3). For each of the types of experiments, we tried to replicate the graphs included in the paper and do some more freeform analysis. Later in the discussion we analyze the results as a whole, with the experiment type as an independent variable.

1.3. Data exclusion criteria.

To determine which points we , we plotted each of the types of graphs as an overlay of a histogram. It was found that there were some major outliers, so based on the graphical representation and looking at the standard deviations, we decided to remove times that were greater than 7,000 milliseconds. These times were also not included in the errors.

2. Results

2.1. H1: CommandMaps is faster than Ribbon

Overall, we found that the mean acquisition times (errors removed) were faster with the CommandMaps interface than with the Ribbon interface with respect to the in-person lab study, the online volunteer study, and the online MTurk study, as displayed in Figures 1(a), 2(a), 3(a), and 1(c), 2(c), and 3(c). Furthermore, we find that mean error rates were lower with the CommandMaps interface than with the Ribbon interface with respect to the in-person lab study, the online volunteer study, and the online MTurk study, as displayed in Figures 1(b), 2(b), and 3(b). Note, the blue indicates ribbon and red indicates CommandMap. The numerical versions of the graphical results can be seen below in Table 2(a) and 2(b). We therefore find support for H1.

Table 2(a). Mean selection time in seconds and standard deviation.

		In-person	Online	MTurk
CommandMaps	Same parent	2018.99 (681.50)	1853.55 (454.48)	2252.75 (841.66)
	Different parent	2023.21 (660.58)	1862.16 (504.04)	2150.23 (726.54)
Ribbon	Same parent	1901.73 (536.15)	1849.29 (557.99)	2084.34 (853.81)
	Different parent	3057.36 (935.14)	3034.06 (840.56)	3491.63 (1135.46)

Table 2(b). Mean error rate and standard deviation.

		In-person	Online	MTurk
CommandMaps	Same parent	0.0027 (0.0083)	0.0042 (0.0064)	0.0022 (0.0054)
	Different parent	0.0260 (0.0194)	0.0130 (0.0117)	0.0052 (0.0070)
Ribbon	Same parent	0.0028 (0.0057)	0.0043 (0.0065)	0.0022 (0.0055)
	Different parent	0.0603 (0.0300)	0.0431 (0.0270)	0.0192 (0.0155)

2.2. H2: Parents interact with interface

We found a parent interaction with parent (same vs. different) based on Figures 1(a), 2(a), and 3(a) for each of the respective interfaces, implying that the CommandMaps overall interface yielded faster results than the Ribbon interface for different parent. In particular, it appears that the slope between same parent and different parent in the Ribbon interface is more pronounced than the slope between same parent and different parent in the CommandMaps interface. This result indicates that the Ribbon interface and CommandMaps interface exhibited similar performance when considering the commands of the same parent condition, but the CommandMaps interface outperformed the Ribbon interface when considering the commands of the different parent condition. We therefore find result for H2.

2.3. H3: Happiness is higher with CommandMaps

Overall, Table 1 demonstrates that user response to the CommandMaps interface was generally positive, with most participants preferring CommandMaps on average. One intriguing point to note is that although the in-person user response to the CommandMaps interface is not negative, it does not demonstrate that the user response overall was more positive than the response to the Ribbon interface, as the NASA-TLX scores were relatively similar between the two interfaces. This surprising result indicates that we do not find support for H3 for the in-person study. For both of the online studies with volunteers and with MTurkers, however, the CommandMaps interface on average was rated as less intensive than the Ribbon interface. We therefore find overall support for H3.

Table 1. Mean (std. dev.) of NASA-TLX scores (1=low, 5=high).

	Person		Online		MTurk	
	Ribbon	CM	Ribbon	CM	Ribbon	CM
Mental Demand	1.6 (0.6)	1.8 (0.9)	2.6 (1.2)	1.5 (0.9)	2.1 (1.1)	1.8 (1.2)
Physical Demand	2.3 (1.4)	2.2 (1.4)	2.1 (1.0)	1.5 (0.7)	1.6 (0.9)	1.4 (0.6)
Temporal Demand	2.3 (0.9)	2.3 (1.0)	2.2 (0.9)	2.2 (0.9)	2.7 (1.5)	2.6 (1.4)
Hard Work	2.2 (1.1)	2.2 (1.1)	2.5 (1.5)	1.8 (1.3)	2.9 (1.5)	2.6 (1.7)
Frustration	2.2 (1.3)	2.2 (1.4)	3.3 (1.5)	2.6 (1.7)	1.8 (0.9)	1.4 (0.8)

3. Discussion

3.1. Aggregate Data.

Similar trends are also seen in the aggregate data analysis, where mean selection time (Figures 4(a) and 4(b)) as well as mean error rate (Figure 4(c)) is graphed using all of the data from all different types of experiments.

Figure 4(a). Mean selection time.

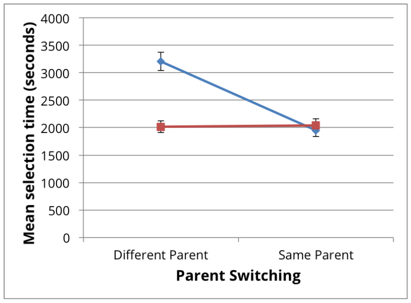


Figure 4(b). Mean selection.

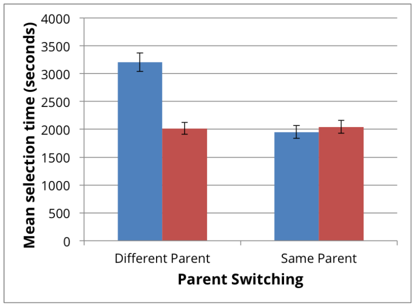
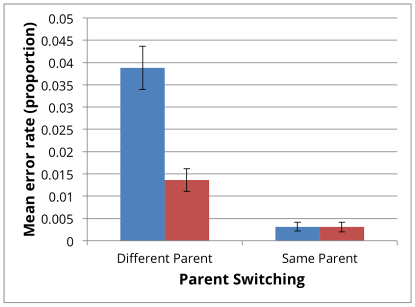


Figure 4(c). Mean error rate.



An ANOVA conducted on the stacked data shows that there is a significant effect of interface, parent and interface and parent on the data, which follows the hypotheses and analysis as presented in the original paper and with our efforts. However, something interesting is that there is also a significant effect of the type of experiment (in-person, online, MTurk) on the timing of the experiment. This would be interesting to further investigate in the future.

Effect	**DFn**	**DFd**	**F**	**p (* p<0.05)**	**ges**
interface	1	136	58.122903191	3.806998e-12 *	0.2994129092
parent	1	136	66.736834804	1.899700e-13 *	0.3291796228
type	2	136	7.095357091	1.171969e-03 *	0.0944846308
interface:parent	1	136	62.618407804	7.830010e-13 *	0.3152699113
interface:type	2	136	0.119140274	8.877758e-01	0.0017489985
parent:type	2	136	0.006269981	9.937499e-01	0.0000921971
interface:parent:type	2	136	0.162816907	8.499121e-01	0.0023886470

3.2. Hypothesis Support.

Overall, we have found support for the following hypothesis:

- H1: Knowledgeable users can select commands faster using CommandMaps than when using Ribbons and menus.
- H2: There is no performance difference between CommandMaps and Ribbons when selecting commands contained in the most recently used tab, but CommandMaps are faster than the Ribbon for tasks requiring switching between different parent tabs.
- H3: Subjectively, users will prefer CommandMaps.

The combination of time and error data is important, as it shows that the CommandMaps interface does not increase errors to achieve its improved temporal performance – it is both faster and more accurate the Ribbon interface.

3.3. Limitations and Assumptions.

Some of the limitations and assumptions we held include:

- There were some changes in the instructions between the in-person studies and the online deployment. For example, the format for the NASA-TLX rating on the happiness form is slightly different. Our assumption is that these do not significantly affect the results.
- We also assume that the random variables we do not measure, such as environment in which the individual is taking the test, do not significantly alter the results.
- During the analysis, we assume that, similar to the original experimental paper, data points with errors do not significantly affect the results.
- We also assume that the data we excluded do not affect the results.
- Some limitations we have include that while this is supposed to be a “general” test of the CommandMap concept, users who have experience using Microsoft Word may perform differently on the Ribbon concept testing.
- Additionally, intrinsic motivation may not be very high for the people doing the experiment, even with the reduction of the number of trials. Simple ennui may affect the performance on later trials for some individuals.

3.4. Interesting findings and future work

Some of the observations about performance across the different *types* of experiments is unique to the way in which we conducted the experiments compared to the original paper. This is prompted by the ANOVA analysis as described in 3.1. and can be seen qualitative in the differences in the averages for each condition. For example, in-person seems to perform “better” qualitatively, with shorter times. And “friends” online, even with less monetary incentive, seem to perform “better” than MTurk workers, potentially because of the friendship bond. Future work could further investigate more of the influence of experiment type on other variables and see if these types of results are pervasive through other experiments done in a similar style across different types.

There were also a few other variables that we tracked and would like to analyze in the future. For example, some of the other demographic variables (age, gender) or experimental setting (operating system, browser, input device, familiarity with Microsoft Word).