Started on	Wednesday, 7 October 2020, 5:20 AM
State	Finished
Completed on	Wednesday, 7 October 2020, 5:06 PM
Time taken	11 hours 45 mins
Grade	Not yet graded

Question **1**

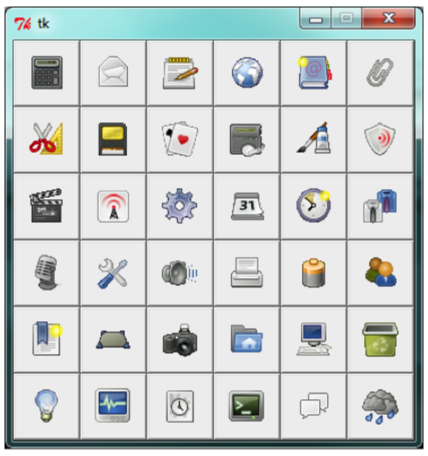
Complete

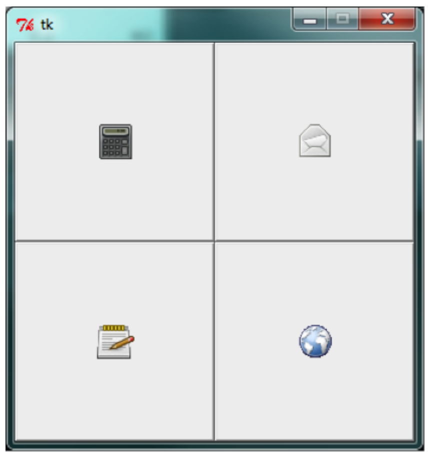
Marked out of 1.00

Evaluation scenario

In Lab 7 you made theoretical predictions for two interface designs, shown below. Your predictions *should* give a good estimate of performance at both novice and expert levels. Predictions, such as these, might be used to help designers hone in on a final design. (Recall that the interfaces varied in the number of items shown at once -- from 36 items at once down to 4 items at once. Users have to 'swipe' left or right to move between pages of icons when there are fewer than 36 items in the display).

Often, however, designers want empirical validation of their theoretical predictions. In this lab you will gather empirical data for novice and expert(ish) performance with each of the designs and you will run statistical tests of the null hypotheses "there is no difference in mean target selection time when using the 36 and 4 item designs”. You will reject the null hypothesis (accepting a difference) only if there is less than a 5% chance that the observed data (or more extreme) should occur if the null hypothesis were true.





1. Generate the data (4-item & 36-item)

Take a copy of the Python program 'gui-36.py' from Learn.

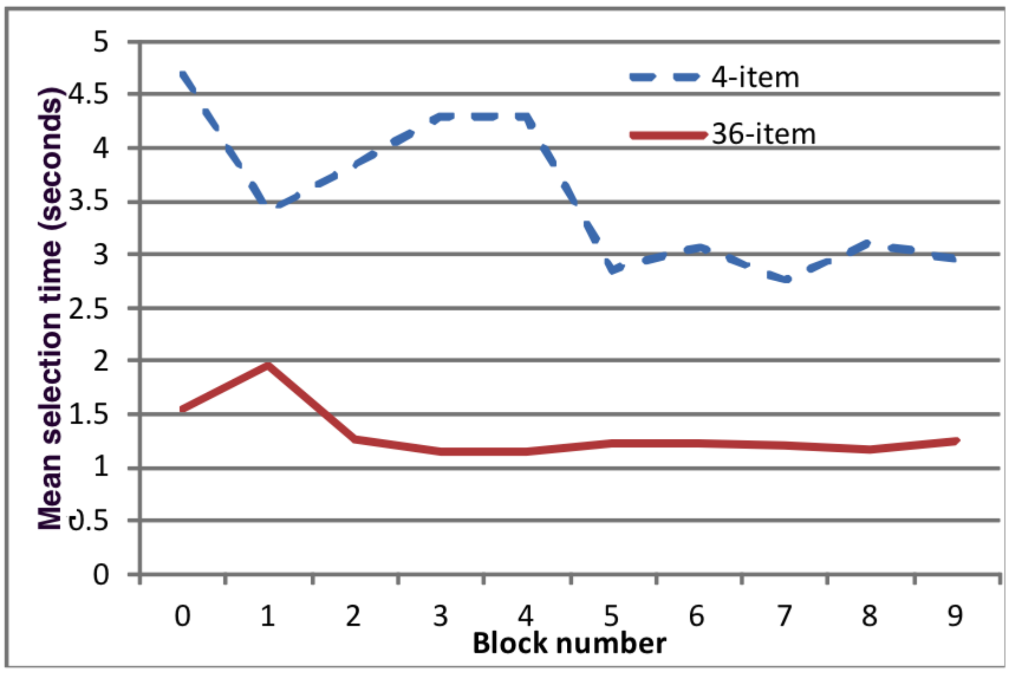
Change the "USER_CODE” constant to your own usercode.

Use the "DESIGN” constant to configure the interface for either the 36 or 4 item condition. If your birthday occurs on an even-numbered day of the month, use the 36-item condition first; otherwise use the 4-item condition first.

The program will randomly select five targets, each of which will be presented 10 times. The items will remain in the same place for each selection, allowing you to build up expertise in locating them. Select the items as quickly and accurately as possible.

2. Plot the data

Import your log file into Excel. Calculate your own mean time per block for each of the interfaces. Plot these values to produce a graph of the format shown below.



Don't be surprised if your plots (like that shown above) do not indicate the classic 'power law of practice' curve. With only one participant's data (your own) the scope for random variation is still extremely large.

3. Share data & plot the combined data

Share your source data files with as many of your classmates as possible. Calculate a grand mean for each block across all participants for each interface. Redraw the plot above using the grand mean. The learning curves should smooth out as you add data from more participants.

4. Can we reject the null hypothesis of no difference for novices?

T-Test in Excel

Data from block zero represents minimally experienced use of the system. (Consider the experimental risks in this statement). Run a statistical test to determine whether we can confidently reject the null hypothesis of no difference between the 36- and 4-item conditions for novices.

All participants completed both conditions, so we can treat their data as "paired", so we should use a paired T-Test to compare means.

Massage the block 0 data in your Excel spreadsheet into the following format:

participant	four	thirtysix
ajc168	4.674	1.534
ppq13	4.123	1.344
...	...	...
foo45	5.442	1.775

Click on a blank cell and use the TTEST formula to compare the 4-item data with the 36-item data. The first 'array' of data is all of the data from the 4-item condition; the second array is the data for 36-items; the number of 'tails' is 2 (meaning we don't care about the direction of the performance difference); and the type value is 1 (meaning we are conducting a paired t-test). If the result of the TTEST function is less than 0.05, then there is less than a 5% chance that data at least as extreme as our sample would occur assuming that the null hypothesis was true; we therefore reject the null hypothesis of no difference.

T-Test in R

You can also conduct a "paired t-test" using "R", probably the leading software package for statistical analysis. Assuming you have a three-column column file named "ttest-data.txt" (first row containing headers), the following script will conduct a paired t-test (you can put this script into a file, say called "ttest.R", then make the script executable "chmod u+x ttest.R", then run it "> ./ttest.R").

```
#!/usr/bin/env Rscript
data <- read.table("ttest-data.txt", header=TRUE)
t.test(data$four, data$thirtysix, paired=TRUE)
```

5. Can we reject the null hypothesis of no difference for experts?

Participants were most expert in the final block. Conduct the same T-Test analysis that you did for novices, but this time use only data from the final block. Can we reject the null hypothesis of no difference between expert mean selection times with the two interfaces?

What limitations are there in our expert analysis? What might we do about them?

6. Generate more data (9-item) & share with others

Generate data for a 9 item (3x3) condition, and gather the output from several other students in the class.

7. One-way ANOVA in R

Imagine that we now want to compare novice and expert performance with each of the three interface conditions. (There's an obvious experimental confound in that all participants completed the 3x3 condition last, so there are substantial risks of learning and fatigue effects. We'll ignore these for now). You might be tempted to conduct a bunch of T-Tests to compare 4 items with 9, 4 with 36, and 9 with 36, but each analysis we do increases the probability of observing a difference by chance (the number of comparisons needed for n conditions is $n \times (n-1)/2$).

Instead, we'll use the main statistical workhorse: ANOVA (analysis of variance).

Fortunately, working with ANOVA in R is relatively simple.

Assuming that the data file "oneway-within.txt" has the following format:

```
subject condition   target                      block   time
ajc168  four        ./icons\user-trash.png.gif  0       4.556
```

The following R script will conduct an ANOVA for the three conditions (presumably named "four", "nine", and "thirtysix"):

```
#!/usr/bin/env Rscript

library(ez)

data <- read.table("oneway-within.txt", header=TRUE)
novice_data <- data[which(data$block == 0),]
expert_data <- data[which(data$block == 9),]

print("NOVICE ANALYSIS")
ezANOVA(data=novice_data, dv=time, within=condition, wid=subject)
ezStats(data=novice_data, dv=time, within=condition, wid=subject)
print("EXPERT ANALYSIS")
ezANOVA(data=expert_data, dv=time, within=condition, wid=subject)
ezStats(data=expert_data, dv=time, within=condition, wid=subject)
```

This script relies on data from many participants, all contained in the file "oneway-within.txt". It reads the file, then assigns to "novice_data" the subset of data from the first block, and assigns to "expert_data" the data from the final block. It then conducts a within-subjects ANOVA, prints some summary statistics. This is repeated for the novice and

expert data. If the p-value is less than 0.05 then you reject the null hypothesis of no difference between conditions (see the lecture notes to find where the p-value you care about is shown).

(Note: if you receive an error that the library(ez) is unavailable, enter the command **install.packages("ez")** at the R prompt.)

8. Plot means & evaluate your results

R includes great tools for producing nice plots, but for now, you can take the mean values output and plot them in Excel.

Use an Excel bar chart to plot mean performance with each of the interface conditions. Can you reject the null hypothesis of no difference for the three conditions when novice and when expert? What does this mean?

Finally, revisit your predictions from Lab 7. How do your empirical measures compare? Can you explain what might have caused any differences?

9. Submit your findings

Submit your Excel file, containing separate sheets for the source data, output from statistical tests, and plots of the results.

lab9.xlsx

◄ Lab 8: Heuristic Evaluation

Jump to...

gui-36.zip ►