Research Methods COMP90044

When it all goes complicated…

…planning research and

recovering from unexpected problems!

Part 1: Planning an Experiment

- We will workshop through how to design an experiment for a particular research question
- This is a real study I did over the last 18 months
- …the fact it took 18 months will perhaps indicate that some problems occurred!
- I don't get to do research full time – personal research like this is perhaps 2-3 hours a week… so in fact rather less total time in a year than most of you have in one semester!

The Research Question: Background

- Most navigation (scrolling) research dates from the 1980s and 1990s…before (modern) touch-screens were available
- Current research recommends not using 2-d scrolling, but rather only using vertical scrolling
- …because users get lost if the scrolling is 2-d
- All the previous research has been on navigating long text….and not on visual workspaces

The Research Question: Motivation

- Recent claims have been made that the old research is outdated
- …and that horizontal scrolling is not a problem
- However, that research looked at *only* horizontal scrolling (no vertical movement allowed)
- …and *only* for swiping pages on a touch-screen

The Research Question

Therefore, we wanted to see if:

- Do the same problems with scrolling on text occur when scrolling over image-rich data?
- Is there an impact on the use of different control mechanisms? (mouse versus touch-screen)

- …if either is true, then the guidelines need to be re-evaluated

Experimental Design

- Previous work times users searching for a particular item (text, or an image) in electronic text
- There was structure to the text (paragraphs, headings, etc.) and it was 'meaningful'
- Content was scrollable in both horizontal and vertical dimensions
- There was a task to perform on the content that was plausible for a wide audience
- …e.g. find the diagram on attaching the hose to a vacuum cleaner (as if assembling it after delivery)

Experimental Design

We needed to come up with a visual task that:

- Has a visual target to find
- In an organised larger collection or document
- Scrollable in both horizontally and vertically
- Plausible task (one most people have done before)

- …that we understood and could predict user behaviour within because we understood it

Experimental Design

- What's something you do regularly where you:
- Look around to find something
- In an organised collection or display
- That involves vertical and horizontal movement?

- … a big challenge is coming up with a meaningful task that actually "makes sense" to a reader

Our options

Had been researching a number of cases of browsing high and low for several years, so could choose from them (i.e. we had plenty of ideas)

Examples include:

- Reading large PDF documents (but mostly text)
- Shopping in a store (shelving)
- Finding content in an image (hard to describe)
- Examining a visualization of data for a pattern (again, hard to describe)

Final Choice

- Ended up with library shelves…which was a good candidate because:
- It's known to be hard to replicate on computers
- …so any insights would give a second benefit
- Easy (if tedious) to get data on
- Had recent data on real-world behavior
- …and we had the expertise to build a real data set

Building the Dataset

- Wanted two presentations (to reduce effects from one or other presentation…and kill two birds…)
- Took photographs of real library shelves
- Gathered the book details on a database
- Ran a program to obtain the book covers from Amazon, Google Books and another source (to make sure we got one that worked for each!)
- Checked the covers (to ensure they were correct)
- Wrote a program to display the photographs or digital shelves

Designing the Study

- Standard previous tests included:
- Balanced order (to avoid ordering effects by always showing a particular thing first – whatever is first does worst…)
- ….for both the mouse and the touch-screen (half used mouse first, half touch-screen)
- …and also the photos versus the shelf display
- Trial-ran the study with pilot users before the real experiment to make sure it all worked

Designing the Study #2

- What measurements to take?
- Time to find the target
- Total movement to find the target (horizontal, vertical)
- Amount of reversals of direction (how many changes of mind)
- …any other ideas?

- …built in the measurements into the program (so we didn't capture it manually – unreliable!)

Designing the Study #3

- How many people to use?
- …how much data to collect

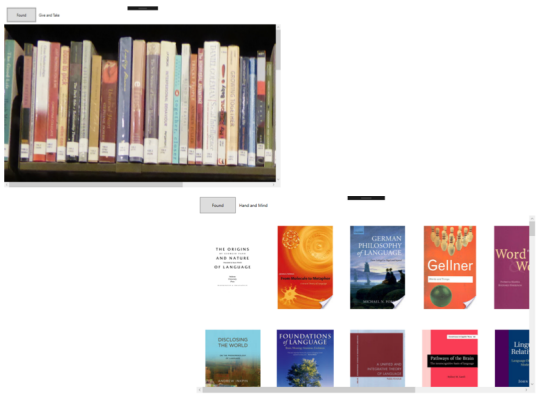
- Always a problem!
- What have previous researchers done?
- …previous ones had used 12-24 participants

- And that was in my experience a viable number *if the differences are substantial*

Designing the Study - Aside

- When using test data to evaluate algorithms, thins are normally complex in other ways…
- You need a lot of data as differences can be small
- But you can retest the data time and time again, as the data doesn't learn

- …in tests that involve people, the problem is that they do!



Basic Tests

- If time to complete the task is normally distributed, then we can use t-test or ANOVA…
- Likewise any movement data
- …success or error rates, chi-squared test (proportion of success, for example)

- The previous researchers had used these tests in their work (c. 10-12 papers)

Problem: Oh dear, the stats

- If you're going to use parametric tests, you need to run a test that the data *is normally distributed*
- ….and it wasn't!

- Was it our data? Did we do something wrong?

- Went back to the previous researchers – and their data had similar characteristics
- Low average (mean) values, and standard deviation greater than the mean… that's not normal!

What would you do?

- Re-run the experiment?
- Ask your supervisor?
- Phone a friend?
- Try a different test?

Find an expert

- Talked to a colleague who is a known statistics guru
- Asked the previous researchers what they had done (and for their reasoning)
- Ran different suggested tests on the data to see if we got different significance results (all $p<0.01$)
- Thankfully they all agreed
- Used the simplest standard test for reporting in the paper…but with a footnote about complications

Statistical Options

- Wilcoxon (non-parametric)
- Trimmed means
  (throw away the highest and lowest 10%)
- Log-linear (good fit test, but rare)
- Just use ANOVA anyway (allegedly “robust”)
- Previous work had used ANOVA but not tested for normality…

The consequences

- The study was put in abeyance for a year…
- Then re-ran the test with other improvements and refinements, focusing on re-finding a target
- …not just finding it the first time

- The second test flowed from knowing how people behaved when searching for an idea
- ….rather than a specific book

Qualitative Data

- Qualitative data reveals people's impressions and feelings about a technology
- And can indicate how much mental effort is required to use it

- This was also part of our goal – which technology is "easier" to use, isn't necessarily the fastest
- Most research involves "proxy values" – testing data that we can get, when we can't directly measure what we want

Qualitative Data

- After their test, each participant was debriefed with five questions about:
- Their preferred control (touch, mouse)
- …and why they preferred that
- Their preferred presentation (shelf, photograph)
- …and why that was preferred
- What they found difficult in the task

Writ

The final submitted paper reported:

- The original test
- The refinding test
- The statistical problem encountered

- …because in fact the latter is likely to get spotted by someone else later – we better warn them!

Summary

- Don't give up – if there's a complication, think about what the data is telling you
- Other researchers may have made a mistake – don't just take it on trust everyone is perfect
- If you encounter a complication, make a virtue of it – maybe you can write a better paper or article
- …that will really help future researchers