New RSP Unit:
Research Design for Human Participants
Friday 13 October – 9:00 in FW26

The unit will provide an overview of quantitative and qualitative data collection and analysis methods, the construction of open and closed research questions, the design of controlled experiments with human participants, and threats to validity of research results.

Research Design for Human Participants

Alan Blackwell

Design of a research study

- Research design: turns a research *question* into a research *project*
- Research designs may be *fixed* or *flexible* (open or closed questions)
- Fixed questions are often associated with collection of *quantitative* data
- Flexible questions are often associated with *qualitative* data
- Fixed questions may involve evaluation on some criterion
- Flexible questions can identify unknowns, and even “unknown unknowns”

Controlled experiments

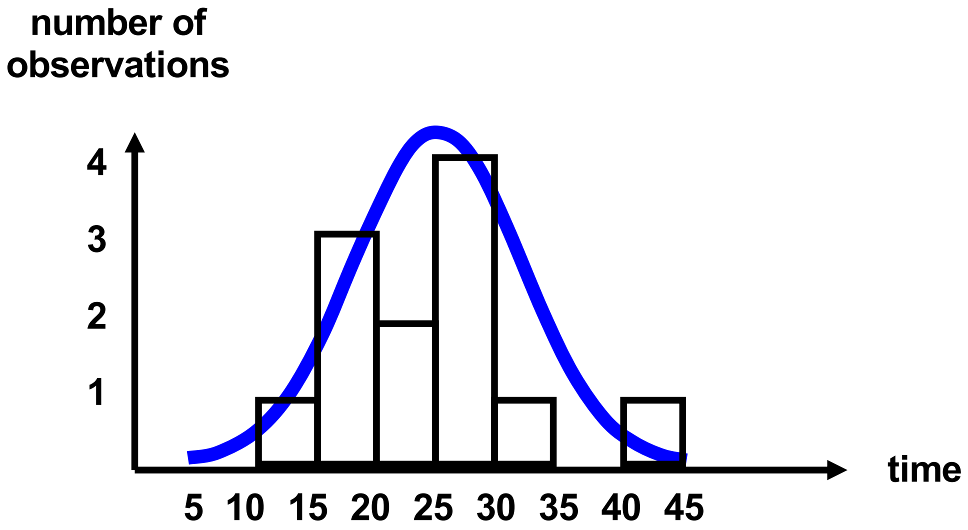
Controlled Experimental Methods

- ***Participants*** (subjects), potentially in ***groups***
- Experimental ***task***
- Performance ***measures*** (speed & accuracy)
- Trials
- ***Conditions*** / Treatments / Manipulations
  - modify the system
  - use alternative systems
  - Use different features of the system
- ***Effect*** of treatments on sample means
  - Within-subjects (each participant uses all versions)
  - Between-subjects (different groups use different versions)

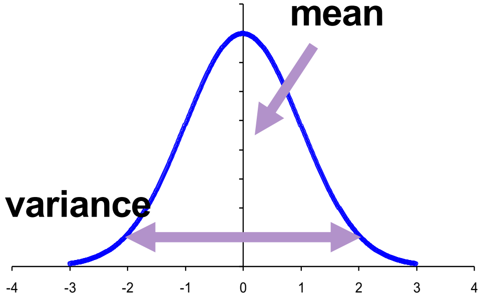
Controlled Experiments in HCI

- Based on a number of observations:
  - How long did Fred take to complete this task?
  - Did he get it right?
- But every observation is different.
- So we compare averages:
  - over a number of trials
  - over a range of people (experimental subjects)
- Results often have a normal distribution

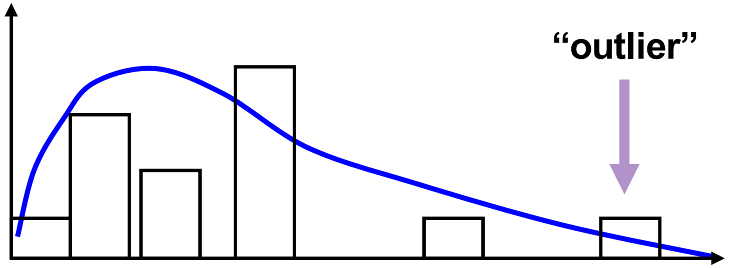
Sample Distribution



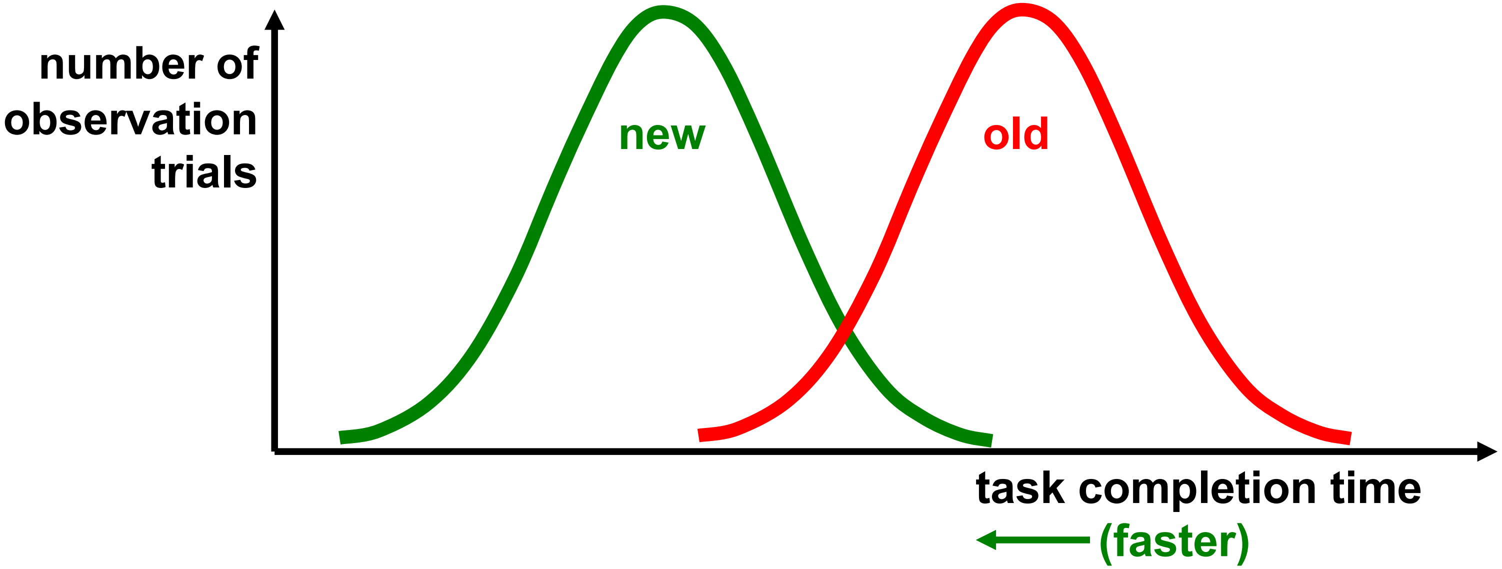
normalisation



log normalisation

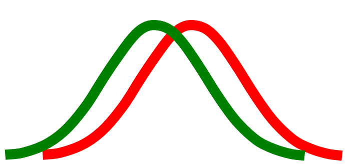


Effect Size

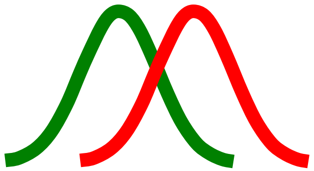

number of observation trials
new
old
task completion time
(faster)

Significance testing

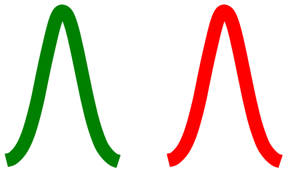
- What is the likelihood that this amount of difference in means could be random variation between samples (null hypothesis)?
- Hopefully very low ($p < 0.01$, or 1%)



only random variation observed



observed effect probably does result from treatment



very significant effect of treatment

Experimental Manipulations

- Compare productivity gains (effect size) of version with new feature to one without?
  - Will system work without the new feature?
  - Will the experimental task be meaningful if the feature is disabled?
  - Must new feature be presented second in a within-subjects comparison (order effect)
  - Is your system sufficiently well-designed for external validity of productivity measure?
- Is full implementation necessary?
  - Can you simulate features with Wizard of Oz technique?

Measurement

- Speed (classically 'reaction time')
  - Time to complete task
- Accuracy (number of (non)errors).
  - Is outcome as expected
- Trade-off between speed and accuracy?
  - Or poor performance on both?
  - Check correlation between them
- Task completion:
  - Stop after a fixed amount of time (ideally < 1 hour)
  - Measure proportion of the overall task completed

Self-Report

- Did you find this easy to use? (Likert scale)
  - applied value: appeal to customers
  - theoretical value: estimate 'cognitive load' – e.g. NASA TLX
- Danger of bias
  - Subjective impressions of performance are often inaccurate
  - Reports may be influenced by experimental demand
    - Participants want to be nice to the experimenter
    - Should disguise which manipulation is the novel one
- May be necessary to capture affect measures:
  - Did you enjoy it, feel creative/enthusiastic, experience a 'flow' state?
  - Use a standardized scale whenever possible for calibration and comparison

Experiment Design

- Arrangement of participants, groups, tasks, trials, conditions, measures, and hypothesized effects of treatments
- Within-subjects designs are preferred
  - because so much variation between individuals, it's more reliable to consider how any one person's responses change
- This leads to order effects:
  - first condition may seem worse, because of learning effect
  - last condition may suffer from fatigue effect
  - task familiarity – can't use the same task twice
- Precautions:
  - Prior training to reduce learning effects
  - Minimise experimental session length to reduce fatigue effects
  - Use different tasks in each condition, but 'balance' with treatment and order
- These are typically combined in a 'latin square' where each participant gets a different combination

Analysis (cookbook statistics – Damon Wischik will describe principles)

- For an easy life, plan your analysis before collecting data!
- Will quantitative data be normally distributed?
  - t-test to compare two groups
  - ANOVA to compare effect of multiple conditions (with latin square of task/order?)
  - Pearson correlation to compare relationship between measures
- Distributions of task times are often skewed:
  - a small number of individuals complete the task quite slowly
  - don’t exclude ‘outliers’ who have difficulty with your system
  - log transform of time often turns out to be normally distributed (e.g. Fitts’ Law)
- Subjective ratings are seldom normally distributed
  - could use chi-square test of categories or permutation tests
  - consider non-parametric comparison of means, though remember that *means* may be normally distributed (central limit theorem)

Flexible designs

Usability evaluation (or "design probe")

- Rather than testing hypothesis, or comparing treatments
  - ask 'is my system usable' (a.k.a. 'fit for purpose', in a user-centric project)?
  - Potentially identify requirements, or register usability 'issues' for bug tracking
- More typical of commercial practice, for short-term rectification of immediate problems, rather than general understanding of design principles
  - Formative evaluation assesses alternatives early in the design process
  - Summative evaluation identifies usability problems in a system you have built
  - Repeated for iterative refinement in user-centred design processes
- Weaker as research, because no direct contribution to theory
  - But applied research venues require *evidence* of claims made for new tools, so doing a usability study may be the cost of publication
  - And a functional "probe" might also operate as an observational instrument to collect data about some context or phenomenon (e.g. Sellen et al "HomeNote")

Think-aloud studies – cognitive science or market research?

- “Tell me everything you are thinking”
  - a.k.a. ‘concurrent verbalisation’
- Problems:
  - Hard tasks become even harder while speaking aloud
  - During the most intense (i.e. interesting) periods, participants simply stop talking
- Alternative:
  - make a screen recording (showing cursor, or even eye-tracking trace?)
  - play this back for participant to narrate
  - ‘retrospective verbal report’

Field Study Methods

- Laboratory studies are not adequate for:
  - understanding context of system deployment (homes, companies, countries …)
  - understanding interactions within a community of users
- Typical methods:
  - ‘contextual inquiry’ interviews
  - ‘focus group’ discussions
  - ‘case studies’ of projects or organisations
  - ‘ethnographic’ field work as participant-observer
- All result in qualitative data, often transcribed, and in HCI research often analysed using *grounded theory* approaches
  - (see video from Part 1B Further HCI: https://youtu.be/xnxrXR3cRPY)

Analysing Qualitative Data

- Protocol analysis methods, e.g.
  - verbal protocol – transcript of recorded verbal data
  - video protocol – recording of actions
- Hypothesis-, or theory-driven
  - Create 'coding frame' for expected/hypothetical categories of behaviour
  - Segment the protocol into episodes, utterances, phrases etc
  - Classify these into relevant categories (considering inter-rater reliability)
  - Compare frequency or order statistically
- Grounded theory
  - Open coding, looking for patterns in the data
  - Stages of thematic grouping and generalization
  - Constant comparison of emerging framework to original data
  - More interpretive, danger of subjective bias

General considerations

Theoretical goal

- What do you expect to learn from conducting your study?
- What contribution will it make to the research literature?
- Where would you publish the results?

- A good starting point is to review contributions that were made in published studies you would like to emulate
  - Warning – be careful of studies done without prior training in HCI, and not published in peer-reviewed HCI venues.

Practical considerations

- Do you wish to carry out a comparison between systems, a (usability) evaluation of one system, or an open exploratory study – perhaps with no existing system?
- If you plan to conduct a controlled experiment, will it be possible to use a within-subjects design to reduce uncertainty resulting from variation between participants?
- What data analysis method will you use?
- What would you need to do in order to complete a pilot study?
- What ethical issues are raised by your planned research?

- A safe starting point is to choose a published study that you would like to emulate.

Choosing tasks and measures

- Identify user activities you plan to observe
  - *either* assigned tasks (controlled experiment)
  - *or* toward the user's own goals (observational study)
- Will these explore an interesting research *question*?
- What *measures* are relevant to that question?
- Will *qualitative* data analysis be necessary?
- Will there be a threat to validity?
  - Potentially resulting from choice of task, choice of measure or approach to analysis

Threats to validity of a study

- Face validity
  - Does the superficial appearance of the study reflect its actual purpose?
- Construct validity
  - Does your data really measure what you say it does?
- Internal validity
  - Did the measured effects actually result from the suggested causes?
- External validity
  - Can your (controlled/sampled) results be applied to other contexts?

Techniques for remote studies (e.g. if required by pandemic 🤞)

- Surveys and questionnaires
- Interviews (e.g. by Zoom, potentially recorded)
- Instrumented remote prototypes (i.e. telemetry)
- Diary studies & experience sampling (see https://www.microsoft.com/en-us/research/project/meetings-during-covid-19/ for a recent example)
- Things that don't work well:
  - prototypes requiring a complicated software setup or low latency interaction
- Paid recruitment tools: UserTesting.com, AMT, Microworkers, Prolific, Gorilla, Sona
- Free recruitment tools: r/SampleSize, friends and family, this class (beware bias)!
- Survey/questionnaire deployment tools: Microsoft Forms, Google Forms, Survey Monkey

Ethical Review of Human Participants Research

- Review the Cambridge Technology Ethics guide
  - What kind of study are you planning?
  - What potential concerns might there be?
  - What will you do to address them?
- Submit a proposal to the Computer Science Ethics committee, giving above details.
  - https://dbwebserver.cl.cam.ac.uk/Administration/Ethics/EthicsRequest.aspx
  - (accessible from department VPN, using department login not Raven)

RSP attendance question:

Remember the ethics application!

Reading suggestions

- Robson and McCartan (4th ed. 2016)
  - *Real World Research*
- Cairns and Cox (2008)
  - *Research Methods for Human-Computer Interaction*
- Cambridge guidance on human participants
  - https://www.tech.cam.ac.uk/research-ethics/school-technology-research-ethics-guidance
- Preece, Rogers and Sharp (6th ed. 2023 - but use an older one!)
  - *Interaction Design beyond HCI*