Evaluation

Plan for today

Controlled experiments

- Not the whole topic! But highlights to keep in mind specifically for controlled experiments with programming interactions.

Final projects group work

Baseline assumptions…

- That this isn’t your first exposure to the scientific method/experiments

The classic

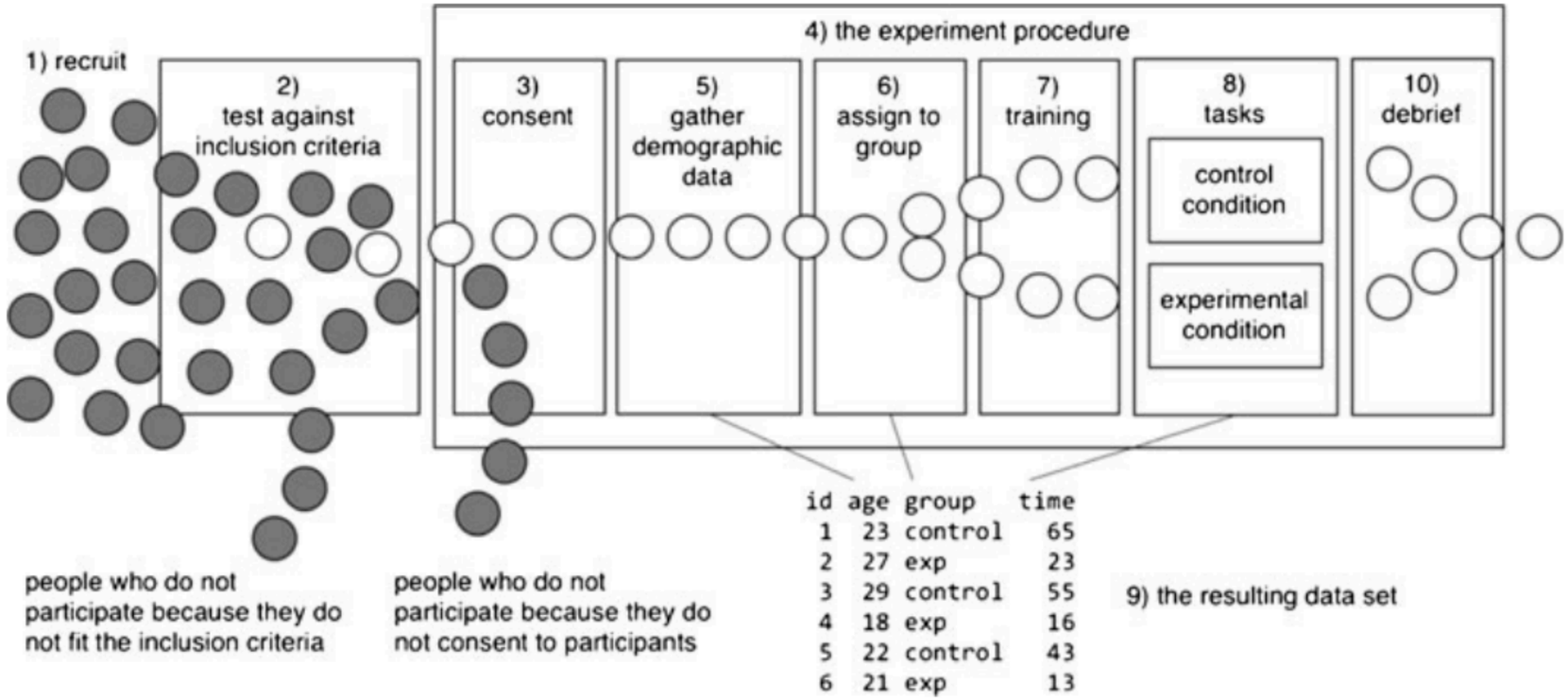


Fig. 1 A canonical design for a tool evaluation experiment with two conditions and a set of tasks. The *circles* represent human participants; the *white circles* are those that satisfy the inclusion criteria. This design includes one independent variable and two conditions. The resulting data set is listed at the bottom

A practical guide to controlled experiments of software engineering tools with human participants, Amy J. Ko et al.

Key Goals

Internal Validity

- Conclusions are warranted within the given setting
- Controlled extraneous variables, eliminated alternative explanations
- Measures are accurate

External Validity

- Conclusions can be generalized to other contexts

Ecological Validity

- Conclusions can be generalized to real-world contexts

External Validity

Different people

Judges with different levels of experience were recruited. This allows the results to be generalised to other judges (different people).

Different places

Different size venues with different size crowds were used. This allows the results to be generalised to different venues and events (different places).

Across different times

The experiment was carried out across a whole year. This allows the results to be generalised to more than just a single point in time (across different times).

Internal Validity

Participants

Random assignment

Judges were randomly assigned to either a 'crowd noise group' or a 'no crowd noise group'.

No Noise

Noise ✓

The 'no crowd noise group' judged in silence, wearing noise-cancelling headphones

The 'crowd noise group', judged while experiencing the full noise of the crowd.

Scorecards were collected after each fight. The scores awarded by the judges in each group were compared to see if the cheering crowd had influenced scores.

Score Card		Compare	Score Card	
Boxer 1	Boxer 2		Boxer 1	Boxer 2
50 points	48 points	⟷	49 points	49 points

CAUSE

EFFECT

Internal Validity: The experiment involved randomly assigning participants (judges in this experiment) to either a crowd noise or a silent no crowd noise condition. Everything else was exactly the same, to see if a noisy crowd influenced the points judges awarded.

External Validity: To be confident that results of the experiment not only applied to people participating in the experiment, we used different size venues and crowds (different places) judges with different levels of experience (different people), across a whole year (different times).

Key Goals: PL edition

Internal Validity

- Did you control for the fact that different programmers have different prior exposure to language X?
- Does your post-test actually assess knowledge of concept Y?
- Did the participants actually use feature Z to complete the task, or did they find some other solution?

External Validity

- Did you study only students in class X at university Y? Will your conclusions apply to class Z at university Y?
- Did you study language A programmers? Will this hold for language B programmers?

Ecological Validity

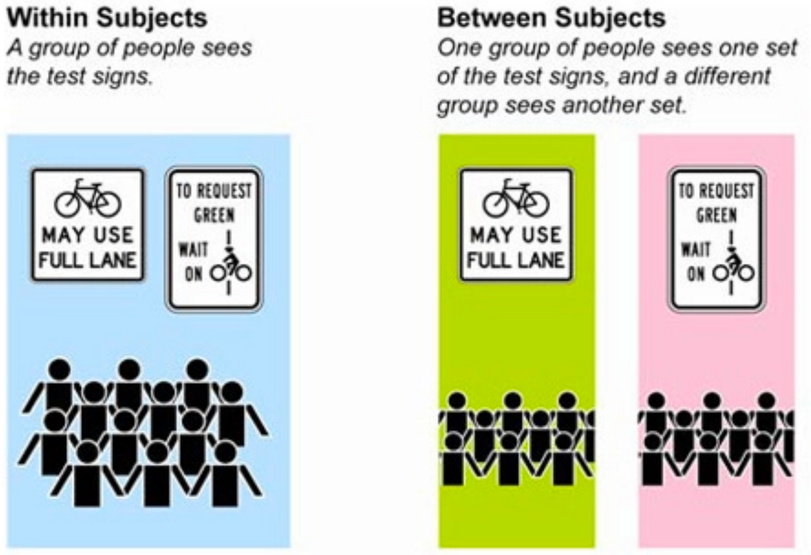
- Is the task codebase like real codebases?
- Is the goal set out in the study reflective of real users' goals?
- Are these participants like the real users?
- Is the study environment like users' real environments?
- Did you let them Google? Can the real users Google?

How can I make my experiment likely to produce a definitive answer?

- Do you expect a big difference when you vary the independent variables?
  - Yes!
    - Likely to get a solid answer even with few participants.
  - Probably not.
    - Are you sure a user study is what you're looking for? Maybe the user experience/performance just isn't the driver of this work?
      - If it's statistically significant, but it's tiny, how important is it to us?
    - Are you sure you're measuring the right element of user experience/performance?
  - How would I know??
    - Well, have you been doing iterative design and checking how users use your innovation throughout?

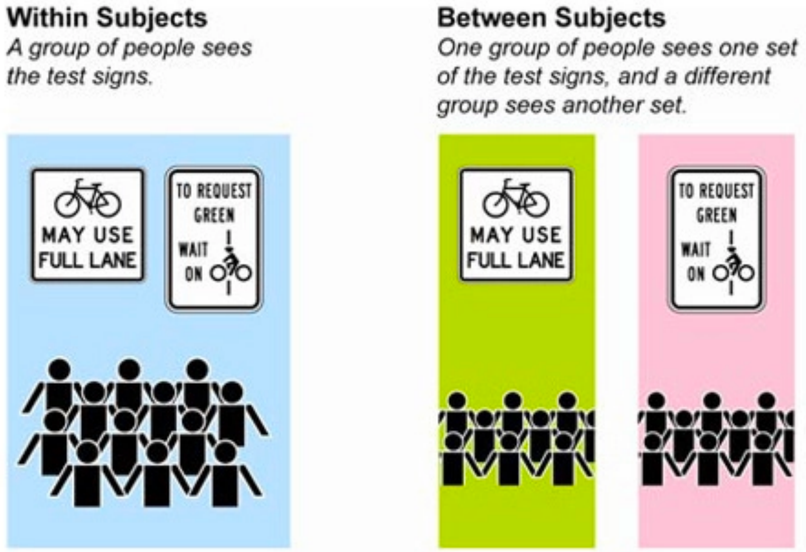
One more answer: Within-Subjects Design

- Controls for variations across individuals
- But some pitfalls…
  - Need to *counterbalance*
    - If everyone sees Tool A before Tool B…
      - *Learning effects*
    - I'm not going to get into *Latin Squares* today, but if you have a within-subjects experiment with more than two conditions, just know that that's the key term to look up!



One more answer: Within-Subjects Design

- Also putting each participant through multiple conditions can make your sessions quite long



Who can participate in my user study?

- See the reading for lots of really useful practical guidance, but we're going to cover one really important rule here
- ~~YOU~~
- You can do all the work in expressivity evaluations, but you gotta stay out of the usability ones

Demand Characteristics

Common demand characteristics include:

- **Rumors of the study** – any information, true or false, circulated about the experiment outside of the experiment itself.
- **Setting of the laboratory** – the location where the experiment is being performed, if it is significant.
- **Explicit or Implicit communication** – any communication between the participant and experimenter, whether it be verbal or non-verbal, that may influence their perception of the experiment.

Some involve the participant taking on a role in the experiment. Roles include:

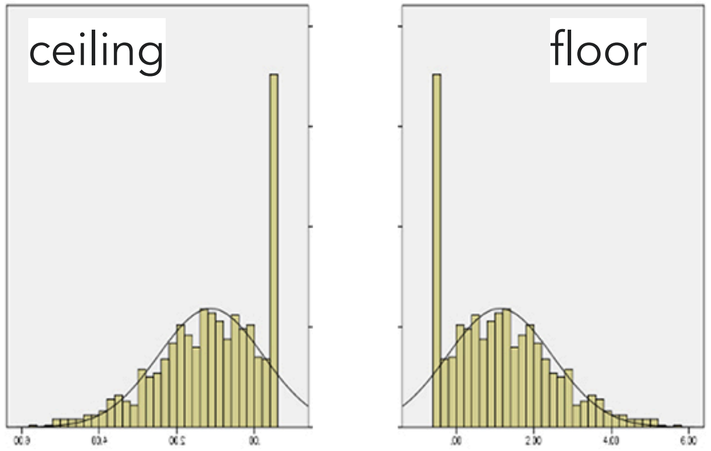
- The **good-participant role** in which the participant attempts to discern the experimenter's hypotheses and to confirm them. The participant does not want to "ruin" the experiment.
- The **negative-participant role** (also known as the **screw-you effect**) in which the participant attempts to discern the experimenter's hypotheses, but only in order to destroy the credibility of the study.
- The **faithful-participant role** in which the participant follows the instructions given by the experimenter to the letter.
- The **apprehensive-participant role** in which the participant is so concerned about how the experimenter might evaluate the responses that the participant behaves in a socially desirable way.

These examples shamelessly lifted from https://en.wikipedia.org/wiki/Demand_characteristics which was actually pretty good considering how short it was.

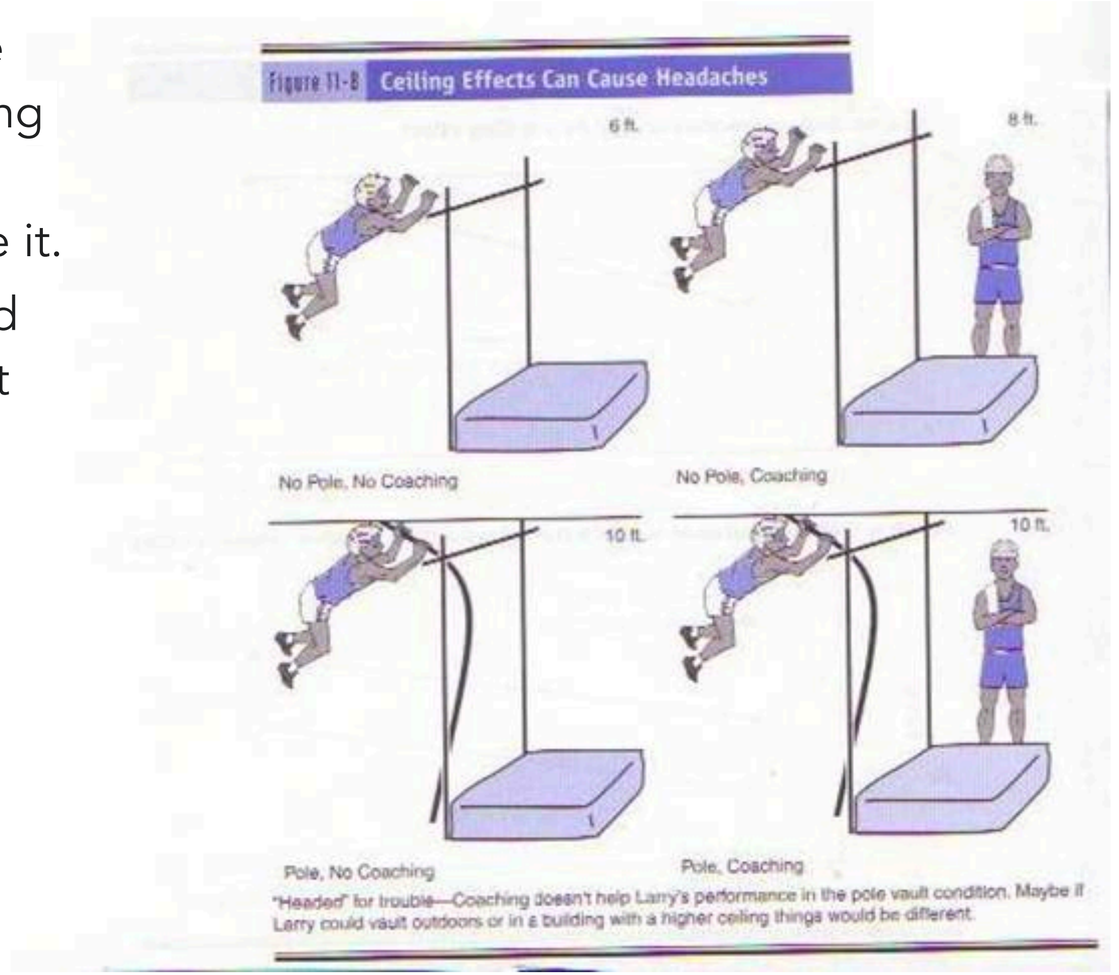
More Effects

Ceiling effects: Everyone's scoring at the top. People could be going higher, but you're not seeing it because you put the ceiling too low. You're artificially putting a lot of the population at the same place (the ceiling), when they should be spread out above it.

Floor effects: Everyone's scoring at the bottom. People should be spread out below the floor of your test, but your test doesn't test for that, so it looks like everyone's at the floor.



* also see right-censored and left-censored data



How do we tradeoff between…

- number of tasks
- study duration
- task difficulty
- between- or within-subjects (or alternative) design
- number of participants

??

The magic solution

- Piloting

What to measure

People measure…

- task completion
- time on task
- failure detection
- search effort
- accuracy
- precision
- correctness
- solution quality
- program comprehension
- confidence
- usability
- utility
- mistakes
- tool-specific metrics

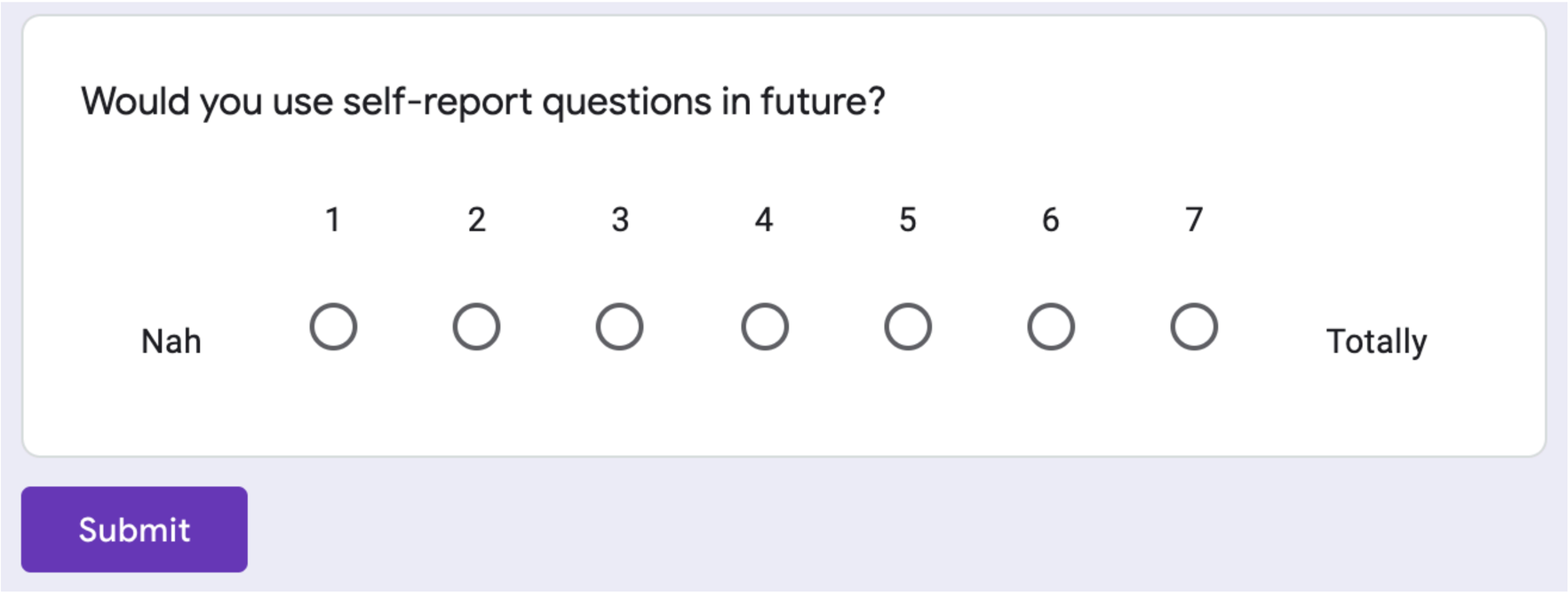
What is “done”?

- You get to decide!
  - And it might be surprisingly hard. Did we mention piloting??

- And once you’ve decided, you still have to decide *when* you’ve reached it.
  - Options:
    - You decide:
      - Via automation
      - Via subjective human decision
        - Inter-rater reliability
    - Participant decides!

Self-report

What do we think about it?



How to do it right

- Ideally you figure out a good domain-specific way to assess usefulness
- But if you must use self-report for usefulness assessment…
  - TAM (Technology Acceptance Model) is validated

Debriefing

- Key reminder: Tell participants how to solve the task if they didn’t get there! Very frustrating to be left hanging like that.
  - And ethicists are insistent on this.
- And remind them not to talk to their friends about it if their friends might do the study too
- Good opportunity to collect info you’ll use for shaping the tool even if it’s not for publication!

Let’s design some evals!

- Final project groups
- HW 11