Controlled Experiments

experimental investigation of a testable hypothesis, in which conditions are set up to isolate the variables of interest ("independent variables") and test how they affect certain measurable outcomes (the "dependent variables")

- good for
  - quantitative analysis of benefits of a particular tool/technique
  - establishing cause-and-effect in a controlled setting
  - (demonstrating how scientific we are!)
- limitations
  - hard to apply if you cannot simulate the right conditions in the lab
  - limited confidence that the laboratory setup reflects the real situation
  - ignores contextual factors (e.g. social/organizational/political factors)
  - extremely time-consuming!

See:

Pfleeger, S.L.; Experimental design and analysis in software engineering. *Annals of Software Engineering* 1, 219-253. 1995

Definitions

- Independent Variables
  - Variables (factors) that are manipulated to measure their effect
  - Typically select specific levels of each variable to test
- Dependent Variables
  - "output" variables - tested to see how the independent variables affect them
- Treatments
  - Each combination of values of the independent variables is a treatment
  - Simplest design: 1 independent variable x 2 levels = 2 treatments
    - E.g. tool A vs. tool B
- Subjects
  - Human participants who perform some task to which the treatments are applied
  - Note: subjects must be assigned to treatments randomly

Hypothesis Testing

- Start with a clear hypothesis, drawn from an explicit theory
  - This guides all steps of the design
  - E.g. Which variables to study, which to ignore
  - E.g. How to measure them
  - E.g. Who the subjects should be
  - E.g. What the task should be
- Set up the experiment to (attempt to) refute the theory
  - H_0 - the null hypothesis - “the theory does not apply”
    - Usually expressed as *no effect* - the independent variable(s) will not cause a difference between the treatments
    - H_0 assumed to be true unless the data says otherwise
  - H_1 - the alternative hypothesis - “the theory predicts…”
    - If H_0 is rejected, that is *evidence* that the alternative hypothesis is correct

Assigning treatments to subjects

- Between-subjects Design
  - Different subjects get different treatments (assigned randomly)
  - Reduces load on each individual subject
  - Increases risk that confounding factors affect results
    - E.g. differences might be caused by subjects varying skill levels, experience, etc
    - Handled through blocking: group subjects into “equivalent” blocks
    - Note: blocking only works if you can identify and measure the relevant confounding factors
- Within-subjects Design
  - Each subject tries all treatments
  - Reduces chance that inter-subject differences impact the results
  - Increases risk of learning effects
    - E.g. if subjects get better from one treatment to the next
    - Handled through balancing: vary order of the treatments
    - Note: balancing only works if learning effects are symmetric

Multiple factors (factorial design)

- Crossed Design
  - Used when factors are independent
  - Randomly assign subjects to each cell in the table
    - Balance numbers in each cell!
  - E.g. 2x2 factorial design:

		Factor B	
		Level 1	Level 2
Factor A	Level 1	A1B1	A1B2
	Level 2	A2B1	A2B2

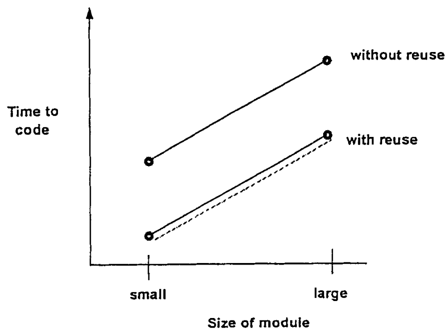
- Nested Design
  - Used when one factor depends on the level of the another
  - E.g. Factor A is the technique, Factor B is expert vs. novice in that technique

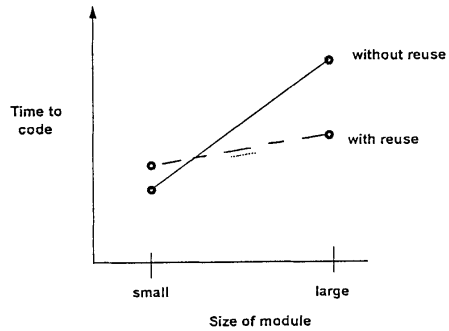
Factor A			
Level 1		Level 2	
Factor B		Factor B	
Level 1	Level 2	Level 1	Level 2
A1B1	A1B2	A2B1	A2B2

Experiments are Positivist

- Relies on reductionism:
  - Assume we can reduce complex phenomena to just a few relevant variables
  - If critical variables are ignored, results may not apply in the wild
  - Other variables may dominate the cause-and-effect shown in the experiment
- Interaction Effects:
  - Two or more variables might together have an effect that none has on its own
  - Reductionist experiments may miss this
    - E.g. A series of experiments, each testing one independent variable at a time
  - Using more than one independent variable is hard:
    - Larger number of treatments - need much bigger sample size!
    - More complex statistical tests

Detecting Interaction Effects



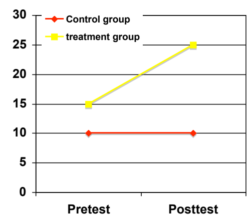


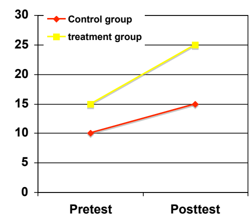
When not to use experiments

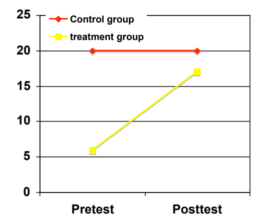
- When you can't control the variables
- When there are many more variables than data points
- When you cannot separate phenomena from context
  - Phenomena that don't occur in a lab setting
  - E.g. large scale, complex software projects
  - Effects can be wide-ranging.
  - Effects can take a long time to appear (weeks, months, years!)
- When the context is important
  - E.g. When you need to know how context affects the phenomena
- When you need to know whether your theory applies to a specific real world setting

Quasi-experiments

- When subjects are not assigned to treatments randomly:
  - Because particular skills/experience needed for some treatments
  - Because ethical reasons dictate that subjects get to choose
  - Because the experiment is conducted on a real project
- e.g. A Non-equivalent Groups Design
  - Pretest-posttest measurements, but without randomized assignment
  - E.g. two pre-existing teams, one using a tool, the other not
  - Compare groups' improvement from pre-test to post-test







Validity (positivist view)

- Construct Validity
  - Are we measuring the construct we intended to measure?
  - Did we translate these constructs correctly into observable measures?
  - Did the metrics we use have suitable discriminatory power?
- Internal Validity
  - Do the results really follow from the data?
  - Have we properly eliminated any confounding variables?
- External Validity
  - Are the findings generalizable beyond the immediate study?
  - Do the results support the claims of generalizability?
- Empirical Reliability
  - If the study was repeated, would we get the same results?
  - Did we eliminate all researcher biases?