Common threats to validity in empirical user research

The 4 aims of empirical research

1. **Reliability:** Results can be replicated by others

2. **Validity (internal):** Results show what we intend them to show
    1. Ability of a research design to test the hypothesis it was designed to test
    2. Measure what we want to measure

3. **Generalizability (external validity):** Results have a wider application than merely the participants and the circumstances of the test

4. **Importance:** Results should be important (subjective).
    1. Results are never important if not reliable, valid and generalizable

Validity

- Any measure/score obtained consists of:
  1) A **true value** for what we measure
  2) A value for "**other things**" that are inadverdently measured
  3) **Systematic**, non-random **bias**
     - Ok, as long as it affects every participant the same
  4) **Non-systematic**, random **bias**
     - Should cancel out over large numbers of observations

- The goal is that our measure should **as close to the true value as possible**

Validity

- Good experimental designs **maximise validity**

- **Internal validity:**
  - Extent to which we can be sure that changes in the DV are due to changes in the IV [meteor kills dinosaurs].
  - Requires confounding variables are eliminated

- **External validity (generalizability):**
  - Extent to which we can generalise from our participants to other groups (e.g. to real-life situations).

Validity

- **Ecological validity**
  - Extent to which research results can be applied to real life situations outside research settings
  - Often used = external validity
  - But focused on the degree to which findings can be **observed in the real world**
  - To have ecological validity, a research design must closely **mimic** the **real life situation under investigation**
  - (Ecology = science of interaction between organism and its environment)

Threats to validity I

Threats to the **internal validity** of an experiment's results:

- **Time threats:**
  Time passage
  History
  Maturation
  Selection-maturation interaction
  Repeated testing
  Instrument change

- **Group threats:**
  Initial non-equivalence of groups
  Regression to the mean
  Control group awareness

Participant reactivity threats:
Experimenter effects
Reactivity
Evaluation apprehension.

Threats to validity II

History threats

- *Extraneous events between pre-test and post-test affect participant's post-test performance.*

Example:

1. Ask participants how often they use condoms
2. Administer advice on safe sexual practices
3. Unrelated, media publicises statistics showing STD's are on the increase
4. Two weeks later: Ask participants how often they use condoms

Threats to validity II

- Changes in reported sexual behaviour may be due to advice, or due to participants' heightened awareness of dangers of unsafe sex due to media coverage. **Confounding factor** in play.

- **Solution:** Add a **control group** that is not given advice on safe sex.
  - Make sure the only factor varying is the IV

- **Note:** This is NOT possible in correlational research -> main challenge in correlation

Threats to validity III

Maturation threats:

- Participants may **change** during the course of the study (e.g. get older, more experienced, fatigued, etc.).

Example: Effects of an educational intervention on reading ability:

1. Children's reading ability tested at age 6.
2. Educational treatment administered.
3. Children's reading ability tested again, at age 9.

- Changes in reading ability may be due to **reading program** and/or **normal developmental changes** with age.

- **Solution:** Add a **control group** who do not receive the reading program, and whose reading ability is tested at ages 6 and 9.

Threats to validity IV

Selection-maturation interaction:

- Different participant groups have different **maturation rates**, that affect how they respond to the experimenter's manipulations.

Threats to validity IV

- Example: Effectiveness of sex education program in Jurassic Park

1. 20-year old dinosaurs in experimental group;
2. 18-year old dinosaurs in control group
3. Pre-test on knowledge about sex
4. Administer sex education program
5. Post-test a year later: Experimental group know more about sex

- But - results may be due to maturational differences (puberty in older group of dinosaurs) and/or exposure to program.

- **Solution:** Ensure **groups differ only on one Independent Variable** (e.g. in this case match groups for age).

Threats to validity V

Time threats: Repeated testing

- Taking a pre-test may alter the results of the post-test.

Example: Effects of fatigue on emergency braking in a simulator:

1. Pre-test: Measure driver's braking RT to an unexpected hazard.
2. Fatigue induction (30 minutes' simulator driving).
3. Post-test: Measure driver's braking RT to an unexpected hazard.

Problem: Pre-test may **alert drivers to possibility of unexpected tests**, and hence maintained concentration at higher levels than otherwise.

Solution: In studies like this, avoid repeated testing or add a control group who get only the post-test.

Threats to validity VI

Instrument change threats:

- E.g. experimenter tests all of one group before testing another, but becomes more practiced/bored/sloppy while running the study
- Now **two systematic differences between conditions**:

Intended experiment:	Actual experiment:
Condition A: drug	Condition A: drug + friendly experimenter
Condition B: no drug	Condition B: no drug + bored experimenter

- A problem for **observational studies** (changes in observer's sophistication affects scoring of behaviours).
- **Solution: Highly standardised procedures**; random allocation of participants to conditions; multiple observers, familiarise oneself with behaviours before formal observations begin.

Threats to validity VII

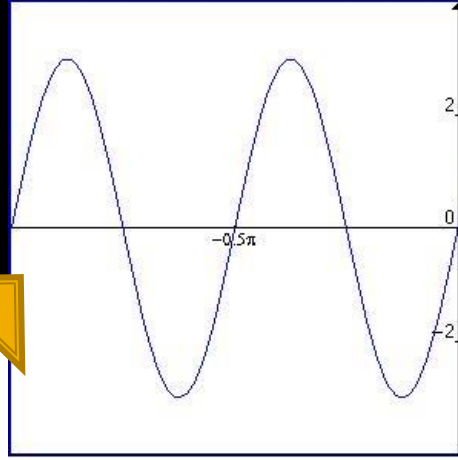
Selection (initial **non-equivalence of groups**):

- Groups differ on many variables other than the one of interest (e.g. gender, age).

Example: Study examines **gender** differences in attitudes to shooting **wookies**

- "Females" are also old ladies, "males" are also stormtroopers. Cannot conclude that observed attitude differences are due **solely** to gender
- **Solution:** Often difficult to fix. Problem of confounding variables.

Threats to validity VIII



Regression to the mean:

- Participants who give very low or very high scores on one occasion tend to give less extreme scores when tested again. **Natural fluctuation**

Example: Testing the effectiveness of a remedial reading program

1. Test children's reading ability;
2. Administer program, but select the *worst* children for it
3. Re-test children - falsely assume that any improvement is due to the reading program and not other factors

Solution:

- Select children randomly, *not* on basis of low scores
- Avoid **floor** and **ceiling effects** with scores (more on those later)

Threats to validity IX

Differential mortality:

- When testing same individuals repeatedly, some may drop out of the study

Example: People in suicide-prevention program.

1. Administer pre-test
2. Provide anti-suicide treatment to group
3. Some participants commit suicide (the treatment did not work)
4. Only survivors in post-test, leading to false positive results of treatment

Solution: Often difficult to fix!

Threats to validity X

Control group problems that stem from social interaction:

Compensatory rivalry:

- If the control group are aware it is **not** receiving the experimental treatment, they may show compensatory rivalry - or resentful demoralisation!

Treatment imitation:

- Control group imitates the experimental group's treatment

Treatment diffusion:

- Benefits from information given to the treatment group is diffused to the control group.

One type of solution: Compensatory equalization of treatments:

- Treatment administrators provide control group with some benefit to compensate them for lacking the experimental treatment (e.g. supply an alternative educational treatment)

Threats to validity XI

Reactivity (Hawthorne Effect):
Practice or fatigue effects in participants, awareness what experiment is about

Example:

- Workers' productivity increased after manipulations of **pay, light levels** and **rest breaks** - regardless of nature of changes made.
- **Problem:** Apparent "productivity" may have been affected by **material factors,** the IVs, - as originally studied, e.g. illumination

But potentially also:

1. **Motivation,** e.g. changes in rewards, piecework pay.
2. **Learning** (practice).
3. **Feedback** on performance.
4. **Attention** and expectations of observers.
5. **Awareness of being studied**

- **Implication: Act of measurement can affect the very thing being measured**

Threats to validity XII

- Finally, but importantly: **Experimenter effects**
- **Expectations** of experimenters (teachers, doctors and managers) may affect performance of the participants
  - Example: Teacher asks students to participate in experiment – teacher chooses their grades, so students try to give teacher what he/she want in the experiment
- **Example: Evaluation apprehension:** People are nervous about being "measured"
- **Example: Placebo effects** - doctors' expectations affect drug effects because patient respond to the expectation.
- **Solution:** "double-blind" procedures if possible - neither doctor nor patient know whether the patient has been assigned to the drug or *placebo condition*

Threats to validity XIII

- Threats to **external validity**
  *Extent to which we can **generalise** from our participants to other groups (e.g. to real-life situations).*

Over-use of participant groups:

- E.g. the overuse of undergraduates in psychology experiments; using volunteers
- i.e.: The groups become **biased** and **not generalizable**

Restricted number of participants

- A threat to reliability but also ability to generalize to the population from the sample.
- Example: Experiments with so few participants we cannot calculate **statistical significance**
- Solution is to control **sampling** (more on this later – basically ensuring sample is representative of the population)