Threats to Validity

Types of threats to validity

- **Construct Validity**

  Are we testing what we intended to test?
- **Internal Validity**

  Are the results solely due to exp manipulations?
- **Conclusion Validity** (statistical validity)

  Are the conclusions that we make justified?
- **External Validity** (generalization)

  How and in what context are the results applicable?

Construct validity

- Are testing what we want or intend to test?
- Similarly to requirements: “Are we building the right system?”
  - If this is wrong, nothing else matters

Construct Validity: *Design threats*

- Inadequate preoperational explication of constructs
  - Constructs are not sufficiently defined before they are translated into measurements and treatments
  - Example: Compare two inspection methods. What is the meaning of better? Find most faults or most faults per hour or most faults per LOC
- Mono-operation bias
  - Is cause-construct under-represented? Single independent variable, case, subject or treatment
  - Does not give a full picture of the theory
- Mono-method bias
  - is single type of measure or observation enough? Or are more needed to cross-check against each other?

Construct Validity: *Design threats*

- Confounding constructs and levels of constructs
  - Sometimes is not the presence or the absence of the construct, but the level of the construct which is important for the outcome
  - Presence of the construct is confounded with the level of the construct
  - Example: Not the presence or the absence of the knowledge of programming language, but the level of experience: 1, 3, or 5 years
- Interaction of different treatments
  - For example a subject involved in more that one study. Is the effect due to either treatment or to a combination of treatments
- Interaction of testing and treatment
  - Testing (i.e., application of treatments) may make subjects more sensitive or receptive to the treatment (e.g. subject awareness)
  - Example: measure number of bugs. Subjects are more careful and make less bugs. Testing becomes treatment

Construct Validity: *Design threats*

- Restricted generalizability across constructs
  - The treatment affects some constructs positively, but unintentionally has negative effect (i.e., side effect) on other constructs
  - Example: A new method increases productivity, but reduces maintainability. If maintainability is not measures, there is a risk of drawing partial or incorrect conclusions

Construct Validity: *Social threats*

- Related to behavior of subjects who may act differently than otherwise, which leads to false results
- Hypothesis guessing
  - Guess what is the purpose and intended result and then act either positively or negatively, depending on their attitude
- Evaluation apprehension
  - Afraid of being evaluated. Look better when being evaluated.
  - Becomes a confounding factor.
- Experimenter expectancies
  - The experimenter can bias the results both consciously or unconsciously. Solution: involve independent people.

Internal Validity

- Influences that can affect the independent variable/measurements without researcher's knowledge
  - Single group threats
    - No control group / sister project. Hard to determine if the treatment or another factor caused the observed effect
  - Multiple group threats
    - Control group and selected group may be affected differently by single group factors
  - Social threats
    - Applicable to single group and multiple group experiments

Internal Validity: Single group

- History
  - If different treatments applied to same object at different times, history may affect the experimental results
- Maturation
  - Subjects can react differently as time passes
    - Negatively: tiered or bored
    - Positively: learn
- Testing
  - if repeated, subjects may respond differently; i.e. from ‘learning’
- Instrumentation
  - effect of artifacts used for experiment execution
  - Example: Instrumentation for profiling adds overhead

Internal Validity: Single group

- Statistical Regression
  - Subjects are classified based on previous experiment or case study
  - May observe improvement, even if no treatment is applied
  - Objects are already ‘similar’ - e.g. hwk1 “winner’s curse”
- Selection
  - Due to variation in human performance. Who and how selected?
  - Example: Volunteers are usually more enthusiastic, and thus may not always be representative of the population
- Mortality
  - Effect of dropping out of case study / experiment
  - Example: All senior reviewers drop out of a case study on effectivness of software inspections
- Ambiguity about direction of causal influence
  - Did A cause B? Did B cause A? Did X cause A and B?

Internal Validity: *Multiple groups*

- Interactions with selection
  - Two groups may mature differently
  - Example: two group use two different methods, one groups learns faster

Internal Validity: *Social threats*

- Diffusion or imitation of treatments
  - control group starts imitating the treatment
- Compensatory equalization of treatments
  - When control group gets compensated
- Compensatory rivalry
  - Underdog effect: “Our old method is great!”
- Resentful demoralization
  - Opposite of the previous. Control group is not motivated: “Old method can’t cut-it anyways.”

Conclusion Validity

- Affects the ability to draw correct conclusions
- Violated assumptions
  - Typical assumption: normality
  - Some test are more sensitive to violating the assumptions
- Low Statistical Power
  - Power: ability of the test to reveal a true pattern in the data (i.e., unable to reject an erroneous hypothesis)
- Fishing & Error rate
  - Searching (i.e., fishing for specific result)
  - Error rate: significance level
- Reliability of measures
  - When the phenomenon is measured twice the outcome should be the same

Conclusion Validity

- Reliability of treatment implementation
  - Standard implementation of treatments over different subjects and occasions
- Random irrelevancies in experimental setting
  - Elements outside of the experimental setting may disturb the results
- Random heterogeneity of subjects
  - Variances due to individual differences may be larger than variances due to the treatment

External Validity

- Limit the ability to generalize the results
- Interaction of **selection** and **treatment**
  - non-representative of population. E.g., wrong people participate in the experiment
- Interaction of **setting** and **treatment**
  - non-representative tools, methods for setting. E.g., case studies/experiments with toy problems
- Interaction of **history** and **treatment**
  - non-representative of regular/normal time