Lecture 5: Introduction to Experimental Design

References

- Donald T. Campbell, Julian Stanley, *Experimental and Quasi Experimental Designs for Research*, Houghton Mifflin Company, **1963**
- John W. Creswell, *Educational research: planning, conducting, and evaluating quantitative and qualitative research, 4th ed.*, Pearson Education, **2012**
- John W. Creswell, J. David Creswell, *Research design: qualitative, quantitative, and mixed methods approaches*, SAGE Publications, **2018**
- Christopher J. Millera, Shawna N. Smith and Marianne Pugatch, *Experimental and quasi-experimental designs in implementation research*, Psychiatry Research, vol. 283, **2020**

Content

- Introductory example
- Definitions
- Characteristics of an experiment
  - Random assignment
  - Control over extraneous variables
  - Manipulation of the treatment conditions
  - Outcome measures
  - Group comparisons
  - Threats to validity

Introduction

Let us assume that we want to improve the quality of the software product and effectiveness of the development process overall.

We can assume that *"Daily code review increase software quality in comparison to weekly code review"*

To confirm or disprove this idea we need to check it in practice.

Experiment is a test under controlled conditions that is made to demonstrate a known truth or examine the validity of a hypothesis.

Experimental Design - Example

The goal of our experiment, or research objectives, is *to test the effect of different conditions of code inspections on a software quality.*

Null hypothesis H_0: *frequent inspections do not increase the number of detected bugs per kLOC.*

Alternative hypothesis H_1: *frequent inspections increase the number of detected bugs per kLOC.*

Experimental Design - Definitions (1/2)

Treatment - the exposure of a group to an experimental variable or event, the effects of which are to be measured.

Subjects - experimental units, who are exposed to treatments.

Observation - the process of measuring the dependent variable within the experiment. Experimenter performs observation(s) before and/or after the treatment intervention.

Experimental Design - Example

We had formed two teams of 9 programmers who conduct inspections for 4 hours at the end of each week. For one of the team we changed the process so they perform daily inspections for 1 hour.
Than we were collecting the data about found defects per kLOC for 4 weeks.

If we give a **treatment** to one group only, we can try isolate whether the treatment and not other factors influence the outcome.

Experimental Design - Definitions (2/2)

An **experimental design** systematically manipulates one or more variables *(frequency and duration of inspections)* in order to evaluate how this manipulation impacts an outcome of interest *(effectiveness of inspection)*. An experiment isolates the effects of this manipulation by holding all other variables constant.

At the end we figured out that first team found 140 defects/kLoC, whereas second team found 50 defects/kLoC.

Our experiment seems to confirm the alternate hypothesis idea: short frequent code inspections help to increase quality of the software product†.
A sound rejection of the null hp requires a statistical test – more later in the course.

† *Read more about it in the paper "'Code Review at Cisco Systems"*

Population and sample (1/2)

Remember:

- **Population** – set of objects that are studied in a task. *It could be all software engineers in the world, employees of XXX company, MSIT students, etc.*
- **Sample** – finite set of objects from the population. *In our example we have selected two teams for the experiment*

Population and sample (2/2)

We have different types of samples:

- **Random sample**: each item in the population is selected randomly, i.e., informally *has an equal probability of being selected.*
- **Convenience sample**: items are chosen based on their convenience and availability (*only few persons agreed to participate in our experiment*)

Characteristics of experiment

A formal, fully fledged experiment has the following key characteristics:

- Random assignment
- Control over extraneous factors
- Manipulation of the treatment conditions
- Outcome measures
- Group comparisons
- Threats to validity
  - internal validity
  - external validity

Random assignment

Random assignment is the process of assigning individuals at random to groups or to different groups in an experiment.

Random assignment $\neq$ Random selection

Random selection is the process of selecting a sample from a population, so that the sample is representative of the population and you can generalize results obtained during the study to the population.

Experiments may not include random selection; however, the most sophisticated type of experiment should involve random assignment.

Control Over Extraneous Factors

Extraneous factors are any influences in the selection of participants, the procedures, the statistics, or the design likely to affect the outcome and provide an alternative explanation for our results than what we expected.

All experiments have some random error that you cannot control, but you can try to control extraneous factors as much as possible by implementing:

- pretests and posttests,
- covariates,
- matching of participants,
- homogeneous samples, and
- blocking variables

Pretests and posttests

A **pretest** provides a measure on some attribute or characteristic that you assess for participants in an experiment before they receive a treatment.

A **posttest** is a measure on some attribute or characteristic that is assessed for participants in an experiment after a treatment.

What are advantages of the pretest? What are disadvantages?

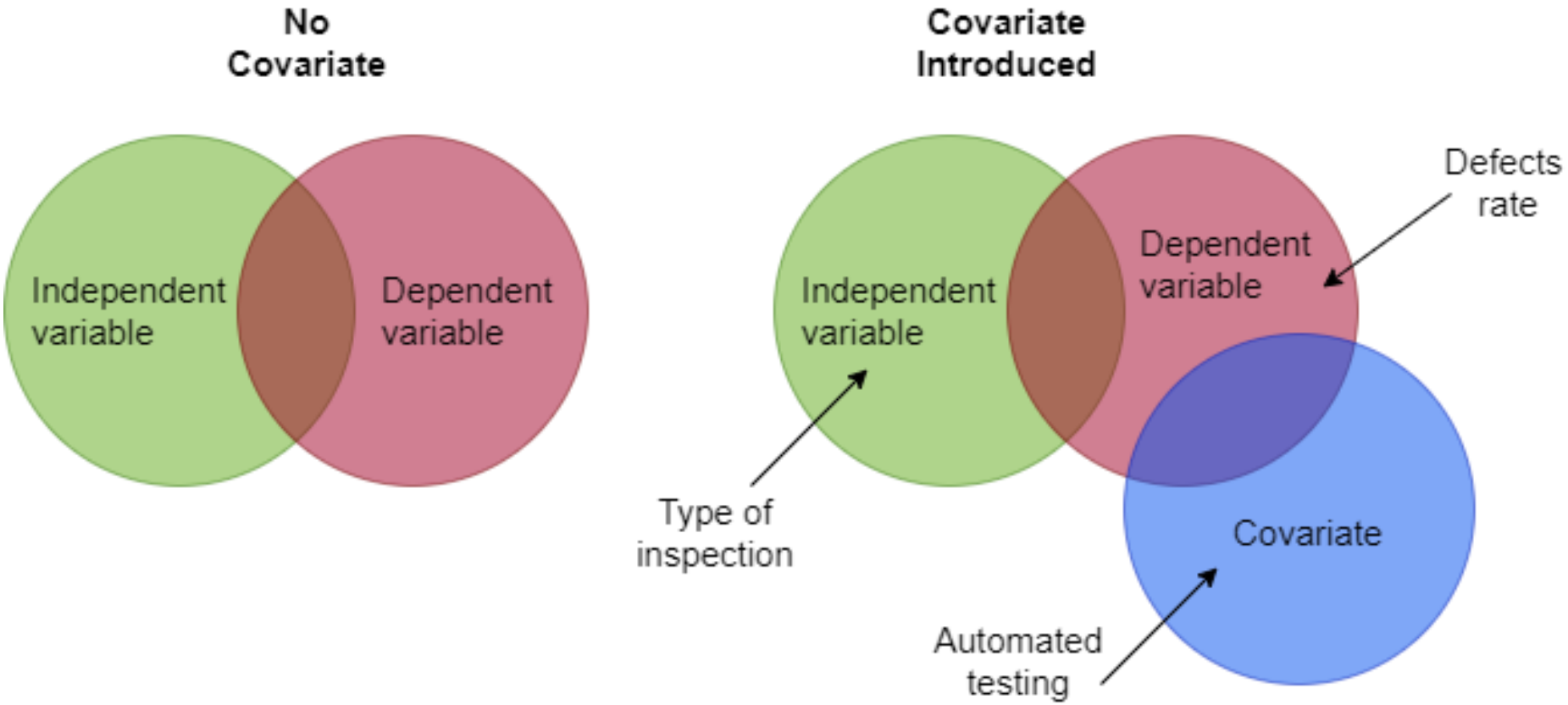
Covariates

Because pretests may affect aspects of the experiment, they are often statistically controlled for by using the procedure of covariance rather than by simply comparing them with posttest scores.

Covariates are variables that the researcher controls for using statistics and that relate to the dependent variable but that do not relate to the independent variable.

The statistical procedure of covariance removes the variance shared by the covariate and the dependent variable, so that the variance between the independent and dependent variable (plus error) is all that remains.

Covariates

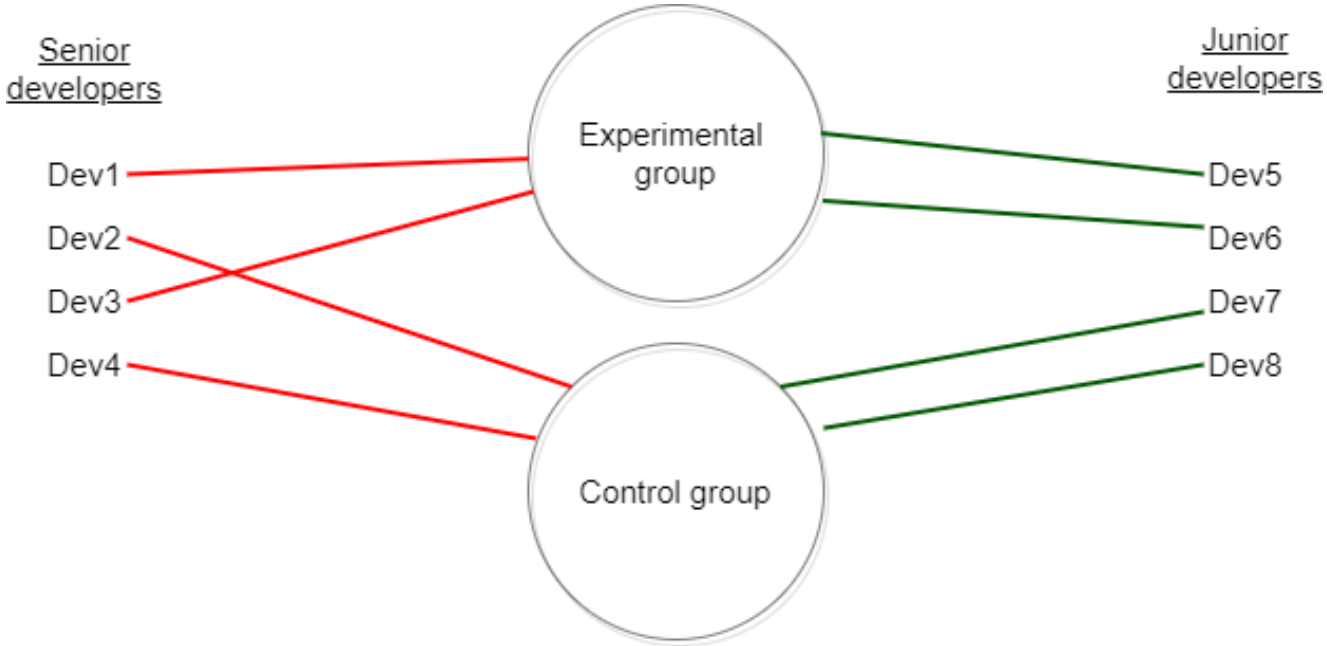


from Creswell pg 299

Matching

Matching is the process of identifying one or more characteristics that influence the outcome and assigning items with that characteristic equally to the experimental and control groups.

Example: matching by experience level



from Creswell pg 300

Homogeneous Samples

Homogeneous sampling is selecting for experimental and control groups items which vary little in their characteristics.

The more similar they are in personal characteristics or attributes, the more these characteristics or attributes are controlled in the experiment.

Example: java developers only, non-international teams

Blocking Variables

A **blocking variable** is a variable the researcher controls before the experiment starts by dividing (or "blocking") the items into subgroups (or categories) and analyzing the impact of each subgroup on the outcome.

In this procedure, the researcher forms homogeneous by choosing a characteristic common to all participants in the study. Then the investigator randomly assigns individuals to the control and experimental groups using each category of the variable.

Manipulation of the treatment conditions

In experiments, you need to focus on the independent variables. These variables influence or affect the dependent variables in a quantitative study.
Treatment variables are independent variables that the researcher manipulates to determine their effect on the outcome, or dependent variable.

In experiments, treatment variables need to have two or more categories, or levels. The experimental researcher manipulates one or more of the treatment variable conditions.

In our example the treatments are different levels of frequency or duration of inspections

Threats to validity

Experiments might lead to wrong conclusion, meaning that the results would not be "valid"
Threads to validity can be divided into:

- internal validity
- external validity
- construct validity
- etc...

which other types of validity do you know?

Internal validity

Internal validity - a measure, that ensures the results and trends seen in an experiment are actually caused by the manipulation (treatment) and not some other factors underlying the process

Internal validity - Examples

In our experiment we can obtain invalid results (let's say, long inspections gives much better results) because teams initially had different experience level (also it can be significantly changed during experiment), or part of team members take annual vacation during the research, or control group followed TTD approach in development.

Another possible cause is the communication across the teams, which can also results in uncontrolled bias.

Which other threads are true for our experiment?

Summary of the threats to internal validity

History

Maturation

Regression

Selection

Mortality

Interactions with selection

Diffusion of treatments

Compensatory equalization

Compensatory rivalry

Resentful demoralization

Testing

Instrumentation

Threats to internal validity (1/5)

- **History**: the specific events occurring between the first and second measurement in addition to the experimental variable
- **Maturation**: processes within the respondents operating as a function of the passage of time per se (not specific to the particular events)
- **Regression**: operating where groups have been selected on the basis of their extreme scores

Threats to internal validity (2/5)

- **Selection**: Biases resulting in differential selection of respondents for the comparison groups
- **Interactions with selection**: Several of the threats mentioned above can interact (or relate) with the selection of participants to add additional threats to an experiment
- **Mortality**: differential loss of respondents from the comparison groups

Threats to internal validity (3/5)

- **Diffusion of treatments**: when the experimental and control groups can communicate with each other, the control group may learn from the experimental group information about the treatment

- **Compensatory equalization**: when only the experimental group receives a treatment, an inequality exists that may threaten the validity of the study. The benefits of the experimental treatment need to be equally distributed among the groups in the study

Threats to internal validity (4/5)

- **Compensatory rivalry**: if you publicly announce assignments to the control and experimental groups, compensatory rivalry may develop between the groups because the control group feels that it is the "underdog"

- **Resentful demoralization**: when a control group is used, individuals in this group may become resentful and demoralized because they perceive that they receive a less desirable treatment than other groups

Threats to internal validity (5/5)

- **Testing**: the effects of taking a test upon the scores of a second testing

- **Instrumentation**: changes in the calibration of a measuring instrument or changes in the observers or scorers used may produce changes in the obtained measurements

External validity

External validity - a measure that shows to which extend the validity of the cause-and-effect relationship is being generalizable to other persons, settings, treatment variables, and measures.

External validity - Examples

The same idea of daily inspections, accepted in company XXX with scientific evidences, led to no effect in company YYY, because there were KPIs and fees based on the number of defects produced by each particular developer.

In our experiment, the researcher needs to be cautious about generalizing results to other companies, other development teams, etc. Which other external threads can arose?

Threats to external validity (1/2)

- **Reactive effect of testing**: pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable
- **Interaction of selection and treatment**: This threat to external validity involves the inability to generalize beyond the groups in the experiment
- **Reactive effects of experimental arrangements**: This threat to external validity arises from the inability to generalize from the setting where the experiment occurred to another setting

Threats to external validity (2/2)

- **Multiple-treatment interference**: likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable
- **Interaction of history and treatment**: This threat to external validity develops when the researcher tries to generalize findings to past and future situations

Construct validity (1/2)

- **Construct validity** refers informally to whether I apply the "right" analysis, that if, for instance, if I apply the mean on data on an ordinal scale.
- Formally, it defines whether the operational definition of a variable actually reflect the true theoretical meaning of a concept. In other words, whether a scale or test measures the construct adequately.

Construct validity (2/2)

It can be divided into:

- *Convergent validity* tests that constructs that are expected to be related are, in fact, related.
- *Divergent validity* tests that constructs that should have no relationship do, in fact, not have any relationship.

Construct validity - Examples

In our experiment, if we select defects found per kLOC to assess the effect of inspections on the software quality. If we rely on Mean time between failures metric, will it be more informative for us in terms of research objectives?

It is not that easy to measure construct validity – several measures are usually required to demonstrate it, including pilot studies and clinical trials.

One approach to assure construct validity is **GQM model**

Threats to validity (recap)

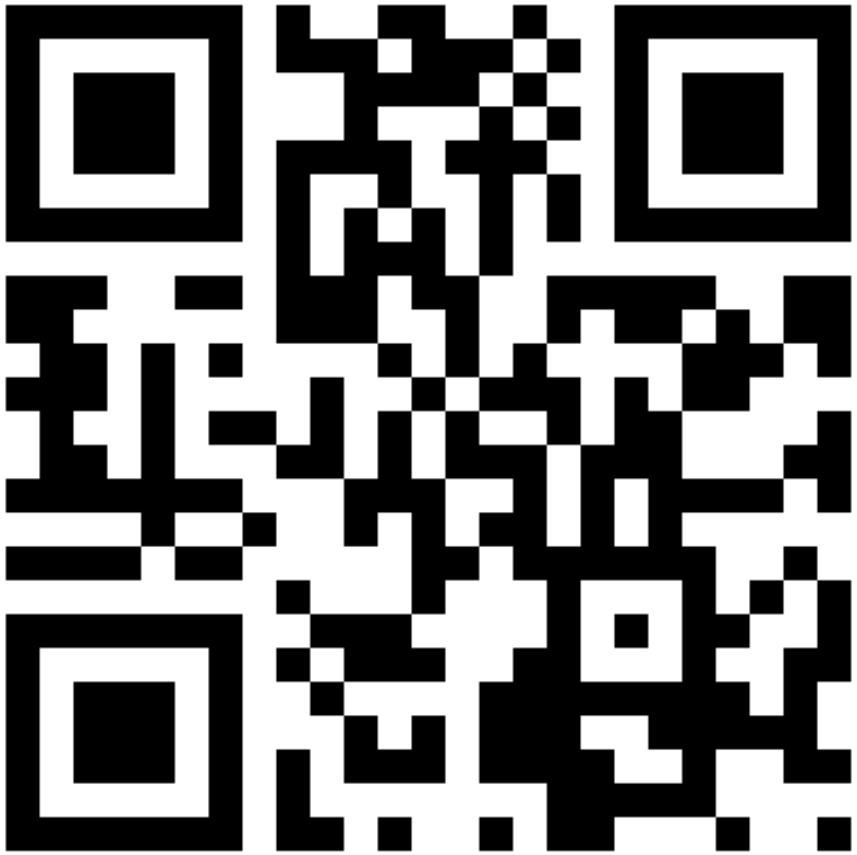
Internal validity	the validity of the cause and effect relationship between the independent and dependent variables
External validity	the validity of the cause and effect relationship being generalizable to other settings, treatment variables, and measures
Construct validity	the validity of inferences about the constructs (or variables) in the study

Proposed exercise (1/2)

- Create a mix team of 3 people of which at least one DS and one SE
- Identify an experimentation that happened in your life from anyone in the team
- For such experimentation determine:
  - Treatment
  - Subjects
  - Population
  - Sample
  - Observation
  - Extraneous Factors
  - Internal Validity
  - External Validity
  - Construct Validity
- Write the results on `http://tiny.cc/IU_EM_F20_L5`, tab: `IntroductionToED`

Proposed exercise (2/2)

`http://tiny.cc/IU_EM_F20_L5`, tab: `IntroductionToED`



End of lecture 5