Experiment Scoping and Planning

Idaho State University | Computer Science

Isaac Griffith

CS 6620
Department of Informatics and Computer Science
Idaho State University

ROAR

Scoping

- Experimentation is a labor-intensive task
  - Thus, we must ensure the intent of the experiment is fulfilled by the experiment execution

- Without a proper foundation, rework will be required
  - Worse, the experiment may not be able to study what was intended to be studied

- Scoping is meant to define the goals of the experiment
  - For this we use the GQM

ROAR
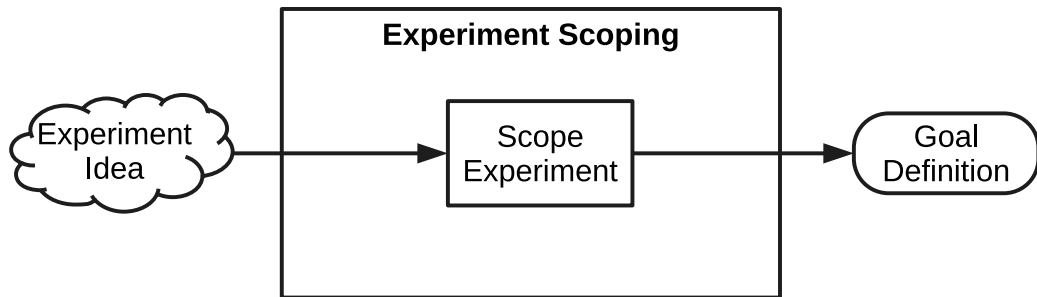
Scoping

Scoping

- A goal template ensures that important aspects of an experiment are defined before the planning and execution takes places.

- **GQM Goal Template**:
  Analyze **<Objects of study>**
  for the purpose of **<Purpose>**
  with respect to their **<Quality focus>**
  in the context of **<Context>**

- **Example**: Analyze the PBR and checklist techniques for the purpose of evaluation with respect to effectiveness and efficiency from the point of view of the research int he context of M.Sc. and Ph.D. students reading requirements documents.

Goal Definition Framework

Object of study	Purpose	Quality Focus	Perspective
Product	Characterize	Effectiveness	Developer
Process	Monitor	Cost	Modifier
Model	Evaluate	Reliability	Maintainer
Metric	Predict	Maintainability	Product Manager
Theory	Control	Portability	Corporate Manager
	Change		Customer
			User
			Researcher

Experimental Context

- Classified in terms of the subjects and objects involved in the study
- **Subjects** (personnel involved) can be characterized by
  - experience
  - team size
  - workload, etc.
- **Objects** (software artifacts) can be characterized by
  - size
  - complexity
  - priority
  - application domain, etc.

Experimental Context

- One Subject
  - One Object
    - Single object study
  - More than One Object
    - Multi-object variation study
- More than One Subject
  - One Object
    - Multi-test within object study
  - More than One Object
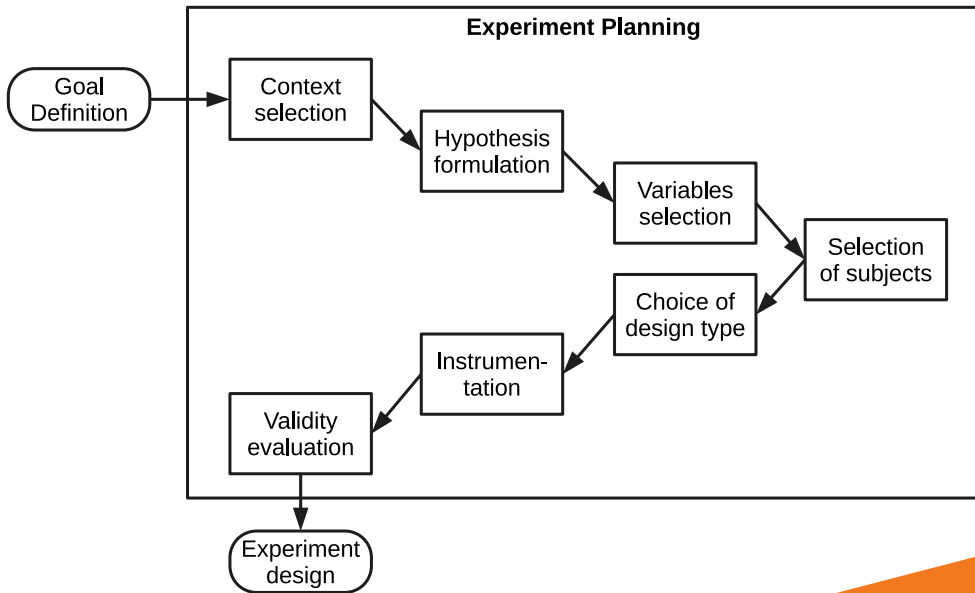   - Blocked subject-object study

ROAR

Planning

- Scoping provides the **why** of the experiment
- Planning provides the **how** of the experiment
- The experiment must be planned and the plans followed in order to control the experiment
- If not planned properly
  – The experiment can be disturbed
  – The experiment can be destroyed

ROAR

Planning

Context Selection

- For the most general results an experiments require projects that are:
  - large
  - real life
  - professionally staffed
- But this involves risk
  - An alternative is to run offline projects in parallel with real projects (reduces risk but increases cost)
  - A cheaper alternative is to run projects staffed by students but this seldom addresses real problems

Context Characterization

- Contexts can be characterized according to four dimensions
  - Off-line vs. on-line
  - Student vs. professional
  - Toy vs. real problems
  - Specific vs. general

- Often we must strike a balance between a specific context and generalizability

Hypothesis Formulation

- **Hypothesis Formulation** is the basis for statistical analysis of an experiment
  - must be stated formally
  - data is then collected in order to, if possible, reject the hypothesis
- If a hypothesis can be rejected, then conclusions may be drawn

Hypothesis Types

- **Null** H_0, stats that there are no real underlying trends or patterns in the experiment setting
  - The only real reason for observed differences is coincidence
  - **EX:** $H_0: \mu_{N_{old}} = \mu_{N_{new}}$

- **Alternative** H_1 or H_A, the hypothesis in favor of which the null hypothesis is rejected.
  - **EX:** $H_A: \mu_{N_{old}} < \mu_{N_{new}}$

- Statistical tests that may be used to analyzes the results of an experiment based on these hypotheses

ROAR

Error

- Hypothesis testing involves risk
- **Type I Error:** occurs when a statistical test indicates a pattern or relationship even if there is no such pattern.
  - A false positive
  - P(type-I-error) = P(reject H_0 | H_0 true)
- **Type II Error:** occurs when a statistical test has not indicated a pattern or relationship even if there actually is a real pattern.
  - A false negative
  - P(type-II-error) = P(fail to reject H_0 | H_0 false)
- Error size depends on different factors

Power

- Power is the ability of a statistical test to reveal a true pattern in the collected data
  - a key factor in calculating error size

- **Power**: the probability that the test will reveal the true pattern if H_0 is false.
  - an experimenter should choose a test with as high a power as possible
  - Power = P(reject H_0 | H_0 is false) = 1 - P(type-II-error)

Variable Selection

- **Independent Variables**: those variables we can control and change
  - Selection requires domain knowledge
  - Should have some effect on dependent variables
  - Choice should include
    - measurement scale
    - value range
    - specific levels at which tests will be made
- **Dependent Variables**: measurement of the effect of treatments
  - derived directly from the hypotheses
  - may require surrogate measures
  - choice should include
    - measurement scale
    - value range
- Selecting a single dependent variable reduces threats to validity

ROAR

Subject Selection

- Closely related to the generalization of the results

- Also called sampling the population
  - probabilistic methods
    - probability of selecting each subject is known
  - non-probabilistic methods
    - probability of selecting each subject is unknown

ROAR

Subject Selection

Probabilistic

- **Simple random sampling**: Subjects are selected from a list of the population, at random.
- **Systematic sampling**: The first subject is selected from the list of the population at random and then every nth person is selected.
- **Stratified random sampling**: The population is divided into a number of groups or strata with a known distribution between the groups. Random sampling is then applied within the strata.

Non-probabilistic

- **Convenience sampling**: Selects the nearest and most convenient persons
- **Quota sampling**: This type of sampling is sued to get subjects from various elements of a population. Convenience sampling within each element.

ROAR

Sample Size

- Sample size impacts generalizability of results
- Larger sample sizes implies lower generalization error
- Sample size also relates to statistical power
- **General principles**
  - Large population variability implies larger required sample
  - Data analysis method influences required sample size

ROAR

Design Choice

- An experiment consists of a series of tests of the treatments
- A design describes how the tests are organized and run
- A design is based on the hypotheses and tests that are required
- During design we must determine the number of tests required (sample size analysis)
  - ensures the treatment effect will be visible
- Design is also necessary to ensure replicability

ROAR

Design Principles

- **Randomization**: ensures that observations are from independent random variables
  - Applicable to:
    - selection of subjects from population
    - allocation of objects and subjects
    - order in which tests are performed
  - Averages out the effect of a factor that may otherwise be present

Design Principles

- **Blocking**: blocks are factors which may have an effect on the response but one which we are not concerned with.
  - Allows for the control of this factor, while eliminating (systematically) the undesired effect in the comparison
  - Within a block the effect is the same, thus we can observe the treatment effect
  - Block effects are eliminated by not studying effects between blocks

- **Balancing**: A balanced design has the same number of subjects per treatment
  - desirable as it simplifies analysis

ROAR

Are there any questions?

ROAR