CSC2130:
Empirical Research Methods for Computer Scientists

Seminar 3: Basics of Empirical Research

Steve Easterbrook sme@cs.toronto.edu

Barbara Neves barbara@bbneves.com

www.cs.toronto.edu/~sme/CSC2130/

Topics for Today

- Recap from last week
- Discussion: Is most published research wrong?
- Study Design Planning Checklist
- Elements of research design
  - Units of Analysis
  - Sampling
  - Measurement
  - Validity
  - Reliability

Planning Checklist

- ✓ Pick a topic
- ❍ Identify the research question(s)
- ❍ Check the literature
- ❍ Identify your philosophical stance
- ❍ Identify appropriate theories
- ❍ Choose the method(s)
- ❍ Design the study
  - Unit of analysis?
  - Target population?
  - Sampling technique?
  - Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- ❍ Critically appraise the design for threats to validity
- ❍ Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- ❍ Recruit subjects / field sites
- ❍ Conduct the study
- ❍ Analyze the data
- ❍ Write up the results and publish them
- ❍ Iterate

What type of question are you asking?

- ❍ Existence:
  - Does X exist?
- ❍ Description & Classification
  - What is X like?
  - What are its properties?
  - How can it be categorized?
  - How can we measure it?
  - What are its components?
- ❍ Descriptive-Comparative
  - How does X differ from Y?
- ❍ Frequency and Distribution
  - How often does X occur?
  - What is an average amount of X?
- ❍ Descriptive-Process
  - How does X normally work?
  - By what process does X happen?
  - What are the steps as X evolves?
- ❍ Relationship
  - Are X and Y related?
  - Do occurrences of X correlate with occurrences of Y?
- ❍ Causality
  - Does X cause Y?
  - Does X prevent Y?
  - What causes X?
  - What effect does X have on Y?
- ❍ Causality-Comparative
  - Does X cause more Y than does Z?
  - Is X better at preventing Y than is Z?
  - Does X cause more Y than does Z under one condition but not others?
- ❍ Design
  - What is an effective way to achieve X?
  - How can we improve X?

What type of question are you asking?

Exploratory

- Existence:
  - Does X exist?
- Description & Classification
  - What is X like?
  - What are its properties?
  - How can it be categorized?
  - How can we measure it?
  - What are its components?
- Descriptive-Comparative
  - How does X differ from Y?

Baserate

- Frequency and Distribution
  - How often does X occur?
  - What is an average amount of X?
- Descriptive-Process
  - How does X normally work?
  - By what process does X happen?
  - What are the steps as X evolves?

Correlation

- Relationship
  - Are X and Y related?
  - Do occurrences of X correlate with occurrences of Y?

Causal Relationship

- Causality
  - Does X cause Y?
  - Does X prevent Y?
  - What causes X?
  - What effect does X have on Y?
- Causality-Comparative
  - Does X cause more Y than does Z?
  - Is X better at preventing Y than is Z?
  - Does X cause more Y than does Z under one condition but not others?

Design

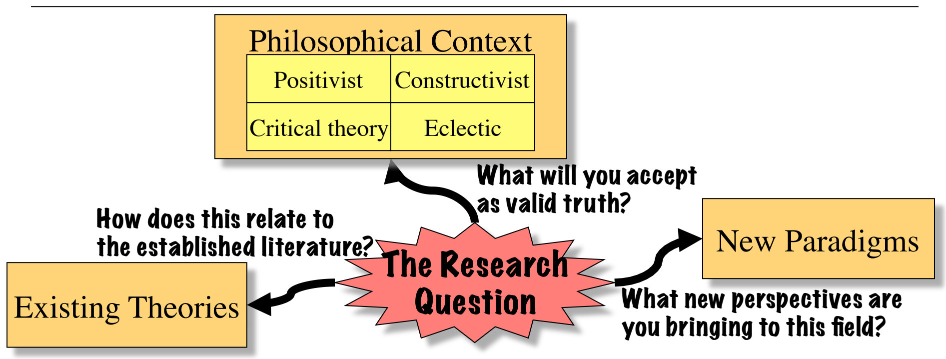
- Design
  - What is an effective way to achieve X?
  - How can we improve X?

Putting the Question in Context

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ❍ Identify your philosophical stance
- ❍ Identify appropriate theories
- ❍ Choose the method(s)
- ❍ Design the study
  - Unit of analysis?
  - Target population?
  - Sampling technique?
  - Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- ❍ Critically appraise the design for threats to validity
- ❍ Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- ❍ Recruit subjects / field sites
- ❍ Conduct the study
- ❍ Analyze the data
- ❍ Write up the results and publish them
- ❍ Iterate

Putting the Question in Context



What will you accept as knowledge?

- Positivist (or "Post-positivist")
  - Knowledge is objective
  - "Causes determine effects/ outcomes"
  - Reductionist: study complex things by breaking down to simpler ones
  - Prefer quantitative approaches
  - Verifying (or Falsifying) theories

- Constructivist/Interpretivist
  - Knowledge is socially constructed
  - Truth is relative to context
  - Theoretical terms are open to interpretation
  - Prefer qualitative approaches
  - Generating "local" theories

- Critical Theorist
  - Research is a political act
  - Knowledge is created to empower groups/individuals
  - Choose what to research based on who it will help
  - Prefer participatory approaches
  - Seeking change in society

- Eclectic/Pragmatist
  - Research is problem-centered
  - "All forms of inquiry are biased"
  - Truth is what works at the time
  - Prefer multiple methods / multiple perspectives
  - seeking practical solutions to problems

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- Identify appropriate theories
- Choose the method(s)
- Design the study
  - Unit of analysis?
  - Target population?
  - Sampling technique?
  - Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- Critically appraise the design for threats to validity
- Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- Recruit subjects / field sites
- Conduct the study
- Analyze the data
- Write up the results and publish them
- Iterate

Identify Appropriate Theories

- Where do theories come from?

The Theoretical Lens

- Our Theories impact how we see the world
  - Real-world phenomena too rich and complex
  - Need a way of filtering our observations
  - The theory guides us, whether it is explicitly stated or not
- In Quantitative Methods:
  - Theoretical lens tells you what variables to measure…
  - …and which to ignore or control
- In Qualitative Methods:
  - Theoretical lens usually applied after data is collected
  - …and used to help with labeling and categorizing the data

Theories are good for generalization…

Statistical Generalization

- Generalize from sample to population
- Can only be used for quantifiable variables
- Based on random sampling:
  - Test whether results on a sample apply to the whole population
- Not useful when:
  - You can' t characterize the population
  - You can' t do random sampling
  - You can' t get enough data points

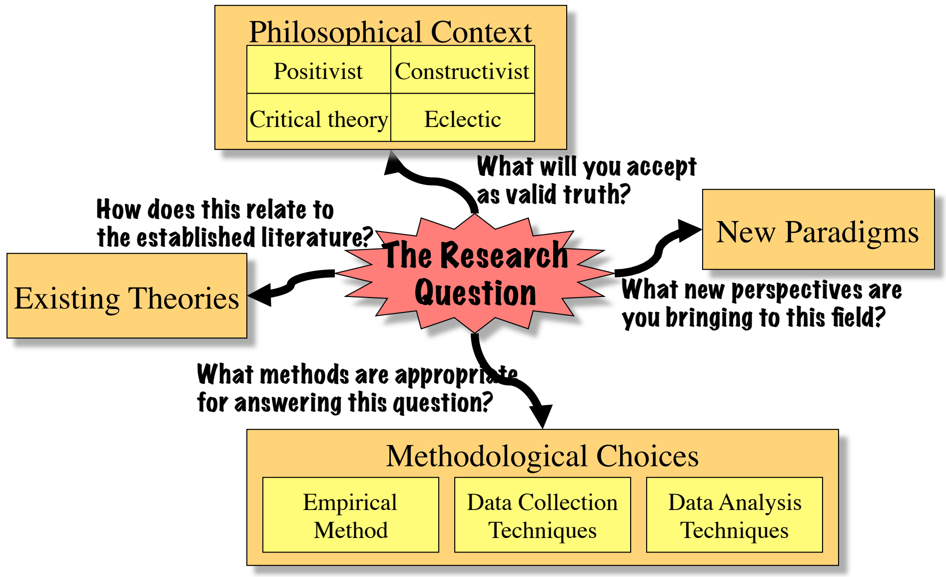
Analytical Generalization

- Generalize from findings to theory
- Applicable to quantitative and qualitative studies
- Compares findings with theory
  - Do the data support/refute the theory?
  - Do they support this theory better than rival theories?
- Supports empirical induction:
  - Evidence builds if subsequent studies also support the theory
- More powerful than stats
  - Doesn' t rely on correlations
  - Examines underlying mechanisms

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- ✓ Identify appropriate theories
- Choose the method(s)
- Design the study
  - Unit of analysis?
  - Target population?
  - Sampling technique?
  - Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- Critically appraise the design for threats to validity
- Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- Recruit subjects / field sites
- Conduct the study
- Analyze the data
- Write up the results and publish them
- Iterate

Putting the Question in Context



 CSC2130: Empirical Research Methods for Computer Scientists 24

Choose a Method…

- Exploratory
  - Used to build new theories where we don't have any yet
    - E.g. What do CMM level 3 organizations have in common?
    - E.g. What are the experiences of developers who have adopted Ruby?
- Causal
  - Determines whether there are causal relationship between phenomena
    - E.g. Does tool X lead to software with fewer defects?
    - E.g. Do requirements traceability tools help programmers find information more rapidly?
- Descriptive
  - Describes sequence of events and underlying mechanisms
    - E.g. How does pair programming actually work?
    - E.g. How do software immigrants naturalize?
- Explanatory
  - Adjudicates between competing explanations (theories)
    - E.g. Why does software inspection work?
    - E.g. Why do people fail to document their requirements?

 CSC2130: Empirical Research Methods for Computer Scientists 28

How will you substantiate your claims?

Common "in the lab" Methods	Common "in the wild" Methods
● Controlled Experiments	● Quasi-Experiments
❍ Rational Reconstructions	● Case Studies
❍ Exemplars	● Survey Research
❍ Benchmarks	● Ethnographies
❍ Simulations	● Action Research

❍ Artifact/Archive Analysis ("mining"!)

Planning Checklist

- ✔ Pick a topic
- ✔ Identify the research question(s)
- ✔ Check the literature
- ✔ Identify your philosophical stance
- ✔ Identify appropriate theories
- ✔ Choose the method(s)
- ❍ Design the study
  - Unit of analysis?
  - Target population?
  - Sampling technique?
  - Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- ❍ Critically appraise the design for threats to validity
- ❍ Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- ❍ Recruit subjects / field sites
- ❍ Conduct the study
- ❍ Analyze the data
- ❍ Write up the results and publish them
- ❍ Iterate

Unit of Analysis

- Defines what phenomena you will analyze
  - Choice depends on the primary research questions
  - Choice affects decisions on data collection and analysis
  - Hard to change once the study has started (but can be done if there are compelling reasons)
  - If possible, use same unit of analysis as previous studies (why?)
- Often many choices:
  - E.g. for an exploratory study of agile programming:
    - Unit of analysis = individual developer (study focuses on a person's participation in the project)
    - Unit of analysis = a team (study focuses on team activities)
    - Unit of analysis = a decision (study focuses on activities around that decision)
    - Unit of analysis = a process (study examines how user stories are collected and prioritized)
    - …

Examples of Units of Analysis

- For a study of how software immigrants naturalize
  - Individuals?
  - … or the Development team?
  - … or the Organization?
- For a study of pair programming
  - Programming episodes?
  - … or Pairs of programmers?
  - … or the Development team?
  - … or the Organization?
- For a study of software evolution
  - A Modification report?
  - … or a File?
  - … or a System?
  - … or a Release?
  - … or a Stable release?

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- ✓ Identify appropriate theories
- ✓ Choose the method(s)
- Design the study
  - ✓ Unit of analysis?
  - Target population?
  - Sampling technique?
  - Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- Critically appraise the design for threats to validity
- Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- Recruit subjects / field sites
- Conduct the study
- Analyze the data
- Write up the results and publish them
- Iterate

Target Population

- Determines scope of applicability of your results
  - If you don't define the target population…
  - …nobody will know whether your results apply to anything at all
- From what population are your units of analysis drawn?
  - UoA = "developer using agile programming"
  - Population =
    - All software developers in the world?
    - All developers who use agile methods?
    - All developers in Canadian Software Industry?
    - All developers in Small Companies in Ontario?
    - All students taking SE courses at U of T?
- Choice closely tied to choice of sampling method…

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- ✓ Identify appropriate theories
- ✓ Choose the method(s)
- ❍ Design the study
  - ✓ Unit of analysis?
  - ✓ Target population?
  - Sampling technique?
  - Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- ❍ Critically appraise the design for threats to validity
- ❍ Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- ❍ Recruit subjects / field sites
- ❍ Conduct the study
- ❍ Analyze the data
- ❍ Write up the results and publish them
- ❍ Iterate

Sampling Method

- ❍ Used to select representative set from a population
  - Simple Random Sampling - choose every kth element
  - Stratified Random Sampling - identify strata and sample each
  - Clustered Random Sampling - choose a representative subpopulation and sample it
  - Purposive Sampling - choose the parts you think are relevant without worrying about statistical issues (see next slide…)
- ❍ Sample Size is important
  - balance between cost of data collection/analysis and required significance
- ❍ Process:
  - Decide what data should be collected
  - Determine the population
  - Choose type of sample
  - Choose sample size

Purposive Sampling

- Typical Case
  - Identify typical, normal, average case
- Extreme or Deviant Case
  - E.g outstanding success/notable failures, exotic events, crises.
- Critical Case
  - if it's true of this one case it's likely to be true of all other cases.
- Intensity
  - Information-rich examples that clearly show the phenomenon (but not extreme)
- Maximum Variation
  - choose a wide range of variation on dimensions of interest
- Homogeneous
  - Instance has little internal variability - simplifies analysis
- Snowball or Chain
  - Select cases that should lead to identification of further good cases
- Criterion
  - All cases that meet some criterion
- Confirming or Disconfirming
  - Exceptions, variations on initial cases
- Opportunistic
  - Rare opportunity where access is normally hard/impossible
- Politically Important Cases
  - Attracts attention to the study
- Convenience sampling
  - Cases that are easy/cheap to study
  - (May reduce credibility)
- …Or any combination of the above

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- ✓ Identify appropriate theories
- ✓ Choose the method(s)
- Design the study
  - ✓ Unit of analysis?
  - ✓ Target population?
  - ✓ Sampling technique?
  - Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- Critically appraise the design for threats to validity
- Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- Recruit subjects / field sites
- Conduct the study
- Analyze the data
- Write up the results and publish them
- Iterate

Data Collection Techniques

- Direct Techniques
  - Brainstorming / Focus Groups
  - Interviews
  - Questionnaires
  - Conceptual Modeling
  - Work Diaries
  - Think-aloud Sessions
  - Shadowing and Observation
  - Participant Observation
- Indirect Techniques
  - Instrumented Systems
  - Fly on the wall
- Independent Techniques
  - Analysis of work databases
  - Analysis of tool usage logs
  - Documentation Analysis
  - Static and Dynamic Analysis

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- ✓ Identify appropriate theories
- ✓ Choose the method(s)
- Design the study
  - ✓ Unit of analysis?
  - ✓ Target population?
  - ✓ Sampling technique?
  - ✓ Data collection techniques?
  - Metrics for key variables?
  - Handle confounding factors
- Critically appraise the design for threats to validity
- Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- Recruit subjects / field sites
- Conduct the study
- Analyze the data
- Write up the results and publish them
- Iterate

How will you measure things?

Type	Meaning	Admissible Operations
Nominal Scale	Unordered classification of objects	=
Ordinal Scale	Ranking of objects into ordered categories	=, <, >
Interval Scale	Differences between points on the scale are meaningful	=, <, >, difference, mean
Ratio Scale	Ratios between points on the scale are meaningful	=, <, >, difference, mean, ratio
Absolute Scale	No units necessary - scale cannot be transformed	=, <, >, difference, mean, ratio

 CSC2130: Empirical Research Methods for Computer Scientists 36

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- ✓ Identify appropriate theories
- ✓ Choose the method(s)
- ❍ Design the study
  - ✓ Unit of analysis?
  - ✓ Target population?
  - ✓ Sampling technique?
  - ✓ Data collection techniques?
  - ✓ Metrics for key variables?
  - Handle confounding factors
- ❍ Critically appraise the design for threats to validity
- ❍ Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- ❍ Recruit subjects / field sites
- ❍ Conduct the study
- ❍ Analyze the data
- ❍ Write up the results and publish them
- ❍ Iterate

 CSC2130: Empirical Research Methods for Computer Scientists 21

What could go wrong?

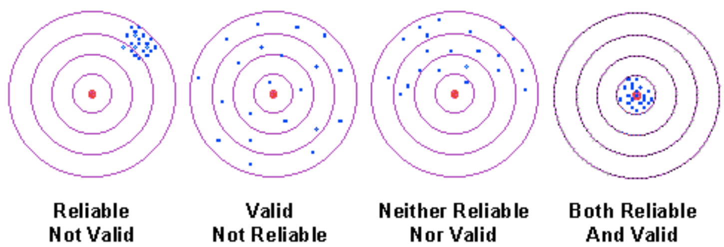
- Many phenomena might affect your results
- Must be able to distinguish:
  - My results follow clearly from the phenomena I observed
  - My results were caused by phenomena that I failed to observe
- Identify all (likely) confounding variables
- For each, decide what to do:
  - Selection/Exclusion
  - Balancing
  - Manipulation
  - Ignore (with justification)

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- ✓ Identify appropriate theories
- ✓ Choose the method(s)
- Design the study
  - ✓ Unit of analysis?
  - ✓ Target population?
  - ✓ Sampling technique?
  - ✓ Data collection techniques?
  - ✓ Metrics for key variables?
  - ✓ Handle confounding factors
- Critically appraise the design for threats to validity
- Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- Recruit subjects / field sites
- Conduct the study
- Analyze the data
- Write up the results and publish them
- Iterate

Validity vs. Reliability

- Reliability: Does the study get consistent results?
- Validity: Does the study get true results?



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?

Typical Problems

- Construct Validity
  - Using things that are easy to measure instead of the intended concept
  - Wrong scale; insufficient discriminatory power
- Internal Validity
  - Confounding variables: Familiarity and learning;
  - Unmeasured variables: time to complete task, quality of result, etc.
- External Validity
  - Task representativeness: toy problem?
  - Subject representativeness: students for professional developers!
- Theoretical Reliability
  - Researcher bias: subjects know what outcome you prefer

Construct Validity

- E.g. Hypothesis: "Inspection meetings are unnecessary"
  - Inspection -> Perspective-based reading of requirements docs
  - Meeting -> Inspectors gather together and report their findings
  - Unnecessary -> find fewer total # errors than inspectors working alone
- But:
  - What's the theory here?
  - E.g. Fagin Inspections:
    - Purpose of inspection is process improvement (not bug fixing!)
    - Many intangible benefits: staff training, morale, knowledge transfer, standard setting,…

Construct Validity

- Are we measuring what we intend to measure?
  - Akin to the requirements problem: are we building the right system?
  - If we don't get this right, the rest doesn't matter
  - Helps if concepts in the theory have been precisely defined!
- Divide construct validity into three parts:
  - Intentional Validity - are we measuring precisely what we intend?
    - E.g. measuring "expertise" as "years of experience"?
  - Representation Validity - do our measurements accurately operationalize the constructs?
    - E.g. is it okay to break "intelligence" down into verbal, spatial & numeric reasoning?
    - Face validity argument – "seems okay on the face of it"
    - Content validity argument – "measures demonstrated to cover the concept"
  - Observation Validity - how good are the measures by themselves?
    - E.g. the short form of a test correlates well with longer form

More on Observation Validity

- Predictive Validity
  - Observed measure predicts what it should predict and nothing else
    - E.g. check that college aptitude tests do predict success in college
- Criterion Validity
  - Observed measure agrees with an independent standard
    - Eg, for college aptitude, GPA or successful first year
- Convergent Validity
  - Observed measure correlates with other observable measures for the same construct
    - I.e. our measure gives a new way of distinguishing a particular trait while correlating with similar measures
- Discriminant Validity
  - Observed measure distinguishes between two groups that differ on the trait in question
    - E.g. Measurement of code quality can distinguish "good" code from "bad"

Internal Validity

- Can we be sure our results really follow from the data?
  - Have we adequately ruled out rival hypotheses?
- Have we eliminated confounding variables?
  - Participant variables
  - Researcher variables
  - Stimulus, procedural and situational variables
  - Instrumentation
  - Nuisance variables
- Confounding sources of internal invalidity
  - H: History
    - events happen during the study (eg, study session was interrupted)
  - M: Maturation
    - older/wiser/better between treatments (or during study)
  - I: Instrumentation
    - change due to observation/measurement instruments
  - S: Selection
    - differing nature of participants
    - effects of choosing participants

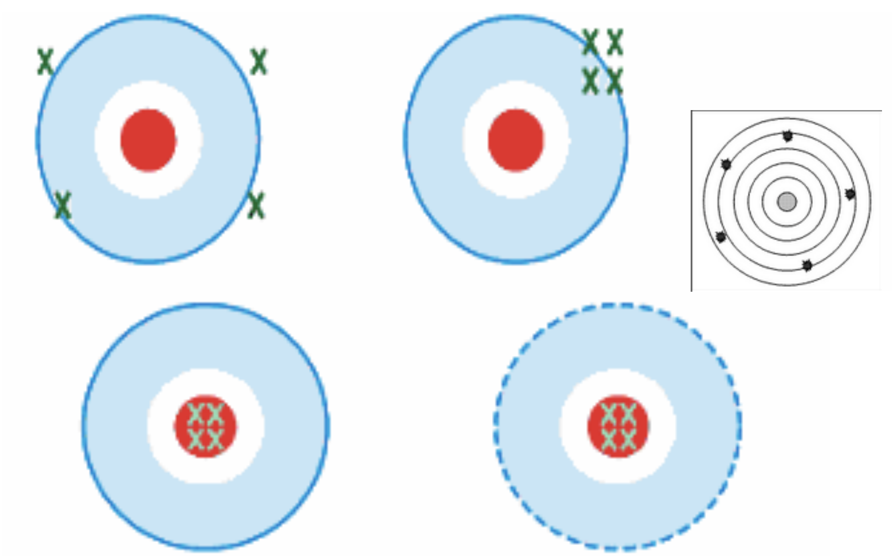
External Validity

- Two issues:
  - Results will generalize beyond the specific situations studied
    - E.g. do results on students generalize to professionals?
  - Do the results *support the claims* of generalizability?
    - E.g. if the effect size is small, will it be swamped/masked in other settings?
    - E.g. will other (unstudied) phenomena dominate?
- Two strategies:
  - Provide arguments in favour of generalizability
  - Replicate the finding in further studies:
    - Literal replication - repeat study using the same design
    - Empirical Induction - related studies test additional aspects of the theory
- Also: Ecological Validity
  - Does the study set-up approximate real-world conditions?
  - (can achieve external validity without this, but it's hard)

Reliability

- Could the study be repeated with the same results?
  - On the same subjects (not a replication!)
- Issues:
  - No mistakes were made in conducting the experiment
  - Steps taken in data collection and analysis were made explicit
  - No biases were introduced by the researchers
- Good practice:
  - Carefully document all procedures used in the study
  - Prepare a “lab package” of all materials and procedures used
  - Conduct the study in such a way that an auditor could follow the documented procedures and arrive at the same results

Validity vs. Reliability



Validity (Constructivist View)

- Repeatability is suspect:
  - Reality is "multiple and constructed", same situation can never recur
  - Researcher objectivity is unattainable
  - E.g. successful replication depends on tacit knowledge
- Focus instead on "trustworthiness":
  - Credibility of researchers and results
  - Transferability of findings
  - Dependability - results are robust across a range of situations
  - Confirmability
- Identify strategies to increase trustworthiness…

Strategies for constructivists

- Triangulation
  - Different sources of data used to confirm findings
- Member checking
  - Research participants confirm that results make sense from their perspective
- Rich, thick descriptions
  - As much detail as possible on the setting and the data collected
- Clarify bias
  - Be honest about researcher's bias
  - Self-reflection when reporting findings
- Report discrepant information
  - Include data that contradicts findings as well as that which confirms
- Prolonged contact with participants
  - Spend long enough to ensure researcher really understands the situation being studied
- Peer debriefing
  - A colleague critically reviews the study and tests assumptions
- External Auditor
  - Independent expert reviews procedures and findings

Validity (Critical theorist's view)

- Validity depends on utility of the knowledge gained
  - Research is intended to challenge perspectives, shift power, etc.
  - Problems tackled are context sensitive…
  - …repeatability not an issue
- Criteria (e.g. for action research)
  - Problem tackled is authentic
  - Intended change is appropriate and adequate
  - Participants are authentic (real problem owners)
  - Researcher has appropriate level of access to the organization
  - Planned exit point
  - Clear knowledge outcomes for participants

Planning Checklist

- ✓ Pick a topic
- ✓ Identify the research question(s)
- ✓ Check the literature
- ✓ Identify your philosophical stance
- ✓ Identify appropriate theories
- ✓ Choose the method(s)
- Design the study
  - ✓ Unit of analysis?
  - ✓ Target population?
  - ✓ Sampling technique?
  - ✓ Data collection techniques?
  - ✓ Metrics for key variables?
  - ✓ Handle confounding factors
- ✓ Critically appraise the design for threats to validity
- Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- Recruit subjects / field sites
- Conduct the study
- Analyze the data
- Write up the results and publish them
- Iterate

