Tutorial T1

Empirical Research Methods in Requirements Engineering

Steve Easterbrook, University of Toronto

http://www.cs.toronto.edu/~sme

Goals of this tutorial

- For researchers:
  - Key questions in selecting empirical methods
  - Basics of research design
  - Understand and avoid common mistakes in empirical studies
- For reviewers:
  - Guidance to judge quality and validity of reported empirical studies.
  - Criteria to assess whether research papers reporting empirical work are suitable for publication
- For practitioners:
  - Awareness of how to interpret the claims made by researchers about new requirements engineering methods and tools.
  - Insight into the roles practitioners can play in empirical studies in RE

Overview

- Session 1:
  Basics of Empirical Research
  - 9:00-9:30 What is Science?
  - 9:30-9:50 Planning a study
  - 9:50-10:10 *Exercise: Planning*
  - 10:10-10:30 Validity and Ethics
  - **10:30-11:00 Coffee break**
- Session 2:
  Quantitative Methods
  - 11:00-11:30 Experiments
  - 11:30-12:00 *Exercise: Design an Experiment in RE*
  - 12:00-12:30 Survey Research
  - **12:30-2:00 Lunch**
- Session 3:
  Qualitative Methods
  - 2:00-2:30 Case Studies
  - 2:30-3:00 *Exercise: Design a Case Study in RE*
  - 3:00-3:15 Ethnographies
  - 3:15-3:30 Action Research
  - **3:30-4:00 Tea break**
- Session 4:
  Strategies & Tactics
  - 4:00-4:15 Mixed Methods
  - 4:15-4:30 *Exercise: Design a Research Strategy*
  - 4:30-5:00 Data Collection / Analysis
  - 5:00-5:15 Publishing
  - 5:15-5:30 Summary/Discussion
  - **5:30 Finish**

www.re07.org

15th IEEE International
Requirements Engineering
Conference

October 15-19th, 2007
India Habitat Center, New Delhi

1. Basics of Empirical Research

9:00-9:30 What is Science?

9:30-9:50 Planning a study

9:50-10:10 *Exercise: Planning*

10:10-10:30 Validity and Ethics

10:30-11:00 Coffee break

Why do you want to do Empirical Research?

- A better understanding of how requirements engineers work?
- Identification of problems with the current state-of-the-art?
- A characterization of the properties of new tools/techniques?
- Evidence that approach A is better than approach B?

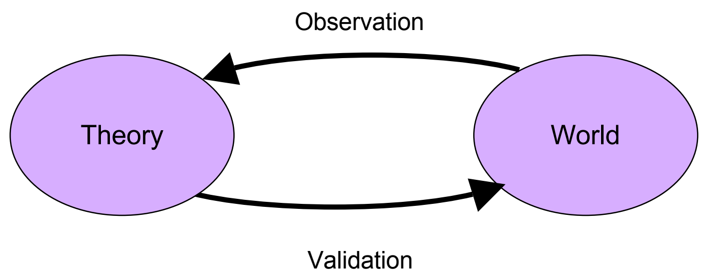
How will you substantiate your claims?

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

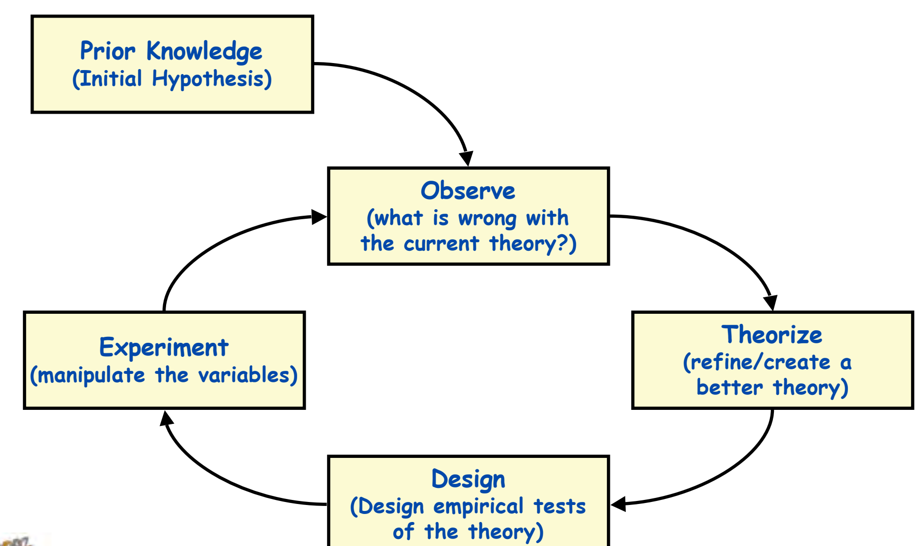
- Artifact/Archive Analysis ("mining"!)

Scientific Method

- No single "official" scientific method
- Somehow, scientists are supposed to do this:



Scientific Inquiry



But sometimes it looks more like this:

The Inductive Method:

1. formulate hypothesis
2. apply for grant
3. perform experiments or gather data to test hypothesis
4. alter data to fit hypothesis
5. publish

The Deductive Method:

1. formulate hypothesis
2. apply for grant
3. perform experiments or gather data to test hypothesis
4. revise hypothesis to fit data
5. backdate revised hypothesis
6. publish

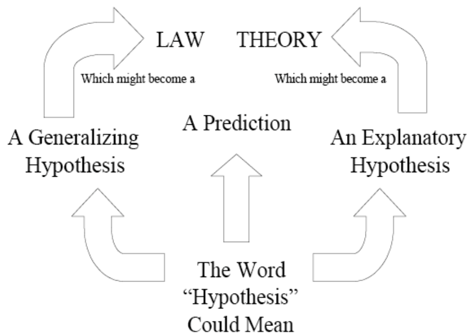
(From "Science Made Stupid", by Tom Well)

Some Characteristics of Science

- Science seeks to improve our understanding of the world.
- Explanations are based on observations
  - Scientific truths must stand up to empirical scrutiny
  - Sometimes "scientific truth" must be thrown out in the face of new findings
- Theory and observation affect one another:
  - Our perceptions of the world affect how we understand it
  - Our understanding of the world affects how we perceive it
- Creativity is as important as in art
  - Theories, hypotheses, experimental designs
  - Search for elegance, simplicity

Myths about Science (I)

- “It’s just a theory”
  - Theory = “best explanation for the available evidence”.
  - Overwhelming evidence doesn’t stop it being a theory…
  - …but lack of evidence does.



Examples:

We have a “law of gravity”
…but no “theory of gravity”

We have a “theory of evolution”
…but no “law of evolution”

Some Definitions

- A model is an abstract representation of a phenomenon or set of related phenomena
  - Some details included, others excluded
- A theory is a set of statements that explain a set of phenomena
  - Ideally, the theory has predictive power too
- A hypothesis is a testable statement derived from a theory
  - A hypothesis is not a theory!

→ In RE (and indeed SE), there are few “Theories”
  - *folk theories vs. scientific theories*

Myths about Science (II)

- "Scientists follow the scientific method"
  - There is no one method
  - Many methods available…
  - …and all of them have known flaws
  - Scientists use imagination, creativity, prior knowledge, perseverance…
- "Scientific knowledge is general and absolute"
  - Empirical Induction used to build evidence
  - Scientists often get it wrong…
  - …but Science (as a process) is self-correcting
  - All scientific laws and theories have limited scope
    - E.g. biological theories probably only apply on our own planet
    - E.g. laws of physics don't apply at the subatomic level

Meta-theories (theories about theory)

- Logical Positivism:
  - Separates discovery from validation
  - Logical deduction, to link theoretical concepts to observable phenomena
  - Scientific truth is absolute, cumulative, and unifiable
- Popper:
  - Theories can be refuted, not proved;
  - only falsifiable theories are scientific
- Campbell:
  - Theories are underdetermined;
  - All observation is theory-laden, biased
- Quine:
  - Terms used in scientific theories have contingent meanings
  - Cannot separate theoretical terms from empirical findings
- Kuhn:
  - Science characterized by dominant paradigms, punctuated by revolution
- Lakatos:
  - Not one paradigm, but many competing research programmes
  - Each has a hard core of assumptions immune to refutation
- Feyerabend:
  - Cannot separate scientific discovery from its historical context
  - All scientific methods are limited;
  - Any method offering new insight is ok
- Toulmin:
  - Evolving Weltanschauung determines what is counted as fact;
  - Scientific theories describe ideals, and explain deviations
- Laudan:
  - Negative evidence is not so significant in evaluating theories.
  - All theories have empirical difficulties
  - New theories seldom explain everything the previous theory did

All Methods are flawed

- E.g. Laboratory Experiments
  - Cannot study large scale software development in the lab!
  - Too many variables to control them all!
- E.g. Case Studies
  - How do we know what's true in one project generalizes to others?
  - Researcher chose what questions to ask, hence biased the study
- E.g. Surveys
  - Self-selection of respondents biases the study
  - Respondents tell you what they think they ought to do, not what they actually do
- …etc...

Strategies to overcome these weaknesses

- Theory-building
  - Testing a hypothesis is pointless (single flawed study!)…
  - …unless it builds evidence for a clearly stated theory
- Empirical Induction
  - Series of studies over time…
  - Each designed to probe more aspects of the theory
  - …together build evidence for a clearly stated theory
- Mixed Methods Research
  - Use multiple methods to investigate the same theory
  - Each method compensates for the flaws of the others
  - …together build evidence for a clearly stated theory

The Role of Theory Building

- Theories allow us to compare similar work
  - Theories include precise definition for the key terms
  - Theories provide a rationale for which phenomena to measure
- Theories support analytical generalization
  - Provide a deeper understanding of our empirical results
  - …and hence how they apply more generally
  - Much more powerful than statistical generalization
- …but in SE we are very bad at stating our theories
  - Our vague principles, guidelines, best practices, etc. could be strengthened into theories
  - Every tool we build represents a theory

Science and Theory

- A (scientific) theory is:
  - more than just a description - it explains and predicts
  - Logically complete, internally consistent, falsifiable
  - Simple and elegant.
- Components
  - Precisely defined terminology
  - Concepts, relationships, causal inferences
  - (operational definitions for theoretical terms)
- Theories lie at the heart of what it means to do science.
  - Production of generalizable knowledge
- Theory provides orientation for data collection
  - Cannot observe the world without a theoretical perspective

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

www.re07.org

15th IEEE International
Requirements Engineering
Conference

October 15-19th, 2007
India Habitat Center, New Delhi

1. Basics of Empirical Research

9:00-9:30 What is Science?

9:30-9:50 Planning a study

9:50-10:10 *Exercise: Planning*

10:10-10:30 Validity and Ethics

10:30-11:00 Coffee break

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?

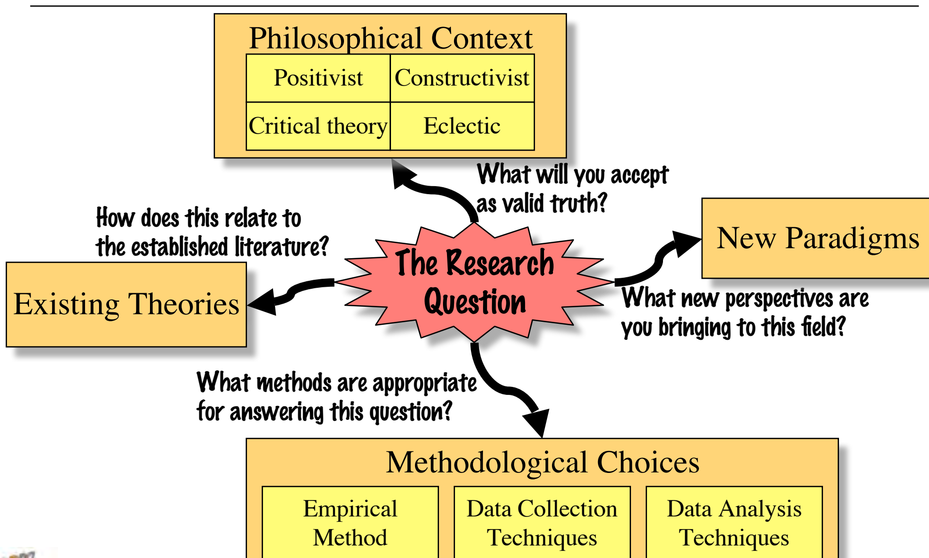
Causal Relationship

- 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



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

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

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?

Many available methods…

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"!)

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 extreme 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?

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 XP"
  - 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…

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
- 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
- …Or any combination of the above

STOP Do not use: Convenience sampling
- Cases that are easy/cheap to study
- low credibility!

Data Collection Techniques

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

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

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,…

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)

1. Basics of Empirical Research

9:00-9:30 What is Science?

9:30-9:50 Planning a study

9:50-10:10 *Exercise: Planning*

10:10-10:30 Validity and Ethics

10:30-11:00 Coffee break

Validity

- In software engineering, we worry about various issues:
  - Validation – is the software doing what is needed?
  - is it doing it in an acceptable or appropriate way?
  - Verification – is it doing what the specification stated?
  - are the structures consistent with the way it should perform?
- In empirical work, worried about similar kinds of things
  - Are we testing what we mean to test
  - Are the results due solely to our manipulations
  - Are our conclusions justified
  - What are the results applicable to
- The questions correspond to different *validity* concerns
  - The logic of demonstrating causal connections
  - The logic of evidence

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

- ❍ 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 "duration 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, company was sold during the project)
  - 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

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 (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

Research Ethics

- Reasons to take ethics seriously:
  - Funding depends on it
  - Relationship with research subjects/organisations depends on it
  - Legal issues (e.g. liability for harm to subjects/organisations)
  - Compliance with privacy and data protection laws
  - …and it's the right thing to do!
- Institutional Review Boards (IRB)
  - Approval usually needed for all studies involving human subjects
  - Every IRB has it's own rules…
    - A study approved at one university may be disallowed at another!
    - Design of the study might have to be altered
  - Institutional research funding may depend on this process!
  - Note: guidelines from other fields may not apply to Software Engineering
    - E.g. use/ownership of source code
    - E.g. effect of process improvement on participants

Informed Consent

- Elements
  - Disclosure - participants have full information about purpose, risks, benefits
  - Comprehension - jargon-free explanation, so participants can understand
  - Competence - participants must be able to make rational informed choice
  - Voluntariness - no coercion or undue influence to participate
  - Consent - usually indicated by signing a form
  - Right to withdraw
    - participant can withdraw from study at any point without having to give reasons
    - Participants can request their data to be excluded (might not be possible!)
- Challenges:
  - Student participants
    - Perception that their grade will be affected if they don't participate
    - Perception that it will please the course instructor if they participate
  - Industrial participants
    - Perception that the boss/company wants them to participate

An Ethical Dilemma..

You are dong a study of how junior analysts use new requirements tool at a leading consultancy company. As part of informed consent, staff are informed that they will remain anonymous. During the study, you notice that many of the analysts are making data entry errors when logging time spent with clients. These errors are causing the company to lose revenue. Company policy states clearly that workers salaries will be docked for clear mistakes leading to loss of revenue.

Questions:

- Would you alter the results of your study to protect the people who helped you in the study?
- How can you report results without causing harm to the participants?
- Would you cancel the study as soon as this conflict of interest is detected?

Should you pay your participants?

- Arguments in favour
  - Can help with recruitment
  - Compensate participants for their time
- Arguments against
  - May induce participants to take risks they otherwise would not take
  - May get expensive (esp if rates are to be more than a token)
- Issues
  - IRB might have standard rate; might be too low for professional SE
- Alternatives:
  - All participants entered into draw for some new gadget

Beneficence

- ❍ Risk of harm to Participants
  - Disrupts participant's work
  - Results of the research may devalue participants' work
  - Publication of study may harm the company's business
- ❍ Benefits of study
  - Scientific value: useful to society?
  - Depends on importance of the research topic!
  - Note: validity is crucial - invalid results means the study has no benefits
  - May also be specific benefits to participants
    - e.g. training, exposure to state-of-the art techniques, etc
- ❍ Beneficence: Benefits should outweigh the risks
  - Understand and justify any tradeoffs in the design of the study

Confidentiality

- ❍ Protecting Anonymity
  - Do not collect any data (e.g names) that allow participants to be identified
  - But you need a signed consent form, so…
  - Sever participants' identity from their data before it is stored and analyzed
  - Researcher-subject interactions should be held in private
- ❍ Protecting the data
  - Consent form states who will have access to the data, and for what purpose
    - Do not stray from this!
  - Raw data should be kept in a secure location
  - Reports should only include aggregate data
- ❍ Exceptions:
  - When it is impossible to identify individuals from the raw data
  - When more harm results from maintaining confidentiality than breaching it

15th IEEE International
Requirements Engineering
Conference

www.re07.org

October 15-19th, 2007
India Habitat Center, New Delhi

2. Quantitative Methods

11:00-11:30 Experiments

11:30-12:00 *Exercise: Design an Experiment in RE*

12:00-12:30 Survey Research

12:30 - 2:00 Lunch

Controlled Experiments

experimental investigation of a testable hypothesis, in which conditions are set up to isolate the variables of interest ("independent variables") and test how they affect certain measurable outcomes (the "dependent variables")

- good for
  - quantitative analysis of benefits of a particular tool/technique
  - establishing cause-and-effect in a controlled setting
  - (demonstrating how scientific we are!)
- limitations
  - hard to apply if you cannot simulate the right conditions in the lab
  - limited confidence that the laboratory setup reflects the real situation
  - ignores contextual factors (e.g. social/organizational/political factors)
  - extremely time-consuming!

See:

Pfleeger, S.L.; Experimental design and analysis in software engineering. *Annals of Software Engineering* 1, 219-253. 1995

Definitions

- Independent Variables
  - Variables (factors) that are manipulated to measure their effect
  - Typically select specific levels of each variable to test
- Dependent Variables
  - "output" variables - tested to see how the independent variables affect them
- Treatments
  - Each combination of values of the independent variables is a treatment
  - Simplest design: 1 independent variable x 2 levels = 2 treatments
    - E.g. tool A vs. tool B
- Subjects
  - Human participants who perform some task to which the treatments are applied
  - Note: subjects must be assigned to treatments randomly

Hypothesis Testing

- Start with a clear hypothesis, drawn from an explicit theory
  - This guides all steps of the design
  - E.g. Which variables to study, which to ignore
  - E.g. How to measure them
  - E.g. Who the subjects should be
  - E.g. What the task should be
- Set up the experiment to (attempt to) refute the theory
  - H_0 - the null hypothesis - "the theory does not apply"
    - Usually expressed as *no effect* - the independent variable(s) will not cause a difference between the treatments
    - H_0 assumed to be true unless the data says otherwise
  - H_1 - the alternative hypothesis - "the theory predicts…"
    - If H_0 is rejected, that is *evidence* that the alternative hypothesis is correct

Assigning treatments to subjects

- Between-subjects Design
  - Different subjects get different treatments (assigned randomly)
  - Reduces load on each individual subject
  - Increases risk that confounding factors affect results
    - E.g. differences might be caused by subjects varying skill levels, experience, etc
    - Handled through blocking: group subjects into "equivalent" blocks
    - Note: blocking only works if you can identify and measure the relevant confounding factors
- Within-subjects Design
  - Each subject tries all treatments
  - Reduces chance that inter-subject differences impact the results
  - Increases risk of learning effects
    - E.g. if subjects get better from one treatment to the next
    - Handled through balancing: vary order of the treatments
    - Note: balancing only works if learning effects are symmetric

Multiple factors (factorial design)

- Crossed Design
  - Used when factors are independent
  - Randomly assign subjects to each cell in the table
    - Balance numbers in each cell!
  - E.g. 2x2 factorial design:

		Factor B	
		Level 1	Level 2
Factor A	Level 1	A1B1	A1B2
	Level 2	A2B1	A2B2

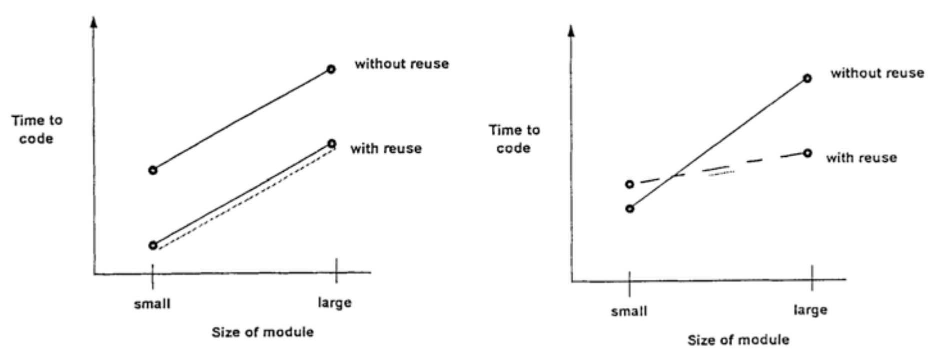
- Nested Design
  - Used when one factor depends on the level of the another
  - E.g. Factor A is the technique, Factor B is expert vs. novice in that technique

Factor A			
Level 1		Level 2	
Factor B		Factor B	
Level 1	Level 2	Level 1	Level 2
A1B1	A1B2	A2B1	A2B2

Experiments are Positivist

- Relies on reductionism:
  - Assume we can reduce complex phenomena to just a few relevant variables
  - If critical variables are ignored, results may not apply in the wild
  - Other variables may dominate the cause-and-effect shown in the experiment
- Interaction Effects:
  - Two or more variables might together have an effect that none has on its own
  - Reductionist experiments may miss this
    - E.g. A series of experiments, each testing one independent variable at a time
  - Using more than one independent variable is hard:
    - Larger number of treatments - need much bigger sample size!
    - More complex statistical tests

Detecting Interaction Effects

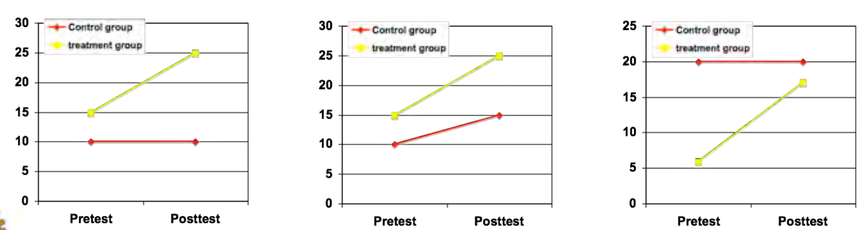


When not to use experiments

- When you can’t control the variables
- When there are many more variables than data points
- When you cannot separate phenomena from context
  - Phenomena that don’t occur in a lab setting
  - E.g. large scale, complex software projects
  - Effects can be wide-ranging.
  - Effects can take a long time to appear (weeks, months, years!)
- When the context is important
  - E.g. When you need to know how context affects the phenomena
- When you need to know whether your theory applies to a specific real world setting

Quasi-experiments

- When subjects are not assigned to treatments randomly:
  - Because particular skills/experience needed for some treatments
  - Because ethical reasons dictate that subjects get to choose
  - Because the experiment is conducted on a real project
- e.g. A Non-equivalent Groups Design
  - Pretest-posttest measurements, but without randomized assignment
  - E.g. two pre-existing teams, one using a tool, the other not
  - Compare groups’ improvement from pre-test to post-test



2. Quantitative Methods

11:00-11:30 Experiments

11:30-12:00 *Exercise: Design an Experiment in RE*

12:00-12:30 Survey Research

12:30 - 2:00 Lunch

Survey Research

"A comprehensive system for collecting information to describe, compare or explain knowledge, attitudes and behaviour over large populations"

- good for
  - Investigating the nature of a large population
  - Testing theories where there is little control over the variables
- limitations
  - Relies on self-reported observations
  - Difficulties of sampling and self-selection
  - Information collected tends to subjective opinion

See:

Shari Lawarence Pfleeger and Barbara A. Kitchenham, "Principles of Survey Research," Software Engineering Notes, (6 parts) Nov 2001 - Mar 2003

What is Survey Research?

- Survey Research ≠ Questionnaires
  - Can use questionnaires in any method
    - E.g. pre- and post- test in experiments
  - Can do survey research without questionnaires
    - E.g. using interviews, data logging, etc
- Distinguishing features:
  - Precondition: a clear research question that asks about the nature of a particular target population
  - selection of a representative sample from a well-defined population
  - data analysis techniques used to generalize from that sample to the population
  - Most suitable for answering base-rate questions

When to use Survey Research

- To evaluate the frequency of some characteristic across a population
  - E.g. how many companies use agile methods?
- To evaluate the severity of some condition that occurs in a population
  - E.g. what's the average cost overrun of software projects?
- To identify factors that influence a characteristic or condition
  - E.g. What factors cause companies to adopt new requirements tools?

Starting point

- Set clear objectives
  - A hypothesis to be tested
  - Any alternative explanations to be investigated
  - Identify a scope for the study appropriate for the objectives
  - Identify resources needed to meet the objectives
- Check that a survey is the right method:
  - Is it clear what population can answer the questions reliably?
  - Is there a way to get a representative sample of that population?
  - Do you have resources to obtain a large enough sample?
  - Is it clear what variables need to be measured?
  - Is it clear how to measure them?

Avoiding Sampling Bias

- Clear definition of the survey sample:
  - Define the U, the unit of analysis
  - Define the P, the target population
  - …such that P = {U}
  - Sample of the entire target population
    - not just the most accessible portion of it!
- Stratified Random Sampling for confounding variables:
  - E.g. U = individual developer,
    P = developers working in Canadian software companies
    - … but what if 80% of our sample comes from a single, dominant company?
  - If we really wanted U = Canadian Software Companies
    - Then change P
  - Alternatively, if company is a confounding variable
    - Group population by company, then sample within each

Survey Study Designs

- Cross-sectional design
  - Used to obtain a snapshot of participants' current activities.
- Case-control design
  - Asks each participant about several related issues
  - Used to establish whether a correlation exists between certain phenomena, across the population.
- Longitudinal study
  - Administer a survey periodically to track changes over time
- Cohort study
  - A longitudinal study that tracks the same participants each time

Avoiding Self-selection Bias

- Sampling the right population might not be enough
  - Low response rates (e.g. under 10%) are common
  - Low response rates may invalidate the sampling method
  - Participants who choose to respond might be unrepresentative:
    - E.g. People who are least busy
    - E.g. People who have a strong opinion on the research topic
- Probe reasons for low response rate
  - E.g. follow up phone calls to non-respondents

Create a survey instrument

- Use/adapt other people's instruments if possible
  - Existing instruments have already been validated
  - Makes it easier to compare research results
- Challenges:
  - Phrase the questions so all participants understand them in the same way
  - Closed questions:
    - Hard to give appropriate choices of answer
    - Hard to ensure all respondents understand the choices in the same way
  - Open questions:
    - Hard to analyse the responses
- Prototyping and validation
  - Test that participants can understand the questions
  - Test how long it takes them to answer
  - Use prototyping results to improve the instrument

Question Design

- Questions must be unambiguous and understandable:
  - Language appropriate to the population
  - Use standard grammar, punctuation, spelling
  - Each question covers exactly one concept
  - Avoid vague or ambiguous qualifiers
  - Balance positive and negative questions
- Typical mistakes:
  - Questions that participants can't answer
    - E.g. asking about decisions they weren't involved in
  - Double edged questions
    - E.g. "have you used RE tools or techniques, or would you consider using them?"
  - Leading questions
    - E.g. "did the project fail because of poorly managed requirements?"
  - Appropriation - reinterpreting participants' responses

Answer Design

- Response Categories
  - Numeric values (e.g number of months on the project)
  - Nominal categories (e.g. type of software being built)
  - Binary (e.g. Yes/No)
  - Ordinal scales (e.g. "how strongly do you agree with this statement…")
- Response options should be:
  - Exhaustive (but not too long!)
    - Include 'other' if you cannot ensure they are exhaustive
  - Mutually exclusive
  - Allow for multiple selections if appropriate
- Using ordinal scales:
  - Use 5 - 7 points on the scale
  - Label the points on the scale with words
  - End points must mean the opposite of one another
  - Intervals must seem to be evenly spaced

Reliability

- Test-Retest Reliability
  - If the same person answers the survey twice, do you get the same answers?
  - Problems:
    - What if the world has changed?
    - What if answering the questionnaire changes their attitude?
    - What if they just remember their answers from last time?
- Alternate Form Reliability
  - Do re-worded or re-ordered questions yield the same results?
- Inter-rater Reliability
  - If someone else administers the questions, do you get the same answers?
  - If someone else codes the responses, do you get the same results?

Interviews

- ❍ Types:
  - Structured - agenda of fairly open questions
  - Open-ended - no pre-set agenda
- ❍ Advantages
  - Rich collection of information
  - Good for uncovering opinions, feelings, goals, as well as hard facts
  - Can probe in depth, & adapt followup questions to what the person tells you
- ❍ Disadvantages
  - Large amount of qualitative data can be hard to analyze
  - Hard to compare different respondents
  - Interviewing is a difficult skill to master
  - Removal from context
  - Hard to elicit tacit knowledge (and post-hoc rationalization)
  - Interviewer's attitude may cause bias (e.g. variable attentiveness)

Interviewing Tips

- ❍ Set interviewees at ease with an innocuous topic to start
  - e.g. the weather, the score in last night's hockey game
  - e.g. comment on an object on the person's desk:
- ❍ Ask if you can record the interview
  - Put the recorder where it is visible
  - Let interviewee know they can turn it off at any time.
- ❍ Ask easy questions first
  - perhaps personal information
    - e.g. "How long have you worked in your present position?"
- ❍ Follow up interesting leads
  - E.g. if you hear something that indicates your plan of action may be wrong,
    - e.g.,"Could we pursue what you just said a little further?"
- ❍ Ask open-ended questions towards the end
    - e.g. "Is there anything else you would like to add?"

Surveys vs. other methods

- Use survey research when:
  - You need to find out what's true across (some part of) the s/w industry
  - Establish what is normal, common or uncommon.
- Use case study when:
  - You want want to understand what developers actually do
  - deeper insights into what happens in a small number of selected cases.
- Use an experiment (or quasi-experiment) when:
  - You want to investigate whether a particular technique has an effect on quality, development time, etc
  - tests for a causal relationship.
- Use an ethnography when:
  - You want to understand the culture and perspective of developers
  - Probes how developers themselves make sense of their context
- Use action research when:
  - You need to solve a pressing problem, and understand whether the solution was effective
  - Focusses on effecting change, and learning from the experience

3. Qualitative Methods

2:00-2:30 Case Studies

2:30-3:00 Exercise: Design a Case Study in RE

3:00-3:15 Ethnographies

3:15-3:30 Action Research

3:30-4:00 Tea break

Case Studies

"A technique for detailed exploratory investigations, both prospectively and retrospectively, that attempt to understand and explain phenomenon or test theories, using primarily qualitative analysis"

- good for
  - Answering detailed how and why questions
  - Gaining deep insights into chains of cause and effect
  - Testing theories in complex settings where there is little control over the variables
- limitations
  - Hard to find appropriate case studies
  - Hard to quantify findings

See:

Flyvbjerg, B.; Five Misunderstandings about Case Study Research. Qualitative Inquiry 12 (2) 219-245, April 2006

Myths about Case Study Research

1. General, theoretical (context-independent) knowledge is more valuable than concrete, practical (context-dependent) knowledge. Not true
2. One cannot generalize on the basis of an individual case; therefore, the case study cannot contribute to scientific development. Incorrect
3. The case study is most useful for generating hypotheses; that is, in the first stage of a total research process, whereas other methods are more suitable for hypothesis testing and theory building. No!
4. The case study contains a bias toward verification, that is, a tendency to confirm the researcher's preconceived notions. Don't believe it!
5. It is often difficult to summarize and develop general propositions and theories on the basis of specific case studies. Wrong!

[**See:** Flyvbjerg, B.; Five Misunderstandings about Case Study Research. Qualitative Inquiry 12 (2) 219-245, April 2006]

When should you use a case study?

- When you can't control the variables
- When there are many more variables than data points
- When you cannot separate phenomena from context
  - Phenomena that don't occur in a lab setting
  - E.g. large scale, complex software projects
  - Effects can be wide-ranging.
  - Effects can take a long time to appear (weeks, months, years!)
- When the context is important
  - E.g. When you need to know how context affects the phenomena
- When you need to know whether your theory applies to a specific real world setting

Why conduct a case study?

- To gain a deep understanding of a phenomenon
  - Example: To understand the capability of a new tool
  - Example: To identify factors affecting communication in code inspections
  - Example: To characterize the process of coming up to speed on a project
- Objective:
  - Exploration - To find what's out there
  - Characterization - To more fully describe
  - Validation - To find out whether a theory/hypothesis is true
- Subject of Investigation
  - An intervention, e.g. tool, technique, method, approach to design, implementation, or organizational structure
  - An existing thing or process, e.g. a team, releases, defects

Misuses of the term "Case Study"

- Not a case history
  - In medicine and law, patients or clients are "cases." Hence sometimes they refer to a review of interesting instance(s) as a "case study".
- Not an exemplar
  - Not a report of something interesting that was tried on a toy problem
- Not an experience report
  - Retrospective report on an experience (typically, industrial) with lessons learned
- Not a quasi-experiment with small n
  - Weaker form of experiment with a small sample size
  - Uses a different logic for designing the study and for generalizing from results

How can I tell it's a case study?

- Has research questions set out from the beginning of the study
- Data is collected in a planned and consistent manner
- Inferences are made from the data to answer the research questions
- Produces an explanation, description, or causal analysis of a phenomenon
  - Can also be exploratory
- Threats to validity are addressed in a systematic way

Research Questions

- Study design always starts with research questions
  - Clarify precisely the nature of the research question
  - Ensure the questions can be answered with a case study
  - Generally, should be “how” and “why” questions.
  - Identify and interpret the relevant theoretical constructs
- Examples:
  - “Why do 2 organizations have a collaborative relationship?”
  - "Why do developers prefer this tool/model/notation?"
  - "How are inspections carried out in practice?“
  - "How does agile development work in practice?"
  - "Why do programmers fail to document their code?“
  - "How does software evolve over time?“
  - "Why have formal methods not been adopted widely for safety-critical software?“
  - "How does a company identify which software projects to start?"

Types of Case Studies

- Explanatory
  - Adjudicates between competing explanations (theories)
  - E.g. How important is implementation bias in requirements engineering?
    - Rival theories: existing architectures are useful for anchoring, vs. existing architectures are over-constraining during RE
- Descriptive
  - Describes sequence of events and underlying mechanisms
  - E.g. How does pair programming actually work?
  - E.g. How do software immigrants naturalize?
- Causal
  - Looks for causal relationship between concepts
  - E.g. How do requirements errors and programming errors affect safety in real time control systems?
    - See study by Robyn Lutz on the Voyager and Galileo spacecraft
- Exploratory
  - Used to build new theories where we don’t have any yet
  - Choose cases that meet particular criteria or parameters
  - E.g. Christopher Columbus’ voyage to the new world
  - E.g. What do CMM level 3 organizations have in common?

Study Propositions

- Propositions are claims about the research question
  - State what you expect to show in the study
  - Direct attention to things that should be examined in the case study
  - E.g. “Organizations collaborate because they derive mutual benefits”
- Propositions will tell you where to look for relevant evidence
  - Example: Define and ascertain the specific benefits to each organization
- Note: exploratory studies might not have propositions
  - …but should lead to propositions for further study
  - …and should still have a clearly-stated purpose and clearly-stated criteria for success
- Analogy: hypotheses in controlled experiments

Unit of Analysis

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

Why Defining your Unit of Analysis matters

- Clearly bounds the case study
  - …and tells you which data to collect
- Makes it easier to compare case studies
  - …incomparable unless you know the units of analysis are the same
- Avoid subjective judgment of scope:
  - e.g. disagreement about the beginning and end points of a process
- Avoids mistakes in inferences from the data
  - E.g. If your study proposition talks about team homogeneity…
  - …Won't be able to say much if your units of analysis are individuals

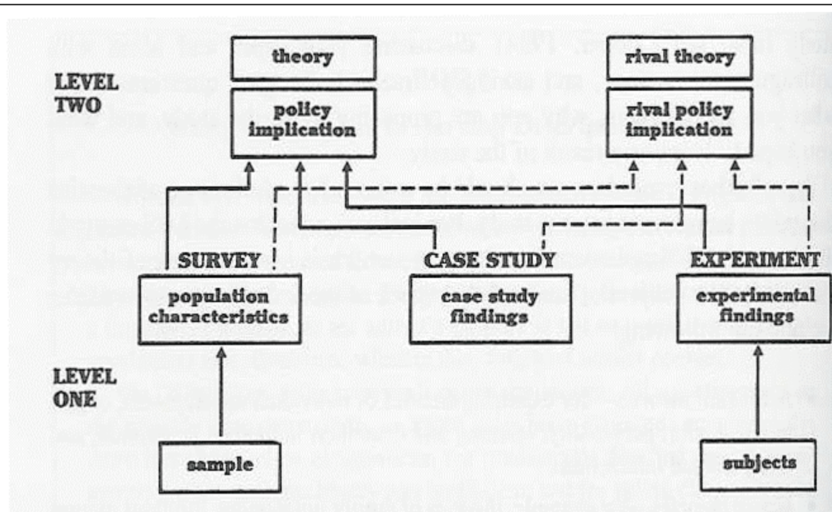
Linking Logic

- Logic or reasoning to link data to propositions
- One of the least well developed components in case studies
- Many ways to perform this
  - …none as precisely defined as the treatment/subject approach used in controlled experiments
- One possibility is pattern matching
  - Describe several potential patterns, then compare the case study data to the patterns and see which one is closer

Interpretation Criteria

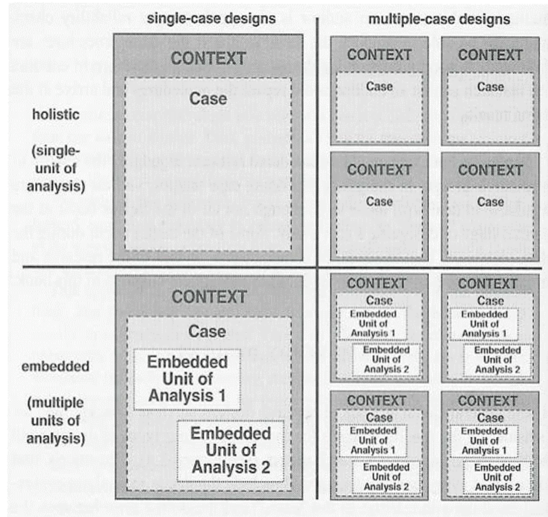
- Criteria for interpreting a study's findings
  - I.e. before you start, know how you will interpret your findings
- A relatively undeveloped component in case studies
  - No general consensus on criteria for interpreting case study findings
  - [Compare with standard statistical tests for controlled experiments]
- Statistical vs. Analytical Generalization
  - Quantitative methods tend to sample over a population
  - Statistical tests then used to generalize to the whole population
  - Qualitative methods cannot use statistical generalization
  - Hence use analytical generalization

Analytical and Statistical Generalization



Case Study Designs

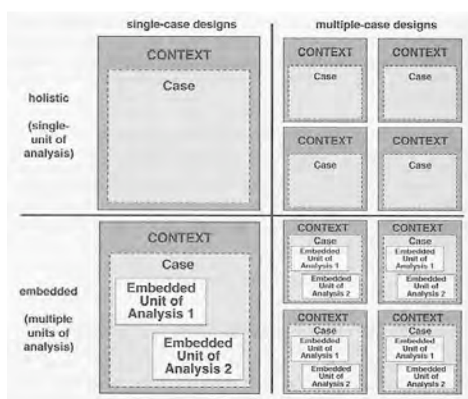
- 4 types of designs
  - Single-case vs. Multiple-case design
  - Holistic vs. Embedded design



Basic Types of Designs for Case Studies (Yin, page 40)

Holistic vs. Embedded Case Studies

- *Holistic* case study:
  - Examines only the global nature of one unit of analysis (not any subunits)
  - E.g: a case study about an organization
- *Embedded* case study:
  - Involves more than one unit of analysis
  - pays attention to subunit(s) within the case
  - E.g: a case study about a single organization may have conclusions about the people (subunits) within the organization



Holistic vs. Embedded?

Holistic Designs

- Strengths
  - Convenient when no logical subunits can be defined
  - Good when the relevant theory underlying the case study is holistic in nature
- Weaknesses
  - Can lead to abstract studies with no clear measures or data
  - Harder to detect when the case study is shifting focus away from initial research questions

Embedded Designs

- Strengths
  - Introduces higher sensitivity to "slippage" from the original research questions
- Weaknesses
  - Can lead to focusing only on the subunit (i.e. a multiple-case study of the subunits) and failure to return to the larger unit of analysis

Replication

- Select each study so that it either:
  - Predicts similar results (literal replication)
  - Predicts contrasting results for predictable reasons (theoretical replication)
- Use Replication logic rather than sampling logic
  - Sampling logic: define a pool of potential respondents, select a subset using a statistical procedure
  - Replication logic: select cases that support empirical induction
- If all results turn out as predicted:
  - That gives strong support for the initial propositions
- Otherwise:
  - the propositions must be revised and re-tested with another set of studies
- The theory should guide the choices of replication cases

How Many Cases?

- How many literal replications?
  - It depends on the certainty you want to have about your results
  - Greater certainty with a larger number of cases
    - Just as with statistical significance measures
    - 2 or 3 may be sufficient if they address very different rival theories and the degree of certainty required is not high
    - 5, 6, or more may be needed for higher degree of certainty
- How many theoretical replications?
  - Consider the complexity of the domain under study
    - If you are uncertain whether external conditions will produce different results, you may want to include more cases that cover those conditions
    - Otherwise, a smaller number of theoretical replications may be used

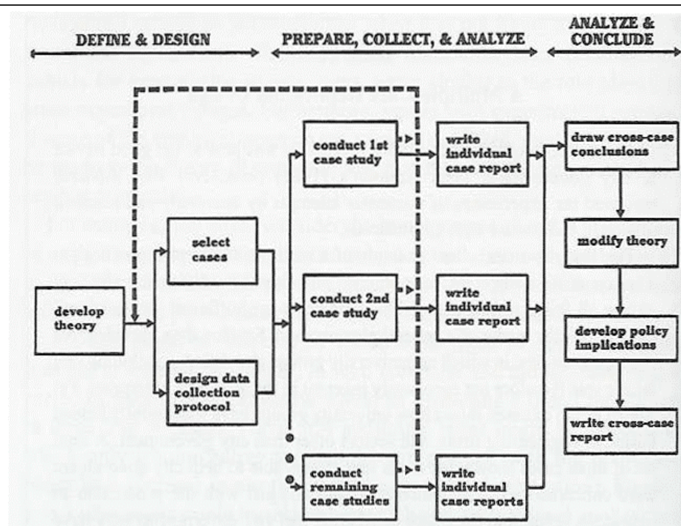
Multiple-Case Designs

- Useful when literal or theoretical replications provide valuable information
- Advantages
  - Evidence from multiple cases is more compelling
  - Overall study is therefore regarded as more robust
  - Differences in context for the cases improves generalizability of the findings
  - Offers opportunity to apply theoretical replications
- Disadvantages
  - Difficulty to find an appropriate number of relevant cases
  - Can require extensive resources and time

When is a single case sufficient?

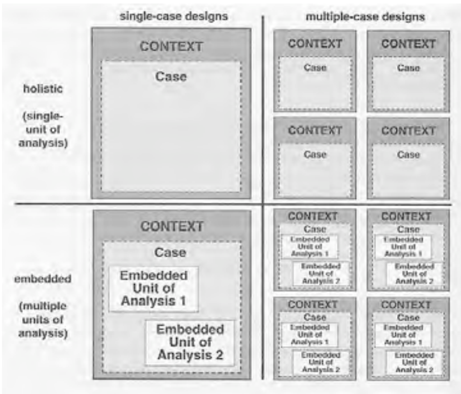
- It is the critical case in testing a well-formulated theory
  - The case meets all of the conditions for testing the theory thoroughly
- It is an extreme or unique case
  - E.g. a case with a rare disorder
- It is a representative or typical case,
  - It will tell us about common situations/experiences
- The case is revelatory
  - a unique opportunity to study something previously inaccessible
  - Opens a new topic for exploration
- The case is longitudinal – it studies the same case at several points in time
  - corresponding theory should deal with the change of conditions over time

Replication Approach for Multiple-Case Studies



Case Study Method (Yin page 50)

Multiple-Case Designs: Holistic or Embedded



- ❍ A multiple-case study can be:
  - multiple holistic cases
  - or multiple embedded cases
  - Cannot mix embedded and holistic cases in the same study!
- ❍ For embedded studies, subunit data are not pooled across cases
  - Used to draw conclusions only for the subunit's case

Selecting Case Study Designs – Closed or Flexible?

- ❍ A case study's design can be modified by new information or discoveries during data collection
  - Your cases might not have the properties you initially thought
  - Surprising, unexpected findings
  - New and lost opportunities
- ❍ If you modify your design, be careful to understand the nature of the alteration:
  - Are you merely selecting different cases, or are you also changing the original theoretical concerns and objectives?
  - Some dangers akin to software development's *feature creep*
  - Flexibility in design does not allow for lack of rigor in design
  - Sometimes the best alternative is to start all over again

3. Qualitative Methods

2:00-2:30 Case Studies

2:30-3:00 Exercise: Design a Case Study in RE

3:00-3:15 Ethnographies

3:15-3:30 Action Research

3:30-4:00 Tea break

Ethnographies

Interpretive, in-depth studies in which the researcher immerses herself in a social group under study to understand phenomena though the meanings that people assign to them

- Good for:
  - Understanding the intertwining of context and meaning
  - Explaining cultures and practices around tool use
  - Deep insights into how people perceive and act in social situations
- Limitations:
  - No generalization, as context is critical
  - Little support for theory building
  - Expensive (labour-intensive)

See:

Klein, H. K.; Myers, M. D.; A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. MIS Quarterly 23(1) 67-93. March 1999.

What is an Ethnography?

- Constructivist study of communities and cultures
  - Understand how people make sense of their (social) context
  - How they create categories and terms that are meaningful to them
  - Understand how social interactions evolve
  - Provides rich and detailed descriptions of participants' culture
- For Requirements Engineering
  - Studies technical work settings
  - E.g. How do teams manage to work collaboratively?
- Data driven rather than theory driven
  - No pre-existing theory
  - Researcher explicitly considers his/her own pre-conceptions

Principles for Ethnographies

- The Hermeneutic Circle
  - The parts only make sense in the context of the whole
  - The whole only makes sense if you understand the parts
  - To study meaning, study inter-dependence of parts and whole
- Contextualization
  - Critical reflection on social and historical background
- Interaction between researcher and subjects
  - Critical reflection on how this interaction shaped the study
  - "Data collection" is a social process too!
- Abstraction and Generalization
  - Hermeneutics and Contextualization link data to theoretical concepts
- Dialogical Reasoning
  - Interplay between preconceptions and data lead to cycles of revision
- Multiple Interpretations
  - No grand narrative
  - Each participant's perspective is valued
- Active Suspicion
  - Sensitivity to "biases" and "distortions" from participants' views

Planning an Ethnography

- Requires:
  - Research question focuses on cultural practices of a particular community
  - Need access to that community!
- Scope:
  - May not know boundaries of the community in advance
    - "membership" and the idea of "becoming a member" are cultural concepts
  - Often uses chain sampling to identify representative members of a community
  - Duration: weeks or months!
- Difficulties
  - Avoiding pre-conceptions
  - Very large volumes of qualitative data
    - Video recordings, field notes, transcripts, diaries, etc.
  - Researcher must be trained in observational techniques

Participant Observation

- Researcher 'becomes a member'
  - Privileged view of being part of the community studied
  - Reveals details that outside observer will miss
  - Allows longitudinal study, useful for very lengthy studies
- Challenges
  - Extremely time consuming
  - Resulting 'rich picture' is hard to analyze
  - Researcher must have the right technical and cultural background
  - Researcher must be trained in observational techniques

Ethnomethodology

- Studies how social order emerges
  - Social order constructed through participants' collective actions
  - Social order only observable when an observer immerses herself in it.
- Members' own Categories
  - Ethnomethodology attempts to use the subjects' own categories
    - What categories (concepts) do they use themselves to order the social world?
  - What methods do people use to make sense of the world around them?
- Techniques:
  - Conversational analysis
  - Measurement of body system functions - e.g. heartbeat
  - Studies of non-verbal behaviour (e.g. gestures, body language)
  - Detailed video analysis
  - Time-motion study - who is where, when?
  - Communication audit - who talks to whom about what?
  - Use of tools - status symbols plus sharing rules

www.re07.org

15th IEEE International
Requirements Engineering
Conference

October 15-19th, 2007
India Habitat Center, New Delhi

3. Qualitative Methods

2:00-2:30 Case Studies

2:30-3:00 Exercise: Design a Case Study in RE

3:00-3:15 Ethnographies

3:15-3:30 Action Research

3:30-4:00 Tea break

Action Research

"research and practice intertwine and shape one another. The researcher mixes research and intervention and involves organizational members as participants in and shapers of the research objectives"

- good for
  - any domain where you cannot isolate {variables, cause from effect, …}
  - ensuring research goals are relevant
  - When effecting a change is as important as discovering new knowledge
- limitations
  - hard to build generalizations (abstractionism vs. contextualism)
  - won't satisfy the positivists!

See:

Lau, F; Towards a framework for action research in information systems studies. Information Technology and People 12 (2) 148-175. 1999.

What is Action Research?

- Mix research with intervention:
  - Work to solve some real world problem
  - Simultaneously study the process of solving it
  - Useful where implementing a change requires a long term commitment
  - Useful when mixing research with professional activities
- Requires:
  - A problem owner, willing to collaborate
    - Sometimes, problem owner = researcher
  - Critical reflection on past, current and planned actions
    - How did these help to solve the problem?
  - An authentic problem
  - Authentic knowledge outcomes for participants
  - A commitment to effect real change

Definitions

- Hult and Lennung's definition:
  - Assists in practical problem solving
  - Expands scientific knowledge
  - Enhances actor competencies
  - Performed collaboratively in an immediate situation
  - Uses data feedback in a cyclic process
  - Aimes at increased understanding of a given social situation
  - Is applicable for understanding change processes
  - Undertaken within a mutually acceptable ethical framework
- Varieties:
  - Participatory Action Research - practitioners as co-researchers
  - Action Science - understand participant's behaviours as theories-in-use
  - Action Learning - focuses on group learning in context

Planning an AR Study

- Clear research aim:
  - E.g. to understand how a change impacts an organisation
  - E.g. to improve the social condition of a community
- Explicit philosophical stance
  - Constructivist - focuses on participant's differing views
  - Critical Theorist - focuses on perspective shift and/or emancipation
  - [Positivist - tests theory about change]
- Make Theoretical Assumptions explicit
  - E.g. preconceptions about the problem and the planned solution
  - E.g. assumptions about the nature of the organisation

Study Design

- Background information
  - Nature of the organisation
  - Nature and extent of the problem to be solved
- The planned change
  - Is it appropriate, adequate and practical?
  - Will it really address the problem?
- Research site
  - Where will the study be carried out?
  - Single or multiple sites?
  - Is the planned site(s) appropriate?
- Participants
  - Authentic problem owners?
- Data sources
  - Decide on collection techniques:
    - Observation, interviews, documents, focus groups, surveys, role play
  - Can we collect credible, dependable data from the research site?
- Duration
  - Length must be adequate to allow diagnosis, action, and reflection
- Degree of openness
  - Action/reflection cycles planned in advance, or emergent?
- Access and Exit
  - How to build mutual trust?
  - How will the researcher access the research site?
  - Define an explicit exit point

4. Strategies and Tactics

4:00-4:15 Mixed Methods

4:15-4:30 Exercise: Design a Research Strategy

4:30-5:00 Data Collection / Analysis

5:00-5:15 Publishing

5:15-5:30 Summary/Discussion

5:30 Finish

Mixed Methods

- Sequential explanatory strategy
  - Quantitative method first to test for a relationship
  - Then qualitative method to find an explanation
  - E.g. Experiment followed by case study
- Sequential exploratory strategy
  - Qualitative method first to develop hypotheses
  - Quantitative method to test the hypotheses
  - E.g. Ethnography followed by survey
- Concurrent triangulation strategy
  - Collect both qualitative and quantitative data in the same study
  - Use both to help confirm the findings
  - E.g. Case study that uses interviews, observations and performance measures

Research Methods

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

- Artifact/Archive Analysis ("mining"!)

Artifact / Archive Analysis

Investigation of the artifacts (documentation, communication logs, etc) of a software development project after the fact, to identify patterns in the behaviour of the development team.

- good for
  - Understanding what really happens in software projects
  - Identifying problems for further research
- limitations
  - Hard to build generalizations (results may be project specific)
  - Incomplete data
  - Ethics: how to get consent from participants

See:

Audris Mockus, Roy T. Fielding, and James Herbsleb. Two case studies of open source software development: Apache and mozilla. ACM Transactions on Software Engineering and Methodology, 11(3):1-38, July 2002.

Simulations

An executable model of the software development process, developed from detailed data collected from past projects, used to test the effect of process innovations

- Good for:
  - Preliminary test of new approaches without risk of project failure
  - [Once the model is built] each test is relatively cheap
- Limitations:
  - Expensive to build and validate the simulation model
  - Model is only as good as the data used to build it
  - Hard to assess scope of applicability of the simulation

See:

Kellner, M. I.; Madachy, R. J.; Raffo, D. M.; Software Process Simulation Modeling: Why? What? How? Journal of Systems and Software 46 (2-3) 91-105, April 1999.

Benchmarks

A test or set of tests used to compare alternative tools or techniques. A benchmark comprises a motivating comparison, a task sample, and a set of performance measures

- good for
  - making detailed comparisons between methods/tools
  - increasing the (scientific) maturity of a research community
  - building consensus over the valid problems and approaches to them
- limitations
  - can only be applied if the community is ready
  - become less useful / redundant as the research paradigm evolves

See:

S. Sim, S. M. Easterbrook and R. C. Holt "Using Benchmarking to Advance Research: A Challenge to Software Engineering". Proceedings, ICSE-2003

4. Strategies and Tactics

4:00-4:15 Mixed Methods

4:15-4:30 Exercise: Design a Research Strategy

4:30-5:00 Data Collection / Analysis

5:00-5:15 Publishing

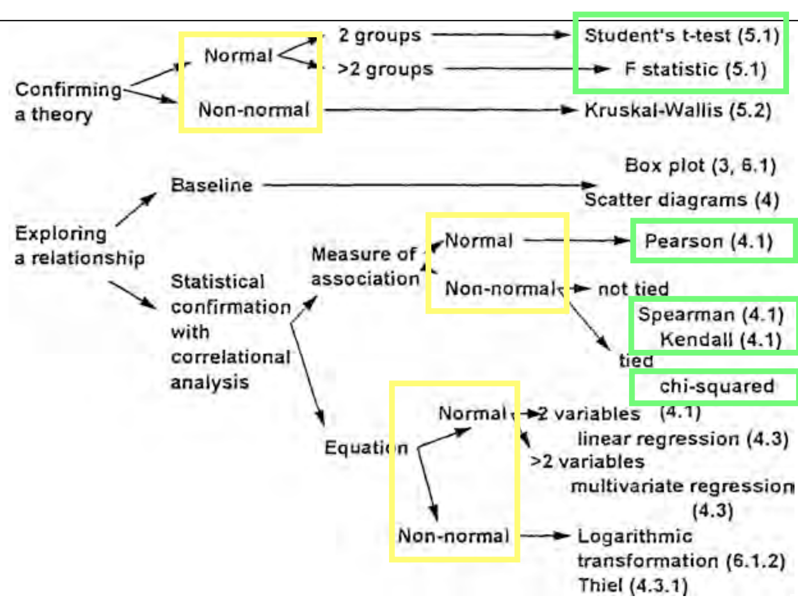
5:15-5:30 Summary/Discussion

5:30 Finish

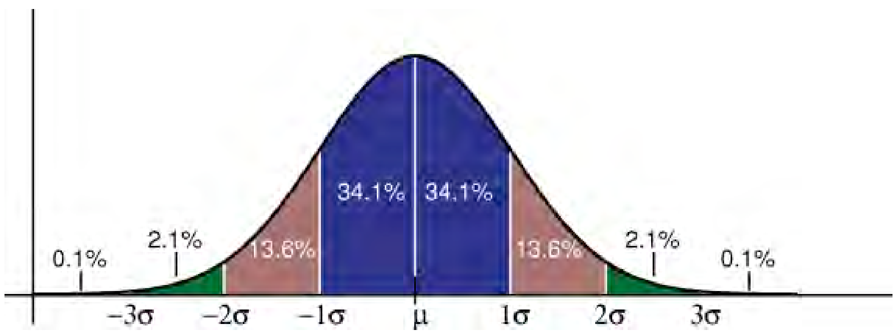
Analyzing Quantitative Data

- Most questions are about relationships between variables:
  - Is there a correlation?
  - Is there a cause-and-effect?
- For each relationship, we'd like to know:
  - Magnitude - how strong is the relationship?
  - Reliability - how well does the relationship in the sample represent the relationship in the population?
  - P value - probability that the relationship happened by chance
- Note:
  - strong relationships can be detected more reliably
  - Larger sample sizes produce more reliable results

Which Statistical Test?



Normal Distribution



Source: wikipedia - http://en.wikipedia.org/wiki/Image:Standard_deviation_diagram.png

Checking your data is normal

- Draw a Histogram
- Compute the mean and standard deviation
- Superimpose the expected normal curve over the histogram

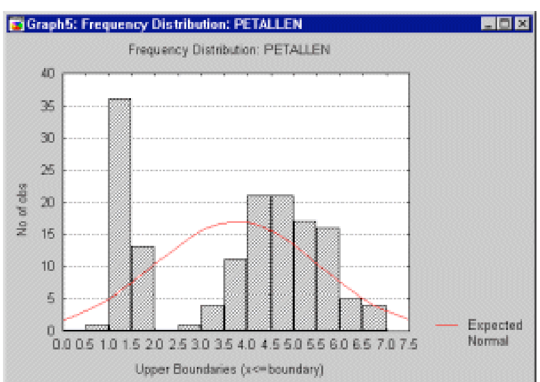
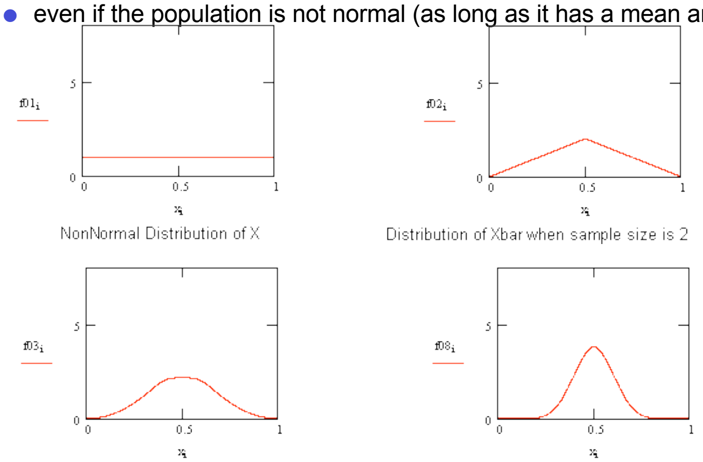


Image source:Statsoft (http://www.statsoft.com/textbook/stathome.html)

Central Limit Theorem

- Average of samples tend to normal distribution
  - …as sample size increases
  - even if the population is not normal (as long as it has a mean and SD)

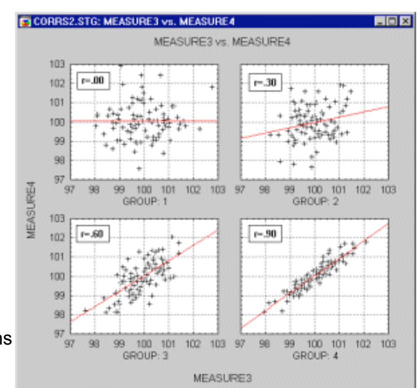


Correlations

- Measure relation between 2 variables:
  - -1 variables are perfect inverses
  - 0 no correlation at all
  - +1 variables are perfectly correlated
    - they appear on a straight line with positive slope
- Pearson's r
  - Computed as:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

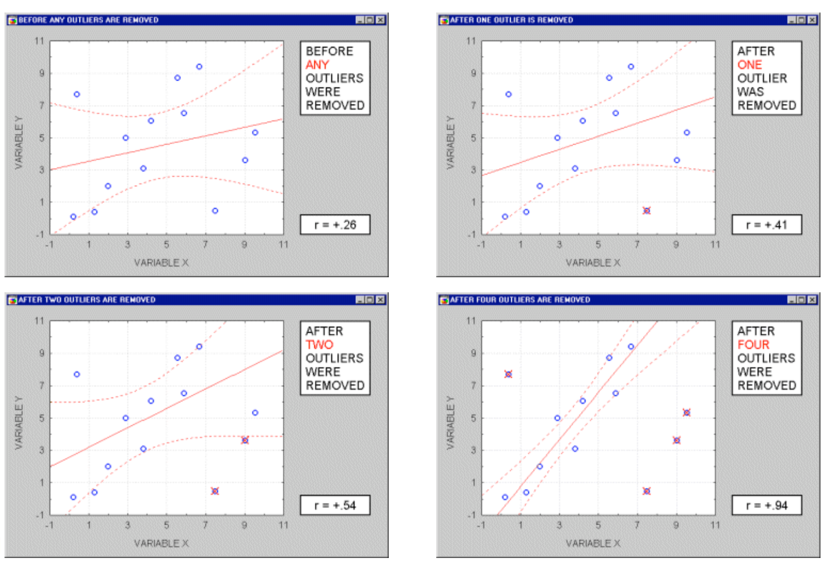
    - $\bar{x}$ and $\bar{y}$ are the sample means
    - s_x and s_y are the sample standard deviations
    - n is the sample size
  - Assumes variables are interval or ratio scale
  - Is independent of the measurement unit



Reminder: Measurement scales

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

Removal of outliers



Correlations for Ordinal Scales

- Spearman's Rank Coefficient (ρ):
  - Convert each variable into a ranked list
  - Compute:
    $$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$
  - D = difference between the ranks for corresponding X and Y values
  - N = Number of pairs of X,Y values
  - Note: assumes no tied ranks
- Kendall's Robust Rank Correlation (τ)
  - n - number of items (X,Y pairs)
  - P - sum (over all items) of the items ranked after the given item by both rankings
    $$\tau = \frac{2P}{\frac{1}{2}n(n-1)} - 1$$
  - Robust in the face of tied rankings

Student's t test

- For testing whether two samples really are different
  - given: two experimental treatments, one dependent variable
  - Assuming:
    - the variables are normally distributed in each treatment
    - the variances for the treatments are similar
    - the sample sizes for the treatments do not differ hugely
  - Basis: difference between the means of samples from two normal distributions is itself normally distributed.
  - The t-test checks whether the treatments are significantly different
- Procedure:
  - H_0: "no difference in population means from which the samples are drawn"
  - Choose a significance level (e.g. 0.05)
  - Calculate t as $t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{(SE_A)^2 + (SE_B)^2}}$ where $SE = \frac{SD}{\sqrt{N}}$
  - Look up the value for t, with degrees of freedom df = $(n_A + n_B) - 2$
  - If calculated value of t is greater than the lookup value, reject H_0

Analysis of Variance (ANOVA)

- Generalization of t-test for >2 treatments
  - given: n experimental treatments, one dependent variable
  - Assuming:
    - the variables are normally distributed in each treatment
    - the variances for the treatments are similar
    - the sample sizes for the treatments do not differ hugely
    - (Okay to deviate slightly from these assumptions for larger samples sizes)
  - Works by analyzing how much of the total variance is due to differences within groups, and how much is due to differences across groups.
- Procedure:
  - H_0: "no difference in the population means across all treatments"
  - Compute the F-statistic:
    - *F*=(found variation of group averages)/(expected variation of group averages)
  - If H_0 is true, we would expect F=1
  - ANOVA tells you whether there is a significant difference, but does not tell you which treatment(s) are different.

Chi-squared test

- "ANOVA for non-interval data"
  - Given: data in an n x m frequency table (e.g. n treatments, m variables)
  - Assuming:
    - Non-parametric, hence no assumption of normality
    - Reasonable sample size (pref >50, although some say >20)
    - Reasonable numbers in each cell
  - Calculates whether the data fits a given distribution
  - Basis: computes the sum of the Observed-Expected values
- Procedure:
  - Calculate an expected value (mean) for each column
  - Calculate χ^2:

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

    - Where O_i is an observed frequency
    - E_i is the expected frequency asserted by the null hypothesis
  - Compare with lookup value for a given significance level and d.f.

Analysing Qualitative Data

- Six Sources of Evidence
  - Documentation
    - Letters, memos, agendas, announcements, minutes, reports, newspaper clippings,…
  - Archival Records
    - Census data, maps, charts, data logs, sevice records, project archive, diaries,…
  - Interviews
  - Direct Observation
    - Notes, audio/video recording
  - Participant-observation
  - Physical Artifacts
    - Tools, devices, user interfaces, …
- Three Principles of Data Collection
  - Use Multiple Sources of Evidence
  - Create a repository for the data
  - Maintain a Chain of Evidence

Coding (based on grounded theory)

1. Open Coding
   - Select and name categories from the data
   - For each line/sentence: "What is this about?"
   - Identify recurring themes or concepts
2. Axial Coding
   - Relate codes to each other
   - Identify causal relationships
   - [Use a general framework for identifying relationships]
3. Selective Coding
   - Choose on category to be core
   - Relate all other categories to the core category
   - I.e. develop an overall storyline

4. Strategies and Tactics

4:00-4:15 Mixed Methods

4:15-4:30 Exercise: Design a Research Strategy

4:30-5:00 Data Collection / Analysis

$\boxed{\text{5:00-5:15 Publishing}}$

5:15-5:30 Summary/Discussion

5:30 Finish

Audience

- diverse audiences:
  - Practitioners
  - Peer Reviewers (make an accept/reject decision)
  - Other researchers working on the same problem
  - Broader research community
  - no single report will satisfy all audiences simultaneously!
- Orient the case study report to an audience
  - preferences of the potential audience should dictate the form of your study report
  - Greatest error is to compose a report from an egocentric perspective
  - Identify the audience before writing a case study report
  - Examine previous study reports that have successfully communicated with the identified audience

Composition Structures

- Linear-Analytic Structures
  - Standard approach
- Comparative Structures
  - E.g. Use key features as basis for comparing several cases
- Chronological Structures
  - Evidence are presented in chronological order
- Theory building Structures
  - Each chapter reveal a new part of a theoretical argument
- “Suspense” Structures
  - The outcome presented in the initial chapter, followed by the “suspenseful” explanation of the outcome
- Unsequenced Structures
  - The sequence of sections or chapters assumes no particular importance
    - make sure that a complete description of the study is presented. Otherwise, may be accused of being biased

Issues in Reporting

- When and How to Start Composing?
  - Start composing early in the analytic process
  - Bibliography, methodological and descriptive data about the studies could be written early in the process
- Participant Identities: Real or Anonymous?
  - Anonymity at two levels: entire organisation and individual person
  - Ideally, disclose of the identities of both the organisation and individuals
  - Anonymity is necessary when:
    - Using the real name will cause harm to the participants
    - The report may affect the subsequent action of those that are studied
  - Compromises
    - Hide individual but identify the organisation
    - Name individuals but avoid attributing any view or comment to a single individual
    - The published report limited to the aggregated evidence
  - Only if these compromises are impossible, make the entire study and the informants anonymous

General Guidelines from SE

Barbara A. Kitchenham, Shari Lawrence Pfleeger, Lesley M. Pickard, Peter W. Jones, David C. Hoaglin, Khaled El Emam, and Jarrett Rosenberg, "Preliminary Guidelines for Empirical Research in Software Engineering," IEEE Transactions on Software Engineering, Vol. 28, No 8, August 2002.

- Empirical Context
- Study Design
- Conducting the Study and Data Collection
- Analysis
- Presentation of Results
- Interpretation of Results

What Makes an Exemplary Study?

- An exemplary empirical study goes beyond the methodological procedures:
- It Must Be Significant
  - The issue are important, either in theory or practical terms
  - Relevant to scientific understanding or to policy decisions
- It Must be "Complete"
  - The boundaries of the study are given explicit attention
  - Exhaustive effort is spent on collecting all the relevant evidence
  - The study was not ended because of non-research constraints
- Considers Alternative Perspectives
  - Include consideration of rival propositions and the analysis of the evidence in terms of such rivals
- Displays Sufficient Evidence
  - Report the most relevant evidence so the reader can reach an independent judgment on the merits of the analysis
  - Use evidence to convince the reader that the investigator "knows" his or her subject
  - Show the validity of the evidence being presented
- The Report is Engaging
  - A well-written study report should entice the reader to continue reading

4. Strategies and Tactics

4:00-4:15 Mixed Methods

4:15-4:30 Exercise: Design a Research Strategy

4:30-5:00 Data Collection / Analysis

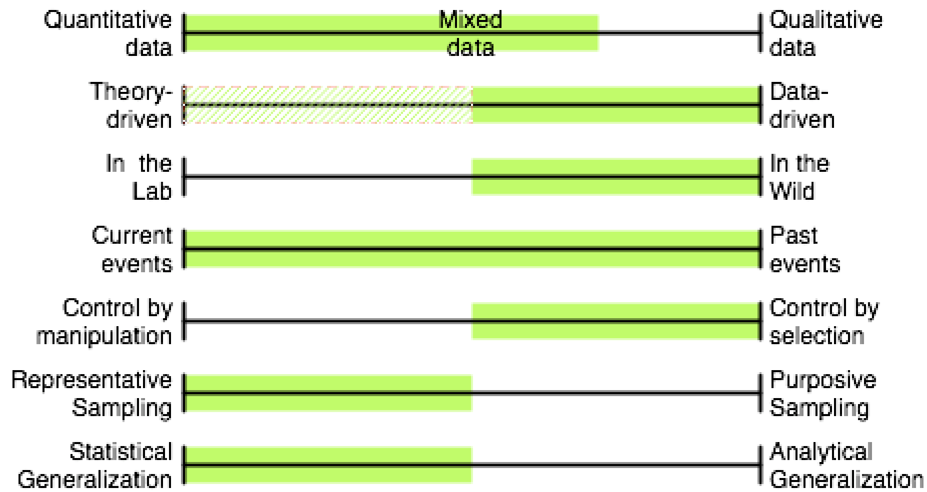
5:00-5:15 Publishing

5:15-5:30 Summary/Discussion

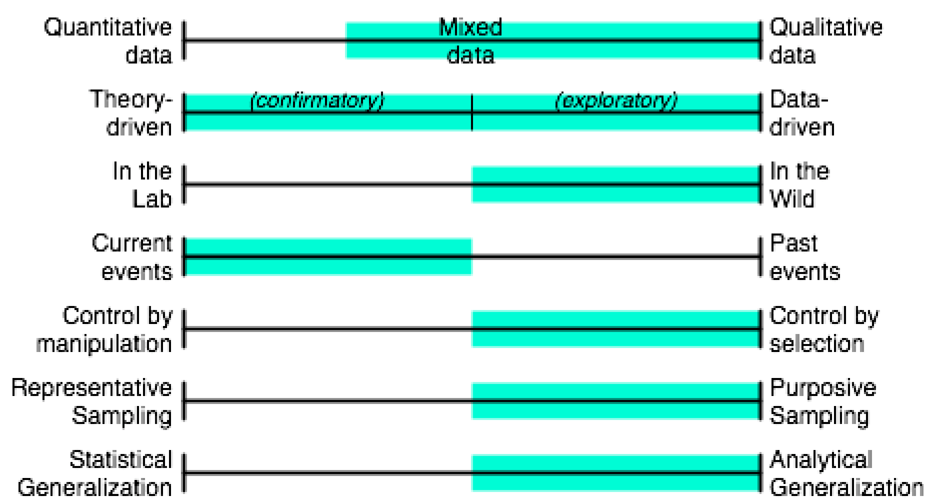
5:30 Finish

Controlled Experiments

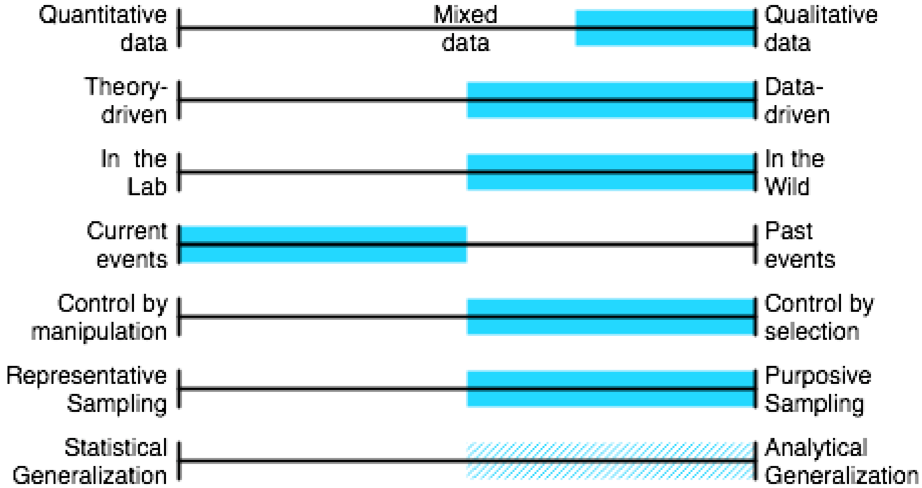
Survey Research



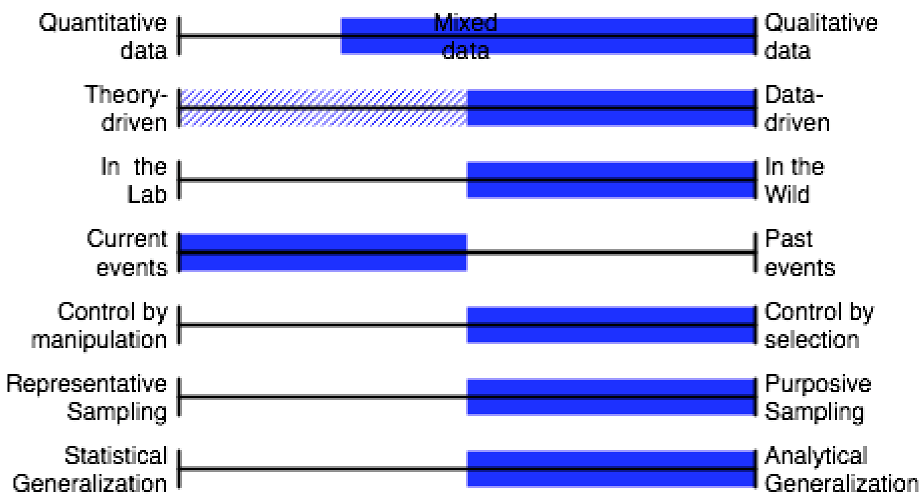
Case Studies



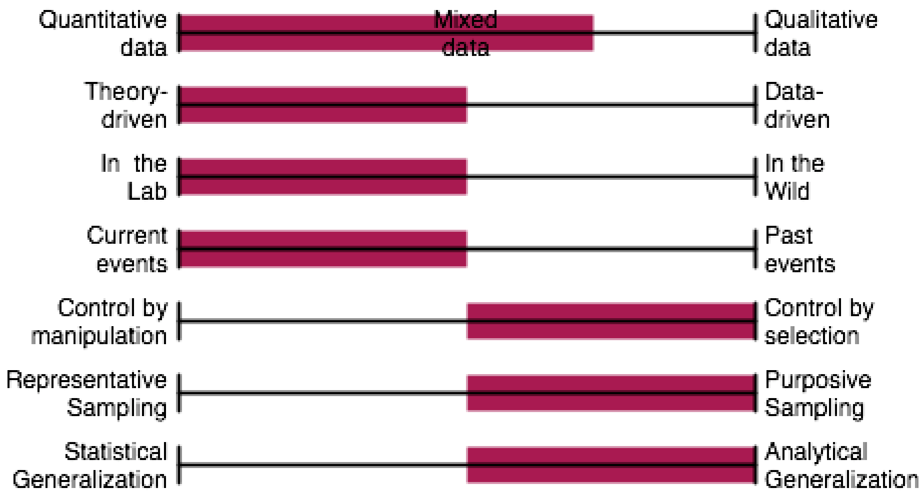
Ethnographies



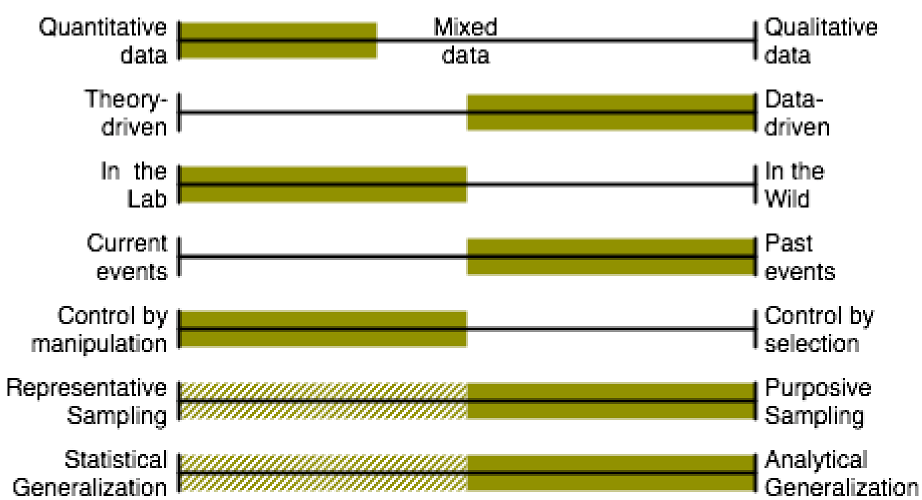
Action Research



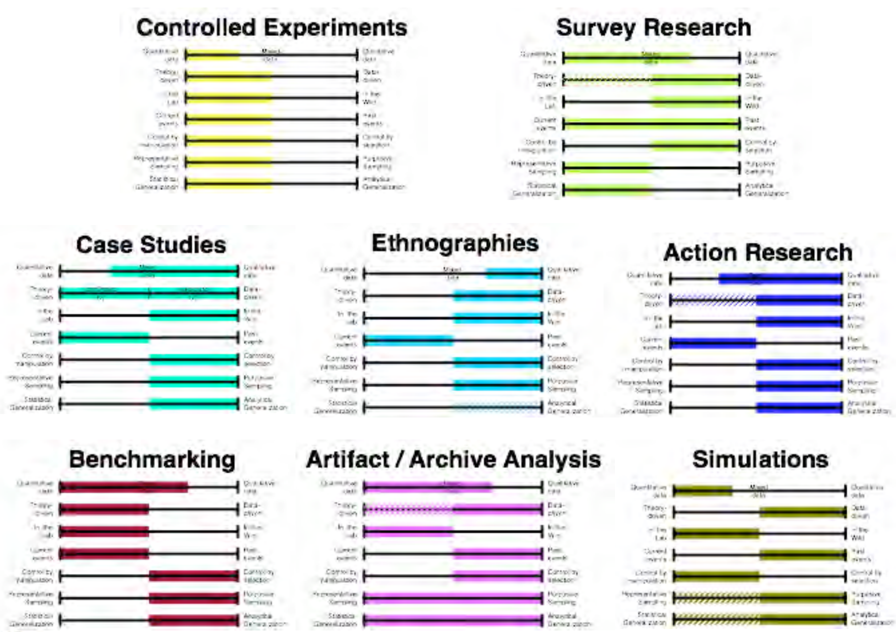
Benchmarking



Simulations



Artifact / Archive Analysis



Warning

No method is perfect

Don't get hung up on methodological purity

Pick something and get on with it

Some knowledge is better than none