MONASH University

Information Technology

FIT5190 Introduction to IT Research Methods

Lecture 7

Experiments

Slides prepared by

David Green, Frada Burstein, Jacques Steyn, Geoff Webb, Chung-Hsing Yeh

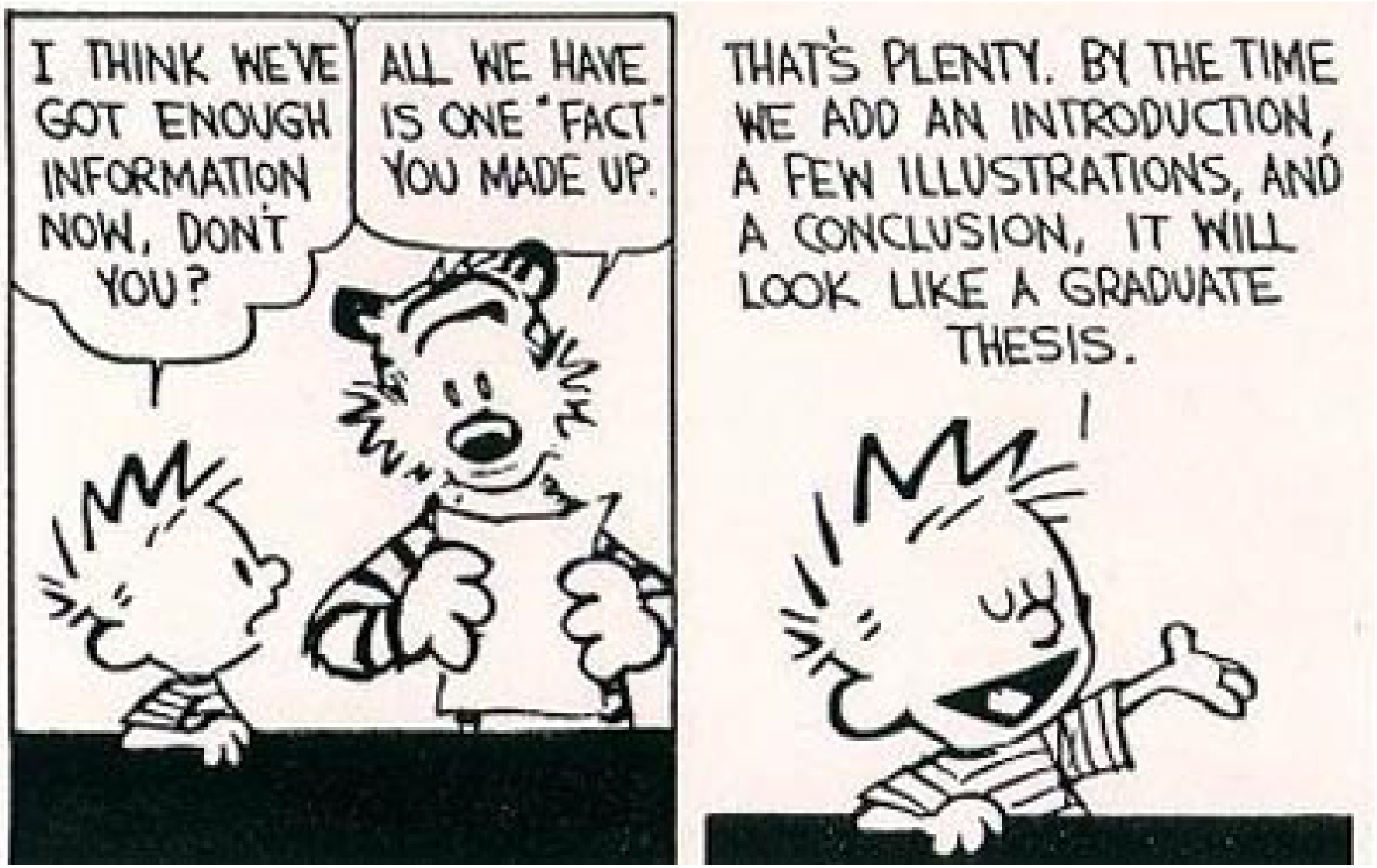
Learning objectives

- Be aware of
  - Experiments as sources of evidence
  - The role of sampling
  - The role of hypotheses in experiments
- Understand
  - Surveys as experiments
  - Causality and independent variables
  - Random and systematic design

Overview

- This lecture introduces the quintessential research method of modern science – the experiment.
- Experiments are systematic procedures by which researchers manipulate a system to gain evidence that will help them to answer their questions.
- Experimental research designs are dominant in the natural sciences but are also popular in medicine, psychology, and business disciplines.
- Experimental research is also central to much IT research.
- The aim of the lecture is to clarify the terminology around experiments and discuss a number of alternative experimental designs.
- At the end of the lecture students should be able to critically read IT papers that use experimental methods and know when to use an experimental design in a research project.

Research needs evidence



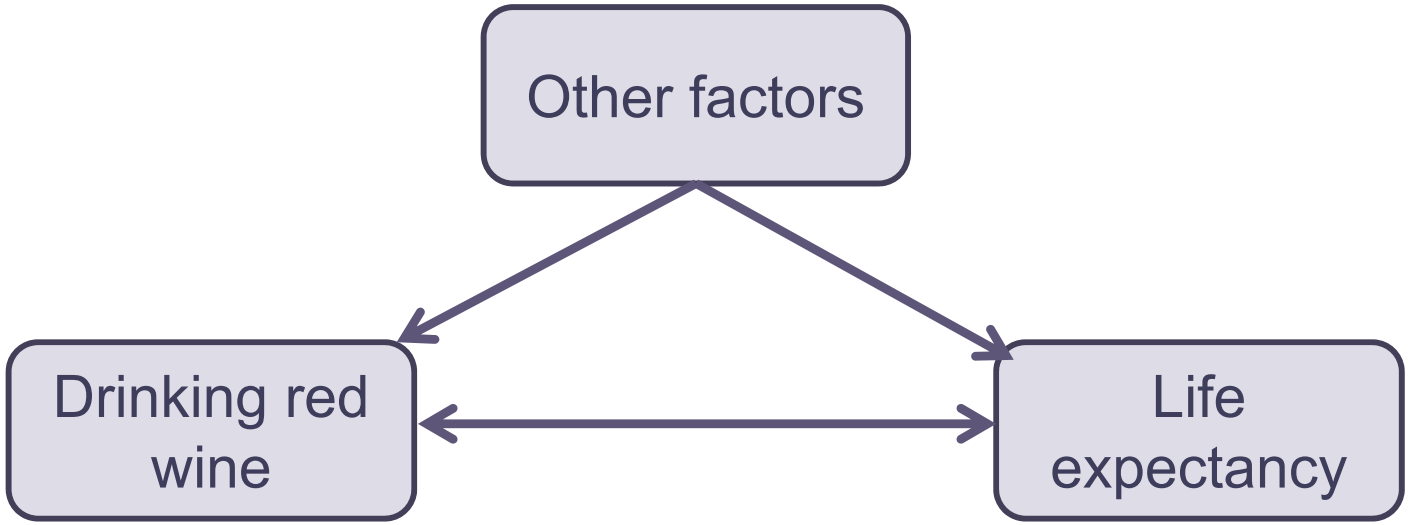
Cartoon from *Calvin and Hobbes* by Bill Watterson

Scientific studies

There are three main classes of investigation:

- Descriptive studies
  - Describe variables or phenomena
- Correlational studies
  - Identify relationships between variables
- Experiments
  - Manipulate and measure variables to infer causality

Reminder: correlation and causation



- Can only be confident that A causes B if
  - systematically manipulating A
  - while keeping all other factors constant
  - results in systematic changes to B
- Need to understand causation if we want to know how to achieve specific outcomes

Experimental methods

"Oh no! Research and development are on a test run."

Hidden assumptions

- All evidence is based on underlying assumptions
- Assumptions are often rooted in research paradigms
- Objectivity
  - Removing observer bias
  - Removing influence of experimenter
  - Heisenberg Uncertainty Principle
    - you can never be sure of everything
  - Double blind experiments
- Positivism
  - There is an objective truth to events, independent of the observer
- Interpretivism
  - Observers interpret the same events in different ways, according to their beliefs, assumptions, experience, etc.

Experiments

- Experiments
  - Methods for gathering evidence by manipulating the system concerned
- Assumption of causality
  - Changes made to independent variables **cause** changes to dependent variables
  - Experiments manipulate independent variables to observe changes caused in dependent variables
  - Example
    - Vary the amount of fertiliser given to plants and observe differences in growth

Subjects, participants, respondents

- Subjects
  - People, animals, flora, compounds, phenomena that are the subject of experiments
- Participants
  - When subjects are people they should be called participants (ARC & NHMRC)
  - Most papers and texts use ‘subjects’
- Respondents
  - In survey based research, a person answering a questionnaire in a survey is often called a respondent.

Groups

- Experimental group
  - Treatment group
  - Group that receives the experimental treatment
- Control group
  - Does not receive treatment
- Groups should be equivalent
  - Control extraneous variables
  - Random assignment
  - Matched pairs

Variables

- Independent variables
  - Variables whose effect we are interested in
  - Manipulated by the researcher
  - Levels - ways manipulated
- Dependent variables
  - The response or behaviour
  - Measures the influence of the independent variables

Variables (continued)

- Extraneous variables
  - Variables, other than the independent variables, capable of affecting the dependent variables
  - Confounding variables or confounds
  - Controlled variables
    - Extraneous variables that are controlled (kept constant across levels of the independent variables).
- Intermediate variables
  - Variables that measure possible constructs that mediate the influence of the independent variables on the dependent variables.

Hypotheses

- Experiments often test hypotheses.
- Hypotheses usually concern causality.
- Predictions about the effect of the independent variable on the dependent variable.
- Research hypotheses
  - Null hypotheses (H_0)
    - What the researcher seeks to reject
  - Alternative hypotheses (H_1)
    - What the researcher seeks to accept
  - Two-tailed, any difference
  - One tailed, direction is important

An example of hypotheses

- Alternative hypothesis
  - In mixed sex couples the woman usually does more housework than the man.
- Null hypothesis
  - In mixed sex couples the man usually does as much or more housework than the woman.

Significant difference

- Not sufficient to simply have a difference between the groups in an experiment to argue that the independent variable can affect the dependent variable in a causal way.
- A difference between two descriptive statistics that is of such magnitude that it is unlikely to have occurred by chance.
- Significance level*
  - 95% or $p \leq 0.05$ is acceptable
  - 99% or $p \leq 0.01$ is a strong result

**As used in statistics, "significance" does not mean important or meaningful.*

Statistical tests

- A statistical hypothesis test assesses the probability that the observed outcomes or more extreme would be observed if the null hypothesis were true.
- In statistical significance testing, the *p*-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.
- One often "rejects the null hypothesis" when the *p*-value is less than the significance level α (Greek alpha), which is often 0.05 or 0.01.
- When the null hypothesis is rejected, the result is said to be statistically significant.
- To be further discussed in the lecture on hypothesis testing.

Experiments

- Investigations where groups are treated identically except for a manipulation of the independent variable.
- Changes in the dependent variable may be attributed to the difference in the independent variable.
- Laboratory experiments
- Quasi-experiments (natural experiments)

Experiments (continued)

- Experiments aim to control a system
  - Vary factors in a controlled way
  - Keep everything else constant
  - Eliminate confounding errors
- Need for a control group
  - Compare difference between control & treated groups
  - Helps to identify the effects of the factors varied
- Example
  - Does fertiliser improve plant growth?
    - Fertilised (experimental or treated group) and non-fertilised plants (control group)
    - Ensure that the two groups are treated identically
      - e.g. Same water, soil, sunlight, etc.

Experimental design

"I was working on developing anti-matter under laboratory conditions, but I couldn't find anything to keep it in."

Experimental design 1

- Between-subjects design
  - Each group receives a different level of the independent variable.
- Within-subjects design
  - Repeated measures design, pretest-posttest designs.
  - Each subject receives each level of the independent variable.

Experimental design 2

- After-only design
  - Dependent variable is only measured after the manipulation of the independent variable.
- Before-after design
  - Dependent variable is measured before and after the manipulation and the effect of the independent variable is the difference between the two measures.

Experimental design 3

- Two group, between-subjects, after-only design
  - Classical experimental design.
- Between-subjects, one independent variable, with many experimental groups.
- Factorial design
  - Study of the effect of two or more independent variables on a dependent variable as well as the interaction between the independent variables.
  - Take samples for different independent variables (factors).
- Within-participants before-after design
- Mixed factorial design
  - Combines between-subjects and within-subjects designs.

Quasi-experiments

- Also called natural experiments.
- Membership of treatment group is determined by factors outside the experimenter's control that are effectively random.
  - e.g. Angrist, J., Evans, W. (1998) Children and their parents' labor supply: Evidence from exogenous variation in family size, *American Economic Review*, 88(3), 450-77.
  - Experimenter cannot control family size.
  - Observed differences may be due to common causes.
  - However, parents whose first two children are the same sex are more likely to have a third child.
  - These families form a group that is equivalent to one formed by the experimenter assigning groups randomly.

Quasi-experiments (continued)

- Between true experiments and correlation studies.
- Don’t meet all requirements for controlling extraneous variables.
- Why?
  - True experiments are not possible
  - Ethics
  - Practicalities

Samples versus census

- Census
  - Complete survey of an entire population
- Samples
  - Sets of observations drawn from a “population”
  - Used when it is not possible to survey the entire population (e.g. too large, or variable)
  - To be representative, the sampling design needs to include the same variations as the whole.

Samples versus census (continued)

- Example – fruit crop experiment
  - Suppose that you are investigating the effect of fertiliser on the size of an apple crop
- Census
  - Measure every apple from every tree
  - Time consuming and expensive
- Samples
  - Gather a sample of apples from a sample of trees.
  - The sample must reflect all variations in the population.

Sampling

- We typically study samples of the populations of interest.
- Aggregate measures gathered from a sample are likely to differ from the statistics gathered from the population as a whole.
- Want to reach conclusions about the population as a whole.
- Have to make allowance for the likely difference between the measures on the sample and the same measures for the population as a whole.
  - e.g., is the average man taller than the average woman?

Control samples

- Samples equivalent to test sample, but with key test features missing.
- Make it possible to detect differences.
- Used to test *alternative* versus *null* hypotheses
  - e.g. To test the effect of a vaccine, compare a sample of treated patients with a sample of untreated.

"It's frightening - according to the test results, the side effects of this designer drug are more beneficial than it's intended use!"

Sampling errors

- Example – fruit crop experiment
- Observer bias
  - Experimenter expects big apples on fertilised trees.
  - Looks for big apples.
- Random errors
  - Unknown factors may affect certain trees or apples.
  - Scales used to weigh apples are not accurate.
- Systematic errors
  - Birds or insects may prefer to attack apples on top of the tree.
  - Light and water may be better in one part of the orchard.
  - Are the apple trees all the same age, type, etc.?

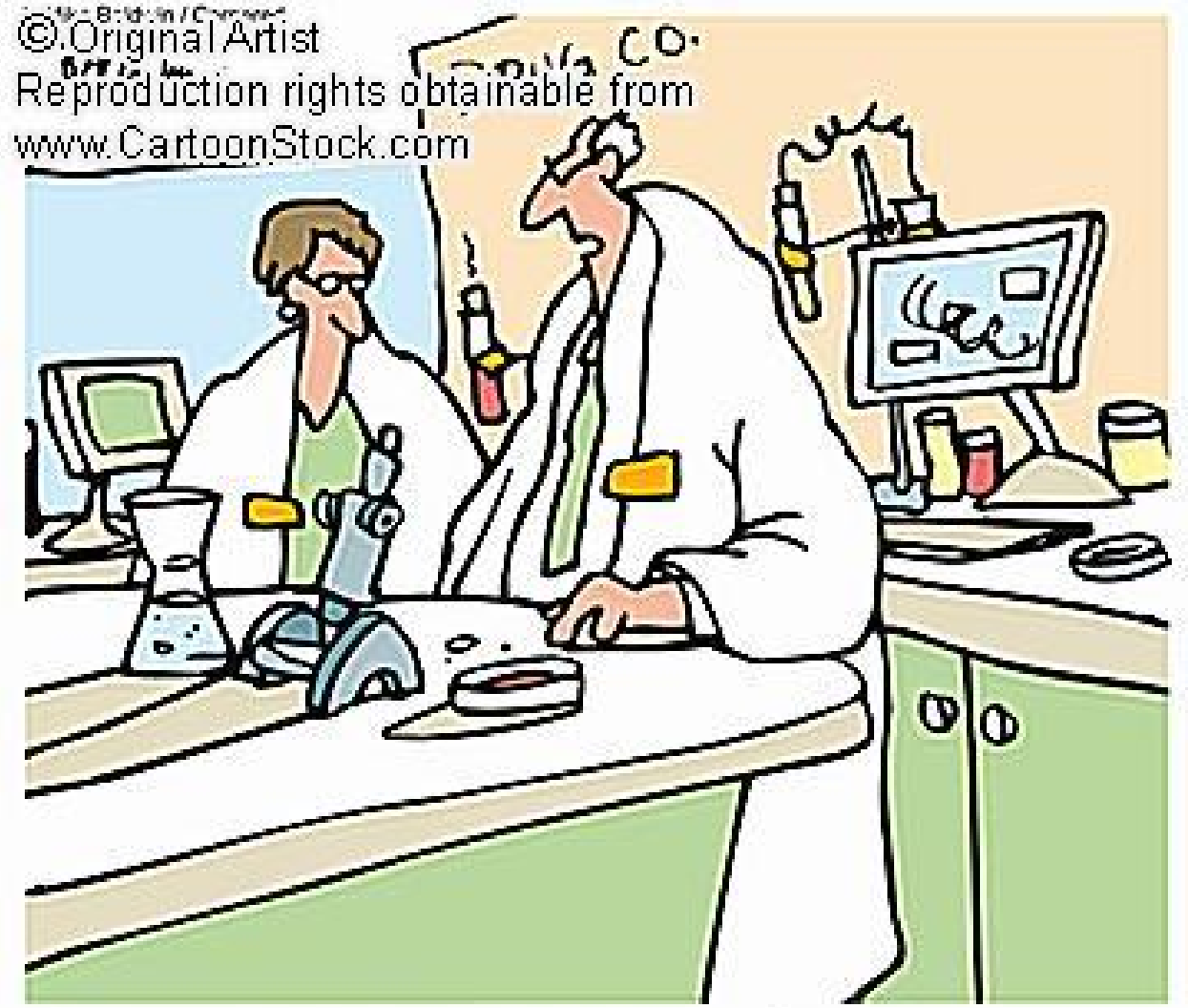
Random sampling

- Used to achieve unbiased results by avoiding bias in the choice of subjects in a sample.
- Uses random numbers to select the subjects from a population to be sampled.
  - To choose 10 from 1,000 objects, generate 10 random numbers in the range 1-1000 and select 10 objects with those numbers.
    - e.g. in Excel you could use “=INT(1000*RAND())+1”

      or “=RANDBETWEEN(1,1000)”
  - To randomly assign objects to 2 groups, randomly assign 0 or 1 to each object.
    - e.g. in Excel you could use “=IF(RAND()<0.5,0,1)”

Double blind sampling

- Used to eliminate bias
  - Neither experimenter nor subject knows which item is which.
  - Subjects may be influenced by the experimenter's body language about which answers to choose.
  - Experimenter may be influenced by knowing which subject is being tested (confirmation bias).
- Examples
  - Judging in Olympic gymnastics, skating etc. is notoriously biased by national interests.
  - French wines always won international competitions, but when double-blind judging was introduced Australian and Californian wines began to win.



"Do a double-blind test. Give the new drug to rich patients and a placebo to the poor. No sense getting their hopes up. They couldn't afford it even if it works."

Categorical versus continuous

- Categorical
  - Does the “stimulus” cause an effect?
    - All or nothing
  - Requires 2 groups
    - Experimental group (treatment applied)
    - Control group (no treatment applied)
  - Example
    - Give medicine to an experimental group, but not to the control group
- Continuous
  - How does effect change in relation to stimulus?
  - Multiple tests with different levels of stimulus
  - Example
    - How much medicine is needed to cause a response?

Example: fertilisation experiment

- Null hypothesis H_0:
  - "Fertilisation makes no difference to plant size"
- Alternative hypothesis H_1:
  - "Fertilised plants grow larger"
- Experiment
  - Grow equal numbers of fertilised and unfertilised plants
  - Gather observations of the resulting plants
- Two kinds of evidence
  - **Categorical**: Classify observed plants by size
  - **Continuous (Measurement)**: Measure the size of observed plants

Plant fertiliser experiment

- Categorical approach
  - Classify each plant as “large” or “small”
  - Use **chi-square** to test hypothesis about size *
- Measurement approach
  - Measure size of each plant
  - Use ***t*-test** to test hypothesis about size *

* To be covered in the lecture on hypothesis testing

Classification → contingency table *

		Treatment		
		No Fertiliser	Fertiliser	*N*
Result	Small plants	24	12	**36**
	Large plants	16	28	**44**
	N	**40**	**40**	**80**

N = total number of cases

Table entries are **observed** numbers of plants

* To be *discussed in detail* in the lecture on hypothesis testing

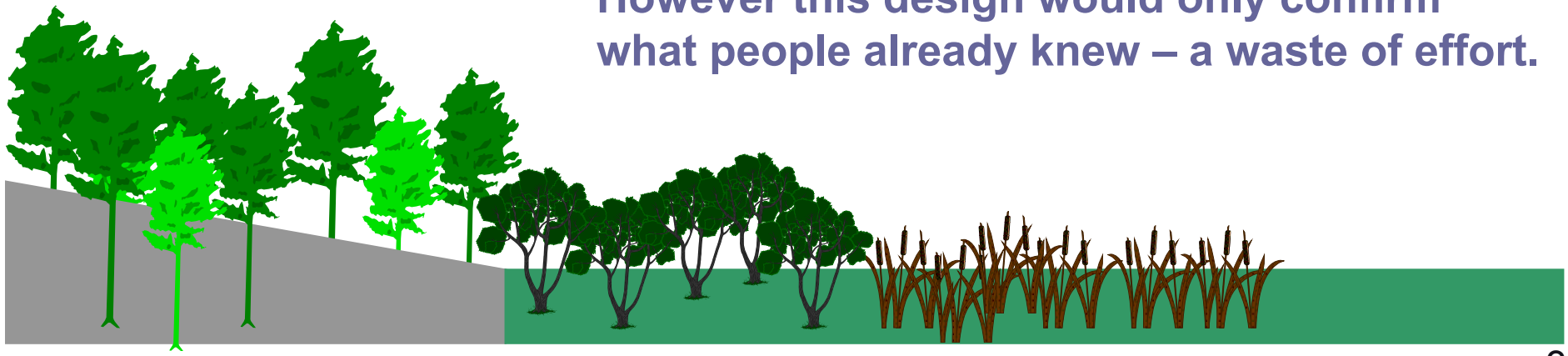
Example - initial design

Question: Does salinity affect growth of mangroves?

To answer this question, the researcher planned to grow 100 mangrove seedlings in pots, with varying amounts of salt in them, and watch how fast each plant grew. At first she opted for a simple, traditional design…

Salinity	Number of plants
Low	50
High	50

However this design would only confirm what people already knew – a waste of effort.

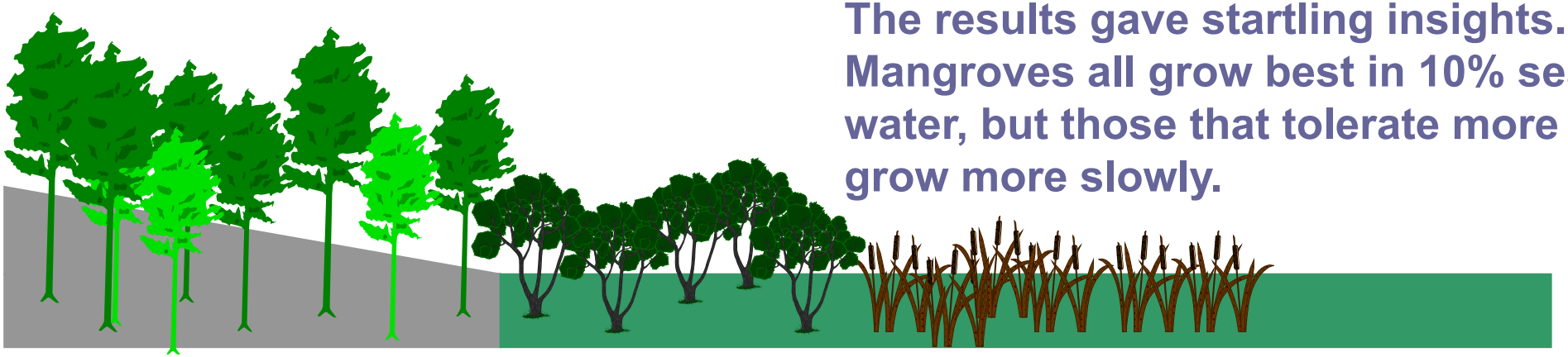


Example - final design

Question: **HOW** does salinity affect mangrove growth?

To answer this revised question, the researcher grew 100 mangrove seedlings in pots with many ***different*** levels of salt. The revised design allowed her to see how mangroves responded as the level of salt increased.

Salinity (% sea water)	0	10	20	30	40	50	60	70	80	90
Number of plants	10	10	10	10	10	10	10	10	10	10



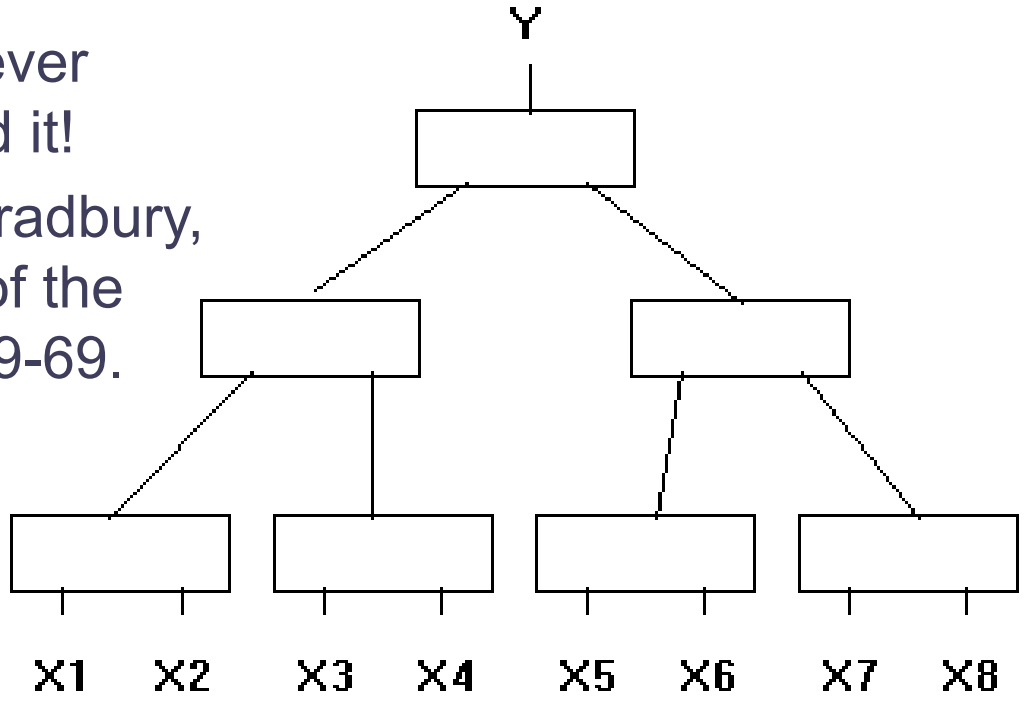
The results gave startling insights. Mangroves all grow best in 10% sea water, but those that tolerate more salt grow more slowly.

Example – evaluate a new algorithm

The Group Method of Data Handling (GMDH)

- Devised by the Russian mathematician Ivakhnenko in 1966
- The algorithm fits models automatically to data presented to it.
- The models are pyramids of polynomials. Inputs (X_1 … X_n) are input at the base and a result Y is output at the top.

- Used in hundreds of studies but never properly tested until we researched it!
- Green, D.G., Reichelt, R.E., and Bradbury, R.H. (1988). Statistical behaviour of the GMDH algorithm. *Biometrics* 44, 49-69.



Example – evaluate a new algorithm

- We asked the question “how well does GMDH work?”
  - We tested its **fidelity**. i.e. Did it reproduce known models? by using artificial data with a range of properties – linear, polynomial, etc.
  - We tested its **robustness**. i.e. Did it always work? by repeating the above tests after adding different levels of noise.
  - Finally, we tested its **utility** on case studies using real data from real systems of different kinds that had already been studied.
- The results showed that …
  - The algorithm performed worst on the very systems where it was supposed to perform best!
  - On basic tests it failed even to match simple regression.
  - It was wildly unstable outside the range of data used to train it.
- Outcome:
  - Researchers promptly abandoned the method.
  - The findings suggested ways to build better algorithms.

Surveys as experiments

- Testing people's attitudes, opinions, etc.
  - Require random and systematic sampling
  - Attempt to reduce observer bias
  - Measurement scales are often used to summarise findings

Likert scale

- Used with survey data
- Expresses the intensity and direction of a variable
  - e.g. Opinion about an issue
- Usually an ordinal scale (1, 2, 3, …)
- Examples
  - Quality of service
  - Trust
  - Happiness

Likert scale

- Procedure
  - Sequence of questions
  - Multiple choice answers
  - Each answer is on a scale, e.g. 5-point scale:
    1. Strongly disagree
    2. Disagree
    3. Neither agree nor disagree
    4. Agree
    5. Strongly agree
- Scale value
  - Sum of values across all people surveyed

Thurstone scale

- Use a set of statements (survey items) to measure an attitude about a particular issue.
- Each statement has a numerical value to indicate how favourable or unfavourable it is judged to be.
- Each of the statements that are favourable is checked, and a mean score is computed, indicating the attitude.
- e.g. attitude about the service quality of a business
  - Delivers promptly and on time
  - Charges reasonable prices
  - Provides friendly service

Guttman scale

- Used to ensure that a questionnaire measures a single trait (issue).
- Guttman's key insight
  - People who agree with more extreme items (e.g. capital punishment) will also agree with all less extreme items (e.g. prison).
- On a Guttman scale, survey items are arranged in an order so that an individual who agrees with a particular item also agrees with items of lower rank-order.
- The survey items have a binary (YES or NO) answer.
- A well-known example of a Guttman scale is the Bogardus Social Distance Scale.

Bogardus Social Distance scale

- Measures “distance” separating social groups
- Cumulative over many survey items
- Example
  1. Are you willing to permit immigrants to live in your country?
  2. Are you willing to permit immigrants to live in your community?
  3. Are you willing to permit immigrants to live in your neighbourhood?
  4. Are you willing to permit immigrants to live next door to you?
  5. Would you permit your child to marry an immigrant?
- Agreement with more extreme item 3 implies agreement with less extreme items 1 and 2.
- The Bogardus social distance scale is a cumulative scale (a Guttman scale), because agreement with any item implies agreement with all preceding items.

Semantic calibration

- In some cases, people react differently to different terms, even though they are interchangeable
  - Good/bad, desirable/undesirable
- Such differences can bias results using a scale
- Several ways to deal with the problem, e.g.
  - Include different versions of the same question

Experiments are not always feasible

- May not be possible to recruit sufficient numbers of subjects (participants or respondents).
- May not be feasible to establish experimental treatments.
  - Factors are not controllable by experimenter.
  - It would be unethical.
- May not be feasible to measure the outcomes.
  - Too expensive
  - Take too long
  - Not measurable

Experimental research

- Important part of IT research.
- Unique contribution is testing causal propositions.
- Experimental design is very difficult.
- Often based on descriptive research.
- Can lead to other styles of research.

Recommended reading

- The Skeptics Dictionary
  - http://skepdic.com/ , *especially entries on*
  - *Science & Philosophy*
    - http://www.skepdic.com/tiscience.html
  - *Critical Thinking*
    - http://www.skepdic.com/ticriticalthinking.html
- Essay on Critical Thinking
  - Facione, P. (2007). *Critical Thinking: What It Is and Why It Counts*. Insight Assessment. http://www.insightassessment.com/pdf_files/what&why2006.pdf

Useful references

- Leady, P.D., Ormrod, J.E. (2013) *Practical Research: Planning and Design.* Pearson, Pearson. Chapter 9.
- Fowler, F.J. (1993). *Survey Research Methods.* 2nd ed. Sage Publications, Newbury Park, CA.
- Galliers, R. (1991). Choosing information systems research approaches. In Robert Galliers (Ed.) (1992) *Information systems research: Issues, methods and practical guidelines.* (Ch. 8). Blackwell Scientific, Oxford.
- Zikmund, W.G., Babin, B.J., Carr, J.C., Griffin, M. (2010). *Business Research Methods*, 8th ed. Mason, OH: South-Western, Cengage Learning.