ICA Commission on Cognitive Visualization

www.geo.uzh.ch/microsite/icacogvis/

Designing Experiments...

Or how many times and ways can I screw that up?!?

Amy L. Griffin

AutoCarto 2012, Columbus, OH

UNSW THE UNIVERSITY OF NEW SOUTH WALES CANBERRA

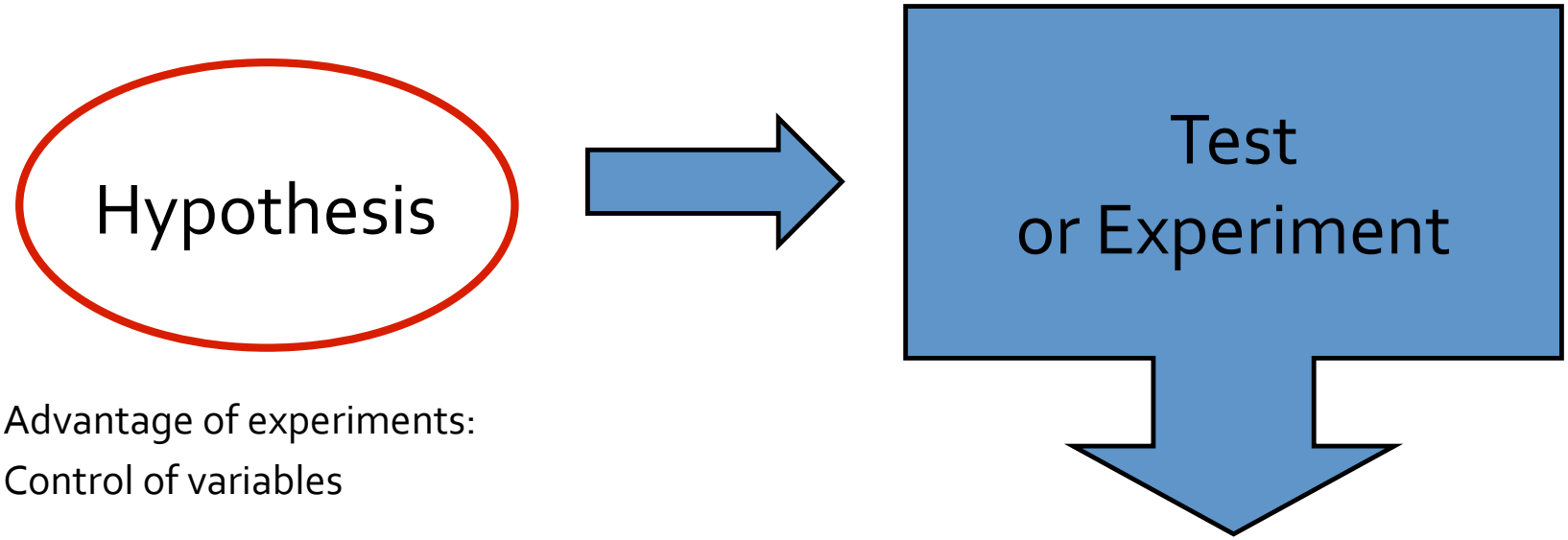
Outline

- When do I need to run an experiment and what kinds of experiments are there?
- How does the way I design my experiment affect what results I get?
- How do I analyze my data?
- What do I need to report when I write up my results?

Methods of Gathering Data on Cognition and Perception

- Introspection
- Naturalistic observation
- Interviews / Surveys
- … many others mentioned by Corné and Kristien
- Experiments

When do I need to run an experiment?



Advantage of experiments:
Control of variables

Disadvantage of experiments:
Real-world work is often variable (uncontrolled)

Independent Variable

Dependent Variable

Different user study goals: (not always mutually exclusive)

1) Improve the design of a geovisual tool
2) Understand perceptual and cognitive processes involved in using tools

Examples from real experiments

Experiment I:

Change detection experiment

Experiment II:

Hypothesis generation / tool use experiment

Experiment III:

Animated / static map visual cluster detection experiment

Experiments - Overview

An independent variable (IV) is manipulated

Single IV or multiple IVs (factorial design)

A dependent variable(s) (DV) is measured

Many basic experiments consist of two levels of the independent variable

- experimental group (e.g., new technique)
- control group (e.g., standard technique)

Control over extraneous variables (sources of error)

- holding constant (e.g, lighting, handedness of participants, etc.)
- randomizing effects

A causal relationship between the independent and dependent variables *can* be established

Choosing an experimental task

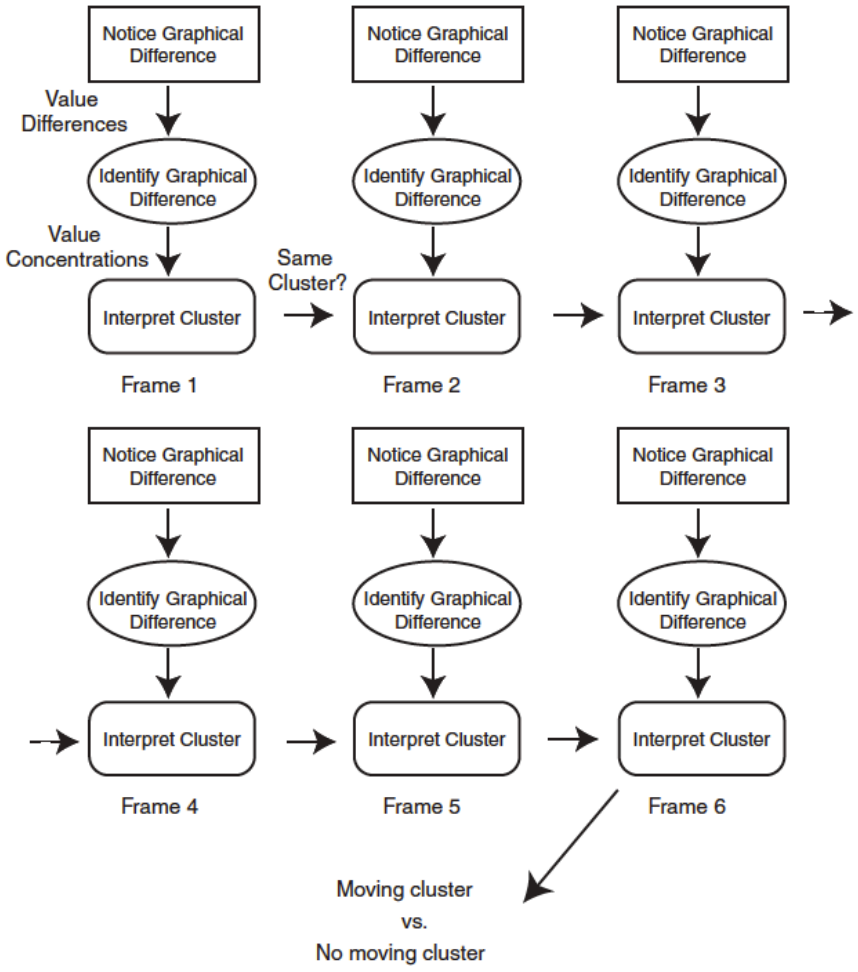
1. You want one that reflects the demands of a real world task.

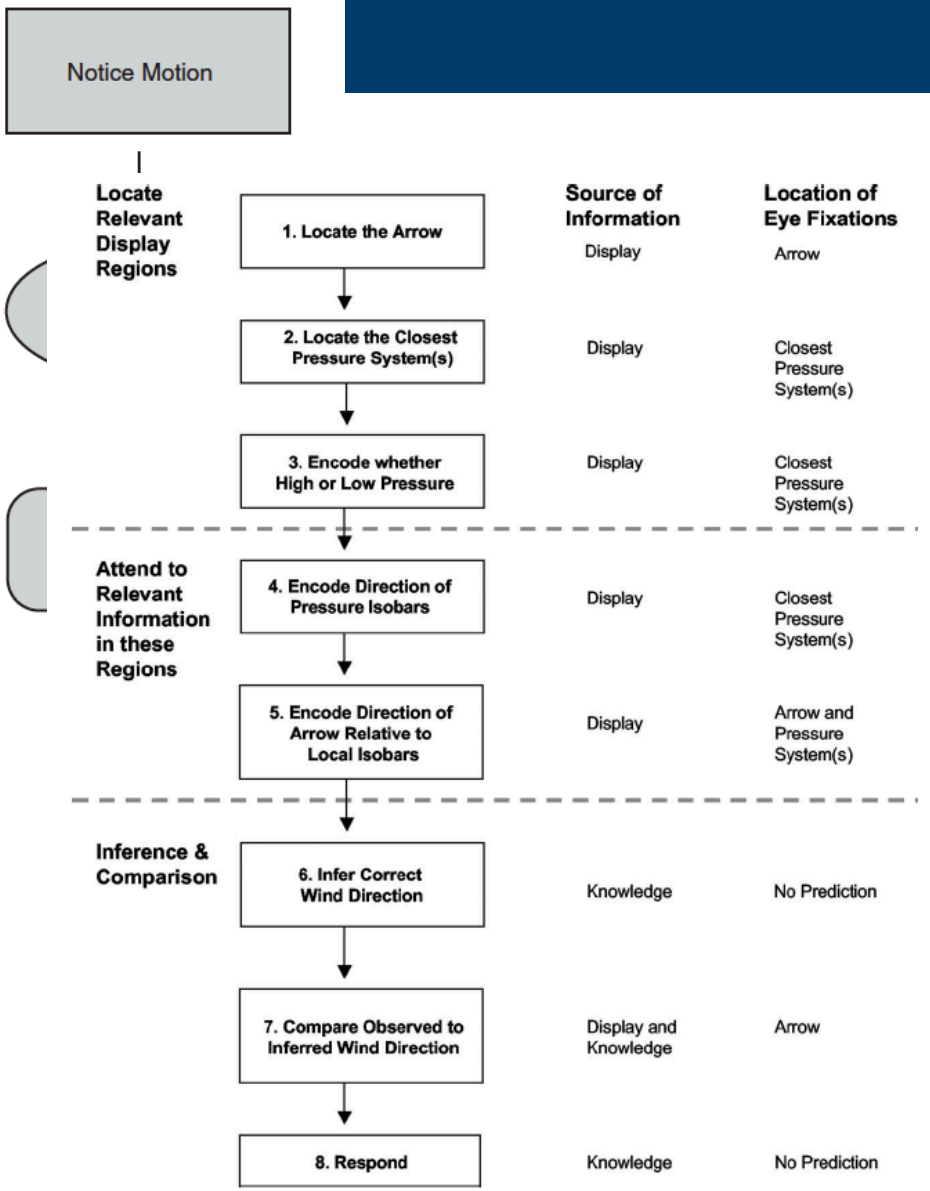
   Otherwise there may be problems of generalizability.

2. But…real world tasks are often complicated.

   So it can be hard to control extraneous variables, especially for complex tasks.

3. Think through the demands of the task to be sure that your experiment lets you say something about the use of your tool/map design/etc (*task analysis*).



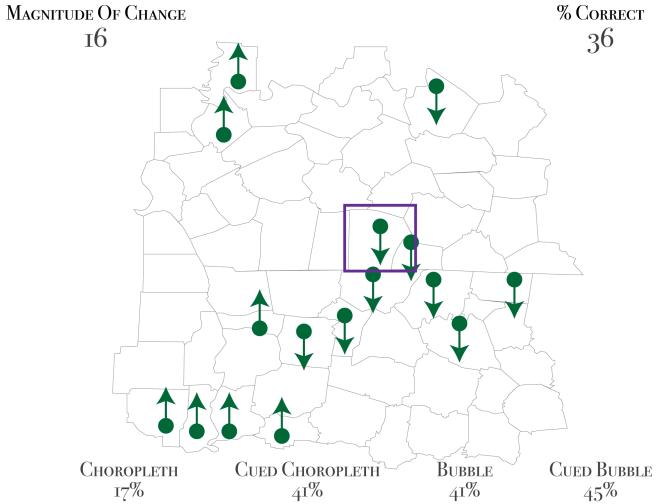


Hegarty et al 2010

Operationalizing Variables (Experiment I)

Sometimes variables can be operationalized in different ways:

IV = Accuracy… could be % correct, or could be % false alarms



How you do this may lead to different conclusions from your experiment, so careful thought here is important.

Image source: Kirk Goldsberry

Control Variables and Confounding Variables

Control variables:

1. It's impossible to control *all* extraneous variables.
2. We don't want to anyway – effects on *external validity*** (generalizability).

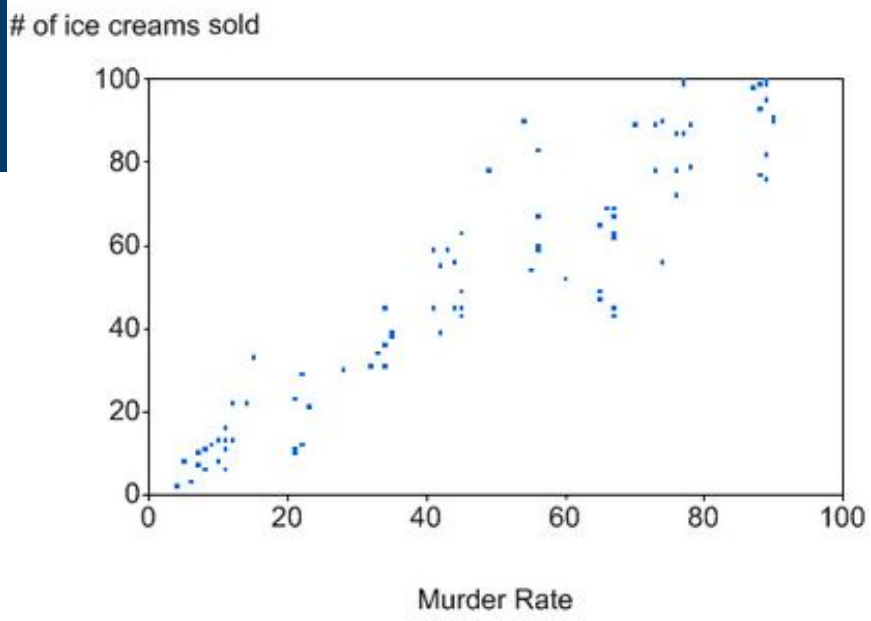
Confounding variables:

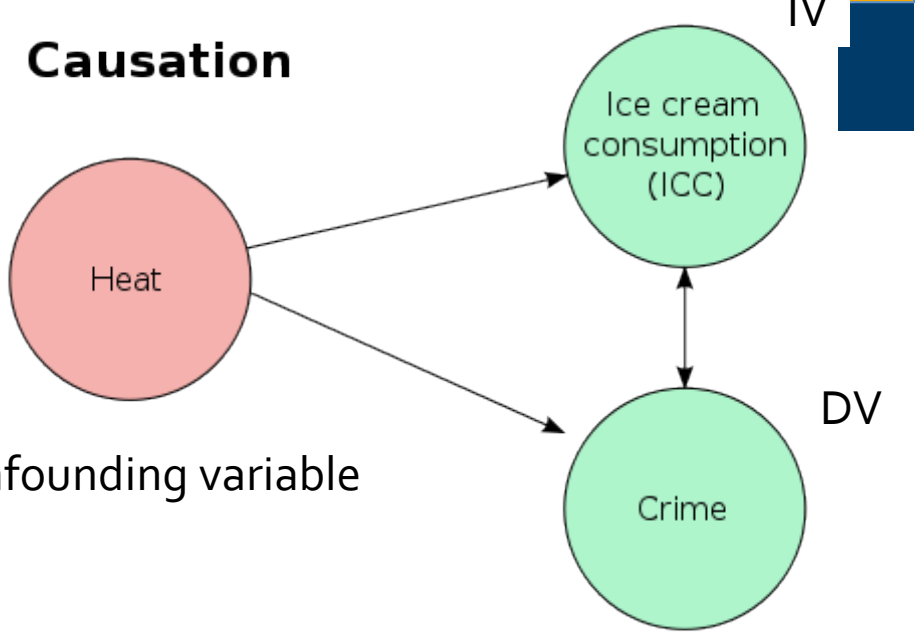
1. Confounders = variables that vary systematically with an independent variable.
2. We really want to avoid these – effects on *internal validity***.
3. Example: labels as confounders.

** More about these in Amy Lobben's module.

Image source: Wikipedia

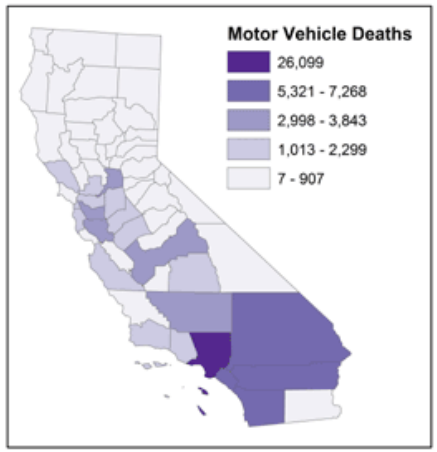
Woolfook, M., Castellan, W., and Brooks, C.I. (1983). Pepsi versus Coke: Labels, not tastes prevail. Psychological Reports, 52(1): 185-186.

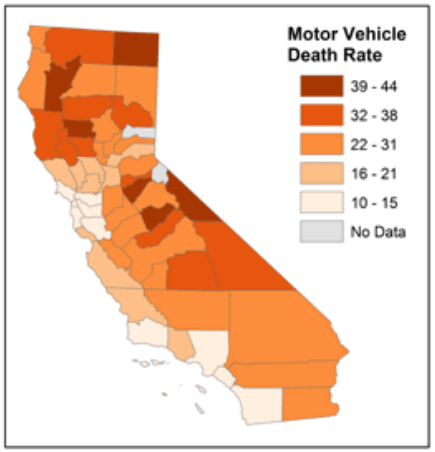




Confounding variable

Image source: WIkimedia





Choropleth maps of absolute counts can confound area with the mapped phenomenon.

Potential problems… reliability and validity (much more in Module 3)

Reliability (consistency):

- If the true value = 2, each time you measure with a method, do you measure a value of 2? (test-retest)
- Reliability is a necessary but not sufficient condition for validity

Validity (accuracy):

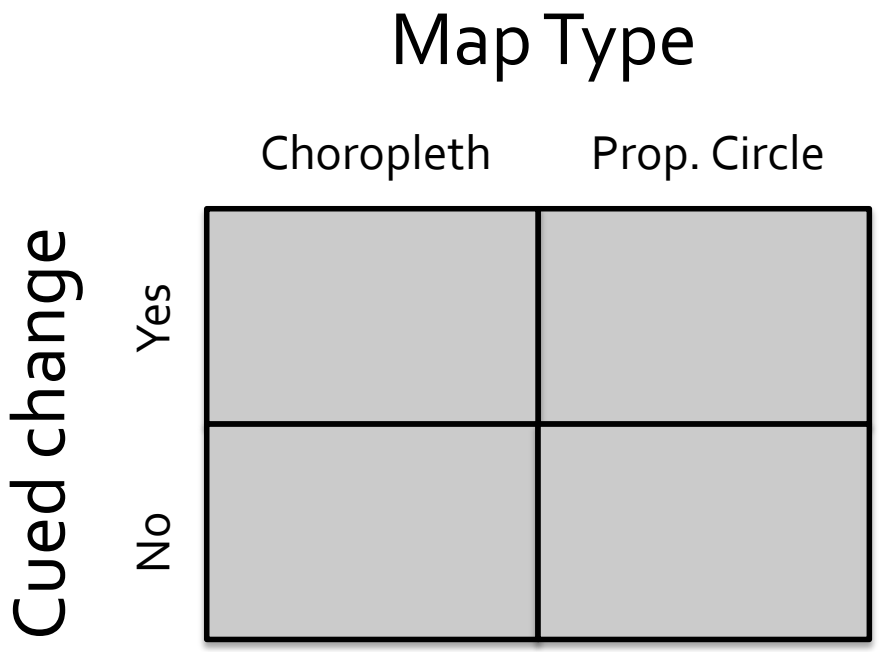
- Construct validity:
  - Did I adequately measure the variable I am looking at?
    - If the true value = 2, did I come up with a measure of value = B?
- Internal validity:
  - Can I infer that x caused y? (does faster pace = fewer correct answers?)
- External validity:
  - Is what you are measuring generalizable to groups beyond the people you are measuring?

Factorial Design

A *factor* = an independent variable

A *level* = an subdivision of a factor

Names refer to both the number of factors and the number of levels.



Experiment I was a 2 x 2 design.

Basic Experimental Design Types

Between subjects

- Two independent variables – 'treatments'
  - e.g. two different types of maps (static vs animated)
- Participant randomly assigned to one condition or another
- Compares separate groups of individuals

Within subjects

- Two independent variables – 'treatments'
- Participant receives both treatments, in sequence
- Compares treatments within one group of individuals

Between Subjects

- Advantages:
  - Participants are not affected by practice on task
  - Not affected by fatigue or boredom
- Disadvantages:
  - Can be differences in other (uncontrolled) variables in the groups (individual differences)
  - You need a large group of participants / more time
  - Differences in environmental variables if groups are tested in different areas

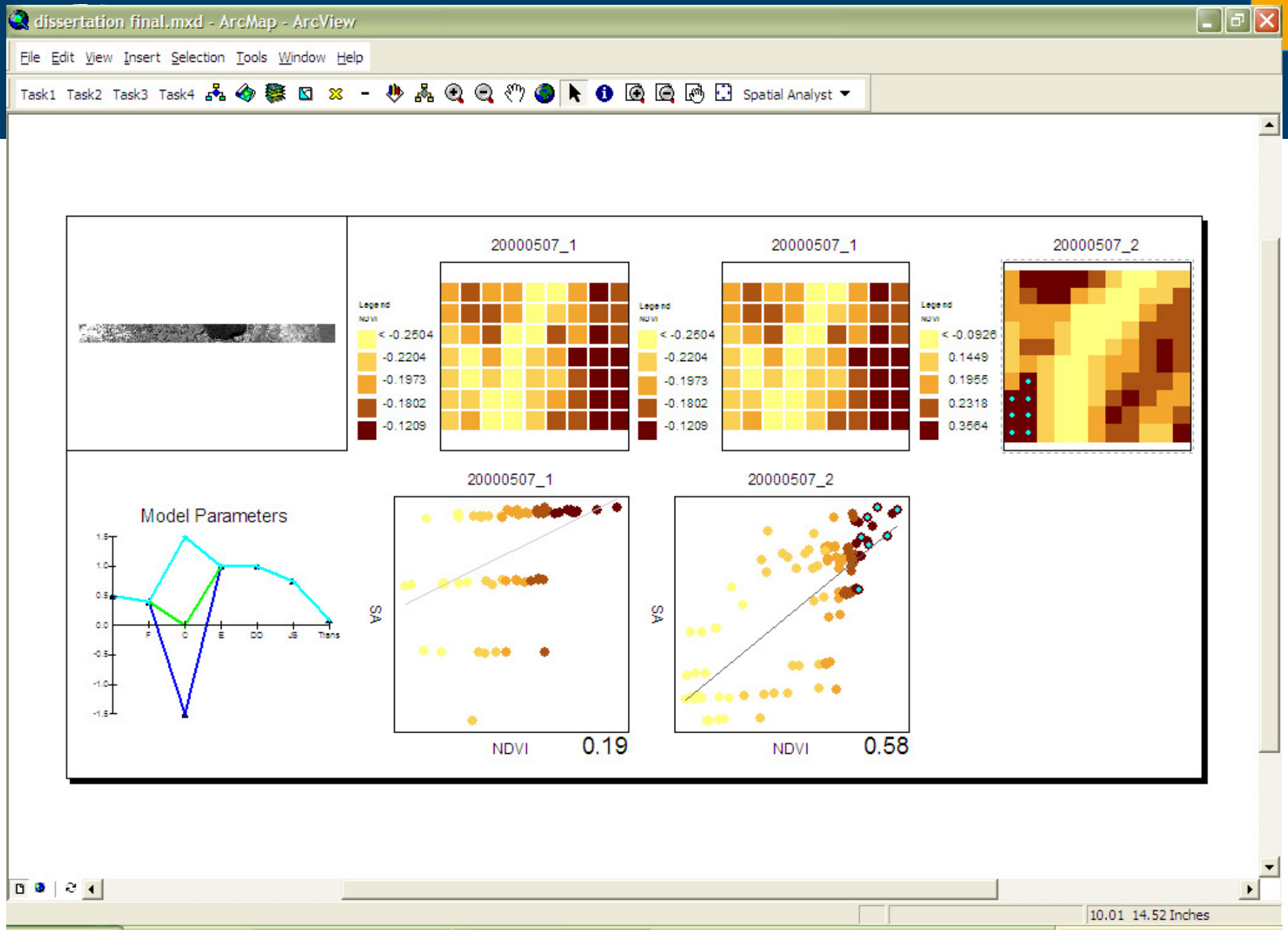
Within Subjects

- Advantages:
  - Controls for individual differences between participants.
  - Reduced environmental effects.
- Disadvantages:
  - Carryover Effects (Practice / Learning):
    - Participants may improve simply through the effect of practice on providing scores.
    - Longer experimental sessions: participants may become tired or bored and their performance may deteriorate.

Experiment II Characteristics

Naturalistic Observation (Quasi-Experimental)

- Participants given real tools
- Participants given a realistic task
  - Understanding the dynamics of hantavirus in a mouse population and its potential effects on human health.
  - Generate hypotheses to explain the patterns seen in the data.


dissertation final.mxd - ArcMap - ArcView
File Edit View Insert Selection Tools Window Help
Task1 Task2 Task3 Task4
Spatial Analyst
20000507_1
20000507_1
20000507_2
Legend
NDVI
< -0.2504
-0.2204
-0.1973
-0.1802
-0.1209
< -0.0926
0.1449
0.1955
0.2318
0.3564
Model Parameters
20000507_1
20000507_2
SA
NDVI
0.19
0.58
10.01 14.52 Inches

Research questions:

1. What do participants do with the system and how do they accomplish their goals?
2. What kinds of information do users attend to in the visual information display?
3. How is the information from the system used (cognitively)?
4. (If the segment is a hypothesis) What kinds of hypotheses are generated and do they differ throughout the process of model exploration?

The problem

Users needed to be introduced to tools

But…

- If they all saw the map first, and then used the map more than other tools…what does that mean?
- That the map is more useful than other tools?
- That the map is the one they remembered best?

The solution: counterbalancing

- Each group of expertise sees each order of tools.
- 18 participants, 3 groups, 3 tools = 3 blocks of 6.

Subject	1	2	3
1(E)	Map	Scatter	Time
2 (E)	Map	Time	Scatter
3 (E)	Scatter	Map	Time
4 (E)	Scatter	Time	Map
5 (E)	Time	Map	Scatter
6 (E)	Time	Scatter	Map
7(G)	Map	Scatter	Time
…			

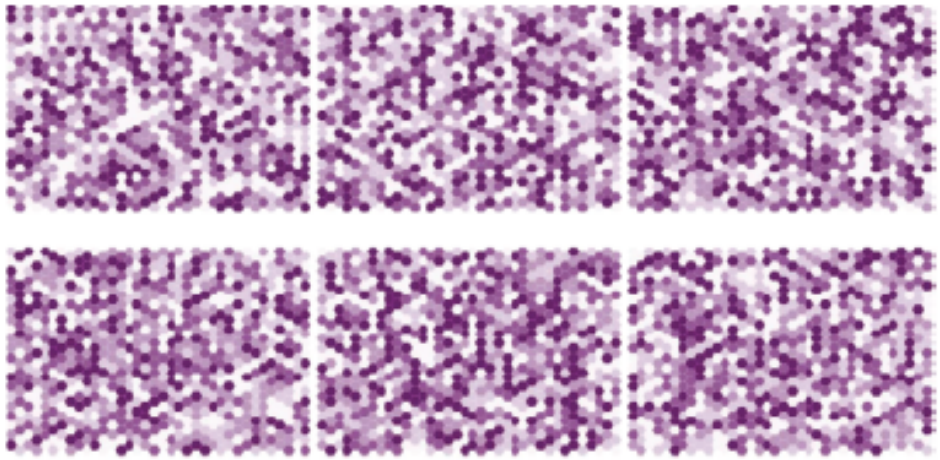
A more complex problem…

Experiment III (factorial design): 2 x 4 x 3 experiment

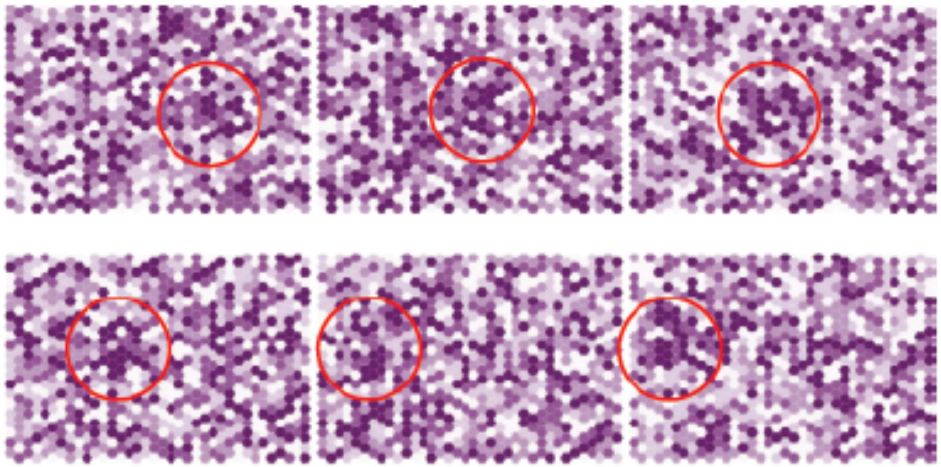
Compares effectiveness of two conditions (animated / static) with a number of other independent variables (IV):

- gender (2), (between-subjects)
- pace (4), (within-subjects)
- coherence (3) (within-subjects)

Potential for learning effects on the results of dependent variables (DV) - % correct and time - if not balanced properly!!

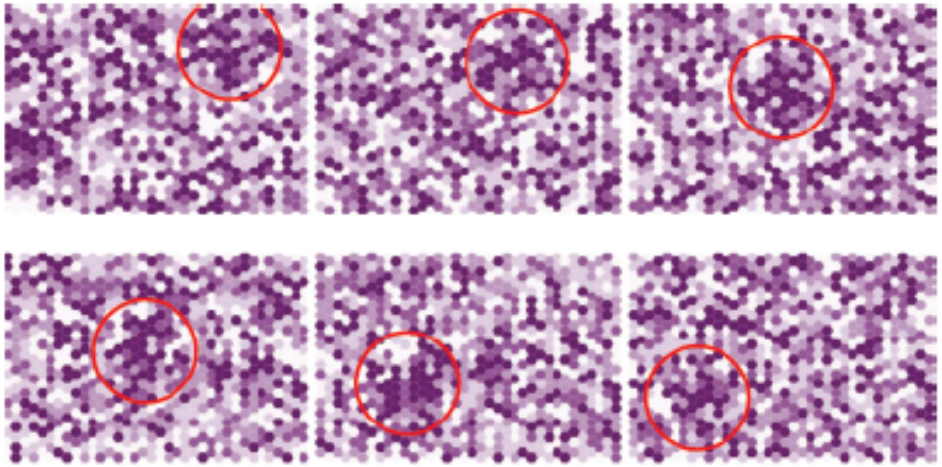


no cluster



subtle cluster

strong cluster



Complete counterbalancing

Study II needed:

- 3 (groups) X 3! (number of conditions)
- So…3X3X2 = 18 subjects for counterbalancing

For this study…

- 2 (conditions) X 2! Gender X 4! Paces X 3! Coherences =
- 2x2x4x3x2x3x2 = 576 participants for full counterbalancing!!!

Clearly another solution is needed…

Balanced Latin Squares method

Each stimulus level is preceded and followed equally often by each other stimulus:

With six conditions:

2 sets of 12 stimuli

- 4 paces x 3 coherence

A	B	F	C	E	D
B	C	A	D	F	E
C	D	B	E	A	F
D	E	C	F	B	A
E	F	D	A	C	B
F	A	E	B	D	C

Ways we screwed it up…

1. Did not program the Latin Squares matrix properly
2. Did not balance the two groups of stimuli and conditions properly
3. Did not balance gender properly (this only needed to rerun a few people).

Message for complicated designs…

- Check and recheck and recheck before starting.
- Don't trust that everyone in the group did their part perfectly. Members of the group check each other's work.
- Create tables that help you check off who comes when (e.g., gender), etc.

Other potential problems…
not enough power (How many participants do I need?)

Power of your experiment

- Can be estimated with programs such as GPower

What does power depend upon?

- Alpha level.
- Sample size.
- Effect size:
  - Association between DV and IV
  - Separation of means relative to error.

Decision	Reality: H_o is true	Reality: H_o is false
Reject H_o	Wrong Type I	Correct
Fail to reject H_o	Correct	Wrong Type II

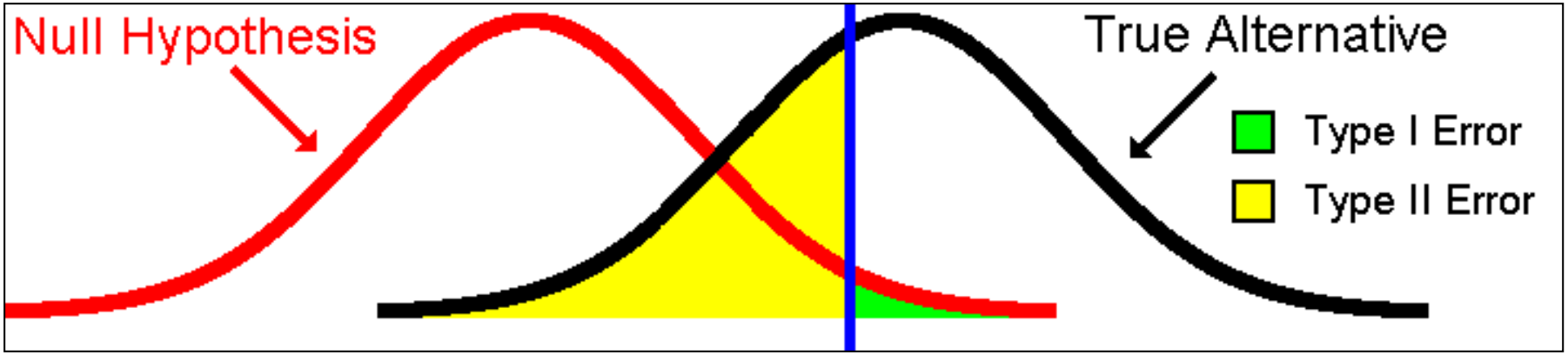


Image source: Bibby and Ferguson (2000)

Statistical Power

Incorrectly failing to reject the null hypothesis:

- a type II error
- There really is an effect but we did not find it

Statistical power: How well can I detect a real effect?

Power can be described using the following equation:

$$1 - \beta$$

where β is the probability of making a type II error

Power = the probability of not making a type II error

Power and alpha

- Relax alpha to increase power. (e.g. $p < 0.05$ instead of 0.01)
- This increases the chance of a Type I error.
- Not a great option.

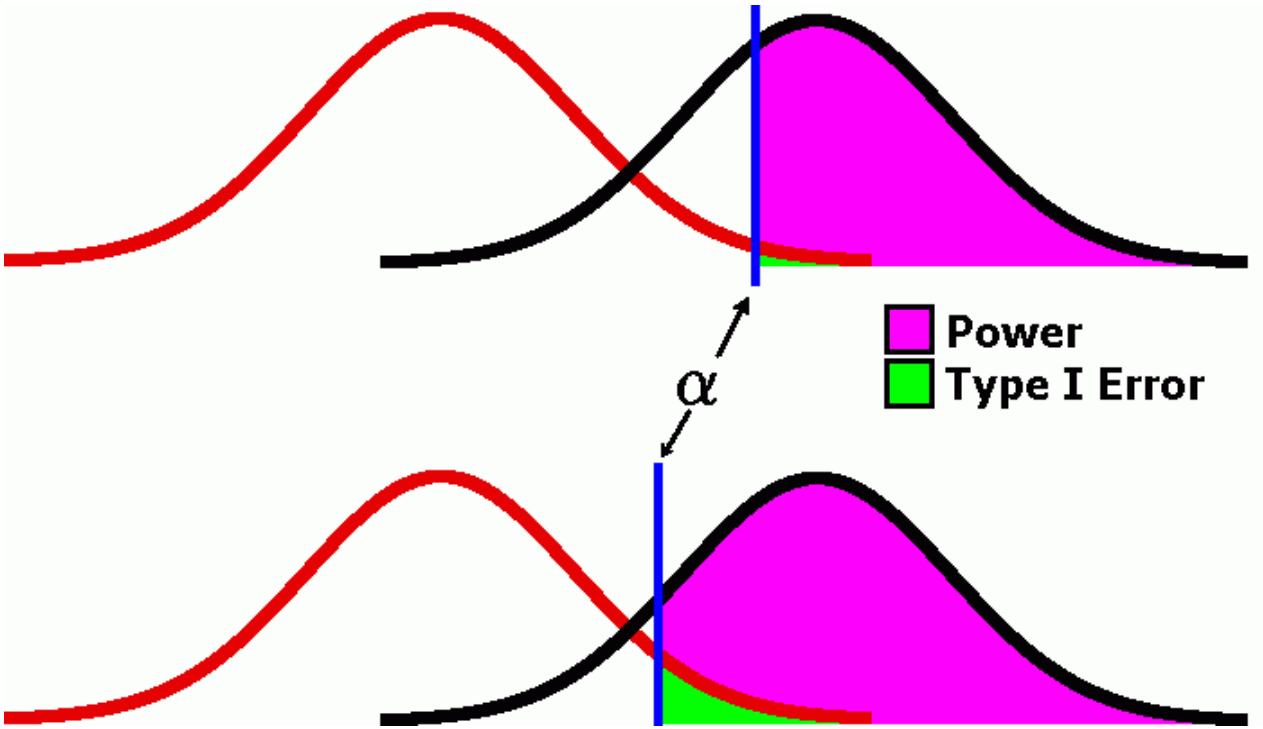


Image source: Bibby and Ferguson (2000)

Power and sample size

Low N = little power.

High N doesn't always produce an increase in power – saturation point

High N = high cost

Goal: minimize N while maximizing power

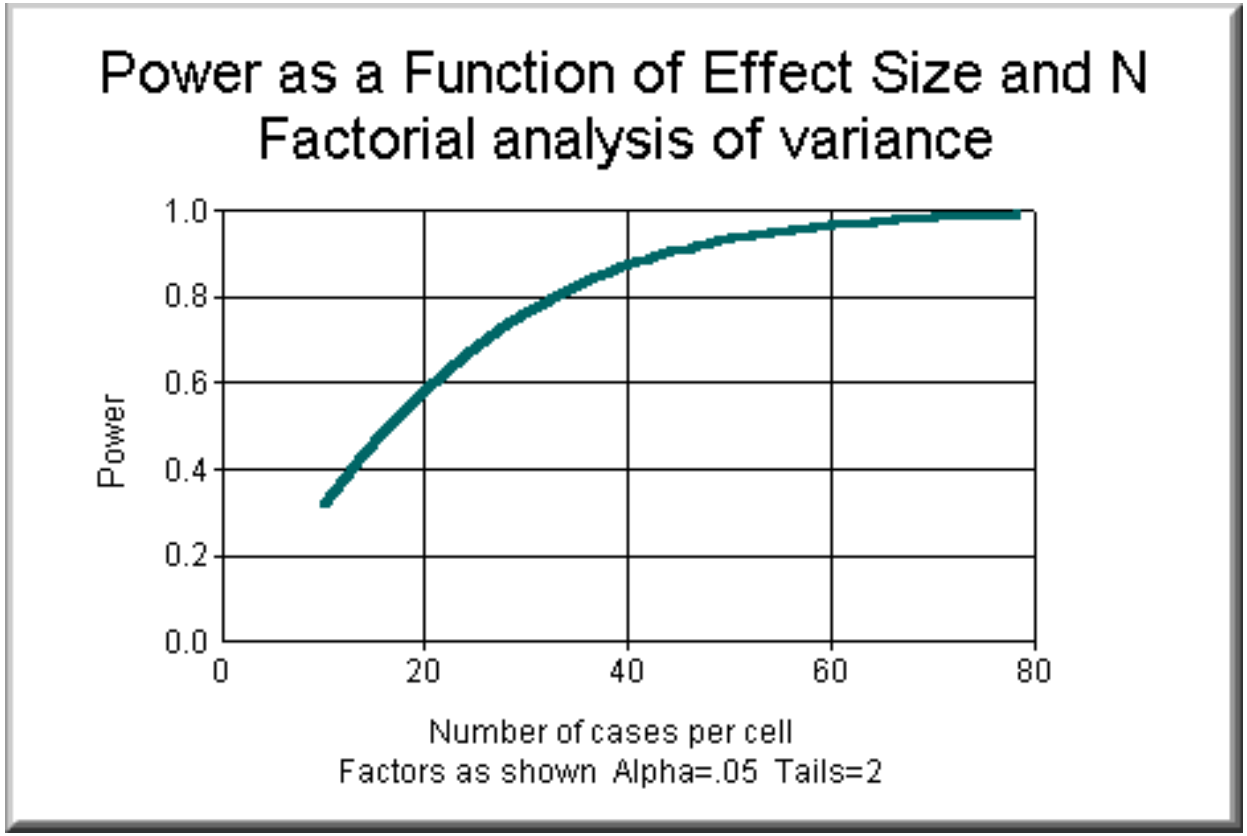


Image source: Bibby and Ferguson (2000)

Power and effect size

As the differences in distributions increases the power also increases

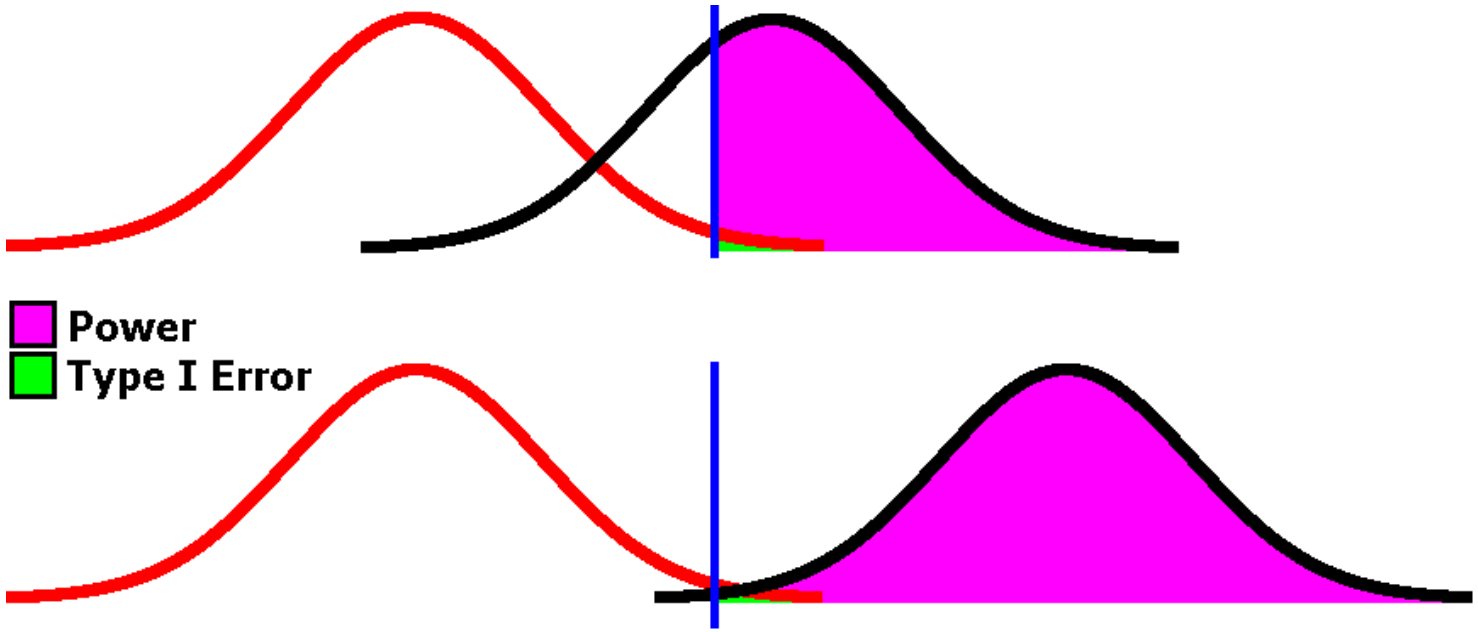


Image source: Bibby and Ferguson (2000)

Power and effect size

As the variance about a mean decreases power also increases

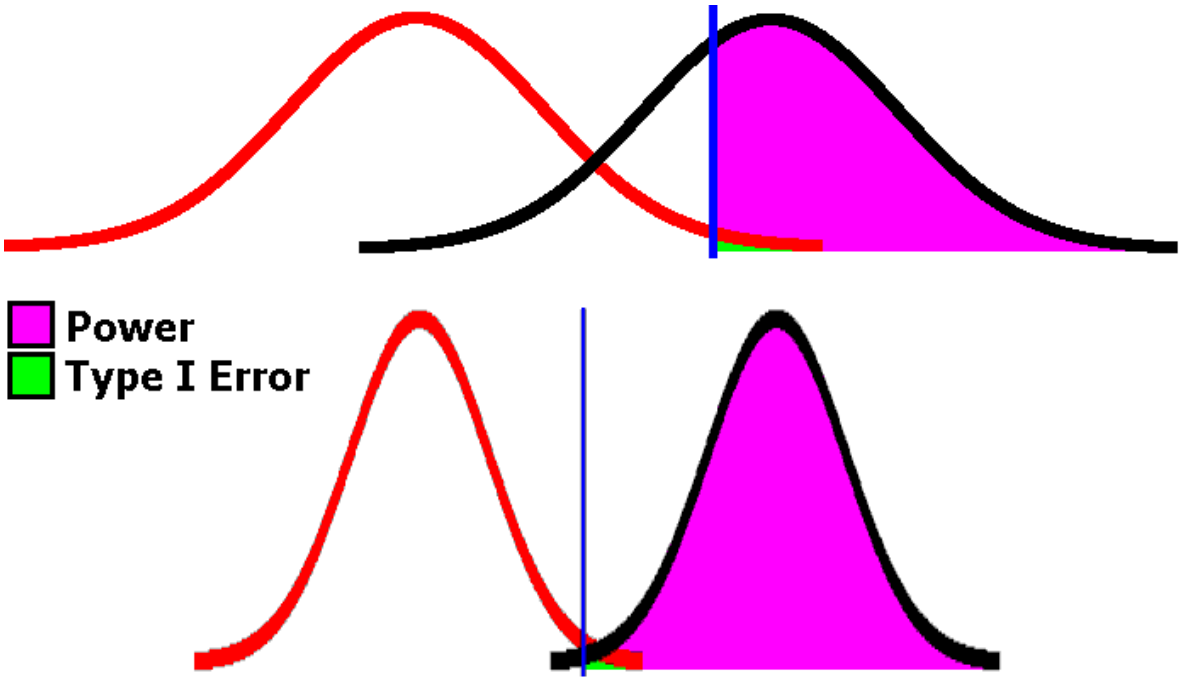


Image source: Bibby and Ferguson (2000)

Practicalities and Experiments

Does my experiment have ethical concerns?

Ethics Panels & Institutional Review Boards

Your panel/board may require training before submission.

Procedures vary by country and by institution, check your local regulations.

Typically our work can get *expedited* approval because it's low risk.

Budget sufficient *time* for this.

Pilot testing

How long will my experiment take?

Should I set *a priori* criteria for eliminating participants?

Is all of my equipment working as anticipated?

How do I analyze my data?

Between-subjects:

Two-way ANOVA

Main effect = effect of one variable holding others constant

Within-subjects:

Two-way repeated measures ANOVA

Interaction effect = effect of multiple variables together

Both within- and between subjects:

Mixed or Split-Plot ANOVA

What do I need to report when I write up my results?

- In addition to the normal sections of a paper:
  - Participants
    - How many, characteristics such as age, gender, experience, vision impairment, etc.
  - Experimental design
    - How many factors are being studied, what are the IVs, DVs, is it within- or between-subjects (or both)? It also includes how you *measured* your variables (e.g., questions asked, etc).

This paper is worth a look:

Olson, J.M. (2009). Issues in human subject testing in cartography and GIS. Proceedings of the 2009 International Cartographic Conference, http://icaci.org/files/documents/ICC_proceedings/ICC2009/html/nonref/17_11.pdf.

What do I need to report when I write up my results?

- Materials/Apparatus
  - **Materials**: should describe what stimuli (e.g., maps, tools, etc) were used (and usually present examples), how the stimuli were created, design decisions made during their creation & rationale for those decisions
  - **Apparatus**: if equipment such as an eye-tracker, colorimeter, iPhone pointing device, etc. is used to capture data, technical details describing this equipment should be provided
- Procedure
  - This description should contain enough detail about how stimuli were deployed/displayed, with what timings, in what order tasks were completed, etc, that someone else could replicate the procedure.

For a good example to follow see the following paper:

Hegarty, M., Canham, M.S., Fabrikant, S.I. (2010). Thinking About the Weather: How Display Salience and Knowledge Affect Performance in a Graphic Inference Task. Journal of Experimental Psychology, 36(1): 37-53.

What do I need to report when I write up my results?

Reporting statistical results:

Conventions:

statistic (degrees of freedom) = statistic value, $p <$ criterion

e.g., $t(18) = 4.7$, $p < .01$

e.g., $F(1, 58) = 26.73$, $p < .01$

Archive of Experimental Stimuli: A proposal

Within the next couple of years, the ICA Commission on Cognitive Visualization proposes to establish an archive of experimental stimuli from research that has been already published to do the following:

1. Help readers of papers more fully understand the design of experiments.
2. Provide researchers with examples of experimental materials that may help them in designing their own materials.
3. Stimulate collaborations and re-use of experimental materials, where appropriate.

We are still thinking through the logistics of this, but watch for updates on our website: https://www.geo.uzh.ch/microsite/icacogvis/mission.html

Summary

Experimental design can be quite complicated.

The more factors you are trying to test, the more ways you can mess things up.

Planning, testing, rechecking is very important.

There are tradeoffs in every experimental design

- The key is finding the right ones for your situation and then executing that design properly.

If you're not sure about decisions you need to make, consult the following sources:

- Experimental psychologists or statisticians at your institution
- Previously published work or textbooks on experimental design

Resources for Further Information

- Field, A., Hole, G. (2003). How to design and report experiments. Thousand Oaks: SAGE.
- Field, A. (2009). Discovering statistics using SPSS, 3rd edition. Thousand Oaks: SAGE.
- Martin, D.W. (2008). Doing psychology experiments, 7th Edition. Belmont, CA: Thomson Wadsworth.

- ICA CogVis website has links to several useful experimental research tools:

https://www.geo.uzh.ch/microsite/icacogvis/resources.html