CPSC 532E — Week 9: Lecture

Experimental Design - I

- The Experimental Approach
- Measurement
- Experimental Conditions
- Data Analysis

0. Why Experiment?

We often need to know if some visual design is or isn't effective

More generally, we need to know:

- what factors do or don't contribute to visual experience
- what methods produce an effective visual display

Often the tendency is to just ask a few friends - but this isn't the best approach…

Problems with simply trying a few things on friends:

1. Friends are often biased
   - will tell you what you want to hear

2. Friends are often experts
   - will not reflect the average used

3. Not generally enough friends for testing
   - need reasonable numbers to get power

4. Right kind of controls aren't generally used
   - need right conditions to validate conclusions

Need a rigorous experimental methodology

Psychophysics

- study of the physical factors that affect psychological experience

- historically, techniques provided basis for
  - brightness perception
  - colorimetry
  - contrast sensitivity
  - motion perception, etc.…

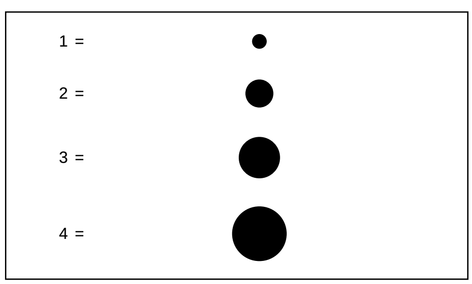
- need to adapt these to more complex displays

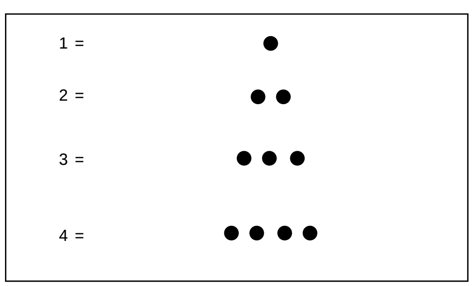
Examples of issues that can be addressed:

1. How similar do two displays look?
   (Which factors are relevant? Which are not?)

2. How effective is a visual attribute in conveying information?

E.g. - How might number be best conveyed?





Examples of issues that can be addressed:

1. How similar do two displays look?
   (Which factors are relevant? Which are not?)

2. How effective is a visual attribute in conveying information?

3. What information is required to carry out a task?

4. What information is needed to create a sense of presence in a simulator?

The answers to such questions can determine, e.g.

- if a task (or design) is feasible
- if a visual display is doing what you think it should
- if a task requires more (or less) resources ***(time, memory, display resolution, etc.)***
- which of two designs is better

1. The Experimental Approach

The goal of an experiment is to

It must discriminate between possibilities
(competing hypotheses)

- at the level of
  (e.g. the need for attention),
- at the level of
  (e.g., the influence of realistic shading)

At the very least, it should determine whether something does or doesn't make a difference.

Approach: Test whether the predictions are true.

1. Assume a hypothesis
   - working hypothesis
   - needed to carry out the experiment
2. Test whether the results contradict the predictions of the hypothesis

If the results go against the predictions,
the working hypothesis is <u>disproven</u>,
and something else must take its place.

If the results don't contradict the predictions,
then the hypothesis is <u>not disproven</u>
-> does not mean it's true (e.g. telepathy)

2. Development of an Experiment

0. Initial hypothesis
1. Observations
2. Experimental conditions
3. Analysis of data
4. Conclusions

2.0. Initial hypothesis

Simply looking for correlations among arbitrary measurements won't provide useful information.

(Anything could be tried, but probably wouldn't mean much).

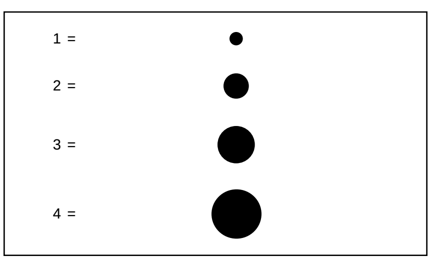
Every experiment is always done against a background set of hypotheses, or framework.

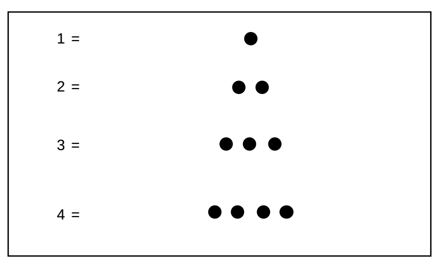
- don't need to assume that everything is true,
  but you do need to assume that most of it is.
  Otherwise, you can't proceed

From this initial framework

- pick out an initial hypothesis to be tested
- design a way to test it

E.g. - How might number be best conveyed?





Initial hypothesis: There is no difference.

(When no difference is expected, this is often called the **null hypothesis**.)

2.1. Observations

- the heart of an experiment is the measurement of some aspect of behaviour
- naive belief: truth can be picked up directly, and "facts" are free of theory.
  (-> no problems with measurements)
- the art of experimenting is then simply observing the correlations

Not so.

Can never observe "things as they really are".

- always need to select some aspect of behaviour
- always need to interpret what the results mean
- thus, every data point is viewed through a theory (hypothesis)

Note: This does not mean that the data are arbitrary. But:

- they may not capture everything
- there is always some uncertainty about what is signal, and what is noise…

Measurements

Ultimately, testing depends on **data**. These are some aspect of the observer's behaviour, e.g.

- response **time**
- response **accuracy**
- subjective impressions
- similarity judgements
- galvanic skin response
- etc., etc

The most reliable indexes of behavior are objective measures. The most common are:

- response **time**
- response **accuracy**

Note that objective ≠ quantitative

- similarity ratings are quantitative, but are not objective

Time vs. Accuracy

E.g. visual search:

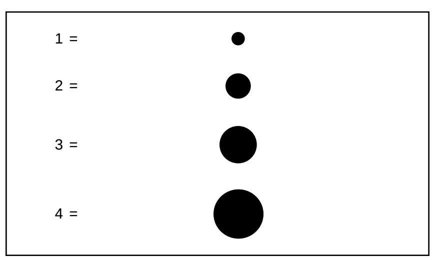
- response time to find target (high accuracy)
- accuracy to detect target (brief display)

Both measures can be used; one is usually dominant

Choice of dominant one depends on the situation.

- accuracy for briefly-presented display is more natural for time-critical situations
- response time to extended display is more natural for safety-critical situations

E.g. - Representation of number



Task: Respond with keypress to displayed item

Dependent variable (aspect of response):

(Note: Could also be accuracy, aesthetic judgement, etc…)

2.2. Experimental Conditions

Observations (along with an initial hypothesis) are the beginnings of an experiment, but…

"It is impossible to discover anything in physics or physiology without envisioning an original experiment, without subjecting the phenomenon of interest to more or less new conditions."

Ramon y Cajal

Experiments always find out how

an observer's response

depends on

a stimulus parameter

Different hypotheses -> different behaviours

Can't do this with just one set of conditions

- need different conditions
- **compare** measurements made in each condition

Independent variable

- one aspect of stimulus (parameter)
  - e.g. size, shape, etc
- based on physical structure

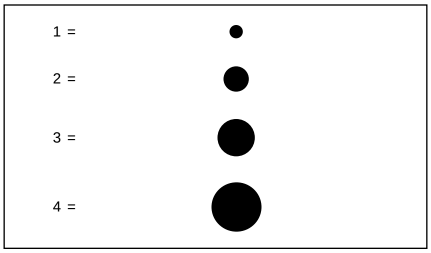
Dependent variable

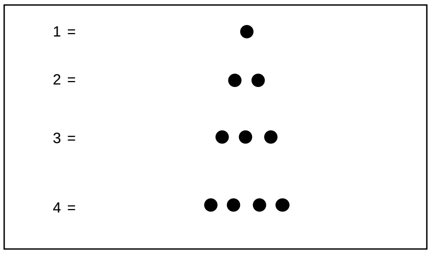
- one aspect of response (measurement)
  - e.g. speed, accuracy
- based on perceptual impression

A key part of an experiment is to determine how the dependent variable varies with the independent one.

These variables should always be kept separate

E.g. - Representation of number





Task: Respond with keypress to displayed item

Independent variable (aspect of stimulus):

Dependent variable (aspect of response):

Experiment: How does response time depend on representation?

One of the most important conditions is where the effect of a stimulus variable is <u>absent</u>

PSYCHIATRIST: Why do you flail your arms around like that?
PATIENT: To keep the wild elephants away.
PSYCHIATRIST: But there aren't any wild elephants here.
PATIENT: That's right. Effective, isn't it?

If the variable is absent, this is called the **<u>control condition</u>**

The "result" is the difference between measurements made on the condition of interest and the control condition

More generally,
experiments are about <u>comparisons</u>

Part of the art of experiment design is to <u>make sure that the comparisons are valid</u>

- e.g. cancer is a consequence of industrialization
- compare death rates in industrialized vs nonindustrialized countries

Is this a valid comparison?

- compare populations that are <u>matched in age</u>

Make sure the right kind of comparison is made

-> need to make sure that the variable you're considering is the only variable that can actually affect the result.

Selection of Conditions

For greatest sensitivity, one condition should have stimuli (displays) that generate balanced responses.

- any disturbance of this balance in a second condition will likely be picked up.

Find best range of conditions via pilot experiments

- look for range where effect is strongest
- try to develop a "feel" (or "affinity") for the phenomenon.

2.3. Analysis of Data

Analysis of data is of two main types:

- determining the existence of an effect
  (does a difference really exist?)
- determining the characteristics of an effect
  (how does the dependent variable
  depend on the independent variable)

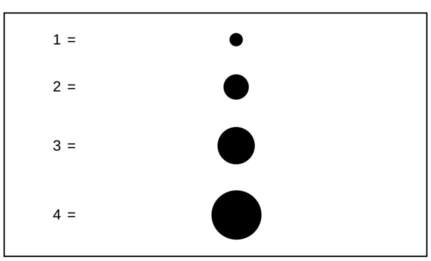
Rigorous analysis of the data strongly depends on the design of the experiment

-> **The design of the experiment should take into account the kind of analysis to be done**

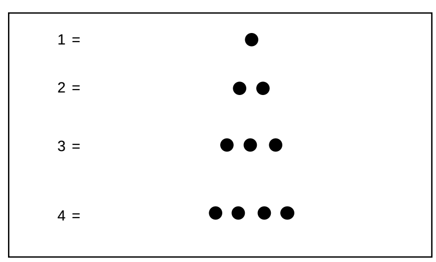
Consider determining if a difference exists

- good way to determine which design is better

E.g. - Representation of number

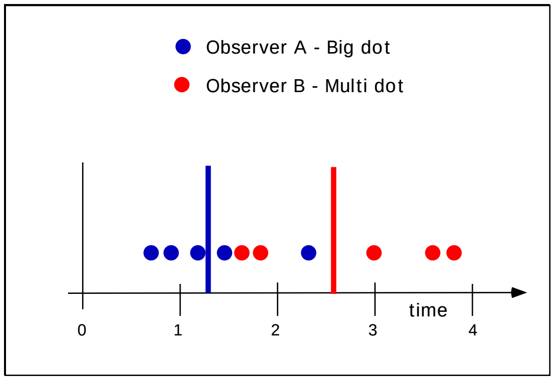


"Big dot" representation



"Multi dot" representation

Is there a difference in the average time to respond?



Can't compare performance on one test alone
- because of randomness, need many measurements from each observer

Obtain an average estimate from each observer
- reliability of difference given by

Interpreting a t-test

The t-test gives the probability that the difference in the means could have arisen by chance.

By convention, $p < .05$ is taken as significant (ie. the null hypothesis is disproven)

Is this the best way to test these designs?

What if the difference were due to the observers?

One solution: Use several different observers,
use the average of all of them
(Between-subject design)

This does work, but requires a lot of observers.
- typically, at least a dozen

Better solution: Use the same observers, but have them each do the same conditions
(Within-subject design)

Observer	Big dot	Multi dot	Diff
A	1.1	1.4	0.3
B	3.1	3.4	0.3

Average difference = 0.3 seconds

Within-subject design is very sensitive, since it can cope with systematic differences between observers

Note: When using multiple observers, it's very important to make sure that they don't run conditions in the same order

E.g., if always run in Big-dot condition first, there could be a practice effect. Performance on subsequent Multi-dot condition would be faster than normal

Solution:

- make sure that all observers are equally likely to run each condition first, second, etc.

When testing for differences, sometimes the measurement is one of **frequency** rather than **quantity**

In this case, the appropriate test is a χ^2**-test**

E.g., compare two monitoring systems in terms of failure rates:

	Failures	Successes
System A	4	43
System B	8	60

Hypothesis:

There is no difference (null hypothesis)

χ^2-test: $p > .85$

The null hypothesis is **not disproven**

(Note: There could be a difference.
However, this test may not have
been sensitive enough to pick it up.)

2.4. Conclusions

Ultimately, end up with a hypothesis that explains the data, and (as much as possible) makes sense
-> final hypothesis

Note 1: An experiment can never prove a hypothesis.
Can only disprove it.

Note 2: Not necessary for a hypothesis to be correct.
The main thing is that something is learned.

"The scientist must never forget that hypotheses must be considered a means, never an end."

—Ramon y Cajal (quoting Huxley)