Understanding and Designing Perceptual Experiments

D.W. Cunningham [†1]

C. Wallraven [‡2]
[1]BTU Cottbus, Germany [2]Korea University, Korea

Abstract

Humans and computer both have limited resources with which they must process the massive amount of information present in the natural world. For over 150 years, physiologists and psychologists have been performing experiments to elucidate what information humans and animals can detect as well as how they extract, represent and process that information. Recently, there has been an increasing trend of computer scientists performing similar experiments, although often with quite different goals. This tutorial will provide a basic background on the design and execution of perceptual experiments for the practicing computer scientist.

1. Tutorial Details

This proposal is for a half-day tutorial (2x90 minutes) on basic experimental design. The focus is on understanding the fundamentals of designing and executing effective experiments as well as helping to understand and judge existing experiments. The tutorial is based on our recent book [CW11].

2. Outline

A century and a half of experience has taught psychologists that designing an experiment whose results can not only uniquely be interpreted, but also will allow us to answer our research question is considerably more difficult than it appears to be at first glance. Just as there are numerous unexpected difficulties that can render experimental data invalid, there are also many helpful short-cuts and tricks that can greatly improve various aspects of an experiment.

The first step in understanding or designing an experiment is knowing what is being investigated. Within computer science, we often will want to evaluate the success of an algorithm, or compare its performance to another similar technique. We might also want to know which range of parameters is optimal. Some studies ask even broader questions, such as examining what sort of information people use for certain tasks (and therefore which information will need to be in a new algorithm). Whether we are investigating something general or very specific, we are more often than not looking for a clear answer. To put it another way, we have a research question and want a definitive, empirical answer. The more clearly specified a research question is, the more obvious it is how one can best answer it. The first secret of experimental design, then, is to devise clearly and precisely formulated research questions.

We can not directly observe perceptual processes. At best, we can try to infer what people see from what they do. Determining what happens inside such a black box by observing its inputs and outputs is formally similar to sampling an unknown function. A closer analysis of sampling with an eye towards its perceptual equivalents yields a mathematical description of experimental design. Surprisingly, the resulting equations are very similar to the Analysis of Variance (ANOVA), which is one of the most common data analysis methods in psychology.

One of most difficult aspects of actually designing an experiment is to decide precisely what the participants should do. Once the participants are sitting in the experimental chamber staring at the stimuli, they have to actually perform some task, but which one? The possibilities are nearly endless. On the one hand, this variety is a very good thing, since it means that there is almost certainly a task that is perfect for any research question. It is also unfortunate, since it makes it very, very difficult even for experienced experimenters to decide which task is the best one for the current experiment.

In addition to explicitly and precisely formulating the

[†] douglas.cunningham@tu.cottbus.de
[‡] wallraven@korea.ac.kr

research question, choosing a task requires knowing what kinds of questions different tasks can answer. One can place all experimental tasks along a continuum (see Figure 1). At the "general" end of the task-continuum are meta-tasks, where the participants are essentially asked how they believe they would act in a given situation. Meta tasks include free description and some forms of rating and forced-choice tasks. These tasks can answer broad, vague questions but are very difficult to interpret uniquely. At the "specific" end of the task continuum are tasks that easily support unique interpretations, but focus on very specific questions (and thus provide very specific answers). The most specific form of task are "physiological tasks", which measure the body's reactions such as heart rate, body temperature, neural firings, etc. These are very useful since they can provide a very direct, unbiased view of what elements of the stimulus the participants really saw or how they really felt about a stimulus. Physiological tasks are, however, exceedingly difficult to use because most research questions involve real-world behavior or subjective experiences, and making solid, definitive connections between physiology and real-world behavior or subjective experiences is an unfinished task, to say the least.

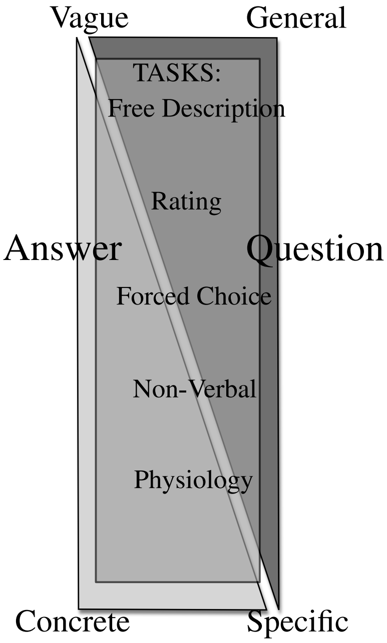


Figure 1: *Experimental tasks can be seen as lying along a continuum. Tasks at one end are very flexible and are good for answering open-ended questions. These questions can provide a substantial amount of information and insight but are very difficult to interpret uniquely. Tasks at the other end of the continuum are much more constrained, and provide very focused answers that are generally unique and easy to interpret.*

The most common form of task is the direct task. Here, participants are required to actually perform some specific act. If we wanted to see which rendering style is better at conveying the meaning of an expression, we could ask the participants to identify a series of expressions, and the style that gave the best scores would be—by definition—the best at conveying the intended meaning. Direct tasks are very useful since they provide direct evidence of how people will respond in certain situations. Unfortunately, direct tasks are difficult to use for precisely the same reason: in order to be useful in answering a real-world question, the situation surrounding a direct task must be as close to the real world as possible. This is never easy, and is sometimes not physically or ethically possible. Direct tasks include some forms of rating and forced-choice tasks, specialized forced-choice tasks, and what we term "Real World" tasks.

A final consideration in designing an experiment is data analysis. Once the data are collected, some method is needed to summarize the data (e.g., looking at mean performance for each condition) and look for trends within the data (e.g., which condition was better). Likewise, methods to determine if a trend is reliable (i.e., would it happen again if I were to run the experiment again) require some form of significance testing. Thinking about which statistics can in principle be used for a given type of task before one collects data can save a lot of time and prevent many problems problems.

3. Relevance

Researchers in computer science are increasingly being required to conduct perceptual experiments. Most of the human experiments performed in computer science are *not* user studies. Instead, they utilize the considerable amount flexibility that perceptual experiments have in terms of choosing what to show, how and when to show it, whom to show it to, and what the users should do. Indeed, many of the perceptual computer science experiments now use a strict psychophysical design, where one tries to obtain a mathematical description of the functional relationship between variations in the physical world and the resulting variations in the psychological (or perceptual) world.

The fact that experimental design can be quite involved is well reflected in the fact that obtaining a PhD in experimental or cognitive psychology requires the successful completion of many years worth of classes—both theoretical and practical—on experimental design. Without some formal, preliminary exposure to the fundamental concepts of experimental design, existing reference works—which usually focus on specific aspects of advanced techniques—are not very accessible to computer scientists. The use of complex, real-world images and situations—as is required in most of computer science—carries an additional set of pitfalls and even violates some of the central assumptions behind many of the traditional experimental designs.

This tutorial will cover the fundamental concepts in experimental design that are needed to understand currently existing study, to be able to read existing reference works on experimental design, and to begin to design new experiments.

4. Target Audience

The target audience for this tutorial includes most researchers in computer graphics and computer vision. More specifically, the tutorial should be useful for any researcher who is required to read and understand perceptual literature or to design and execute perceptual experiments. The tutorial is appropriate for beginner and intermediate levels, and requires only basic math and logic skills.

5. Topics and Syllabus

1. Introduction (10 Minutes)
   In this section, we will define what constitutes a single experiment and how this is different from experimentation in general. This definition revolves heavily around the concept of a well-defined research question. The relationship between research questions and formal hypotheses will also be addressed.
   - What is an experiment?
   - What types of experiment are there?
   - Why would a computer scientist want to do an experiment?
   - What are research questions and hypotheses?
2. The foundations of experimental design (30 Minutes)
   In this part, we will define and explain the typical components of a single experiment. We will also address the often criticized issue of why perceptual experiments take so long to arrive at an answer to a seemingly simple question. A critical part of this understanding the central role of systematic sampling and the balance between specificity and generality.
   - What are the elements of an experiment?
   - An analogy to sampling an unknown function
   - Balancing specificity and generality
3. A mathematical model of experimental design (40 Minutes)
   In this section, we will derive a formal, mathematical description of experimental design. The core idea starts with sampling an unknown function, and extends to include sampling and processing error. Once the basic formulation is in place, it is extended to include more advanced concepts like confounds, repeated measures, trial order, randomization, and factorial combination.
   - The importance of errors.
   - Repeated Measures
   - Two or more conditions
   - Confounds and trial order
4. The Task Taxonomy (20 Minutes)
   The task given to the participants is perhaps the central element in an experiment, rivaled only by the stimuli. In this section, we look at a few issues that all tasks must address. These include making sure participants are doing what we think they are, the use of deception and other ethical issues, and the use of practice. Other general issues such as participant selection and design of the experimental chamber will be discussed.
   - Different types of tasks.
   - Response Bias
   - Deception
   - Ethics
   - Practice trials
   - the experimental chamber
5. Qualitative Tasks (10 Minutes)
   Qualitative tasks are the easiest to implement and the hardest to interpret. The rely heavily on the participants' verbal skills, and are always a meta-task.
   - What kind of questions do they answer?
   - What are the advantages?
   - What are the limitations?
   - Guidelines
   - Specific variants (Free Description, Interviews, Questionnaires, Partial Report).
6. Rating Tasks (10 Minutes)
   Rating tasks are increasingly common, and have a wide variety of applications. There are also a few specific variants that when executed and analyzed properly, can yield surprisingly detailed, quantitative insights into perceptual structures.
   - What kind of questions do they answer?
   - What are the advantages?
   - What are the limitations?
   - Guidelines
   - Specific variants (Ordered ranking, Likert scales, semantic differentials)
7. Forced Choice Tasks (20 Minutes)
   Forced choice tasks are the main task used by traditional psychophysics. The participant is given a number of alternative and must chose one of them. There are a number of variants, each with its own advantages and disadvantages. Forced choice tasks provide a nice trade-off between specificity and generality.
   - What kind of questions do they answer?
   - What are the advantages?
   - What are the limitations?
   - Guidelines
   - Specific variants (two alternative/interval forced choice, Go/no go, Matching-to-sample, Visual Search).
8. Basic Data Analysis (15 minutes). Statistical analysis is a

core part of any experiment. It is strongly recommended that one things thoroughly about how the data can and will be analyzed before one gathers them. In this section, we will talk about a few of the more common forms of analysis and their assumptions that must be met before one can use them

9. General Guidelines and Rules of Thumb (10 Minutes)
10. Discussion (15 Minutes)

6. Related Courses

In 2005 at the IEEE Virtual Reality conference, Douglas Cunningham along with Katerina Mania, Heinrich H. BÃijlthoff, Bernard Adelstein, Nick Mourkoussis and J. Edward Swan II offered a full-day course that looked at "Human-centered fidelity metrics for virtual environment simulations". That course focused explicitly on research results and problems specific to VR. Since the, Christian Wallraven and I have formally modeled experimental design and written a book on general design issues. This proposal is the first tutorial on from Cunningham and Wallraven on general experimental design issues, and is based extensively on their book.

7. Author Biographies

Douglas W. Cunningham Brandenburg Technical University Cottbus, Cottbus, Germany
Email: douglas.cunningham@tu-cottbus.de
URL:http://www.tu-cottbus.de/fakultaet1/en/graphical-systems/department/overview.html

Douglas Cunningham's research focuses on the conceptual and practical integration of computer graphics and perception research, with particular emphasis on the perception and synthesis of natural conversations as well as on image statistics. Douglas received a Ph.D. in Cognitive Psychology from Temple University in 1997 and a habilitation in Cognition Science from the University of TÃijbingen in 2007. Douglas is the lead author of the 2011 book âĂIJExperimental Design: From user studies to psychophysicsâĂİ. He was program co-chair for Computational Aesthetics in 2008, 2009, 2011 and 2012. He is currently a professor for Graphical Systems at the Brandenburg Technical University Cottbus.

Christian Wallraven Korea University, Seoul, Korea
Email: wallraven@korea.ac.kr
URL: http://cogsys.korea.ac.kr

Christian Wallraven received his PhD in Physics in 2007 from the Eberhard-Karls-UniversitÃd't TÃijbingen for work conducted at the Max Planck Institute for Biological Cybernetics on creating a perceptually motivated computer vision algorithm. In 2010, he joined Korea University as Assistant Professor and head of the Cognitive Systems Lab. His current research interests lie in the interdisciplinary intersection between computer graphics, computer vision, and the cognitive sciences. Within this, his work focuses on the cognitive and computational study of face recognition, facial expression processing, multisensory object recognition, and evaluation of computer graphics and visualization algorithms. Christian Wallraven has co-organized an international workshop on Biologically Motivated Computer Vision (BMCV 2002) and the 2007 ACM Applied Perception, Graphics, and Visualization Conference. Christian is co-author of the book âĂIJExperimental Design: From user studies to psychophysicsâĂİ.

References

[CW11] CUNNINGHAM D. W., WALLRAVEN C.: *Experimental Design: From user studies to psychophysics*. A.K. Peters, 2011. 1

Understanding and Designing Perceptual Experiments

Prof. Dr. Douglas W. Cunningham
Prof. Dr. Christian Wallraven

Eurographics Tutorial 2013

What is an experiment?

First:

What is experimentation?

Perform an action – or a series of actions that are variants of each other – in order to answer a question.

Can be thought of as “controlled playing with something the aim of understanding it better”

What is an experiment?

Second:

What question?

The research question.

Simply put:
If we do not know what question we are trying to answer, we can not go about finding an answer for it!

What is an experiment?

Third:

What do you mean by...?

Any question or statement will tend to involve undefined or vague terms and implicit assumptions. Be aware of them.

Example:
“Is my technique 'better' than the others?”

- What **precisely** does “better” mean?
- Which others?
- What elements of the technique?

What is an experiment?

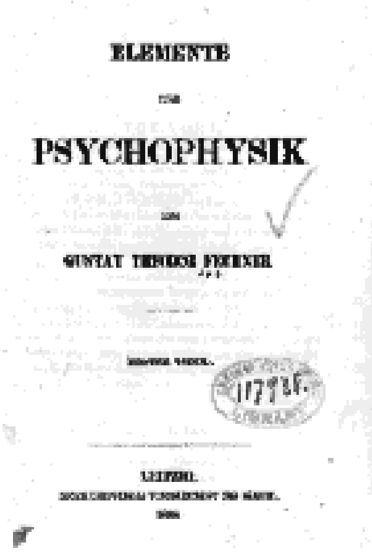
The more precisely the research question (and its assumptions) is formulated, the clearer it will be what we must do to answer the question.

What is an experiment?

What is an Experiment?

The collection of actions (which are systematic variants of each other) done in specific circumstances that are necessary in order to answer a specific question.

Why Experiment?



ELEMENTE

DER

PSYCHOPHYSIK

VON

GUSTAV THEODOR FECHNER

ERSTER THEIL.

LEIPZIG.
DRUCK UND VERLAG VON BREITKOPF UND HÄRTEL.
1860.

Physiologists and psychologists want to know:

- What signals can organisms extract?
- How are those signals are transduced?
- How is the transduced signal represented?
- How is the transduced signal processed?
- [How does the processed signal affect action?]

Why Experiment?

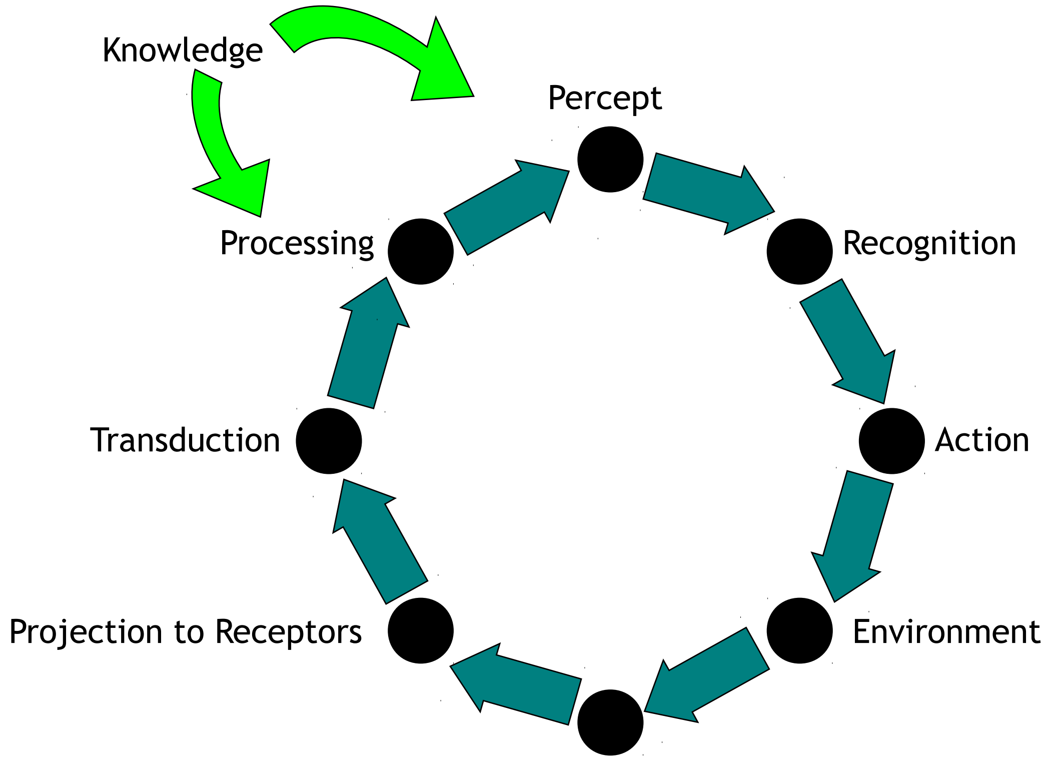
Computer Scientists want to know:

- What information ***must*** be in my technique?
- Which range of parameters is optimal?
- Which elements of my technique are “good enough”?
- Does my technique improve performance over SoA?
- Is my technique “better” than the others?

Why is this hard?

Why can't I just show stuff and ask people what they see?

Why is this hard?



Knowledge

Percept

Processing

Recognition

Transduction

Action

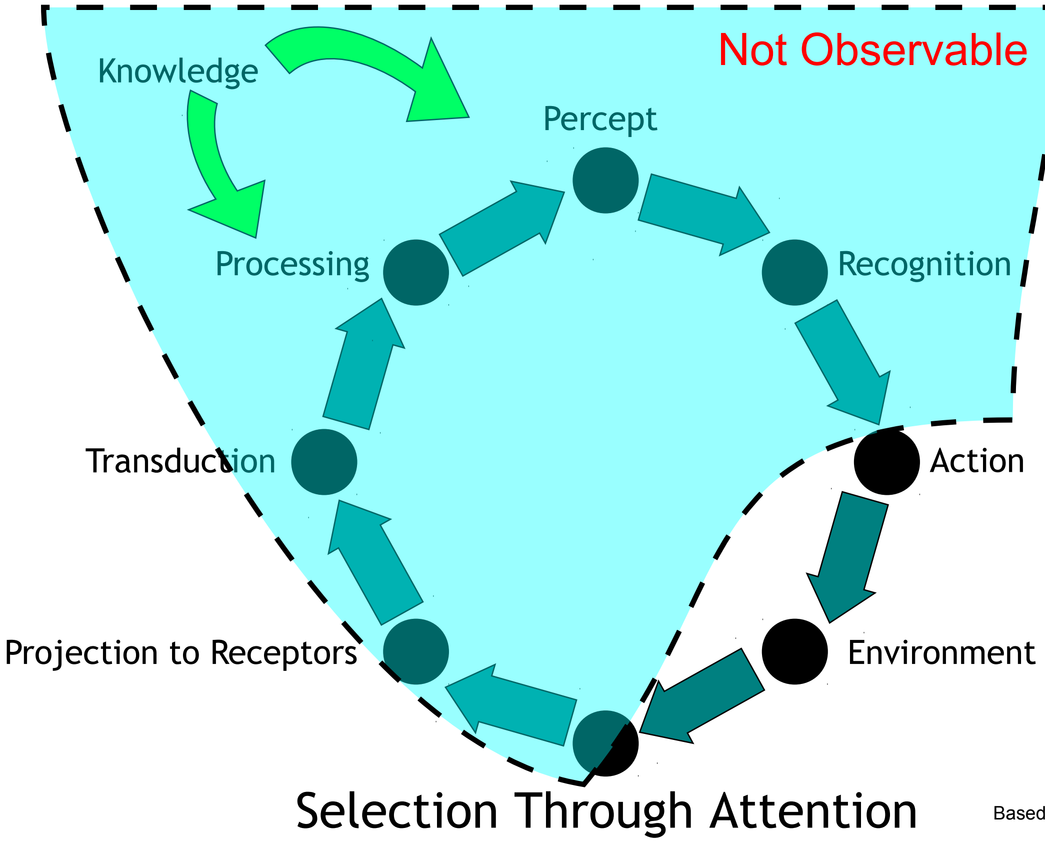
Projection to Receptors

Environment

Selection Through Attention

Based on B. Goldstein 2002

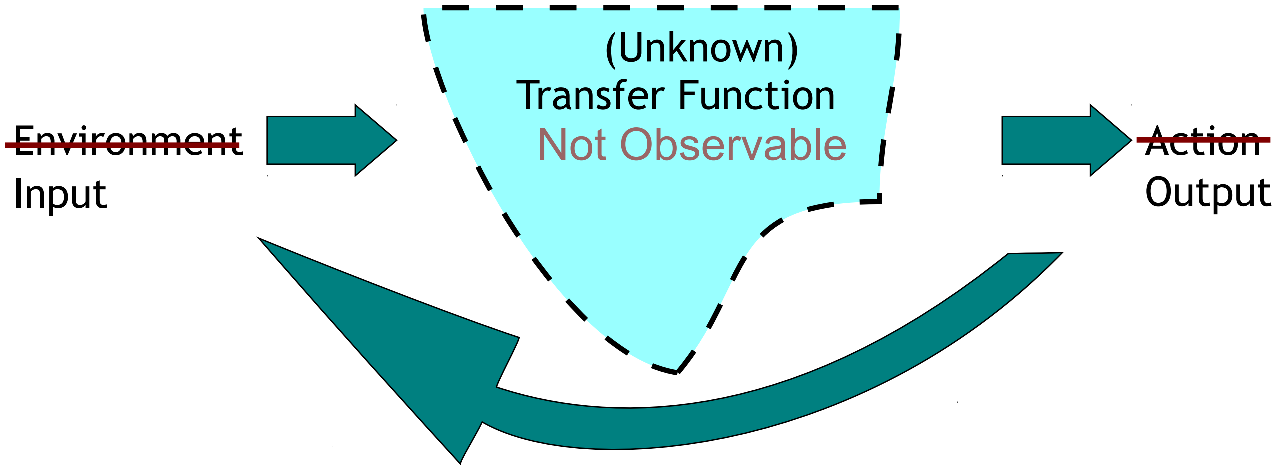
Why is this hard?



Based on B. Goldstein 2002

The classic black box

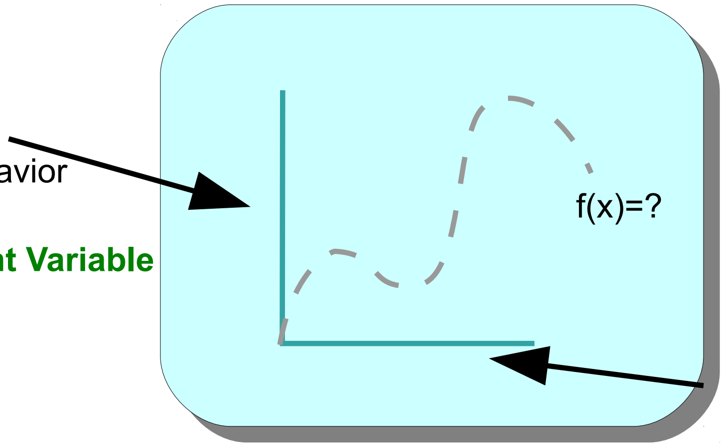
(naja, the light cyan box)



Dynamically adaptable, non-linear system
with feedback
(and probably feed-forward...)

A different (black) box

Goal: Estimate an unknown function



“Action” or
Output or
measured behavior

Dependent Variable

“Environment” or
Input or
Parameter Values or
Stimulus

Independent variable

Method: ??

Based on Cunningham & Wallraven, 2011

Degree of Control

To interfere or not?

- No control: ***Observational*** Research
  - examine things as they happen
  - Examples: Astronomy, Anthropology, Zoology, ...
- Complete control: ***Experimental*** or controlled research.
  - repeatedly and reliably produce a specific event in order to examine it.
  - Examples: Physics, Chemistry, Perceptual Psychology, Informatics, ...

Degree of Control

Type of experimental studies

User Study:

- also called usability testing
- a class of human factors experimentation
- examines whether a **finished, end-to-end, system** meets its **design goals**.
- ...so the question is fixed and there is little to no control over WHAT is presented (stimuli) or WHAT is presented (task).

Degree of Control

Type of study

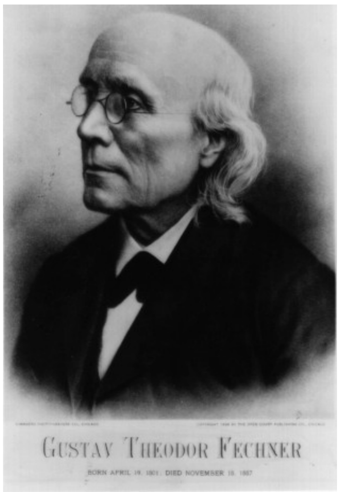
Perceptual experiment:

- Examines more general questions (about the underlying parameters of the system and its influence on the participant)
- Requires control over
  - what is shown (stimuli)
  - how and when the stimuli are shown (experimental procedure)
  - what the participants should do (task)

Degree of Control

Type of study

Psychophysics:



GUSTAV THEODOR FECHNER

- set of experimental methodology invented by Gustav Fechner in 1860 (and since extended by lots of people).
- Provides mathematical descriptions of the functional relationship between variations in the physical world and the resulting variations in the psychological (or perceptual) world.
- Requires **very fine** control over
  - what is shown (stimuli)
  - how and when the stimuli are shown (experimental procedure)
  - what the participants should do (task)

Degree of Control

Perceptual and psychophysical experiments can be thought of as **“exploring the parameter space” of a technique, procedure, or algorithm in order to determine the function relationship between parameter values and perceptual effects (possibly, in order to optimize the technique, etc.)**

Control

WARNING: The increased flexibility in answering questions that perceptual and psychophysical experiments offer comes at the cost of a need for increased vigilance, rigor, and expertise.

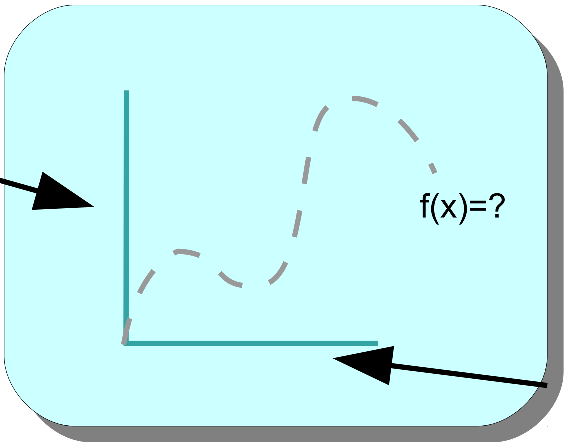
Why is this so hard again?

(FLASHBACK!)

Goal: Estimate an unknown function

“Action” or
Output or
measured behavior

Dependent Variable



“Environment” or
Input or
Parameter Values or
Stimulus

Independent variable

Method: Systematically sample the function

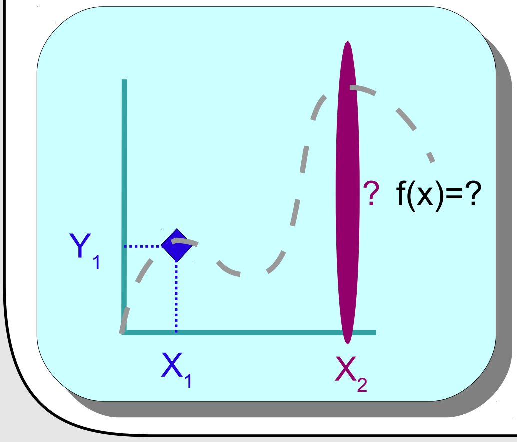
Based on Cunningham & Wallraven, 2011

Balancing Act

Specificity

- Can only talk about what you measured

Generality



Based on Cunningham & Wallraven, 2011

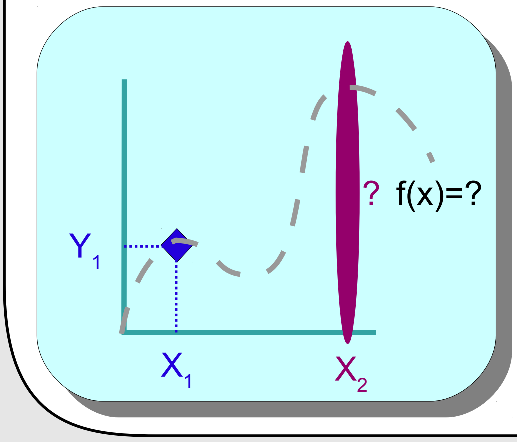
Balancing Act

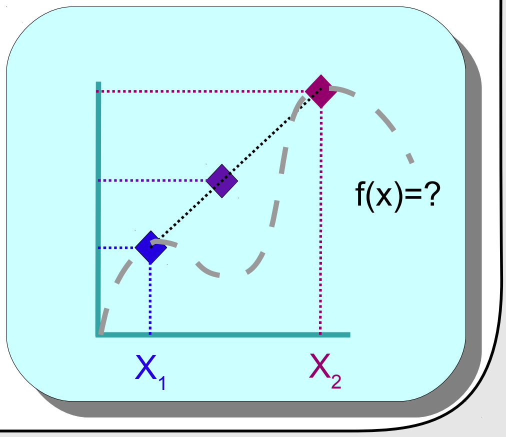
Specificity

- Can only talk about what you measured

Generality

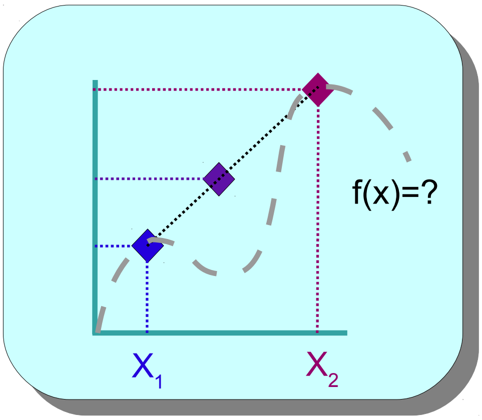
- Make broad statements without measuring every point...**Interpolate!**

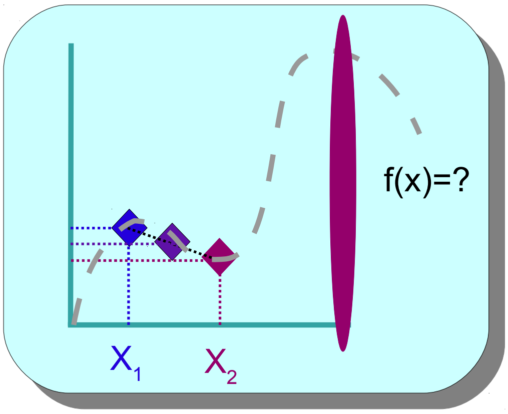




Based on Cunningham & Wallraven, 2011

A quick word on interpolation





Based on Cunningham & Wallraven, 2011

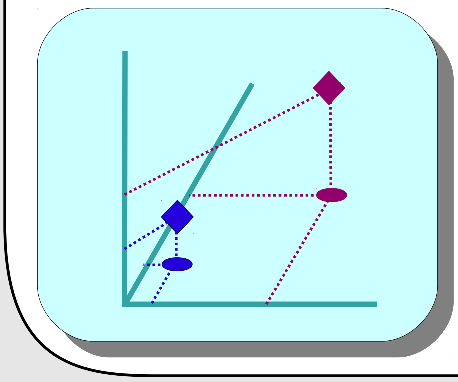
Balancing Act

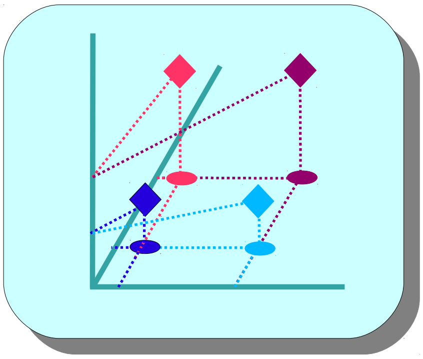
Specificity

- Can only talk about what you measured
- If too much varies at once, you cannot say what **caused** any differences

Generality

- Make broad statements without measuring every point...**Interpolate!**
- Systematically vary dimensions





Based on Cunningham & Wallraven, 2011

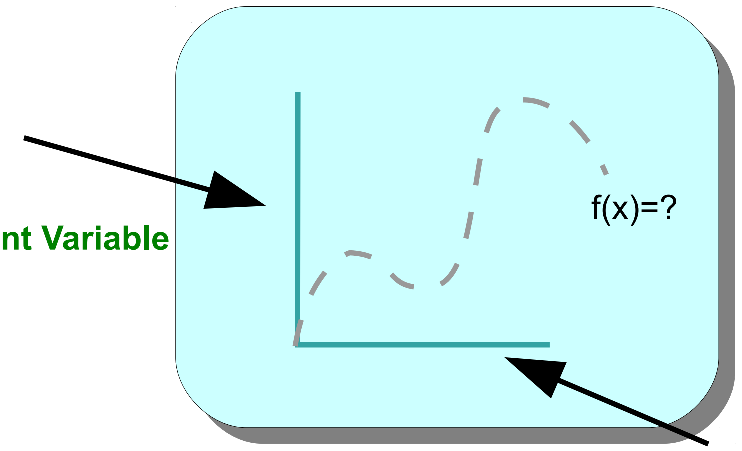
Some Terms

(these will become clearer as the talk progresses)

- Each dimension that we manipulate is called a **factor.**
- Each value that we used from a factor is called a **level**
- Usually, all combinations of factors are used. Any given combination is called a **condition**.
- Each single execution of a condition is a **trial**.
- Since only an examination of what happens under all relevant conditions can answer our question conclusively, the full collection of trials that addresses the current research question is **an experiment**.

Experimental design?

Goal: Estimate an unknown function



Dependent Variable

Independent variable

Method: Systematically sample the function

Based on Cunningham & Wallraven, 2011

Experiment I

Research Question:

How accurately can people point to a target?

Methods

Stimulus: Blue bullseye target

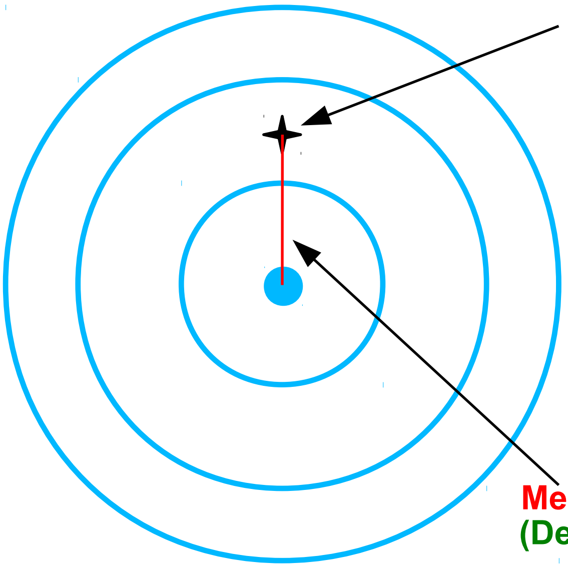
Participants: One

Task: Point (once) as quickly and accurately as possible to the center

(total number of trials: 1)

Experiment I

Let B(x) be the perception-action loop for pointing accuracy (for this situation)



Response

$$M(x) = B(x)\ ?$$

If I repeat the experiment, keeping *x* constant, will I get the same result?

Measurement M(x)
(Dependent Variable)

Situation: x (participant, task, stimulus, time of day, ...)
(includes the Independent Variable)

Experiment Ib

Repeated Measures

Research Question:

How accurately can people point to a target?

Methods

Stimulus: Blue bullseye target

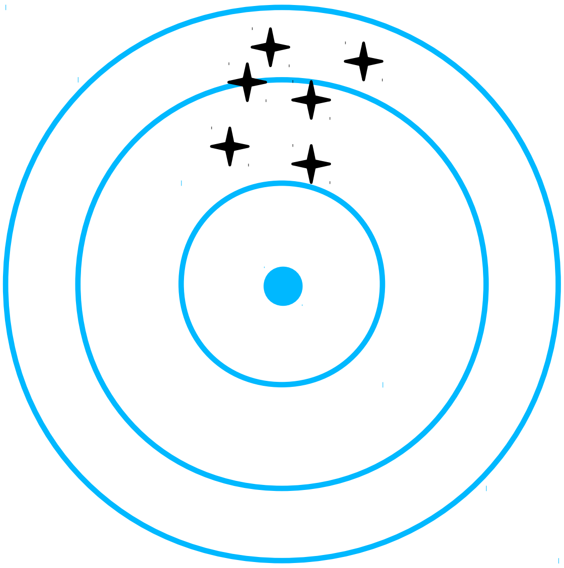
Participants: One

Task: Point (***n times***) as quickly and accurately as possible to the center

(total number of trials: 1 x n)

Experiment Ib

Repeated Measures



$$M(x) = B(x) + \epsilon_w$$

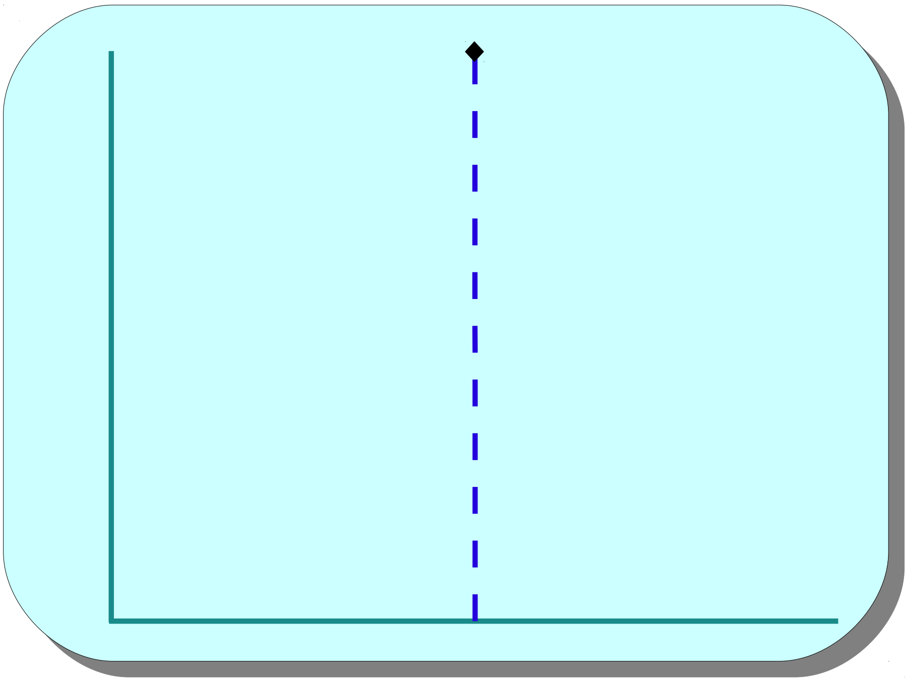
Situation: x (participant, task, stimulus, time of day, ..)

- People **cannot** exactly repeat **any** performance
- There is some inherent, unintentional variation/noise

Experiment Ib

Repeated Measures: Noise

for $\epsilon_w = 0$, $M(x) = B(x)$



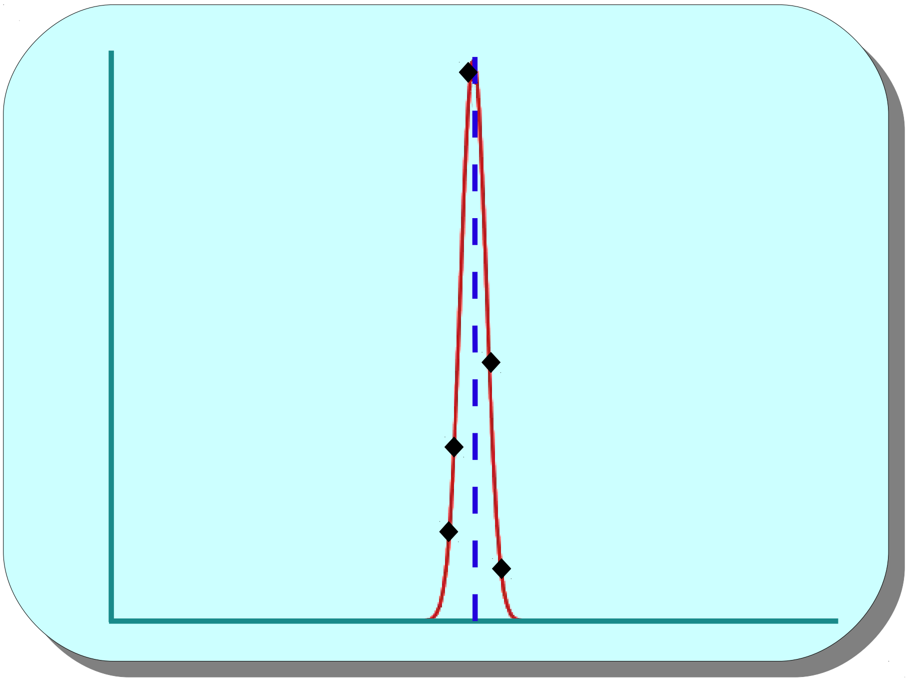
$$M(x) = B(x) + \epsilon_w$$

(note: Distributions have been rescaled)

Experiment Ib

Repeated Measures: Noise

For small ϵ_w each measurement is close to $B(x)$



$$M(x) = B(x) + \epsilon_w$$

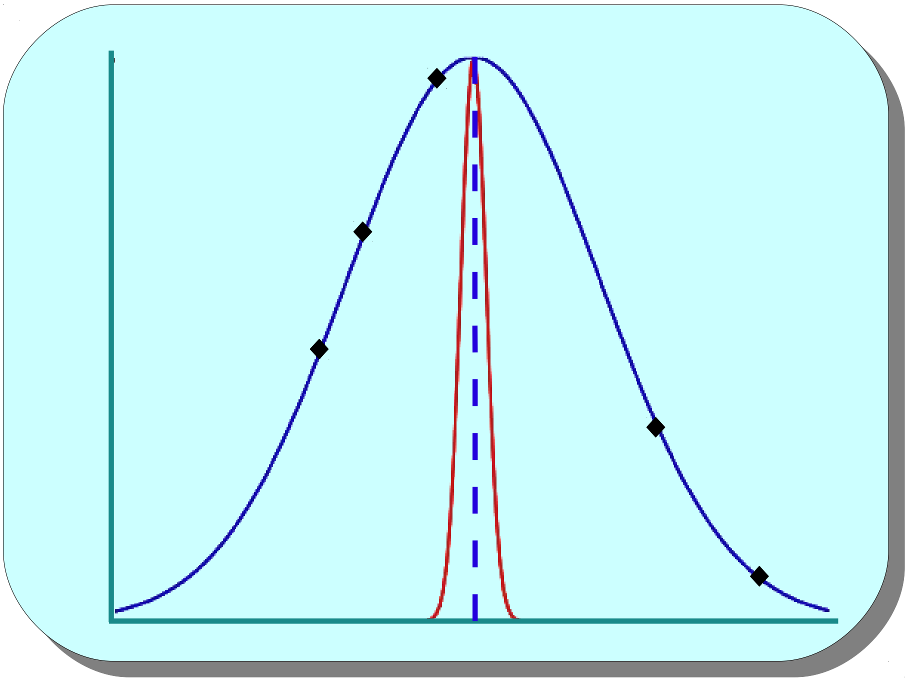
(note: Distributions have been rescaled)

Experiment Ib

Repeated Measures: Noise

For small ϵ_w each measurement is close to $B(x)$

For large ϵ_w each measurement can be very far from $B(x)$



$$M(x)=B(x)+\epsilon_w$$

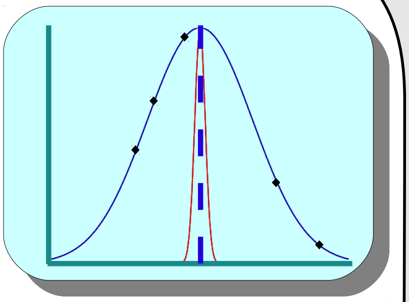
(note: Distributions have been rescaled)

Experiment Ib

Repeated Measures: Conclusions

The average approximates B(x)

- B(x) is constant
- ϵ_{w1} is different every time we measure M(x)
- Bias from ϵ_{w1} is sometimes positive, sometimes negative
- With enough trials, we can estimate the error, and factor it out



$$\text{trial 1} \;: m_1(x)=B(x)+\epsilon_{w1}$$

$$\text{trial 2} \;: m_2(x)=B(x)+\epsilon_{w2}$$

...

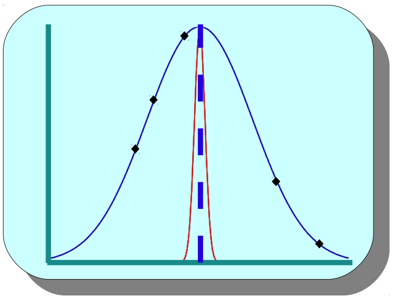
$$\text{trial n} \;: m_n(x)=B(x)+\epsilon_{wn}$$

$$\text{average} \;: \bar{M}(x)=\frac{\sum_{i=1}^{n} m_i}{n}=\frac{\sum_{i=1}^{n} B(x)+\epsilon_{wi}}{n}\approx B(x)$$

Experiment Ib

Repeated Measures: Conclusions

- Anything that can reduce noise improves the approximation
- Increasing number of samples (trials) improves approximation
- How many repetitions?
  - There are equations for calculating this, based in part on
    - expected effect size
    - noise size
  - Rule of thumb: More than 5 less than 20



Experiment Ib

Repeated Measures

Research Question:

How accurately can people point to a target?

Methods

Stimulus: Blue bullseye target

Participants: *One*

Task: Point (***5 times***) as quickly and accurately as possible to the center

(total number of trials: 1x 5 = 5)

Experiment Ib

Repeated Measures

Research Question:

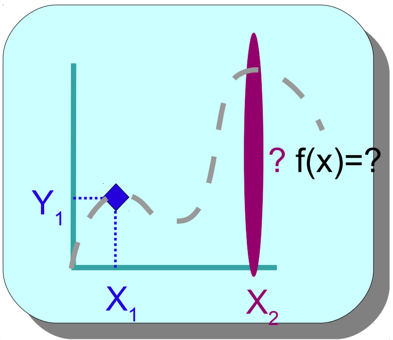
*How accurately can **people** point to a target?*

Methods

Stimulus: Blue bullseye target

Participants: ***One***

Task: Point (***5 times***) as quickly and accurately as possible to the center



- The results will be ***specific*** for this person
- B(x) might be different for different people
- To ***generalize*** to the population, we need more people!

Experiment Ic

Multiple Participants

Research Question:

*How accurately can **people** point to a target?*

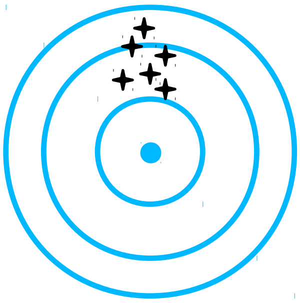
Methods

Stimulus: Blue bullseye target

Participants: **Several (p)**

Task: Point (***once each***) as quickly and accurately as possible to the center

(total number of trials: 1 per person)

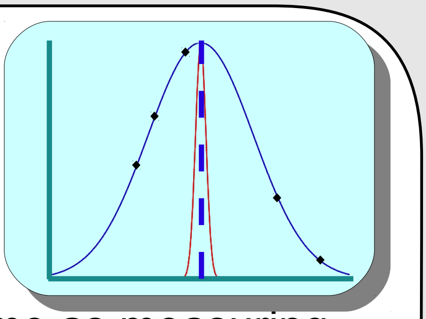


Experiment Ic

Multiple Participants

- We can calculate the error as above

$$\bar{M}(x)=\frac{\sum_{i=1}^{p} m_i}{p}=\frac{\sum_{i=1}^{p} B(x)+\epsilon_{wi}}{p}$$



- Is measuring many people once **really** the same as measuring one person multiple times?
- Why are people different?
  - Fundamentally different action-perception loops?
  - Constant (population) action-perception loop with everyone having a minor variation of that (e.g., noise)?

Experiment Ic

Multiple Participants

- Per **person**, we assumed a constant effect B(x) plus internal noise ϵ_w within person noise
- We can likewise assume a **globally** constant effect B(x) and additional noise between people ϵ_b between person noise

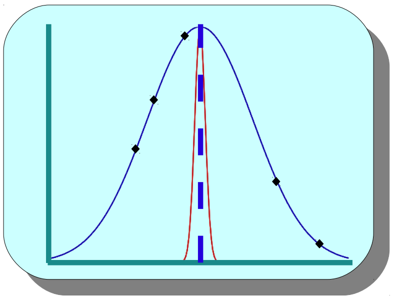
$$\bar{M}(x)=\frac{\sum_{i=1}^{p} m_i}{p}=\frac{\sum_{i=1}^{p} B(x)+\epsilon_{wi}+\epsilon_{bi}}{p}$$

- Might be wise to sample each of the two error functions separately! (*n* trials for each of *p* participants)

$$\bar{M}(x)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{p} m_i}{n\cdot p}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{p} B(x)+\epsilon_{wi}+\epsilon_{bj}}{n\cdot p}$$

How Many Samples?

- We are sampling unknown (noise) functions
- Multiple samples are needed
  - Per population
    - How many participants?
      - Again, there are equations for this
      - As a rule of thumb,
        - for large effects, 10 is sufficient
        - for smaller effects, more
  - Per Person (5+ reps per person)
  - Two error terms, sample both: Both Population and Person!



Experiment Id

Multiple Participants and Repetitions

Research Question:

*How accurately can **people** point to a target?*

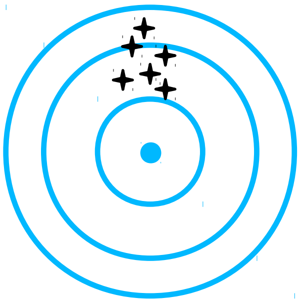
Methods

Stimulus: Blue bullseye target

Participants: **10**

Task: Point (**5 times**) as quickly and accurately as possible to the center

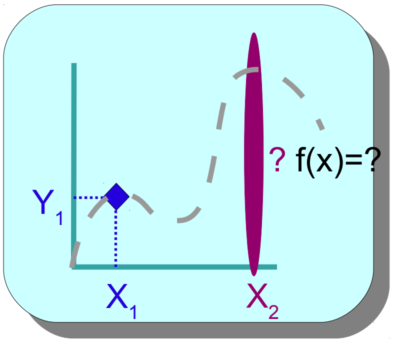
(total number of trials: 1 x 5 = 5 per person)

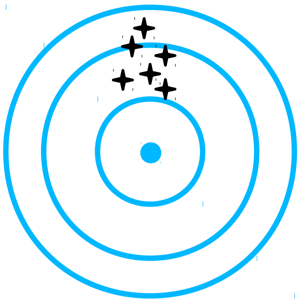


$$M(x) = B(x) + \epsilon_{wi} + \epsilon_{bi}$$

Experiment Id
Multiple Participants and Repetitions

- Why is performance so bad?
- Despite the 50 measurements, we have, effectively, one data point.
- To figure out why performance is “bad”, we need to find factors that affect performance
  - By systematically varying aspects of the situation *x*
  - Helps to ask why we think performance **should** have been better





$$M(x)=B(x)+\epsilon_{wi}+\epsilon_{bi}$$

Experiment II
Color

Research Question:

How do changes in color affect pointing?

Methods

Stimulus: Red and Blue bullseye targets

Participants: 10

Task: Point (5 times) as quickly and accurately as possible to the center

(total number of trials: 2 x 5 = 10 per person)

Experiment II

Color

$$
\begin{aligned}
M(x) &= B(x)+\epsilon_w+\epsilon_b \\
M(x+\Delta c) &= B(x+\Delta c)+\epsilon_w+\epsilon_b \\
&= B(x)+B(\Delta c)+\epsilon_w+\epsilon_b
\end{aligned}
$$

Effect of color change

Change in stimulus color

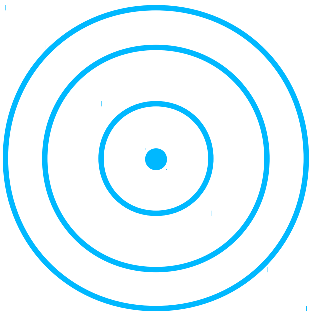
$$M(x+\Delta c)-M(x)=\{(B(x)+B(\Delta c)+\epsilon_w+\epsilon_b)\}-\{B(x)+\epsilon_w+\epsilon_b\}$$

$$M(x+c)-M(x)\approx B(\Delta c)+B(x)-B(x)+\epsilon_w-\epsilon_w+\epsilon_b-\epsilon_b$$

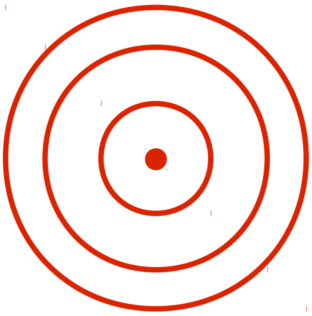
$$M(x+c)-M(x)\approx B(\Delta c)$$

*The difference between conditions is the effect of changing color **from blue to red***

specificity



Situation: x

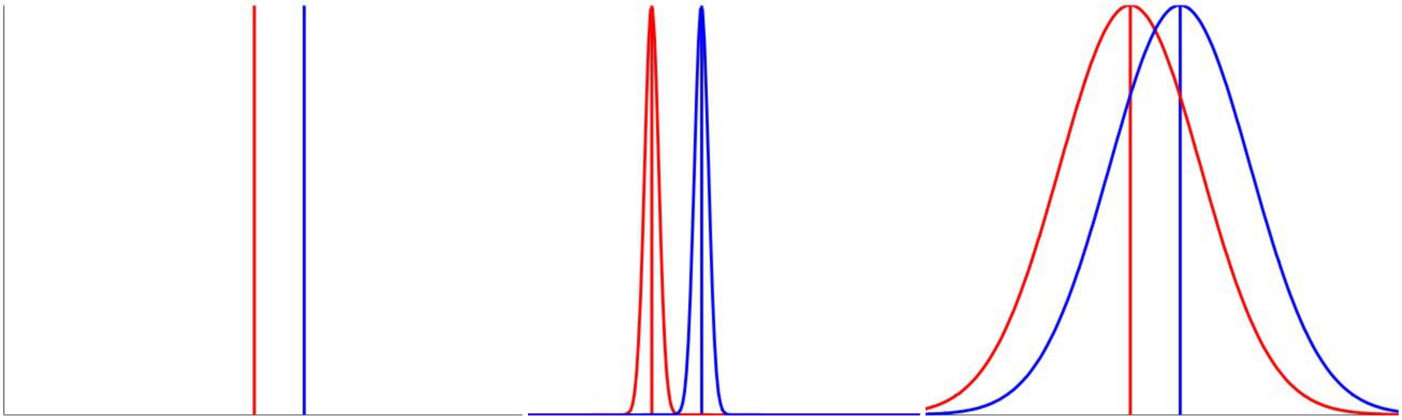


Situation: $x+\Delta c$

Note that splitting the function B(x+Δc) into its component parts (B(x) and B(Δc)) requires that the function B be homomorphic. Linear functions satisfy this property. Since we have assumed that the elements of x are independent of each other and can be modeled with as a linear, weighted sum, B is homomorphic.

Experiment II

Color: Are they different?



Are the means *really* different?

- Noise may “swamp” effect
- Control noise
- Identify and remove unwanted variance
- Run (complicated) statistics

Experiment III

Color and Size

Research Question:

How do changes in color and size affect pointing?

Methods

Stimulus: Large and small, Red and Blue bullseye targets

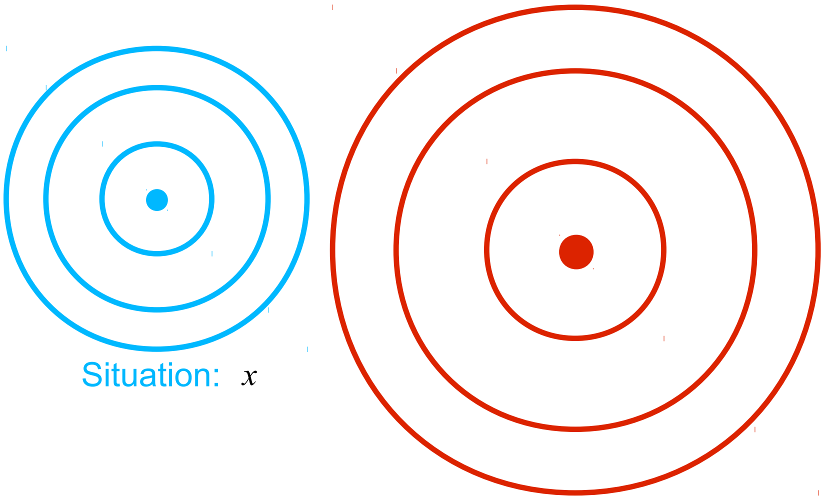
Participants: 10

Task: Point (5 times) as quickly and accurately as possible to the center

(total number of trials: 2 x 5 = 10 per person)

Experiment IIIa

Color and Size



Situation: x

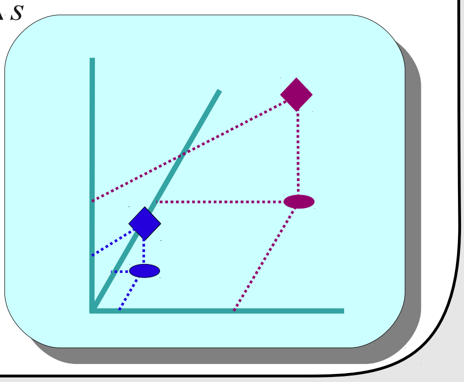
Situation: $x+\Delta c+\Delta s$

$$M(x)=B(x)+\epsilon_w+\epsilon_b$$

$$M(x+c+s)=B(x)+B(\Delta c)+B(\Delta s)+\epsilon_w+\epsilon_b$$

$$M(x+\Delta c+\Delta s)-M(x)\approx\boxed{B(\Delta c)+B(\Delta s)}$$

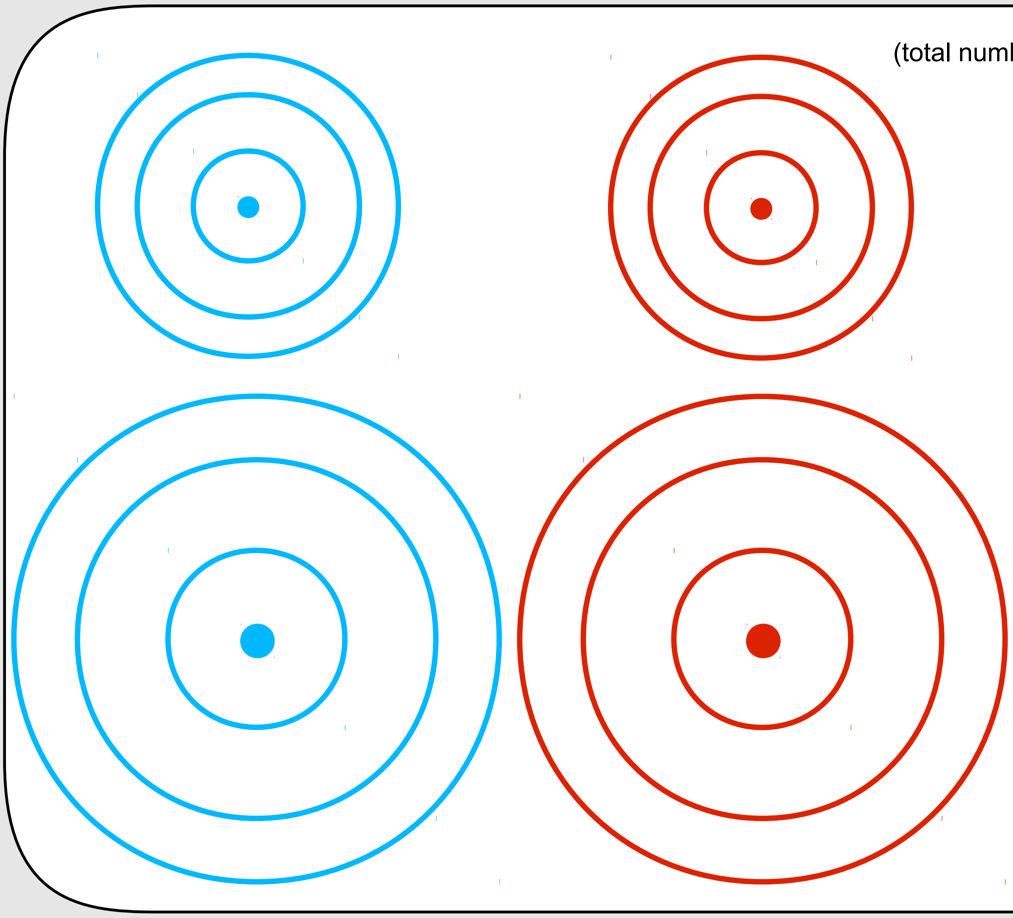
Can no longer **conclusively** say what caused the difference between conditions

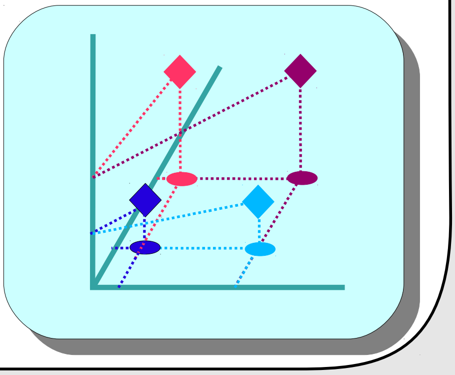


Experiment IIIb

Color and Size

(total number of trials: 4 x 5 = 20 per person)





Experiment IV

Contrast

Research Question:

How do changes in contrast affect pointing?

Methods

Stimulus: High and low contrast bullseye targets

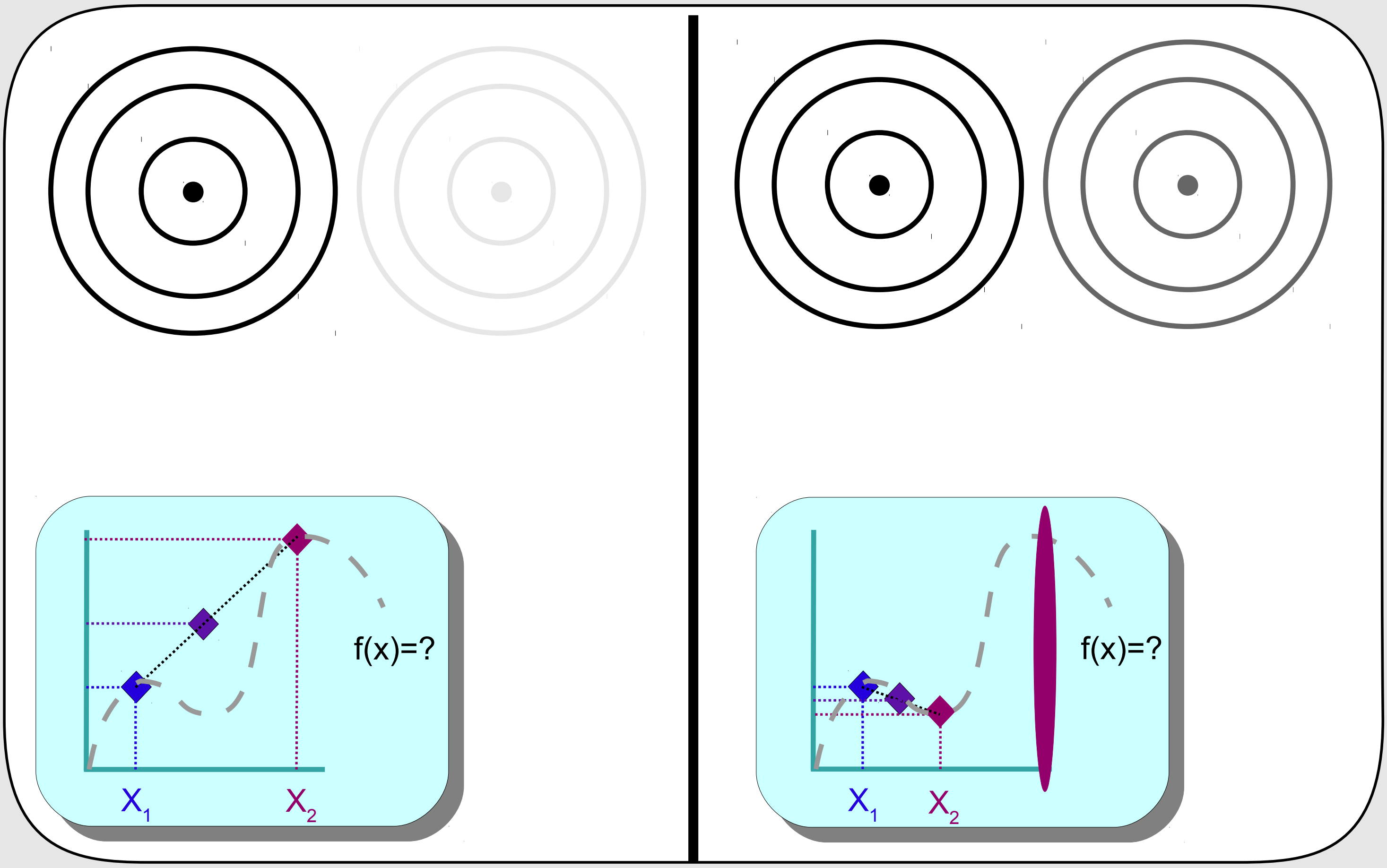
Participants: 10

Task: Point (5 times) as quickly and accurately as possible to the center

(total number of trials: 2 x 5 = 10 per person)

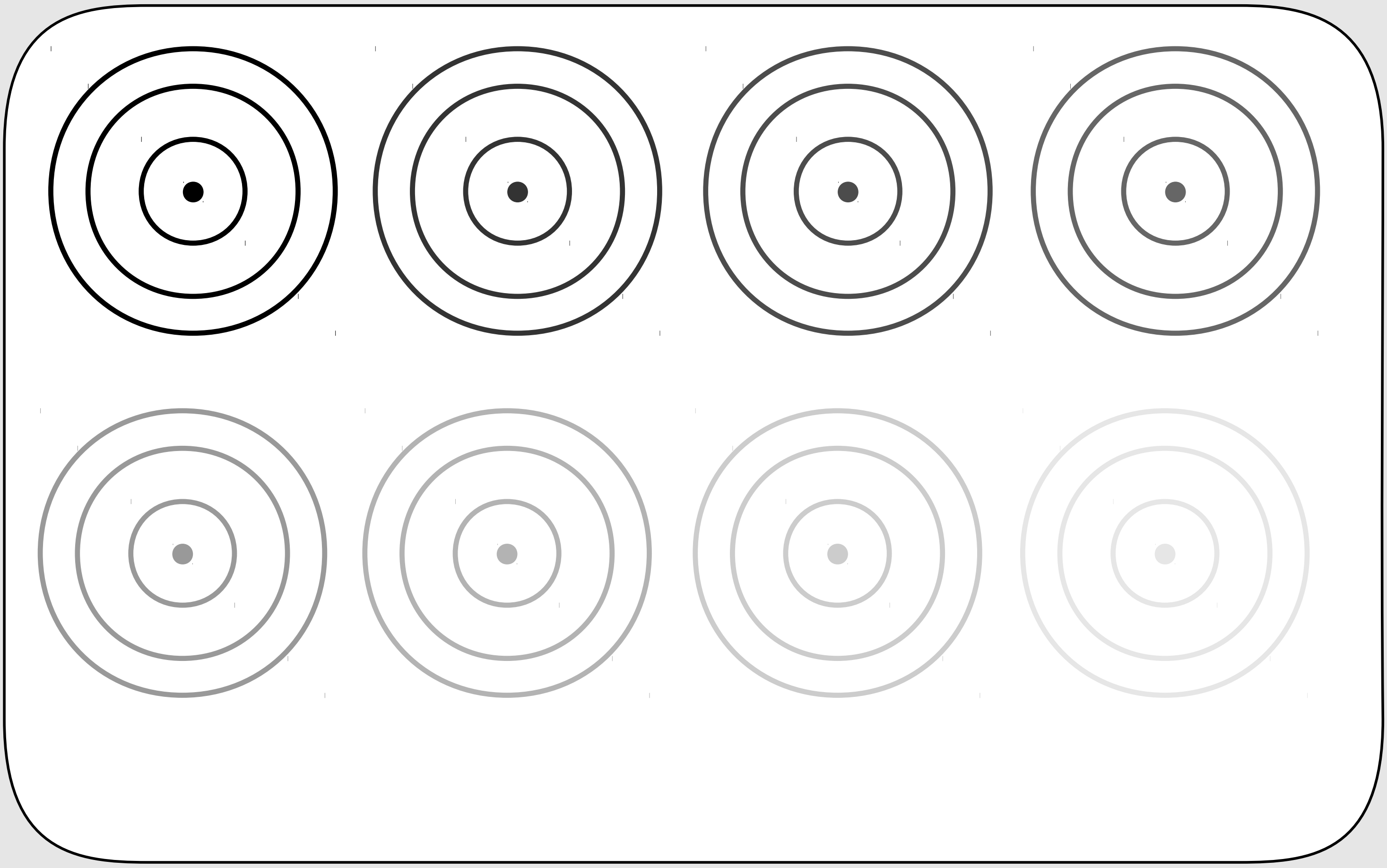
Experiment IV

Contrast



Experiment IV

Contrast



Experiment IV

Contrast

Research Question:

How do changes in contrast affect pointing?

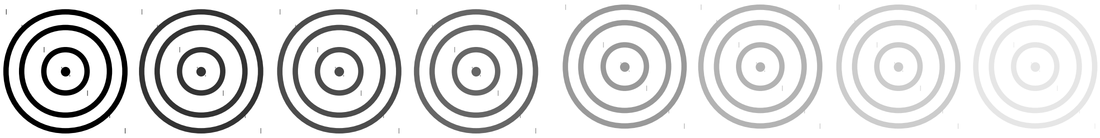
Methods

Stimulus: 8 bullseye targets, systematically varying contrast in equal steps

Participants: 10

Task: Point (5 times) as quickly and accurately as possible to the center

(total number of trials: 8 x 5 = 40 per person)



Experiment IV

Contrast

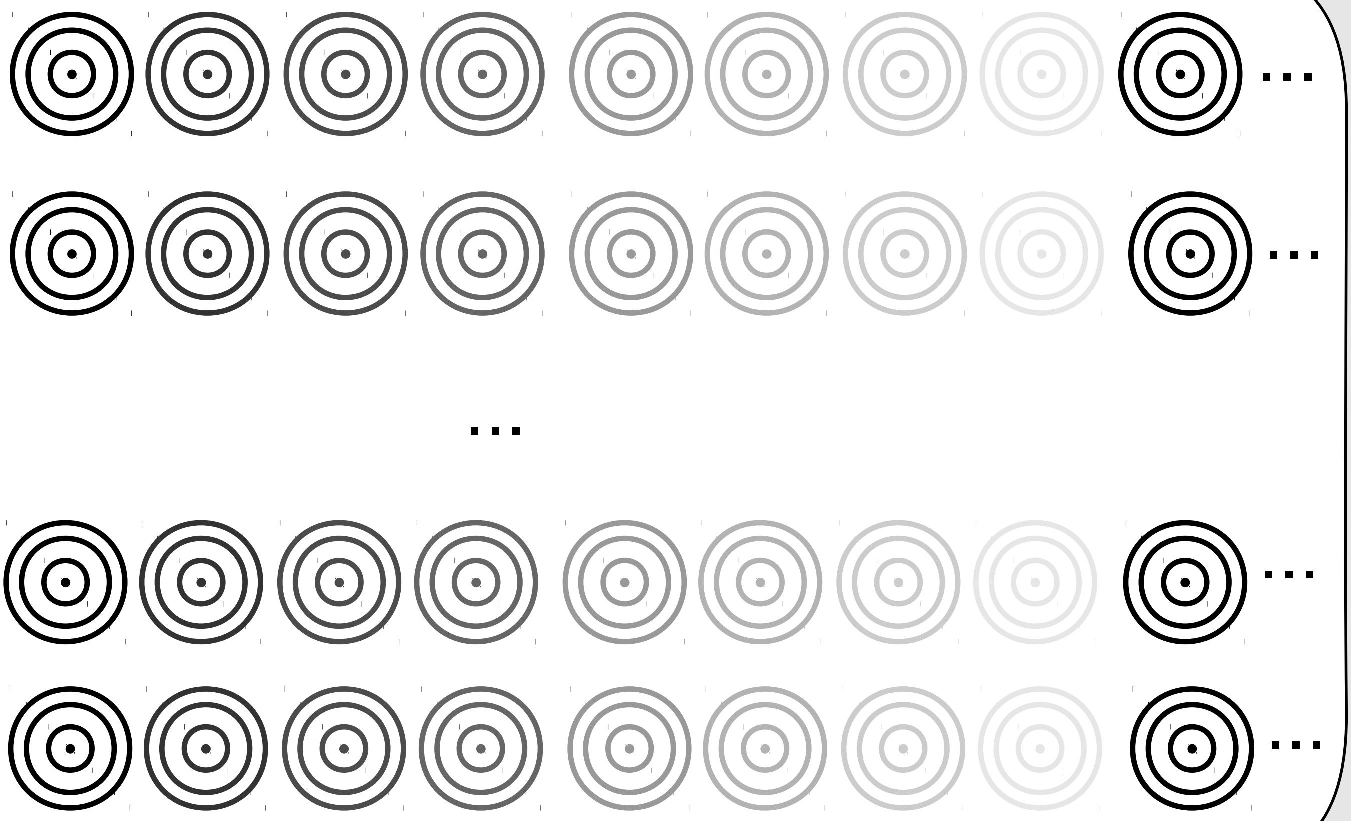
$$
\begin{aligned}
&M(x) = B(x) + \epsilon_w + \epsilon_b \\
&M(x - p) = B(x - A) + \epsilon_w + \epsilon_b \\
&M(x - 2\mathrm{p}) = B(x - B) + \epsilon_w + \epsilon_b \\
&\ldots \\
&M(x - 9\mathrm{p}) = B(x - I) + \epsilon_w + \epsilon_b
\end{aligned}
$$

Note:
Measurements are in terms of a base condition x and multiples of 10% contrast change.

The underlying perception-action loop does **not** use this periodic representation.

Experiment IV

Contrast: Trial Order

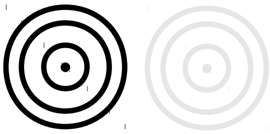


Experiment IV

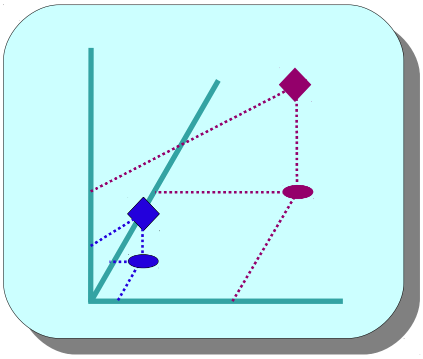
Contrast: Trial Order

Is trial 1 identical to trial 8?

- Different contrasts



- Difference in practice!
- So, any difference between trial 1 and trial 8 **might** be due to
  - Contrast
  - order/practice effect



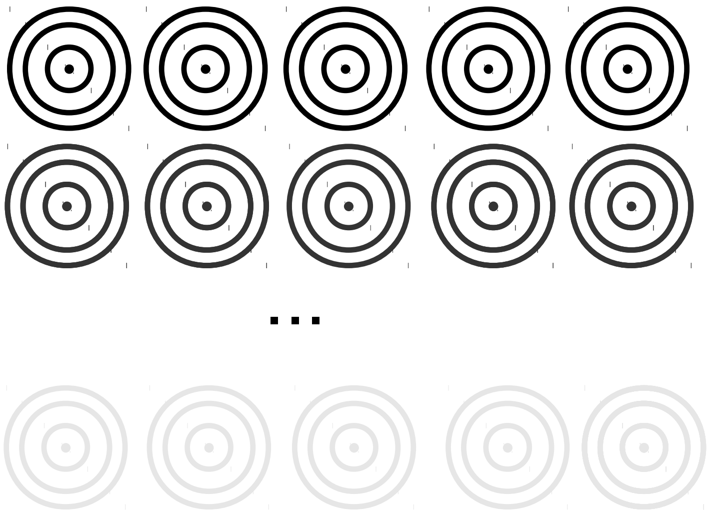
Experiment IV

Contrast: Trial Order

Solution 1: Eliminate order!

(total number of trials: 1 x 5 = 5 per person)

- Everyone sees only one contrast
- Each contrast is seen by one person
- Between-participants design



- Difference between high and low might be due to
  - contrast
  - participant (P1 vs P8)
- So, multiple people per contrast.
- The higher ϵ_b is, the more participants one will need for **each** contrast condition
- 10 people for 8 groups= 80 people!

Experiment IV

Contrast: Trial Order

Solution 2: Control for order! (hybrid design)

- Everyone sees every contrast (within-participant factor)
  - Each person acts as their own control or baseline condition.
- Each group sees a different order (between-participant factor)
- Every possible order is used
  - So, need many people per order condition
- How many different orders are there?
  - For a two condition experiment (A versus B): AB and BA (2 orders)
  - For three conditions: ABC, ACB, BAC, BCA, CAB, CBA (six orders)
  - For four conditions: ABCD, ABDC, ACBD, ACDB, ... (24 orders)
  - In general: N! (with N being the total number of conditions)

total number of trials: 8 x 5 = 40 per person
8 conditions, so 40,320 orders,
10 people per group= 400,000+ participants!

Experiment IV

Contrast: Trial Order

Solution 2: Control for order! (hybrid design)

- Everyone sees every contrast (within-participant factor)
  - Each person acts as their own control or baseline condition.
- Everyone gets a different (random order)
- Order as a noise term: With enough participants, order will average out.

Experiment IV

Contrast: Trial Order

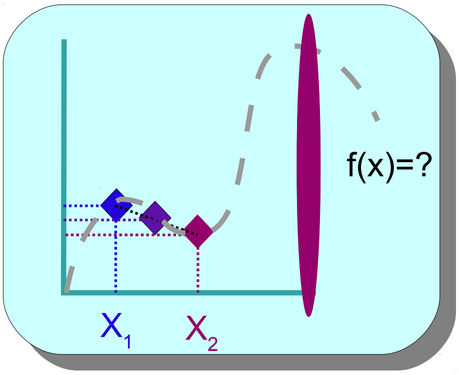
Generally,

- Order effects are generally smaller than individual differences
- Explicitly control some order: Latin squares (use a subset of possible orders)
- Full randomize: (any trial could be any condition)
- Blockwise randomize (Each condition seen at least once before any condition is seen a second time)

Special Factors
Participants

Why do people differ?

- *Natural talent*
- *Expectations*
- *Motivation*
- *Fatigue*
- *Physical differences (e.g., eye sight)*
- *Experience*



What if we only used the authors?

- *Authors are not naive, can* ***Bias***

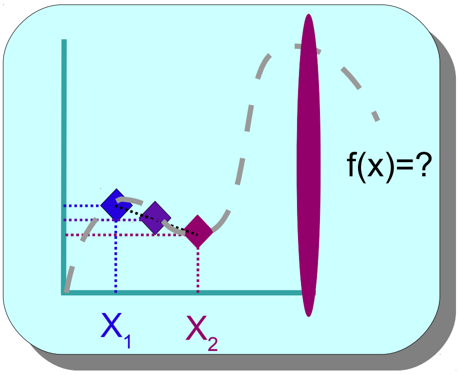
What if we only used experts?

- Experts/Novices have **different** skill levels, can **Bias**

Special Factors
Participants

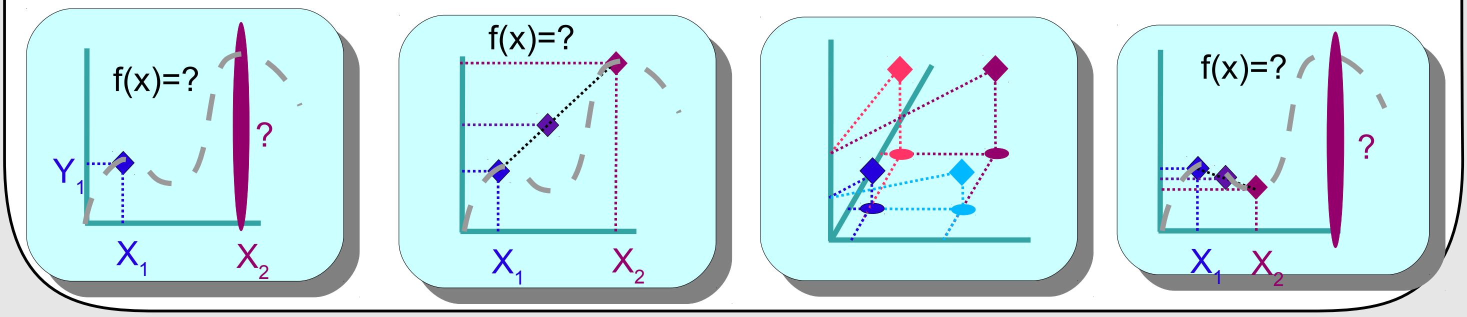
Generally,

- Sampling a distribution!
- Participants should be representative
- If you want to understand people in general, use people in general
- Use naive participants, unless
  - sure knowledge cannot affect results
  - desired population consists solely of experts
- If you want to understand how a technique will affect surgery, use surgeons!



Summary so far

- An experiment seeks to estimate an unknown function by **systematically** sampling the function.
- Can only talk in detail about what was measured
- Can interpolate between measured points, but with less accuracy
- Vary one thing at a time (or several factorially combined)
- Samples should be representative
  - People of the appropriate population
  - Stimuli of the dimension



Summary so far

- Anything that randomly varies between conditions adds noise
- Anything that reduces noise, makes it easier to find your effect
- *Keep conditions and trials as similar as possible*
  - Instructions
  - Experimenter
  - Your answer to participants' questions (before experiment)
  - Time of day of experiment
  - …
- Try to remove variance through proper experimental design, and **not** through complex statistics
- The more complex your statistics become, the fewer the number of people that understand what you did or why becomes (i.e., you loose your audience)

General Guidelines

- Once we are clear about our research question, we need to go about trying to answer it. In general, that means:
  - Show something (stimuli)
  - somehow (stimulus presentation)
  - to someone (participants)
  - And ask them to do something (the task)
  - (and then analyze the results)

The Task

What do we ask the participant to do?

We might ask them

- to describe what they see.
- to rate some specific aspect of what they see
- to interact with the stimuli (driving a virtual car,...).

We might even measure how their brains, hearts, or sweat-glands respond

The Task

So, which task is the right one for me?

The more clearly and explicitly the research question is defined, the more obvious it is which tasks will be

The Task

Once it is clear what we want to know the next step is to decide what would serve as an answer.

Example 1: we want to know which value for a specific parameter in a new visualization technique provides the fastest and most accurate identification of brain tumors

Suggestion: the answer should tell us something about response speed as well as identification accuracy (and maybe localization accuracy).

Which Excludes:

Free Description
Physiology

The Task

Once it is clear what we want to know the next step is to decide what would serve as an answer.

Example 2: Do people see a facial expression in the motion of a collection of dots---and if so which expression?

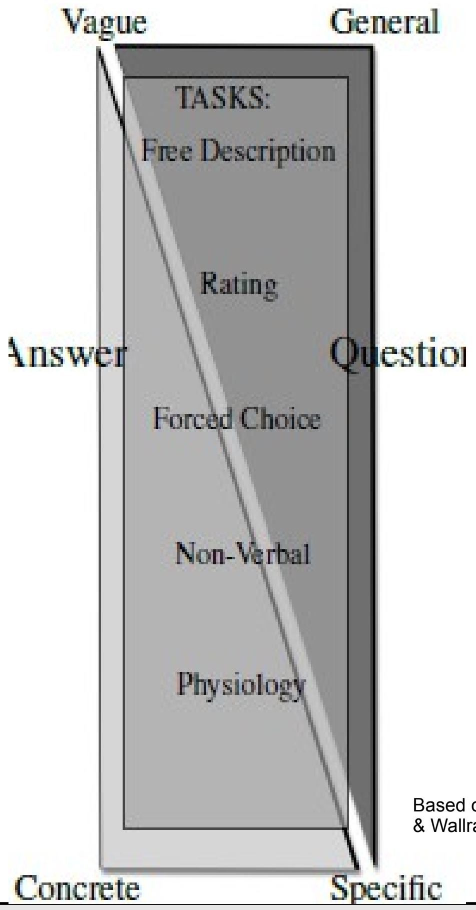
Suggestion: Asking people what they see, provides a lot of leeway but does not influence people unduly

The Task

Meta-tasks

Direct Tasks

Physiological Tasks



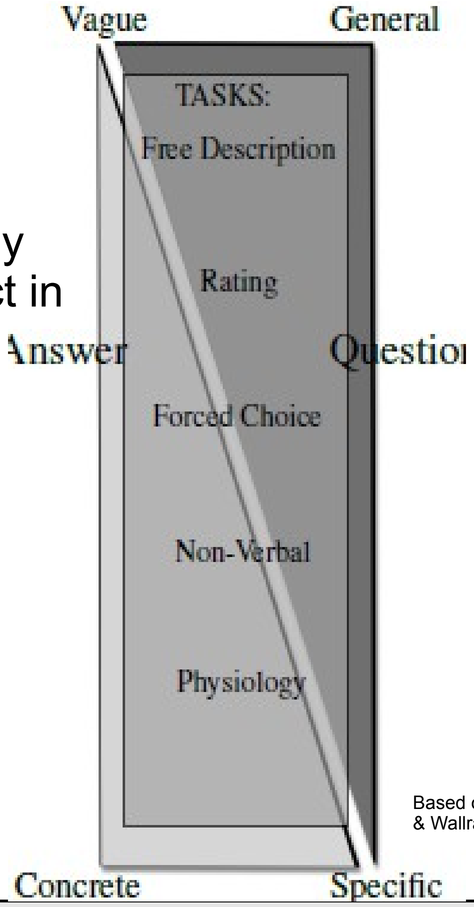
Based on Cunningham & Wallraven, 2011

The Task

Also called Qualitative Tasks

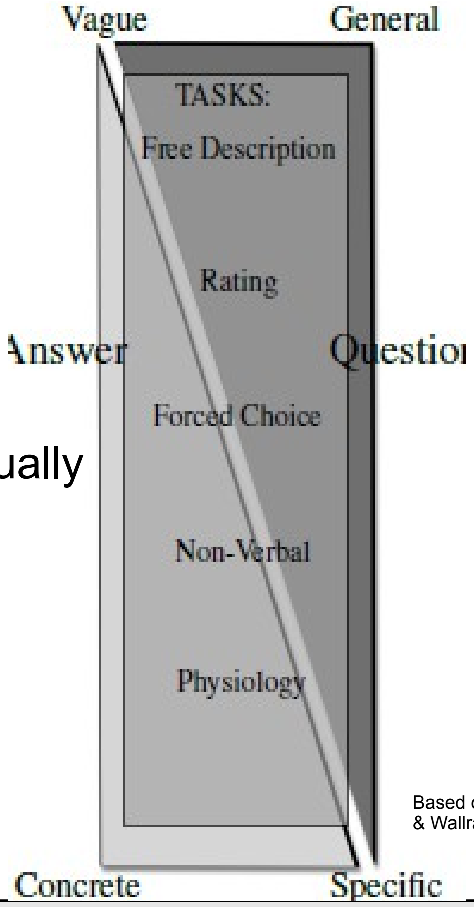
Meta-tasks

participants are asked how they **think** or **believe** they would act in a given situation.



Based on Cunningham & Wallraven, 2011

The Task

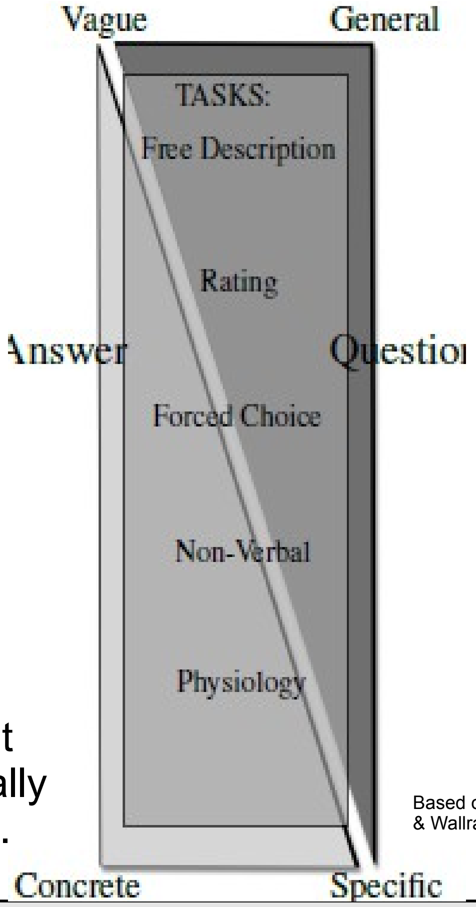


Direct Tasks

participants are asked to actually act in that given situation

Based on Cunningham & Wallraven, 2011

The Task



Physiological Tasks

provide a very direct, unbiased view of what elements of the stimulus the participants really saw or how they really felt about a stimulus.

Based on Cunningham & Wallraven, 2011

The Task

So, which task is the right one for me?

There is no “best” method.

Some tasks, however, are **more** appropriate than others for certain types of questions.

Naive Participants

How much can participants know about the experiment before and during the experiment?

Question: Can the participants intentionally influence how the task is performed?

For low-level processes: NO.
So, we can use an **overt** task.

CAUTION: participants can not affect low-level processes, but **CAN** affect their response!

Response Bias

Example:
Question: Brightness Thresholds

Participant A: Has *a* job requiring perfect vision (e.g., fighter pilot) and does not want to admit that he/she does not see something that is there.

Participant B: Has *a* job requiring high accuracy (e.g., flight controller) vision, and will only report that he/she sees something when he/she is really certain.

So, **motivation** and **strategies** will influence the **response criteria**, and therefore the pattern of results, even in the study of low-level processes.

Response Bias

Solution 1: Conceal expectations
Participants should **never** know (before or during an experiment) what answers we expect, and should not know the research question (if possible)

Solution 2: Preserve Anonymity
Participants should be convinced that no one (not even the experimenter can connect them to their data. (Use numbers for data files, not names, each person performs the experiment alone, etc.).

Response Bias

Solution 3: Use Statistics
Statistics (e.g., from signal detection theory) can be used to figure out the response bias (often requires a specific experimental design)

Solution 4: Alter the response Criteria
Add a reward structure, etc.. Tends to make the experiment longer, requires giving the participant feedback...which have problems.

Solution 5: Use a Covert Task
Hide the true task inside another. Note that this requires deception, which has ethical issues. This solution is not really recommended.

Response Bias

More can be said...for example about Instructions, catch trials, filler trials, ethics, practice trials, feedback, the experimental chamber, etc. Please see the book for more details!

Free Description

- Present a participant with a question, usually written.
- Ask people to describe their beliefs and opinions.
- Yields an explicit, word-based answer (as opposed to a numerical answer). The responses are usually writte, can be video or audio taped (get permission!)
- The answers will vary wildly in length, quality, informativeness, and relevance.
- Not common in perceptual research. Often not viewed positively.

Free Description

Best at answering broad, general questions that seek broad, general, and vague answers.

Useful at the beginning of a new line of experiments, where it is unclear what type of responses participants might give.

- What words would people use to describe this facial expression (or painting, etc.)?
- What is the most dominant aspect of this display?
- What do people notice first in this display?
- What kind of observations would people make about this situation, concept, or scene and how often does each observation occur?

Variants

- Interview
- Questionnaire
- Long Answer
- Short Answer
- Partial Report

Clear, Unique Interpretation?

Of all tasks, this is the hardest to cleanly and uniquely interpret.

The greatest advantage: Can provide a wealth of information.

Example: Ask someone which painting he/she prefers

Answer 1: the name of a painting
Answer 2: Long, rambling personal history
Answer 3: Vague description of a painting
Answer 4: “ I once owned a copy of ...” …?

Clear, Unique Interpretation?

Example 2: ``describe the display as carefully as you can – as if you are trying to describe what you see to someone who is not in the room, and has never seen these displays." (from Cunningham et al., 1998).

Display: consisted of a black screen with a random array of small white dots on it. A black triangle moved over the white and dots a second field of dots was superimposed over this display, reducing the overall contrast between the figure (the triangle) and the background (the dot fields).

Answer 1: I see Picasso's *Nude Descending a Staircase*.
Answer 2: I see a happy face, no wait now it is sad.

Why written?

Social Factors

- **Follow-up questions**
- **Conversational Goals (leading questions, hypothesis confirmation, “socially appropriate answers”, etc.)**

Guidelines

- **Anonymity**

The more participants are convinced that no one can connect their answers to them, the more likely they are to be completely honest.

- **Clarity**

Each and every participant should know precisely what is being asked of them. Also be careful with categories creation in the data analysis.

- **Relevance**

The task should not only address the research question, but should also reflect the real-world situation

Guidelines

- **Preparation**

Have everything ready **before** the participants show up (including category creation for data analysis!).

- **Written questions and answers**

- **Follow-up experiment**

check the validity of any interpretations of the results – and the degree to which they reflect the real world – in a more objective, reliable experiment.

Rating Scales

Measures: numerical value determining how each stimulus compares other stimuli.

Questions:

Give an insight into how elements of a class of stimuli (e.g., expressions, paintings, cities) vary along a given dimension (e.g., sincerity, aesthetic value, size).

- What are people's preferences among the following paintings?
- Do people tend to prefer cubist, surrealist, impressionist, or pop-art paintings?
- Which of the follow expressions do people find to be more attractive or appealing?
- How do the following computer generated animations compare in terms of realism?

Variants

- **Ordered ranking.**

Participants list the stimuli in order along the relevant dimension.

- **Magnitude estimation.**

Participants are asked to assign any number they want to the stimuli. These numbers should represent a more or less intuitive indication of precisely how much of the dimension each stimulus has.

- **Likert ratings.**

Participants are given a range of numbers and are asked to assign each stimulus the number that represents its location within he allowed range.

- **Semantic differentials.**

A special variant of the Likert task where the endpoints of the fixed range of numbers are assigned bipolar opposite terms (e.g., good and bad, fast and slow).

Guidelines

- **The underlying dimension**

The scale dimension should be clear and as representative as possible of the research question.

- **Anchoring**

Participants must understand where along that dimension the specific values on the scale are.

- **Resolution**

Trying to extract more information than exists in the results will only lead to inaccurate conclusions. People tend to use no more than ten points on any scale, and if they do, the results will not be reliable.

Guidelines

- **Scale usage**

Be aware of the type of information present in the scales.

- **Cultural bias**

Rating scales are essentially a self-report task.

Forced Choice

Measures: which of a limited number of potential answers participants choose for different stimuli. This is a discrimination task that shows how well participants can perceive a specific difference between several stimuli

Questions: They give a qualitative measurement of how distinct or discriminable several stimuli are from one another.

- How well can people recognize, identify, etc the following paintings?
- Which style of painting do people prefer?
- Which of the follow expressions do people find to be most attractive?
- How much do these specific stylization methods affect the recognition of facial expressions?

Variants

- **N-alternative forced-choice**

N alternatives are given and the participants must choose one of them for each stimulus. The alternatives might be absolute descriptions or relative to some standard.

- **N+1-alternative non-forced-choice**

Same, but with an alternative that allows the participant to refuse to make a choice (e.g., ``none of the above'').

- **N-interval forced-choice**

A special variant of the N-alternative forced-choice task in which $N$ stimuli are shown sequentially and the participants are required to choose one interval based on some criterion

- **N+1-interval non-forced choice**

The same, but with an alternative that allows the participant to refuse to make a choice.

Variants

- **Go/no go**

“Gö” Trial: the stimulus currently present meets certain criteria, then the participants issue a simple response
“No Go” If the stimulus does **not** meet the criteria, the participant does nothing. Serve as a **gateway** task.

- **Matching-to-sample**:

A standard stimulus is shown together with more than one other stimuli. The participant chooses the comparison stimulus that most closely matches the sample.

- **Visual search**:

A specific **target** is presented simultaneously with a series of other items called distractors. The distractors are similar to the target along the dimension of interest. The number of distractors is manipulated across trials systematically. The participant is to indicate whether the target is present or not (usually the target is present on half of the trials).

Variants

- **Rapid Serial Visual Presentation (RSVP)**

A number of stimuli are presented one at a time in a very rapid sequence. One of the stimuli is the target and the rest are distractors. The participant's task is to indicate when the target stimulus is present.

- **Free grouping**

Participants are presented with large set of stimuli and are asked to arrange them into groups.

Guidelines

- **Non-mutually exhaustive alternatives**

If the alternatives are not mutually exhaustive, the results might not be uniquely identifiable.

- **Asymmetry**

Participants will usually assume that the alternatives show up in the stimulus set, and that the alternatives occur equally often. If the frequency with which the different alternatives occur varies, participants will guess what the relative frequency is.

- **Order**

People are willing to accept a short sequence of identical stimuli, but not a long one.

Guidelines

Category level

We could present several pictures of animals and ask participants to indicate if the animal is dog or a Siamese cat. The two alternatives are in different category levels. Dog is in an **entry-level** category, while Siamese cat is in a **subordinate-level** category.

Chance level and task resolution

Chance performance is the percent of the time that any given alternative would be chosen if participants had no knowledge of what the alternatives or the stimuli are.

Identification versus discrimination

Quite often, forced choice tasks are incorrectly considered to be identification tasks.

Guidelines

Clarity of the alternatives

Not only must the participants understand the alternatives, they must understand what we mean by the alternatives.

Uncertainty.

A forced-choice task measures a participant's judgment, and thus includes a measurement of uncertainty. If uncertain, some of the time they will choose one alternative, and some of the time they will choose a different alternative. The frequency distribution is related to their uncertainty

Statistics?

- You need statistics in order to make concrete statements about data that has a random component to it
  - you asked random people
  - you repeated a question multiple times
  - your measurements contain noise
- There are two different kinds of statistics:
  - descriptive
  - inferential

Statistics?

- Descriptive statistics describe the data
  - obviously used in “Descriptive” experiments (surveys, etc.)
  - BUT: use it for every data you gather! plot it!
- Inferential statistics test hypothesis about the data
  - Are means different?
  - Do the two variables correlate?
  - Used to answer specific questions

What can it do for you?

- Provide objective criteria for testing hypotheses
  - by following statistical procedures, observer bias, for example, is greatly reduced, if not eliminated
- Provides a way to critically assess conclusions of other people
- When used beforehand, can save you a lot of time and effort
  - how many samples do you need to test?
  - what kinds of answers can you expect?

Where does it fail?

- Only tells you about probability of a hypothesis being rejected!
  - you cannot prove that something is TRUE
- Bad input = bad output
  - YOU need to worry about the quality of the data and its analysis
- YOU need to provide the hypothesis, statistics can then take over
- YOU need to provide the interpretation and worry about its significance!

A few terms

- **Population**: the set of all possible items under question
- **Sample**: the subset of the population that is tested
  - obviously, the sample should be representative of the whole population
  - this is often NOT the case!
- **Sample size**: the number of data points
- **Dependent variable:** what you measure
- **Independent variable**: what you manipulate

What variable types are there?

- Nominal
  - categories
- Ordinal (magnitude)
  - also called rank-ordered variables
- Interval (magnitude, interval)
  - temperature in °C / °F
- Ratio (magnitude, interval, rational zero)
  - reaction time data, accuracy, confidence
- Note: there are various problems with these definitions, there are listed here as guidance!

Hypothesis testing

- There are a series of statistical tests that are used to test specific hypotheses
- As all good scientific experiments should be hypothesis-driven, these tests are rather important
- There is a huge number of these tests, navigating through them is difficult

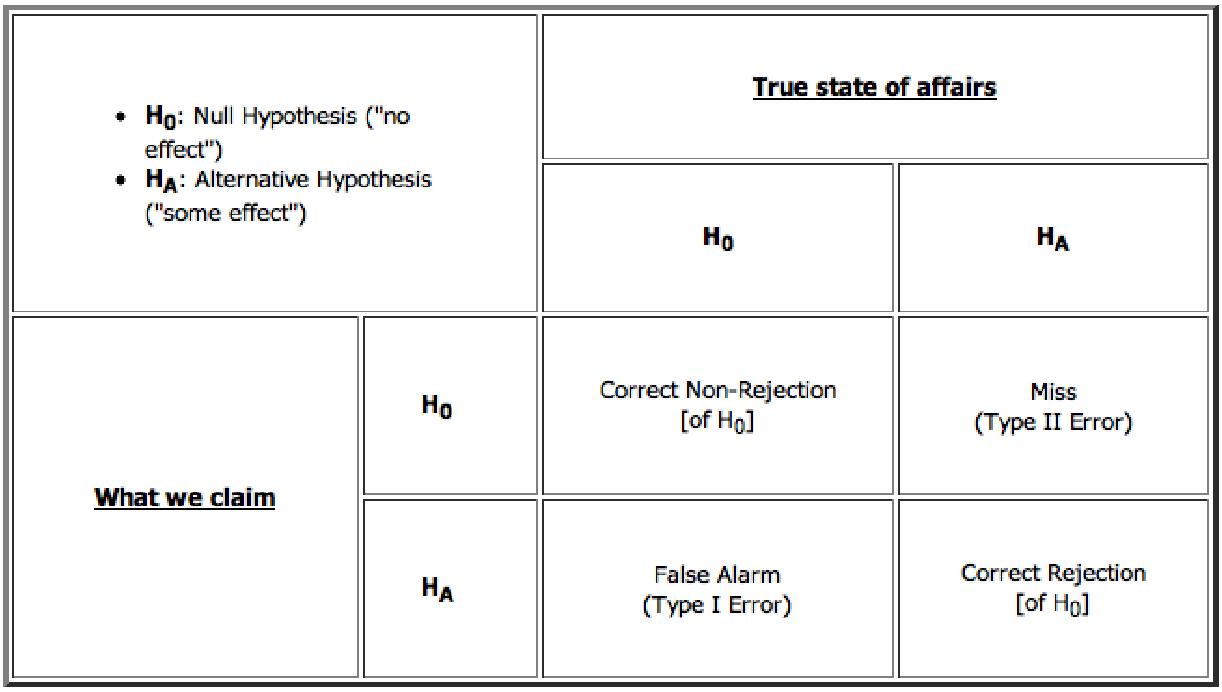
Alternate & Null Hypotheses

- Alternate Hypothesis (HA): testable statement induced from specific observations: it states that “there is an effect”
  - Experimental condition is different from control condition
  - Variable y is related to variable x
- Null Hypothesis (H0): the hypothesis that **no** effect exists
  - **statistical tests evaluate the null hypothesis**
  - that is, H0 is assumed to be true, unless experimental evidence suggests that the probability of H0 being true is very low

Example: HA and H0

- Experiment: Participants are required to search for a target face among distractor faces
  - Distractor faces always are neutral faces
  - Experiment 1: target faces are either sad or neutral
  - Experiment 2: target faces are either happy or neutral
  - You measure the time it takes to find the target face
- HA: Emotional faces have different response times from neutral faces
- H0: Emotional faces do not have different response times from neutral faces

Alternate & Null Hypotheses



• **H_0**: Null Hypothesis ("no effect") • **H_A**: Alternative Hypothesis ("some effect")		**True state of affairs**	
		H_0	**H_A**
What we claim	**H_0**	Correct Non-Rejection [of H_0]	Miss (Type II Error)
	H_A	False Alarm (Type I Error)	Correct Rejection [of H_0]

Probabilities

- In the experiment, you usually test for the probability p that your findings happened by random chance
  - Small probability means that it is unlikely that the observed effects are due to random chance, but rather due to some driving mechanism, it is likely that you will observe similar responses again
  - Large probability means that is likely that the observed effects are due to random chance, and it is unlikely that you will observe the exact pattern of results again

Meaning of small & large p

- Experimental data for Experiment 1 (sad vs. neutral faces):
  - Mean response times for sad faces: 2.5 s
  - Mean response times for neutral faces: 2.7 s
- H0: no difference in response times for the two conditions
- This is equivalent to asking:
  - If H0 were true, what is the probability p of observing a difference of 0.2 s based on the sample data?
- **Small p: conclude observed effects are real**
  - **H0 is rejected (i.e., unlikely that sad = neutral faces)**
- **Large p: conclude observed effects are due to random chance**
  - **H0 is accepted (i.e., likely that sad = neutral faces)**

How small should p be?

- By convention, $p < 0.05$
- This means that
  - there is a smaller than 5% chance that H0 is true
  - H0 should be rejected, HA can be accepted
  - experimental results are statistically significant
  - the sample reflects the true nature of population
  - results are likely due to the experimental manipulation
- This also means that 1 out of 20 experiments would result in a different outcome!
- By convention, if $p<0.001$ results are sometimes said to be “highly significant”

How Large is Large?

- Conversely, also by convention, if $p > 0.05$
  - there is a larger than 5% chance that H0 is true
  - you need to accept H0, reject HA
  - experimental results are statistically non-significant
  - the sample does most likely not reflect the true nature of population
  - results are likely due to random chance

Are Accepted Hypotheses True?

- Accepted means that the hypothesis is “probable” or “likely true”, it does NOT mean that it is “true”
- Again, if p=0.05, in one out of 20 experiments, your accepted HA will turn out to be false
- Statistics cannot “prove” a result, only provide support/evidence for the hypothesis

Are Rejected Hypotheses False?

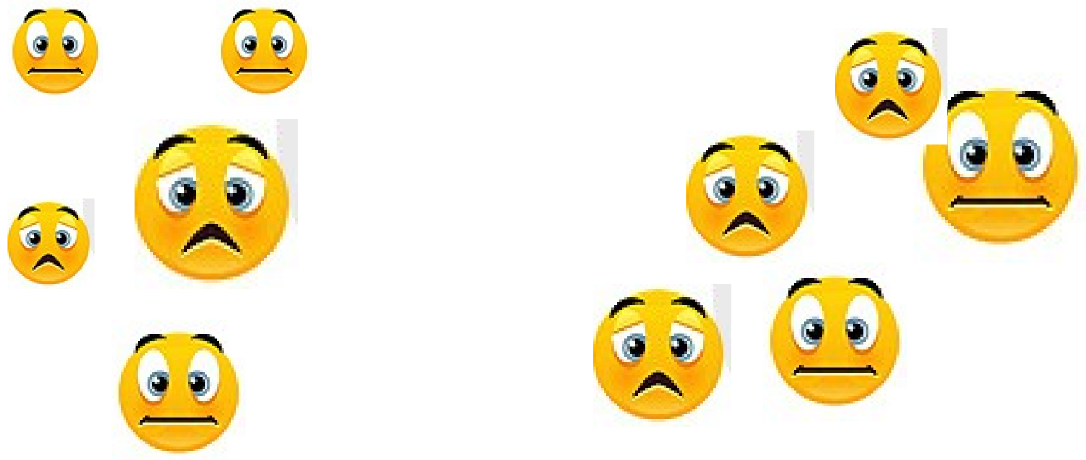
- Rejected means that it is “not probable” or “most likely not true”, but again, does NOT mean that the hypothesis is “false”
- Rate of error is determined by “power” of test
- Even if p>0.05, there could be an effect

Experiment 1:
sad versus neutral faces

- sad: 2.7, 2.5, 3.2, 1.7, 2.2 s (mean = 2.5)
- neutral: 2.0, 3.0, 2.8, 3.6, 2.0 s (mean = 2.7)

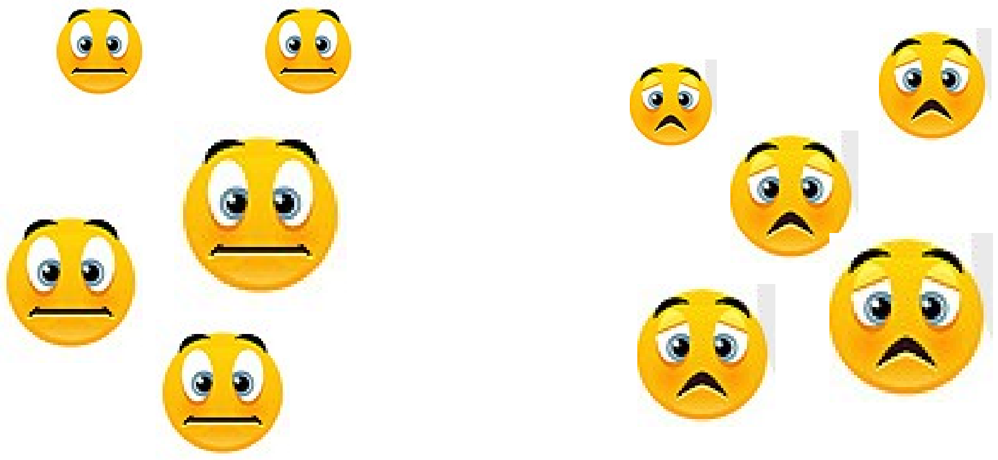
- Are these two group means the same of different?
- Is the mean difference of 0.2 s significant?
- What is the probability of observing this magnitude of difference by random chance?
  - Doing the test yields p=0.59

Main idea behind testing for sample differences



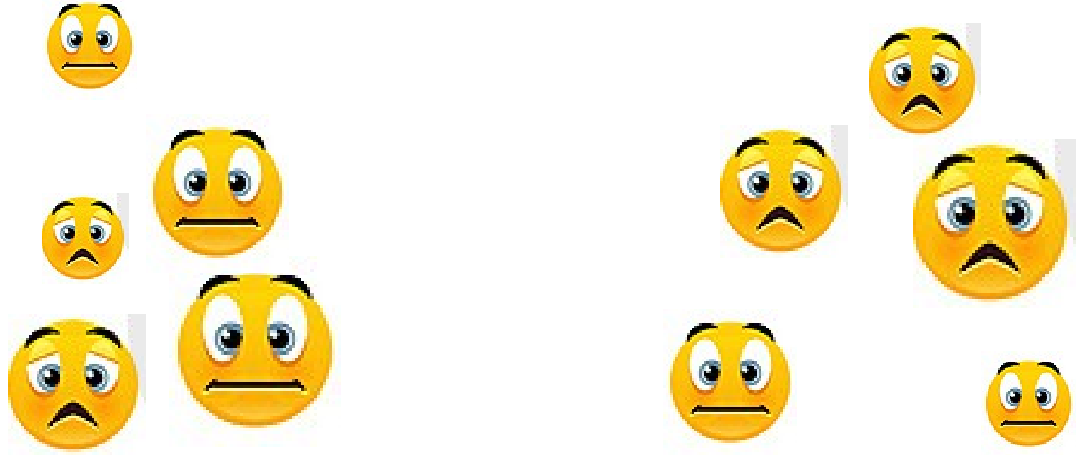
- Now randomly re-assign neutral and sad faces into two groups and determine new means
- group1: mean = 2.5s & group2: mean = 3.0s
- group1 < group2

Main idea behind testing for sample differences



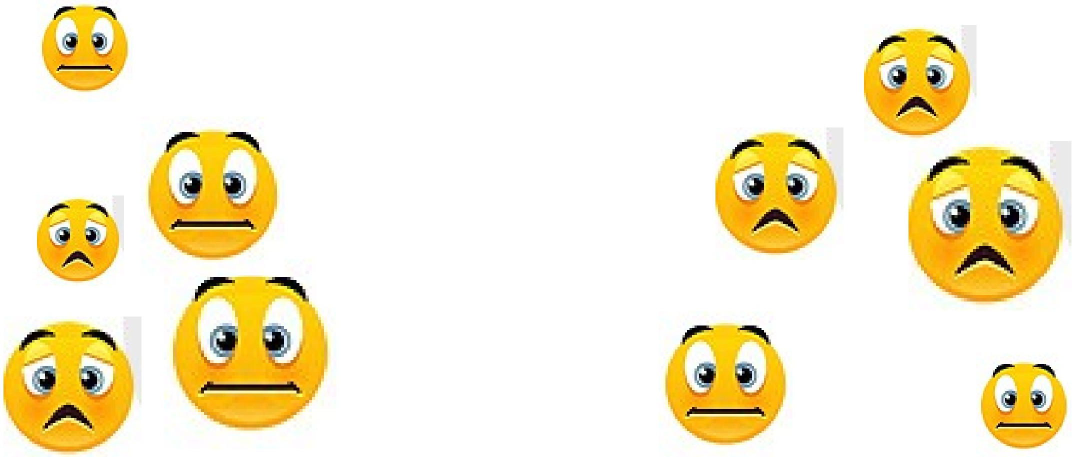
- neutral (group1): mean = 2.7s
- sad (group2): mean = 2.5s
- group1 > group2

Main idea behind testing for sample differences



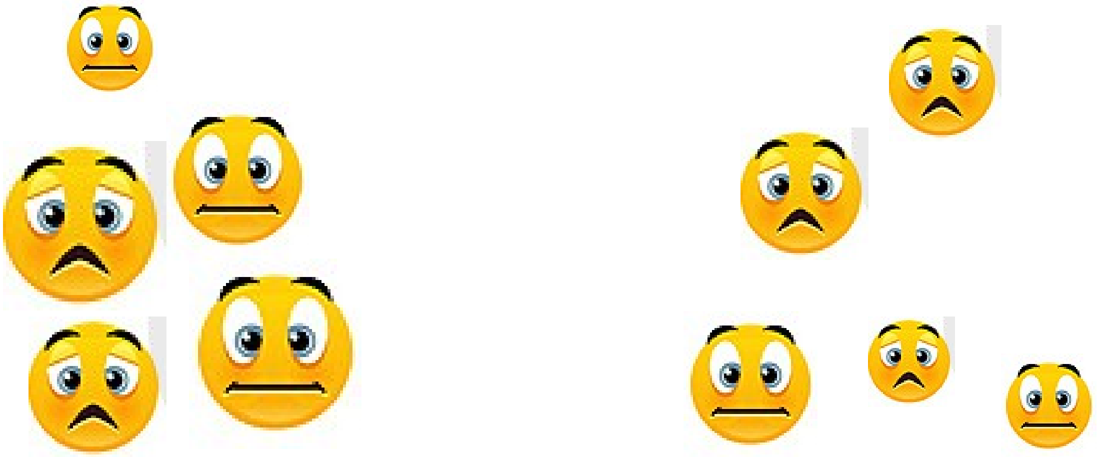
- Now randomly re-assign neutral and sad faces into two groups and determine new means
- group1: mean = 2.8s & group2: mean = 2.9s
- group1 < group2

Main idea behind testing for sample differences



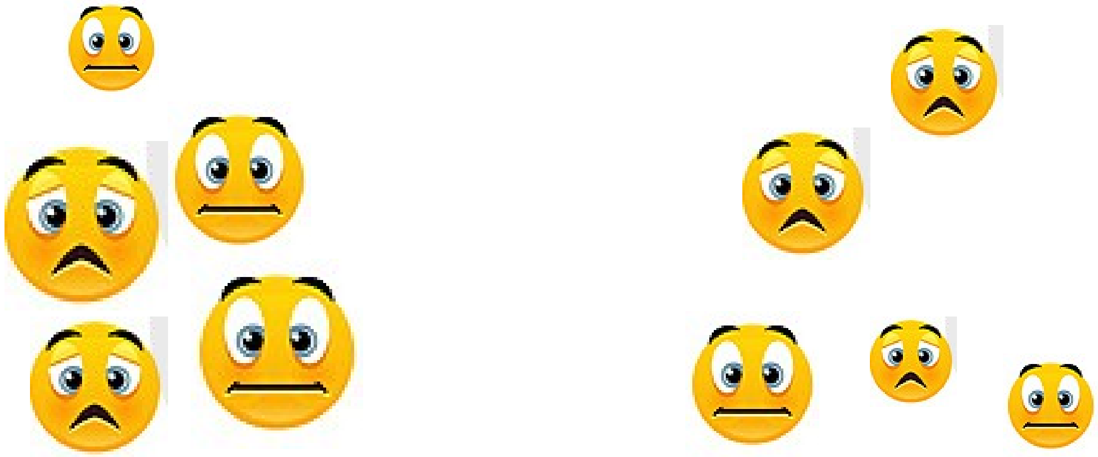
- Now randomly re-assign neutral and sad faces into two groups and determine new means
- group1: mean = 2.8s & group2: mean = 2.9s
- group1 < group2

Main idea behind testing for sample differences



- Now randomly re-assign neutral and sad faces into two groups and determine new means
- group1: mean = 3.0s & group2: mean = 2.8s
- group1 > group2

Main idea behind testing for sample differences

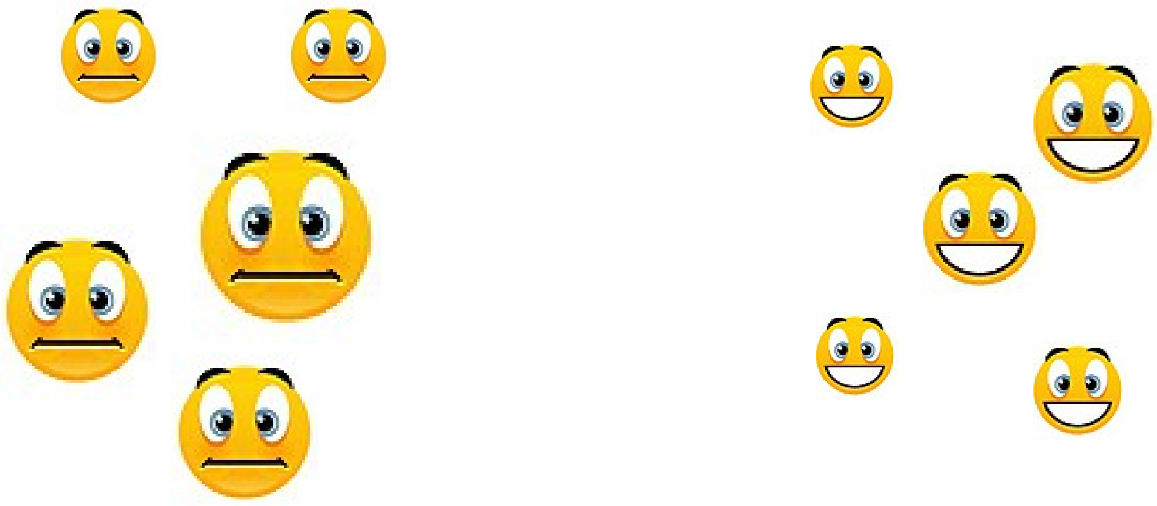


- Repeat many times and observe the differences between the group means

Resampling

- If you had resampled 100 times, you would observe differences between 0.05s to almost 1.0s in the group means
- Now, simply count how many differences are greater or equal to the original 0.2s
  - In our case, 61 out of a 100
  - This is close to the theoretical p-Value of 0.59 we get from the proper t-test

Experiment 2:
neutral versus happy

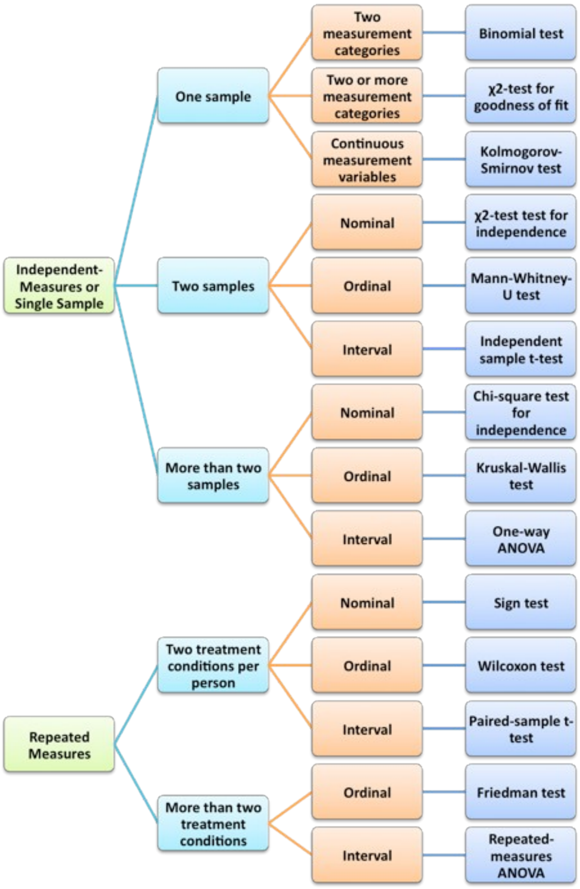


- neutral faces = 2.0, 3.0, 2.8, 3.6, 2.0 s : mean = 2.7s
- happy faces = 2.2 2.3 1.7 1.5 1.6 s; mean = 1.9s
- group1 > group2

Resampling

- If you had resampled 100 times, you would observe differences between 0.1s to 1.5s in the group means
- Now, simply count how many differences are greater or equal to the original 0.8s
  - In our case, 5 out of a 100
  - This is close to the theoretical p-Value of 0.046 we get from the proper t-test

How to choose the right test for your data



Independent-Measures or Single Sample

One sample

Two measurement categories

Binomial test

Two or more measurement categories

χ2-test for goodness of fit

Continuous measurement variables

Kolmogorov-Smirnov test

Two samples

Nominal

χ2-test test for independence

Ordinal

Mann-Whitney-U test

Interval

Independent sample t-test

More than two samples

Nominal

Chi-square test for independence

Ordinal

Kruskal-Wallis test

Interval

One-way ANOVA

Repeated Measures

Two treatment conditions per person

Nominal

Sign test

Ordinal

Wilcoxon test

Interval

Paired-sample t-test

More than two treatment conditions

Ordinal

Friedman test

Interval

Repeated-measures ANOVA