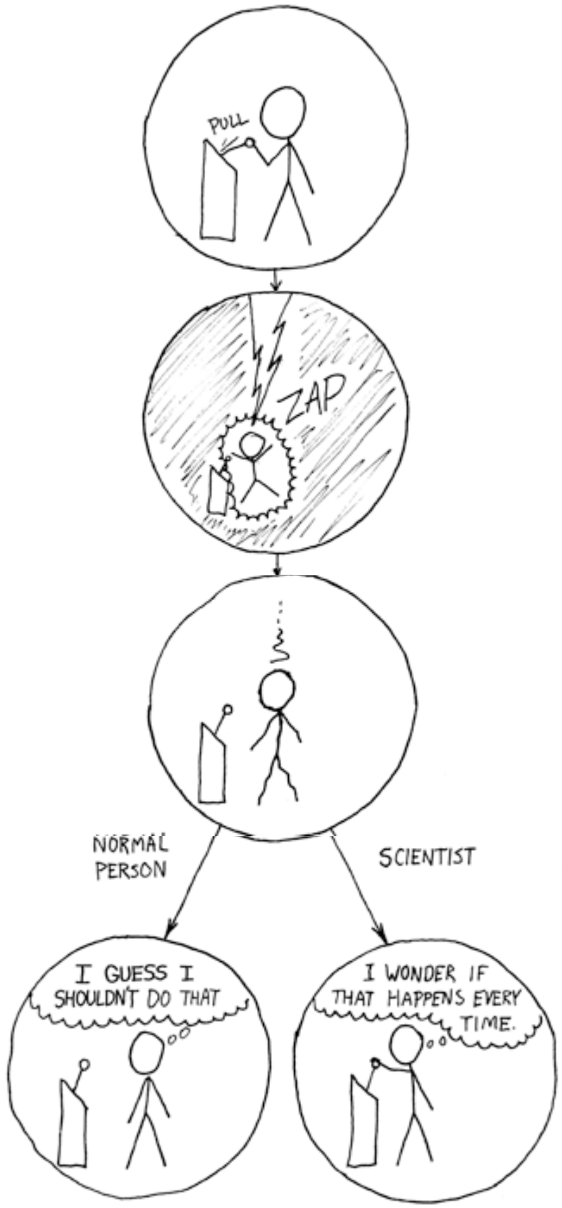
Design of Experiments

Psych 251

October 9th, 2017

Why do we do experiments?



To figure out how things work!

(in generalizable ways)

Why do we do experiments?

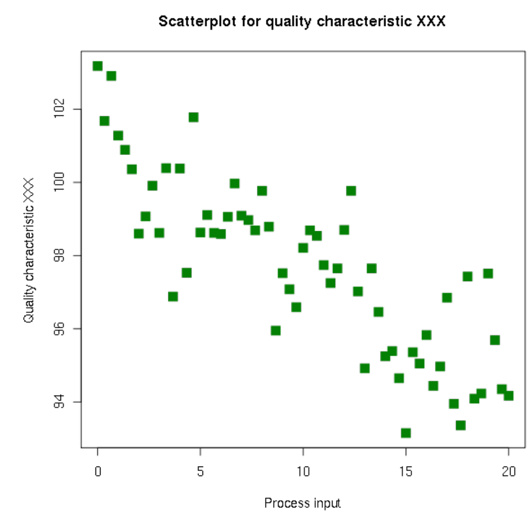
- Full discussion of this is outside the scope of this course
  - Philosophy of science issue
- Pragmatic approach to this issue: experiments help us measure theories of **causal relationships**
- Relationships between what?
  - In physics: abstractions like matter, energy, etc.
  - In psychology: abstract constructs like emotions, knowledge, language, etc.

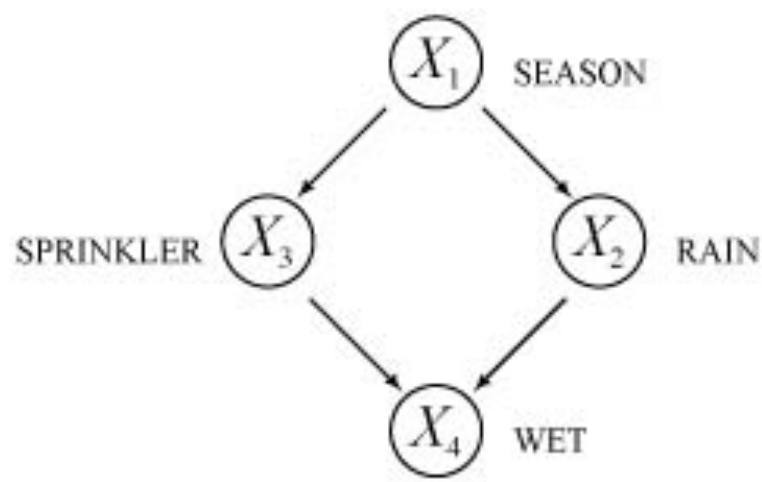
Operationalization and measurement

- The key skill of the psychologist is **operationalizing** theoretical abstractions: turning them into measurable entities
- How do we measure:
  - How angry you feel
  - How social a person you are
  - How well you know a word
  - How risk-seeking you are
  - How smart you are

Causal inference

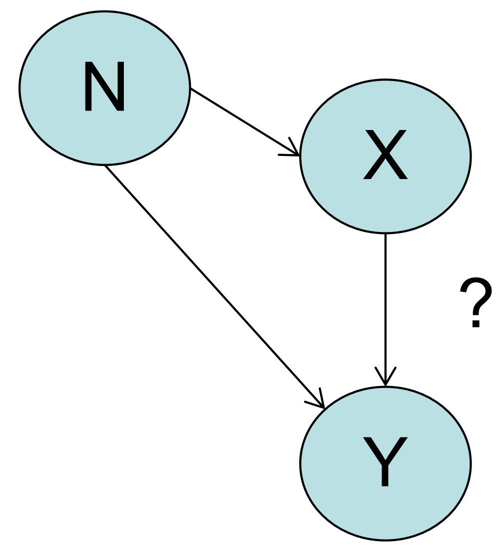
- Simplified version:
  - **Observation** allows for the measurement of correlations between two constructs
  - **Experiment** (with control) allows for an inference of causality

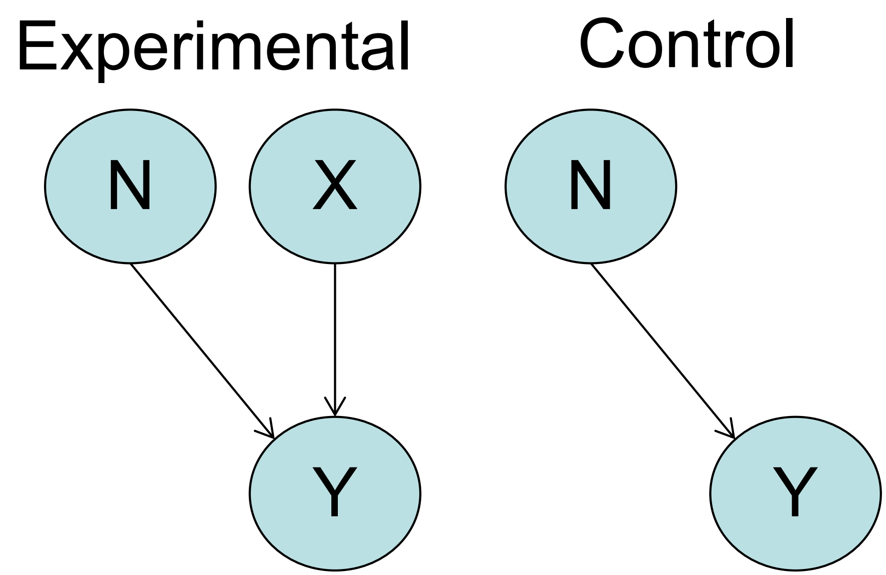




Causal inference (revised)

- Correlations provide evidence for a relationship
  - Without correlation there is no causation
- Experiments provide further evidence
  - by allowing many (not always all) variables to be held constant
  - Different designs allow control of different variables (today's topic)





Outline

- Design of experiments
  - Basic factorial designs
  - Within- vs. between-subjects designs
- Confounding and design
  - Basic counterbalancing
  - Latin square designs and randomization

Outline

- **Design of experiments**
  - **Basic factorial designs**
  - **Within- vs. between-subjects designs**
- Confounding and design
  - Basic counterbalancing
  - Latin square designs and randomization

But first…

- We did it this way out of adherence to folklore and superstition. #overlyhonestmethods
- All experimenters—and authors—were blind to the study's hypotheses. #overlyhonestmethods
- The first author didn't write this Methods section and doesn't understand half of it. #overlyhonestmethods
- Our sampling locations happen to match tropical resort towns because field work doesn't have to be mud and agony. #overlyhonestmethods
- We did experiment 2 because we didn't know WTF to make of experiment 1. #overlyhonestmethods

Terminology

- **DV = Response variable**
  - Measured output value
- **IV = Factors**
  - Input variables that can be changed
- **Levels**
  - Specific values of factors (inputs)
  - Can be continuous or discrete
- **Interaction**
  - **Effect** of one input factor depends on **level** of another input factor

Design of Experiments

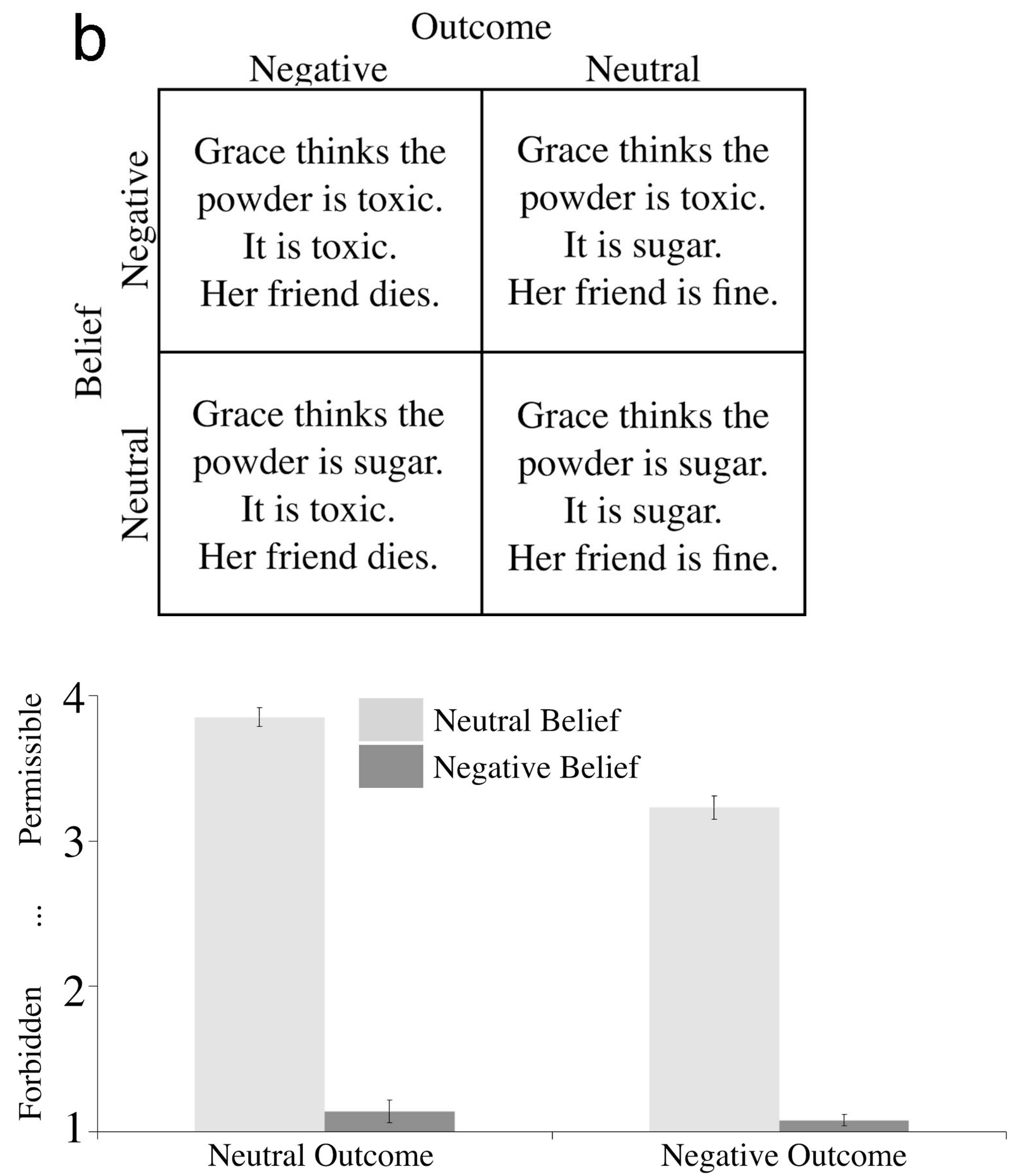
- Separates total variation observed in a set of measurements into
  - Systematic variability due to experimental manipulations
  - Variability to measurement error
- Goals
  - Isolate effects of each input variable
  - Determine effects of interactions
  - Determine magnitude of experimental error
  - Obtain maximum information for given effort

Note basic congruence between “ANOVA” framework and measurement framework!

Two-factor Experiments

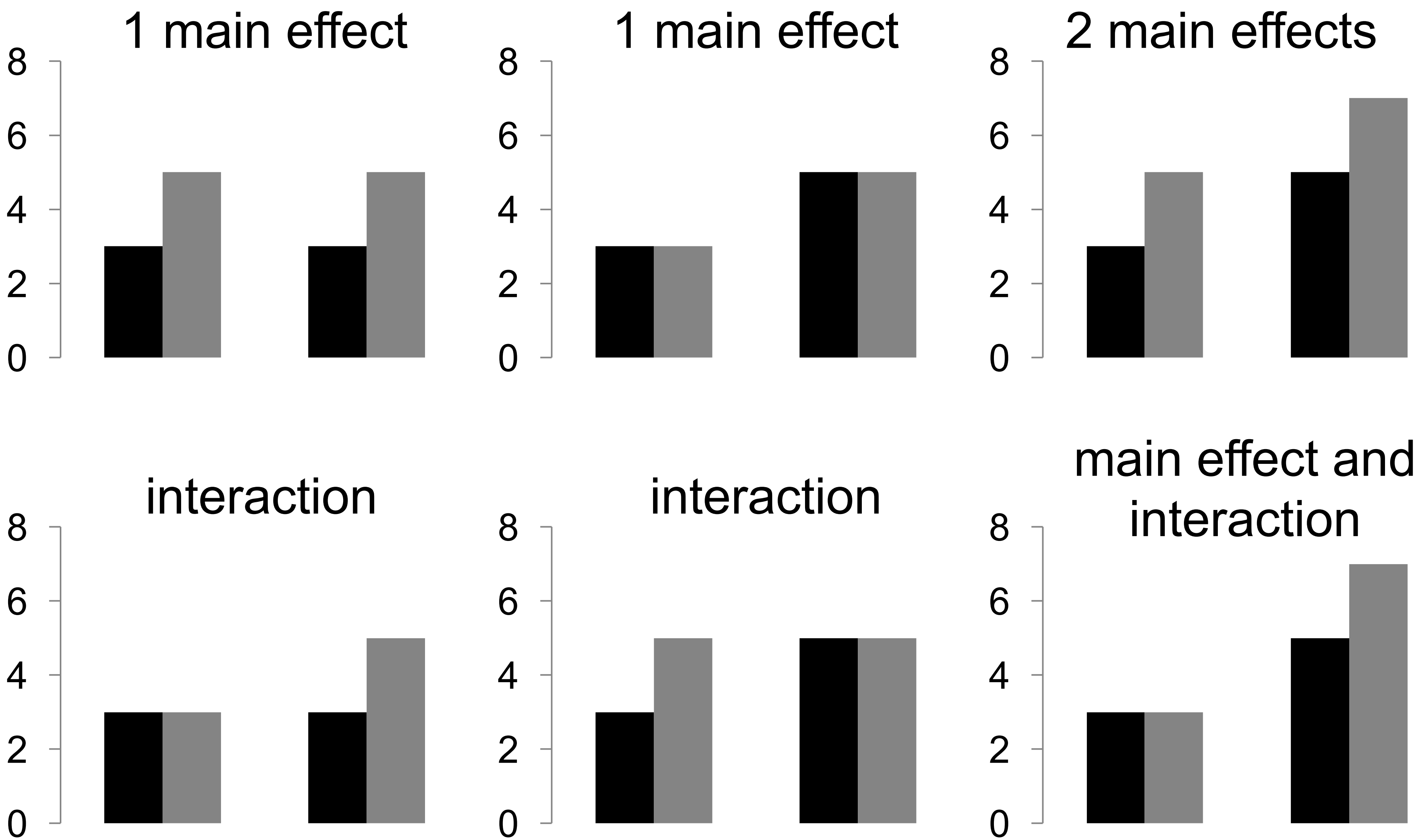
- Two factors (inputs)
  - A, B
- Separate total variation in output values into:
  - Effect due to A
  - Effect due to B
  - Effect due to interaction of A and B (AB)
  - Experimental error

Great example of such a design



Young et al. (2007)

Terminology for factorial designs



Generalized *m*-factor Experiments

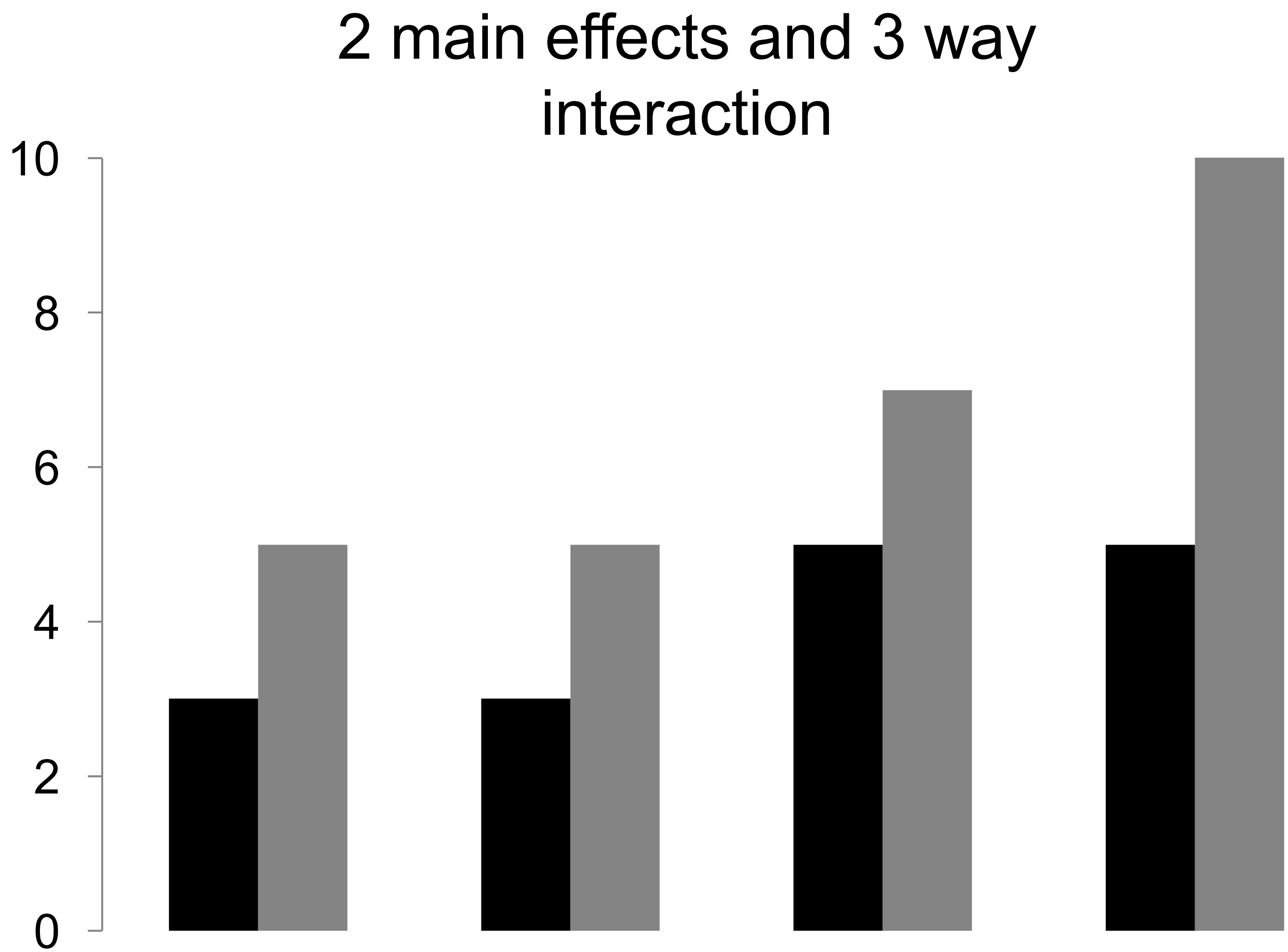
Effects for 3 factors:

A
B
C
A*B
A*C
B*C
A*B*C

Effects for 4 factors:

A
B
C
D
A*B
A*C
A*D
B*C
B*D
C*D
A*B*C
A*B*D
...

Higher-order interpretation



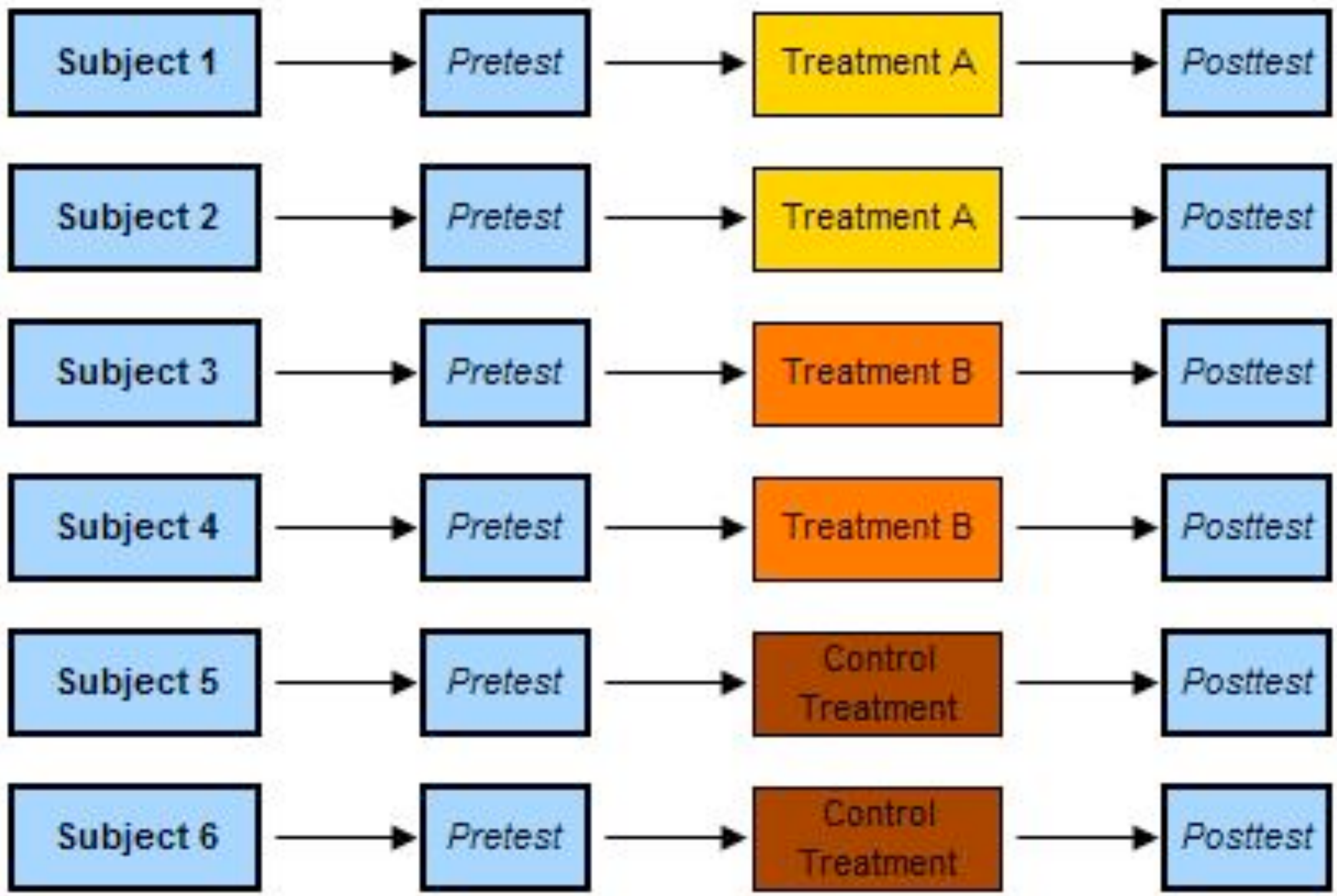
A Problem

- Full factorial design
  - Measure response with all possible input combinations
- *m* factors, *v* levels
  - v^m experiments
  - 5 input factors, 4 levels = 1,024 conditions!
  - Very difficult to interpret n-way interactions
- Solutions
  - Can choose only important factors/levels
  - Can vary separately and not estimate interactions
  - Many more innovative solutions in DOE literature

Between vs. within?

- **Between-subjects** design: participants only exposed to a single level of factors in the design
- **Within-subjects** design: participants exposed to multiple levels of factors in the design
- **Mixed** design: participants exposed to single level in some factors, multiple levels in other factors

Between-subjects

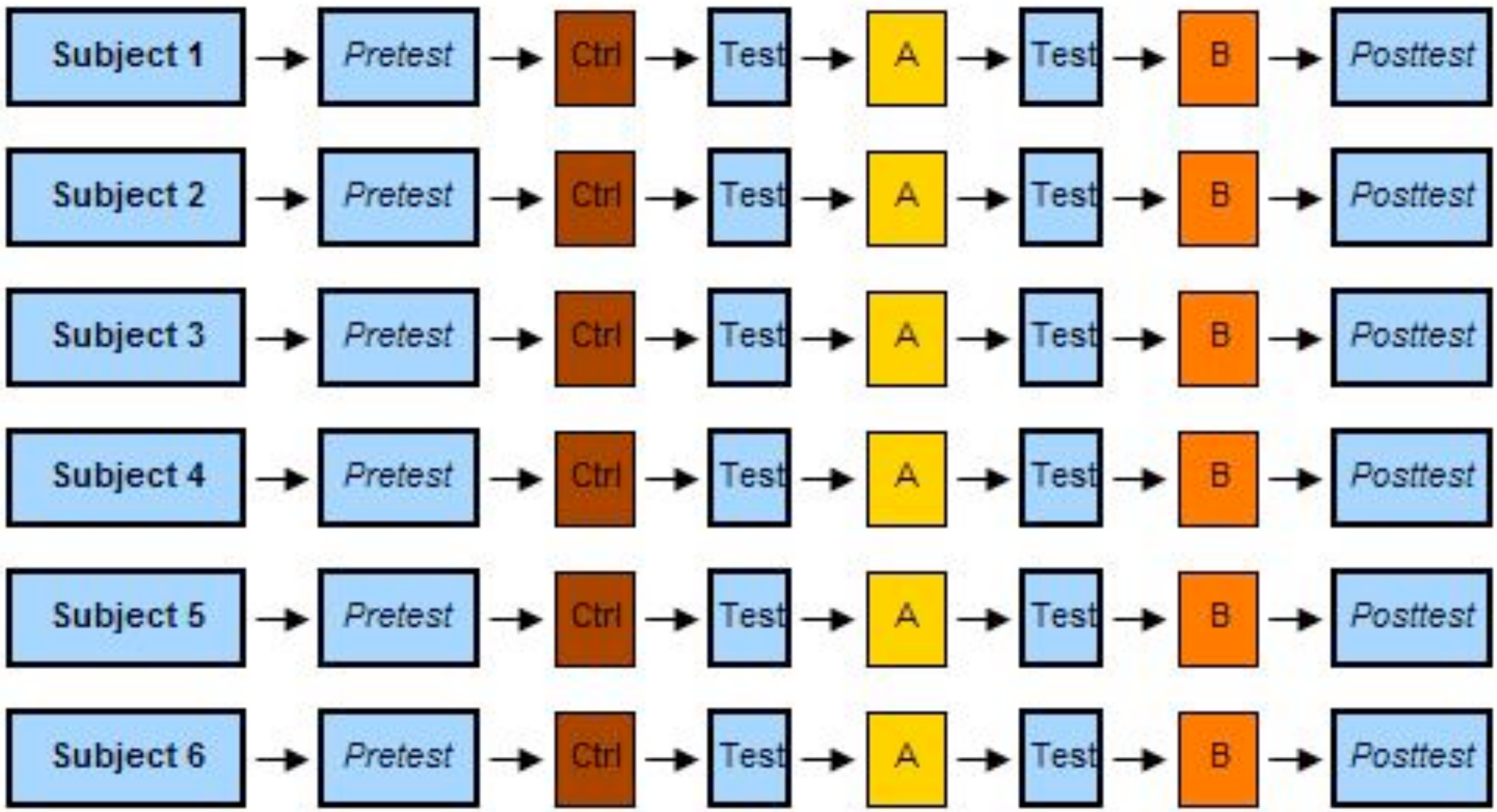


Between: pros & cons

- Main advantage
  - No contamination by other exposure to experimental materials
- Disadvantages
  - Requires many participants
  - Individual differences create a lot of variability in groups
  - **Assignment bias**: need to control for differences between groups
  - Other environmental group differences

Within-subjects

Sometimes called "repeated measures"



Within: pros & cons

- Main advantage
  - Eliminates subject variability
  - Relatively few participants needed, because of this lack of variability
- Disadvantages
  - Carryover effects mean that ordering of conditions can be problematic
  - Not always possible: imagine trying a within-subjects design for surgery

Note: when is it a factor?

- Generally we call something a factor if it's of interest to the study
- And an "item" e.g. when it's not
- Examples
  - "drug treatment vs. not" is definitely a factor
  - "scenario" is on the edge
  - "right vs. left for correct answer" is definitely not a factor

Case study – one expt or two?

Ann thinks that negation comprehension is affected by

- construction (“has no” / “doesn’t have”)
- context (expecting apples / oranges)
- referent (nothing vs. oranges)
- age of child

Questions:

- How many experiments should she do?
- Which factors should be between/within subjects?

Bob has no apples.

Outline

- Design of experiments
  - Basic factorial designs
  - Within- vs. between-subjects designs
- **Confounding and design**
  - **Basic counterbalancing**
  - **Latin square designs and randomization**

What is a confound?

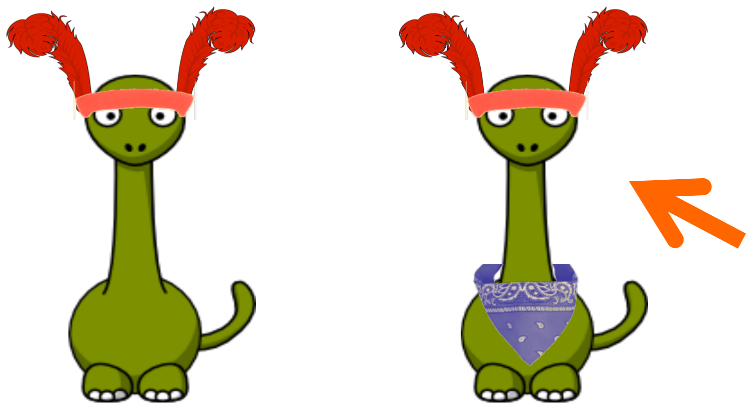
Examples?

Alternate cause that cannot be distinguished by experiment

Nuisance variable that also varies with the IV

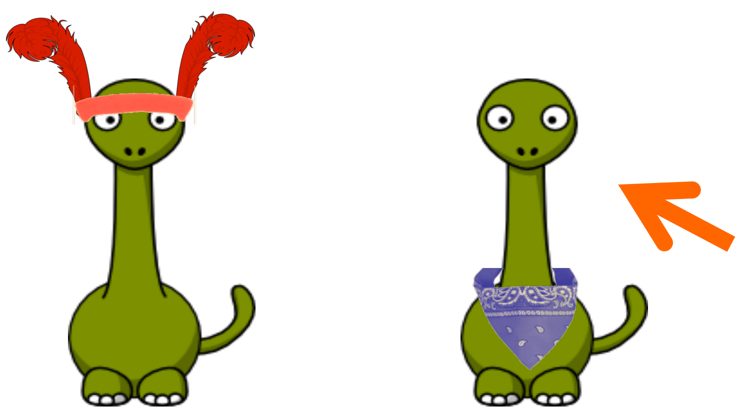
Simple mixed-design example

Inference trial

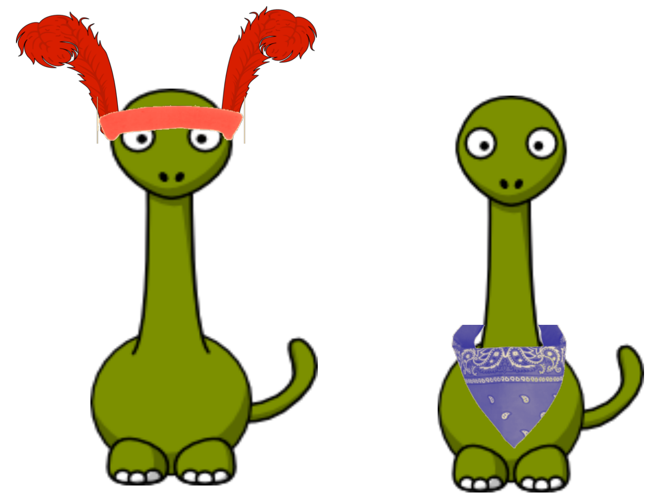


This dinosaur has a blicket!

Control trial



This dinosaur has a blicket!



Which of these has a blicket?

Within: inference/control
Between: age

Frank & Goodman (2014)

The confounds

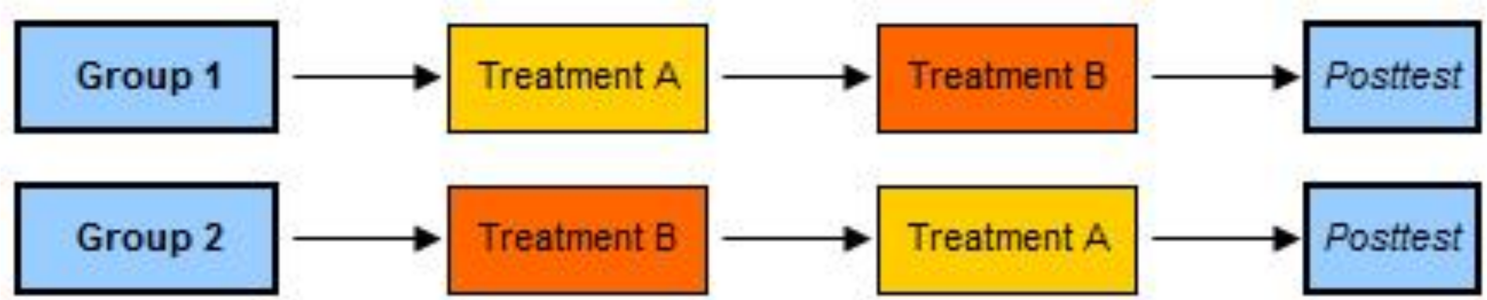
- How to tell that kids really learned the word “blicket”?
  - Could have had a preference for bandanas
  - Could have had a preference for the right-hand side at test
  - Could think that “blicket” sounds more like a bandana than a headdress
  - …

In general, need to control these

- We do this by **counterbalancing**:
  - Creating different conditions/orders/stimulus sets
  - That **deconfound** the nuisance variables
  - Ensuring that if an irrelevant variable is set one way for one subject, it is set another way for a different one

Simple example

Counterbalance for order



Dylan and reading example

Simple, elegant counterbalance

- Can learners take advantage of the statistical structure of utterances?

  golabupadotitupirobidakugolabu

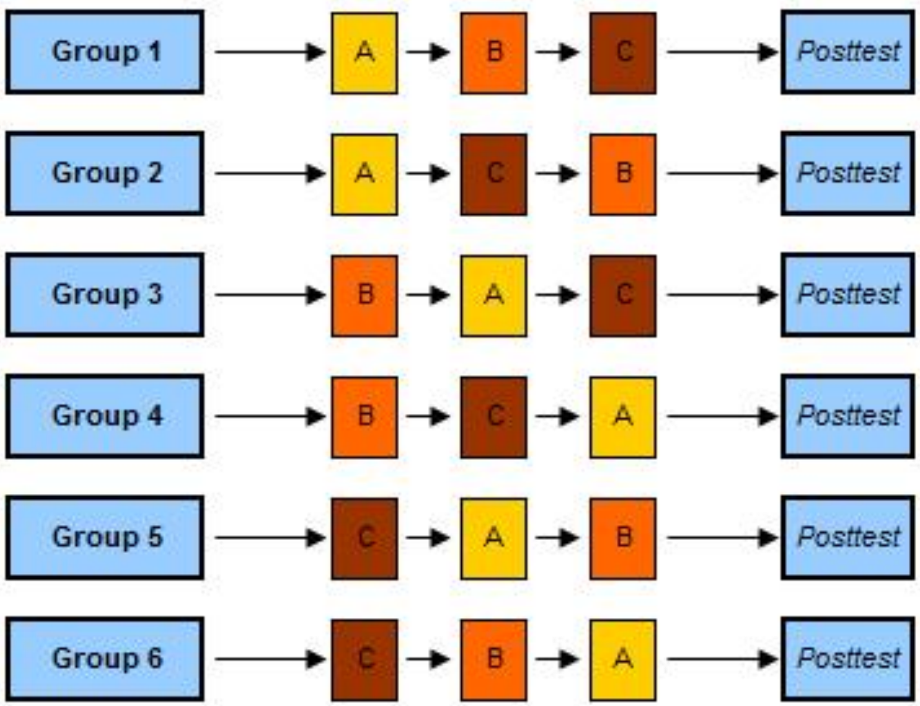
- Then test on golabu vs. bupado
- Have to counterbalance to make sure no golabu preference
- So create another language where

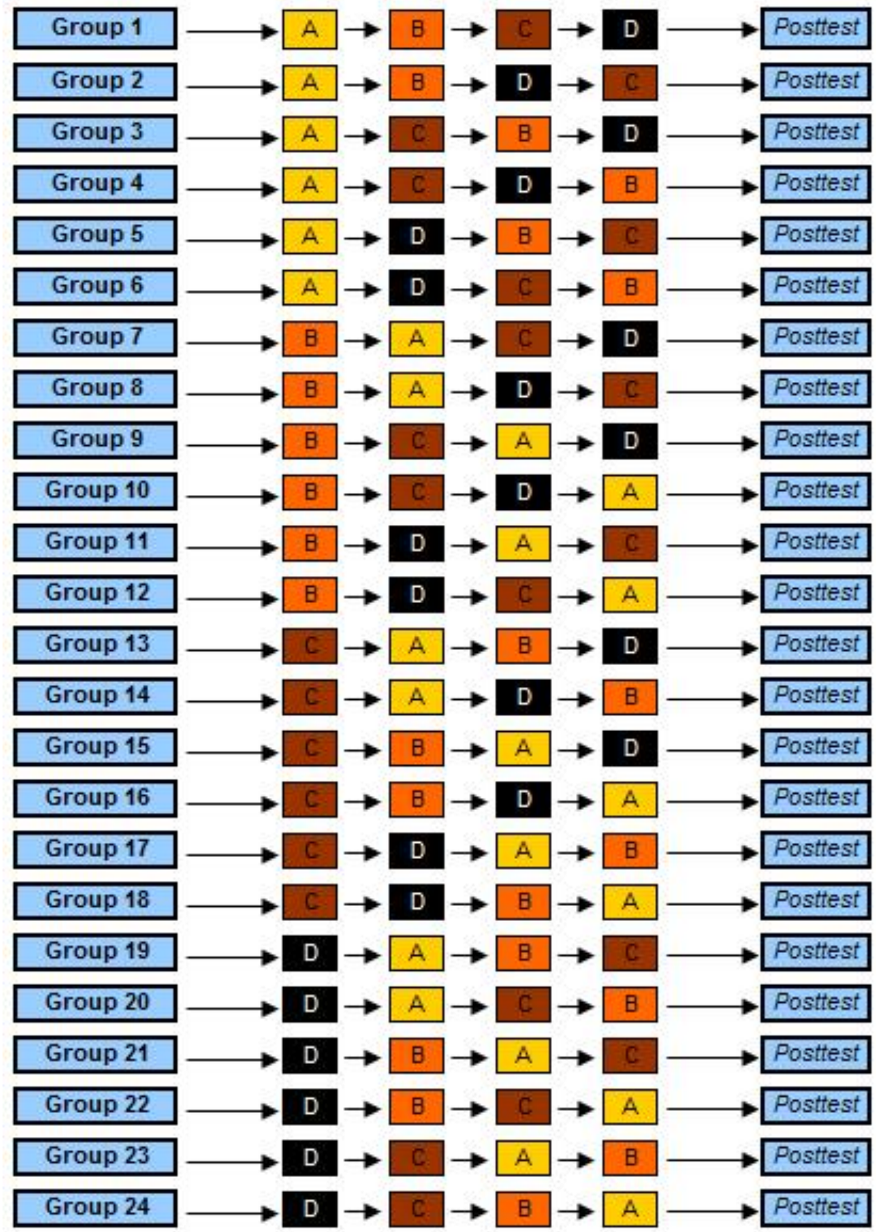
  bupadotitupirobidakugolabupado

- Then test on golabu vs. bupado

Saffran, Aslin, & Newport (1996)

Multiple factors





What can we do?

- Incomplete designs
  - Latin squares
  - Other incomplete designs
- Randomization
  - Interesting new alternative
  - Increasingly more practical as experiments are designed programmatically
  - Better with larger N (e.g. good for web expts.)

Latin squares for ordering

Latin square is an n × n array filled with n different symbols, each occurring exactly once in each row and exactly once in each column

	Position 1	Position 2	Position 3	Position 4	Position 5
Order 1	A	B	C	D	E
Order 2	B	C	D	E	A
Order 3	C	D	E	A	B
Order 4	D	E	A	B	C
Order 5	E	A	B	C	D

B always still follows A here, though

Balanced Latin Square

Subjects	1st	2nd	3rd	4th	5th	6th
A	1	2	6	3	5	4
B	2	3	1	4	6	5
C	3	4	2	5	1	6
D	4	5	3	6	2	1
E	5	6	4	1	3	2
F	6	1	5	2	4	3

Now every condition follows every other one once.
Hard to design very large Latin squares, though.

Incomplete designs

- Imagine counterbalancing item and side for 4 items
- For each item, 4 possible combos
  - Bandana, left
  - Bandana, right
  - Fascinator, left
  - Fascinator, right
- Can be in 24 orders
- If each combo treated separately
  - And each subject sees only one for each item
  - 24*23*22*21 orders!

Incomplete designs

- One possible solution: assume independence between dimensions
  - Side (R/L) counterbalanced separately from order
  - Label ("blicket") and feature ("bandana") also counterbalanced separately
- Principle: unlikely to have interactions between side and label
  - "blicket on the left" effect?

Incomplete designs

Use Latin Square to create orders

Dino	Robot	Bear	Rocket
Rocket	Dino	Robot	Bear
Bear	Rocket	Dino	Robot
Robot	Bear	Rocket	Dino

Then for each order, assign features so that they are counterbalanced

One possible method

Order	Trial	Item	Train Side	Item
1	1	Dino	L	A
1	2	Robot	R	B
1	3	Bear	L	A
1	4	Rocket	R	B
2	1	Rocket	L	B
2	2	Dino	R	A
2	3	Robot	L	B
2	4	Bear	R	A
3	1	Bear	L	B
3	2	Rocket	R	A
3	3	Dino	R	B
3	4	Robot	L	A
4	1	Robot	R	A
4	2	Bear	R	B
4	3	Rocket	L	A
4	4	Dino	L	B

Randomization

- "golabupadotitupirobidakugolabu" method does not generalize to larger languages (and construction is difficult and time-consuming)
- A simple random algorithm
  - Pick 12 syllables
  - Randomly string them together
  - Split them into four, three-syllable words
  - String them together randomly into sentences
  - Compare words to non-words

Randomization pro/con

- Pro: Simple and easy rule
- Pro: Viable for very complex designs
- Con: Factors will not be exactly balanced
  - Smaller samples may have significant balance issues
  - Meaning you are betting these factors are not large enough to need to be balanced
- Con: Completely random assignment to condition can backfire
  - The law of large numbers is your friend
  - 4 conditions, 32 subjects:
    - [9 7 9 7], [7 11 7 7], [8 11 10 3]
  - 4 conditions, 100 subjects:
    - [22 24 22 32], [31 25 20 24], [24 27 27 22]