experimental design for linguists

HPSG 2012

PHILIP HOFMEISTER
UNIVERSITY OF ESSEX

EXPERIMENTS

- Many types of linguistic research involve a form of experimentation

EXPERIMENTS

- Linguistic intuitions about syntax, semantics, prosody
- Eliciting responses to questions and questionnaires
- Testing pronunciation
- Etc.

EXPERIMENTS

- These ‘informal’ methods diverge from common practices in other social sciences (e.g. cognitive psychology, sociology)

CHOMSKY (2011)

> A related criticism, also widespread, is that linguistic research resorts to idealization and abstraction, relying on invented examples as in the cases I mentioned, not keeping to unanalyzed data but rather creating evidence by design.

CHOMSKY (2011)

> A related criticism, also widespread, is that linguistic research resorts to idealization and abstraction, relying on invented examples as in the cases I mentioned, not keeping to unanalyzed data but rather creating evidence by design. In other words, linguistic research is like the sciences generally. The sciences typically rely on experiments, highly idealized and abstract, and theory-internal—and even on thought experiments, including classic discoveries.

CHOMSKY (2011)

(1) He wondered whether the mechanics fixed the cars.
(2) How many cars did he wonder whether the mechanics fixed? (answer, "3 cars")
(3) How many mechanics did he wonder whether fixed the cars? (answer, "3 mechanics")

- "Sentences (2) and (3) clearly differ in status: unlike (2), (3) is severely deviant"

CHOMSKY (2011)

> A related criticism, also widespread, is that linguistic research resorts to idealization and abstraction, relying on invented examples as in the cases I mentioned, not keeping to unanalyzed data but rather creating evidence by design. In other words, linguistic research is like the sciences generally. The sciences typically rely on experiments, highly idealized and abstract, and theory-internal—and even on thought experiments, including classic discoveries.
>
> The observation about (1)–(3) is an experiment, much like the study of perceptual illusions,the foundation of much perceptual psychology. One might argue that better experimentation is required in this and other cases—though in reality the facts are so clear in this case that an experiment would be a test of the experiment, not an investigation of the facts: as any scientist knows, it is easy to design experiments that yield noise and hard to design ones that yield meaningful results, a task that often requires determining whether the experimental method proposed gives the right results in clear cases.

WHY BOTHER?

- Experimenter bias

WHY BOTHER?

- Experimenter bias
- Variation in participants and items

WHY BOTHER?

- Experimenter bias
- Variation in participants and items
- Scientific community standards

WHY BOTHER?

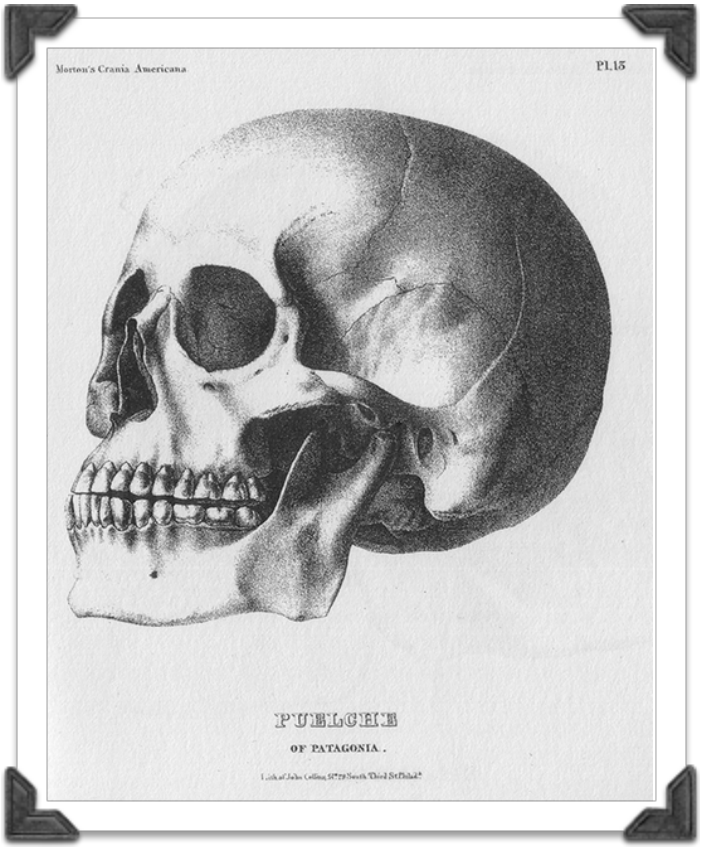
- Experimenter bias
- Variation in participants and items
- Scientific community standards
- Intuitions may be insensitive to various processes

WHY BOTHER?

- Experimenter bias
- Variation in participants and items
- Scientific community standards
- Intuitions may be insensitive to various processes
- Replicability & posterity

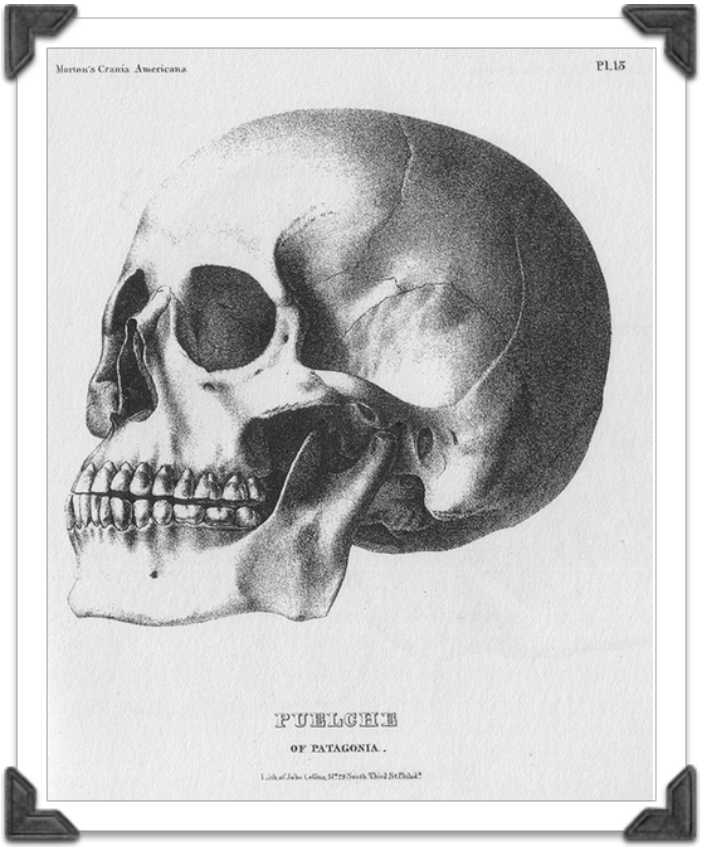
EXPERIMENTER BIAS

- Craniometry (Samuel George Morton)



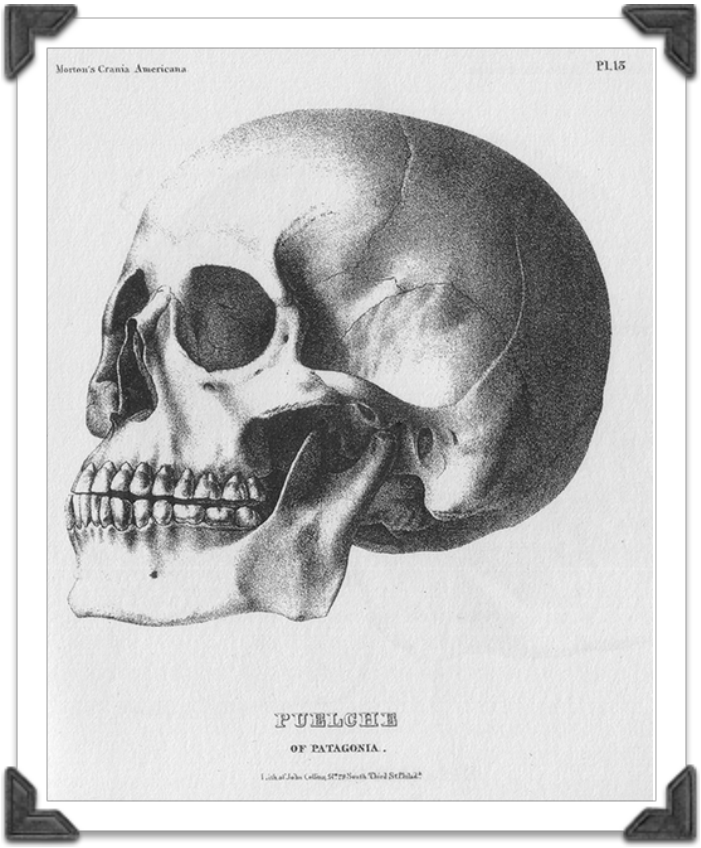
EXPERIMENTER BIAS

- Craniometry (Samuel George Morton)
  - Assumption that cranium size correlated with intelligence



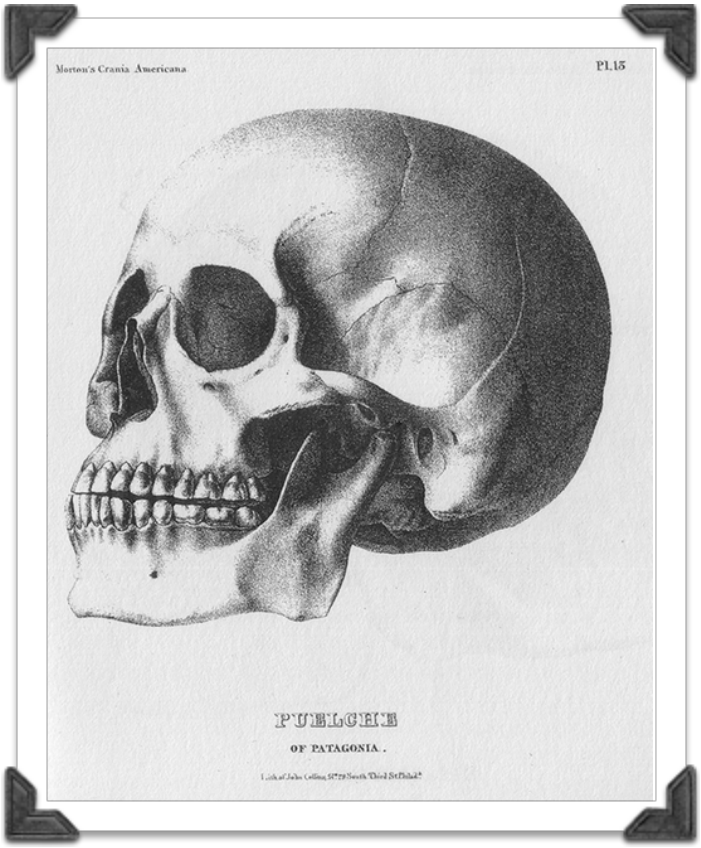
EXPERIMENTER BIAS

- Craniometry (Samuel George Morton)
  - Assumption that cranium size correlated with intelligence
  - Measured the quantity of BBs the skull would hold



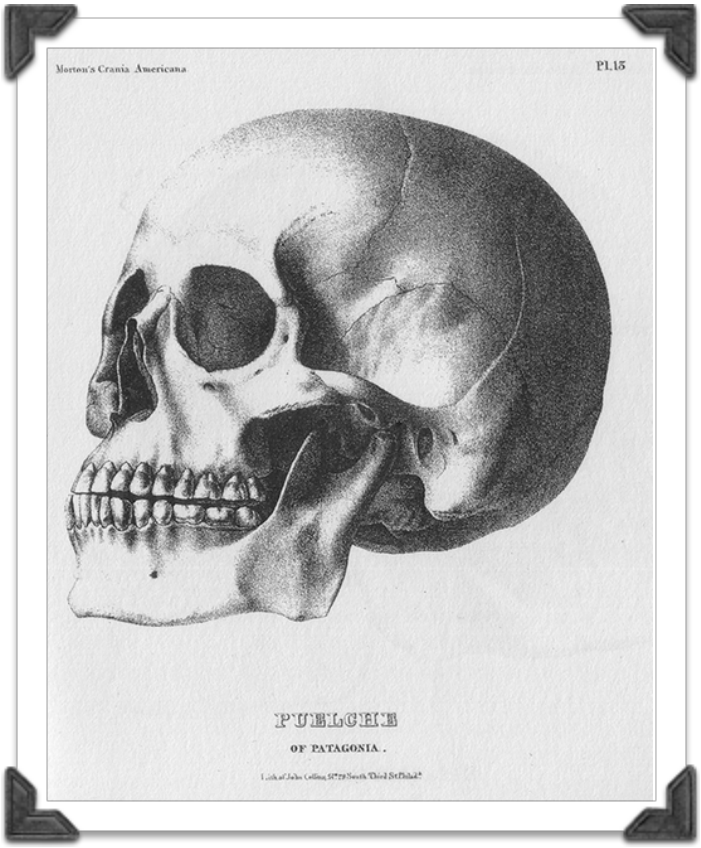
EXPERIMENTER BIAS

- Craniometry (Samuel George Morton)
  - Assumption that cranium size correlated with intelligence
  - Measured the quantity of BBs the skull would hold
  - Concluded that the English & Germans were more intelligent than Jewish people who were more intelligent than Hindus



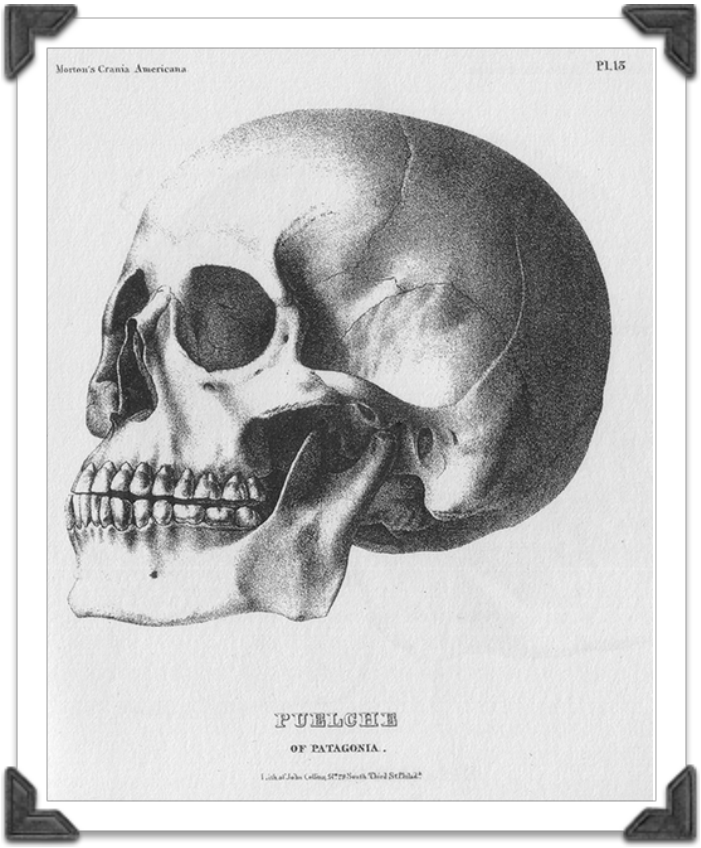
EXPERIMENTER BIAS

- Craniometry (Samuel George Morton)
  - Assumption that cranium size correlated with intelligence
  - Measured the quantity of BBs the skull would hold
  - Concluded that the English & Germans were more intelligent than Jewish people who were more intelligent than Hindus
  - Stephen Jay Gould suggested that Morton had "unconsciously" selected his samples in a way to confirm his hypothesis



EXPERIMENTER BIAS

- Lewis et al (2011)
  - It was actually Gould who was biased
  - Gould didn't (re-) measure any of Morton's original skulls
  - Gould reports erroneous values and suggested the existence of computational errors that did not exist



EXPERIMENTER BIAS

- Some types of experimenter bias in linguistics

EXPERIMENTER BIAS

- Some types of experimenter bias in linguistics
  - Stimulus framing

EXPERIMENTER BIAS

- Some types of experimenter bias in linguistics
  - Stimulus framing
  - "Repeated exposure" effects

EXPERIMENTER BIAS

- Some types of experimenter bias in linguistics
  - Stimulus framing
  - "Repeated exposure" effects
  - Linguistic authority

EXPERIMENTER BIAS

- Some types of experimenter bias in linguistics
  - Stimulus framing
  - "Repeated exposure" effects
  - Linguistic authority
  - Ignoring / dismissing contradictory evidence

EXPERIMENTER BIAS

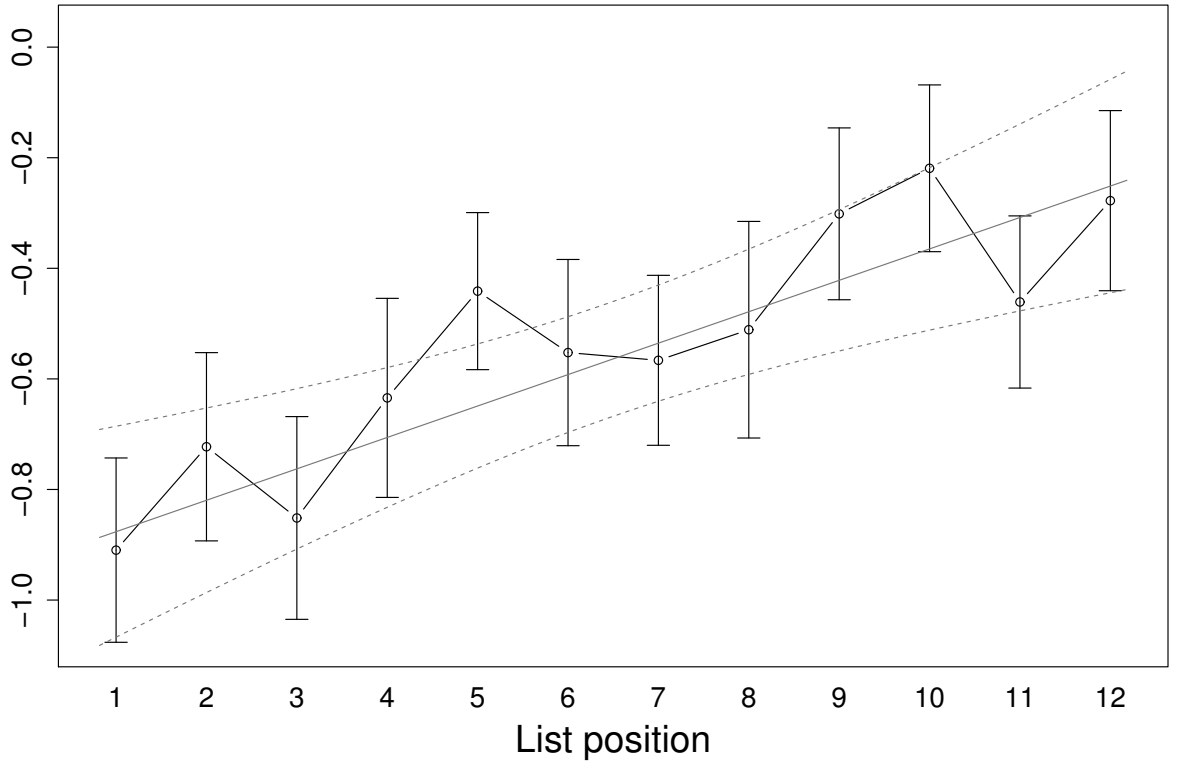
- Stimulus framing
  - "Hey, tell me this is grammatical!"
  - "You're my friend, and if you want to stay that way, I think you'll agree you can't say this in English . . ."

EXPERIMENTER BIAS

- Repeated exposure effects even in extremely "ungrammatical"utterances
  - *Iran has gun-control strict laws that bar people from private firearms carrying*

EXPERIMENTER BIAS

- Repeated exposure effects even in extremely "ungrammatical" utterances
  - *Iran has gun-control strict laws that bar people from private firearms carrying*



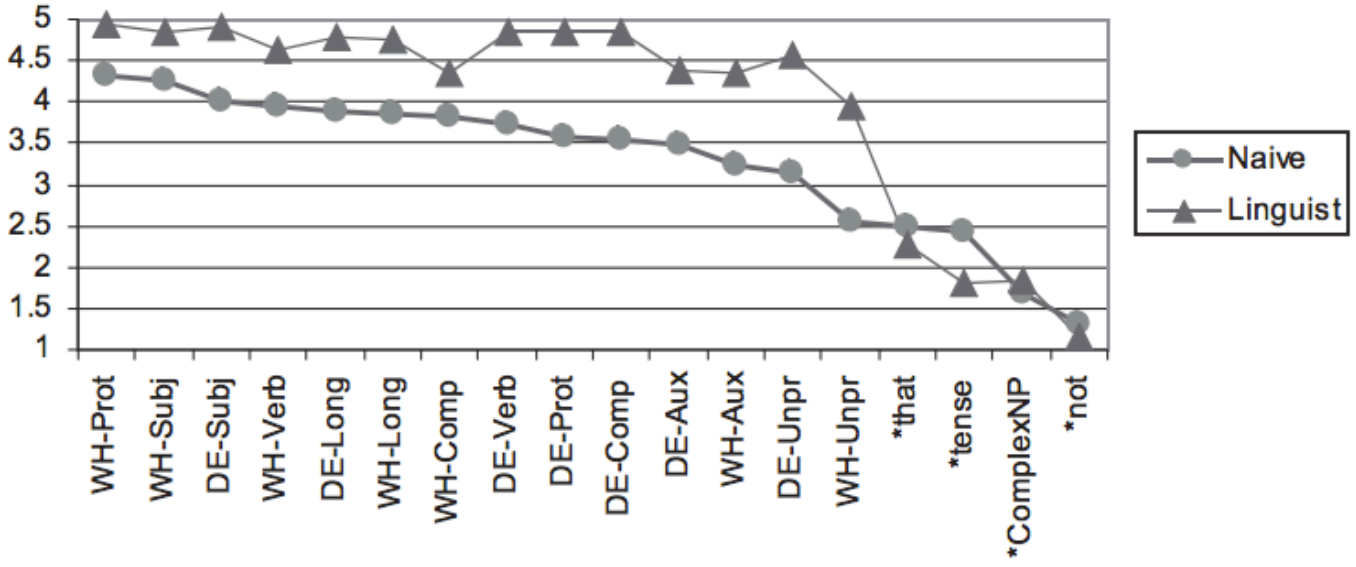
EXPERIMENTER BIAS

Some speakers seem to accept such forms as *What did he wonder whether John saw? What crimes did he wonder how they solved?* For me, these are unacceptable. It would be possible to add special rules to allow for these examples by a complication of the particular grammar, given the suggested interpretation of the conditions. (Chomsky 1973: 244)

GIBSON, PIANTADOSI, & FEDORENKO, IN PRESS

“We view expert linguistic judgments as expert *predictions* . . .”

EXPERT VS. 'NAIVE' KNOWLEDGE (DABROWKSA 2010)



RANDOM SAMPLING

- Assumption of external validity
  - The participants and items you test represents a random sample
  - Non-random samples decrease the likelihood that results will *generalize*

RANDOM SAMPLING

- Assumption of external validity
  - Language research confronts the problem of random sampling of language
  - In making materials, high frequency words probably come to mind first, as well as the ubiquitous John & Mary
  - It may be hard to imagine appropriate examples

SENSITIVITY

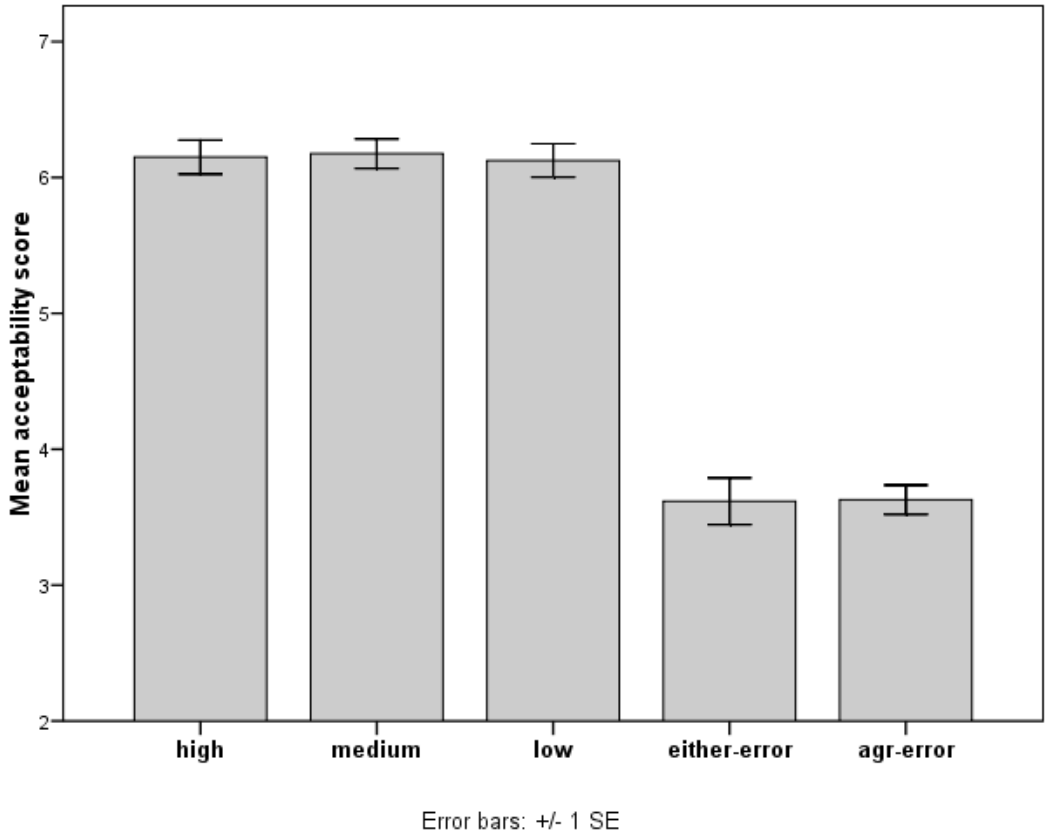
- Expert intuitions may be either too sensitive or too insensitive, especially in subtle contrasts

DEN DIKKEN (2006)

- *John either said that he would eat rice or beans.
- John said that he would eat either rice or beans.

ATTESTED EXAMPLE

- There are a lot of people who either think that Iraq was a doable proposition that was botched or a project destined for failure.



- There are a lot of people who <either> think that <either> Iraq was <either> a doable proposition that was botched <either> or a project destined for failure.

NOBODY'S PERFECT

- Superiority violations improve with a third wh-phrase (Bolinger 1978; Kayne 1983)
  - **Julius tried to remember what who carried.*
  - *Julius tried to remember what who carried when.*

FEDORENKO & GIBSON (2010)

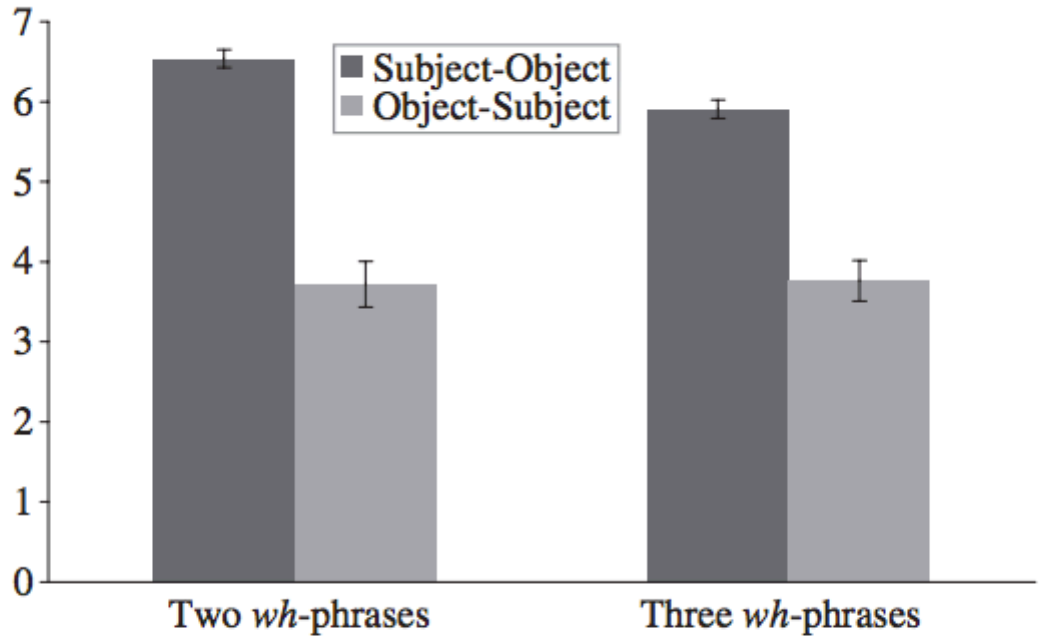
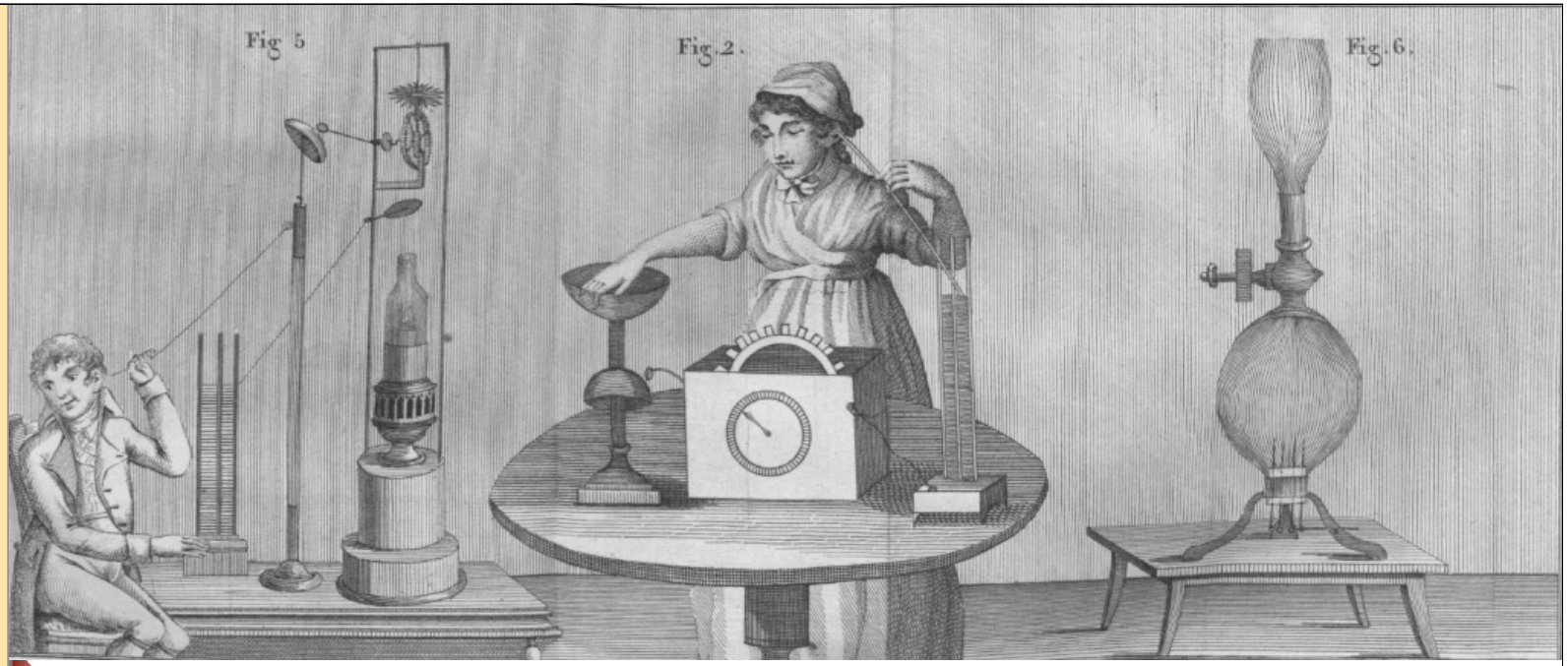


Figure 1: Acceptability ratings as a function of the *wh*-phrase order and the number of *wh*-phrases. The error bars indicate standard errors of the mean.

SURPRISES: GIBSON & THOMAS (1999)

- Center-embedded sentences with a verb missing is more acceptable than its grammatical counterpart
  - *The apartment that the maid who the service had sent over was cleaning every week was well-decorated*
  - **The apartment that the maid who the service had sent over was well-decorated.*



REPLICABILITY

- An advantage of formal experiments is that a recipe accompanies the data

STANDARDS IN THE SOCIAL SCIENCES

> . . . there is no other field of science where the intuitions of the investigators are treated as admissible data for evaluating theories . . . Science, in short, seeks objectivity

A COUNTER-ARGUMENT

- Expert accuracy
  - Expert judgments for individual contrasts are replicated with a high degree of success (Sprouse & Almedia, in press; Sprouse & Almedia submitted)

A COUNTER-ARGUMENT

- Expert accuracy
  - 98% of judgments from Adger's Core *Syntax* confirmed
  - 95% of phenomena from *LI 2001-2010* replicated

- Replicating individual data points does not increase generalization
  - *What does John doubt whether you bought?*
  - *What does John doubt that you bought?*

- While there are many beneficial aspects to child adoption, there are a number of disadvantages that you should consider and decide whether you are comfortable with before committing time, energy and resources to the process.
- Insul-knife is one of those time-saving tools that you will wonder how you ever lived without.

STANDARDS IN THE SOCIAL SCIENCES

- To be clear, the message here is NOT that every single judgment contrast needs to be tested experimentally

STANDARDS IN THE SOCIAL SCIENCES

- *The was student arrested
- The student was arrested.

GIBSON, PIANTADOSI, AND FEDORENKO (IN PRESS)

“

- the relevant examples are ones that can *distinguish among current theories*

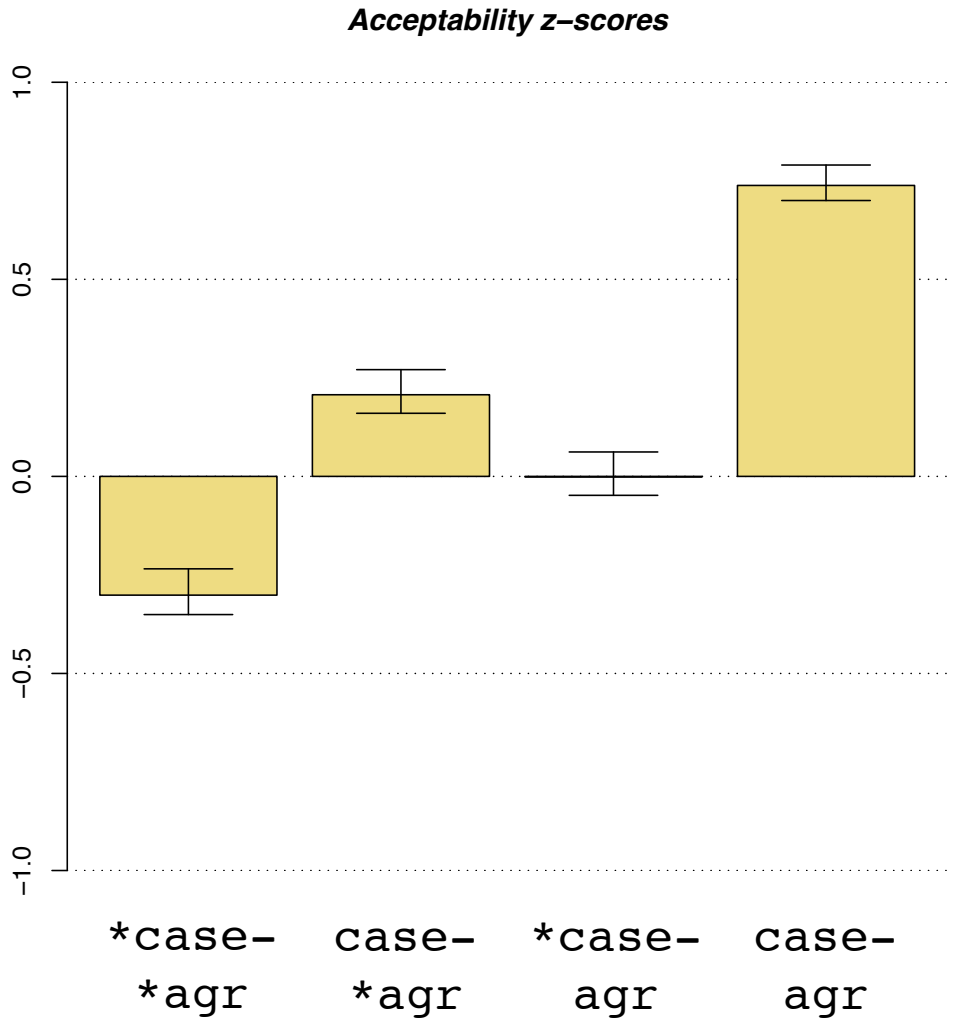
”

INFORMATIONAL RICHNESS

- While formal and traditional methods of experimentation may often lead to the same conclusions, formal methods produce richer databases of information

INFORMATIONAL RICHNESS

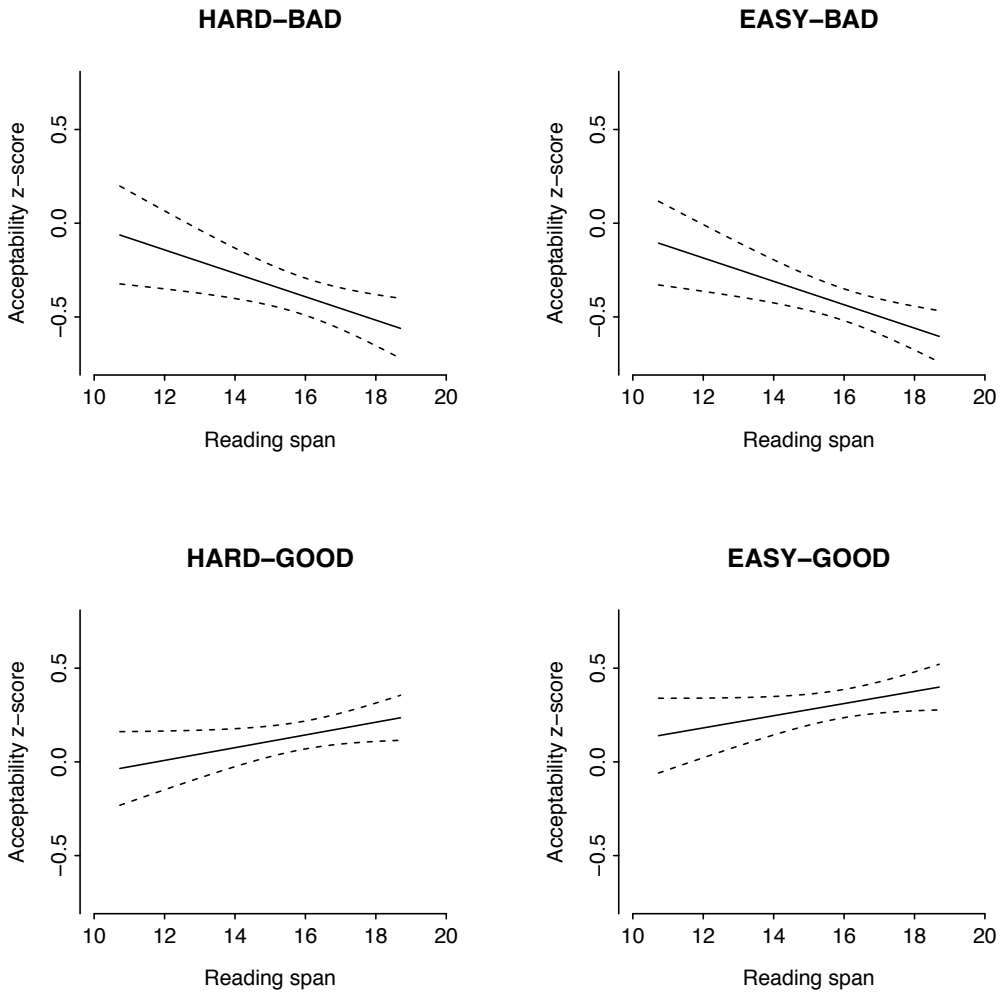
- Gradient effects of different types of violations



HPSG2012

INFORMATIONAL RICHNESS

Judgments and other response types vary with participant and item variables



SUMMARY

- There are numerous reasons to opt for formal experiments where possible:

SUMMARY

- There are numerous reasons to opt for formal experiments where possible:
  - objectivity

SUMMARY

- There are numerous reasons to opt for formal experiments where possible:
  - objectivity
  - replicability & posterity

SUMMARY

- There are numerous reasons to opt for formal experiments where possible:
  - objectivity
  - replicability & posterity
  - sensitivity

SUMMARY

- There are numerous reasons to opt for formal experiments where possible:
  - objectivity
  - replicability & posterity
  - sensitivity
  - scientific standards

SUMMARY

- There are numerous reasons to opt for formal experiments where possible:
  - objectivity
  - replicability & posterity
  - sensitivity
  - scientific standards
  - information richness

SUMMARY

- There are numerous reasons to opt for formal experiments where possible:
  - objectivity
  - replicability & posterity
  - sensitivity
  - scientific standards
  - information richness
  - increasingly easy and cheap

OUTLINE FOR TODAY

OUTLINE FOR TODAY

- Basics of experimental design

OUTLINE FOR TODAY

- Basics of experimental design
- Experimental control

OUTLINE FOR TODAY

- Basics of experimental design
- Experimental control
- Understanding your data

OUTLINE FOR TODAY

- Basics of experimental design
- Experimental control
- Understanding your data
- Experimenting with acceptability judgments

OUTLINE FOR TODAY

- Basics of experimental design
- Experimental control
- Understanding your data
- Experimenting with acceptability judgments
- Mechanical Turk

Experimental Design

STAGES

HPSG2012

STAGES

- Develop a simple hypothesis

STAGES

- Develop a simple hypothesis
- Select dependent and independent variable(s)

STAGES

- Develop a simple hypothesis
- Select dependent and independent variable(s)
- Select design type

STAGES

- Develop a simple hypothesis
- Select dependent and independent variable(s)
- Select design type
- Control materials and check for confounds

STAGES

- Develop a simple hypothesis
- Select dependent and independent variable(s)
- Select design type
- Control materials and check for confounds
- Visualize and analyze data

HYPOTHESIS TESTING

- Keep hypotheses simple; build upon your work progressively

- Let's take a few examples:
  - Superiority violations
  - Resumptive pronouns

SUPERIORITY (KUNO & ROBINSON 1972; CHOMSKY 1973)

- I know who read what.
- *I know what who read.

SUPERIORITY (KARTUNNEN 1977; PESETSKY 1987)

- *I know what who read.
- I know what <u>which student</u> read.
- I know <u>which book</u> who read.

HYPOTHESIS TESTING

- Determine both the null hypothesis (H_0) & positive hypothesis in advance of the study
  - H_0 : There is no difference between bare wh-words and complex wh-phrases in Superiority violations
  - H_1 : Complex wh-phrases raise the acceptability of Superiority violations

- Resumptive pronouns
  - There was a prisoner that the guard helped him/___ to make a daring escape.
  - There was a prisoner that the officials confirmed that the guard helped him/___ to make a daring escape.

HYPOTHESIS TESTING

- Determine both the null hypothesis (H_0) & positive hypothesis in advance of the study
  - H_0 : The acceptability difference between resumptives & gaps does not differ with levels of embedding
  - H_1 : The acceptability difference between resumptives & gaps differs with levels of embedding

HYPOTHESIS TESTING

- Compare hypotheses that make opposite predictions
- It's much more challenging to test hypotheses that make predictions in the same direction but to differing degrees

- Developing a hypothesis will often give you an idea of what you want to manipulate and what cognitive outcome you want to assess

OPERATIONAL DEFINITIONS

- Often, the process we want to measure has to be operationalized
  - processing difficulty is operationalized as the time it takes to press a button and move to the next stimulus
  - grammaticality is operationalized as a rating on a scale or an up-or-down vote

PITFALLS

- Because we are operationalizing, we do not have a direct window onto a cognitive process

- Self-paced reading methodology

- Self-paced reading methodology

This -- --- -- -----

- Self-paced reading methodology

---- is --- -- -----

- Self-paced reading methodology

---- -- how -- -----

- Self-paced reading methodology

---- -- --- it -----

- Self-paced reading methodology

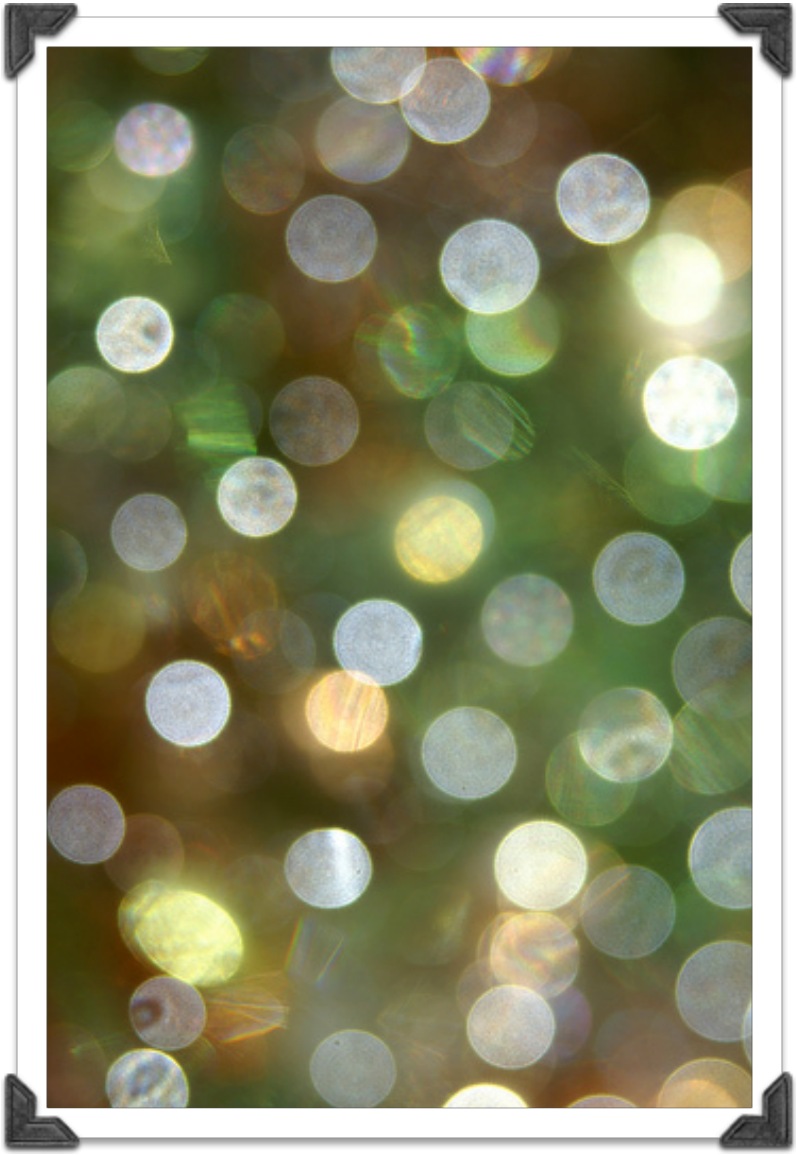
---- -- --- -- looks

PITFALLS

- Self-paced reading methodology
  - The time intervals between button presses are taken as an indication of processing difficulty

PITFALLS

- Assumption that delay = difficulty
  - Other things could affect button presses
    - Rhythmic responses
    - Coordination
    - Fatigue
    - Something shiny



JUDGMENTS

- Often, linguistic acceptability is operationalized as a choice between *, **, ?, #, or ✓

JUDGMENTS

- Differences between these levels are not necessarily equivalent
- Diacritics probably do not correspond to any particular cognitive states

TERMS

- Factors = Variables to be manipulated
  - e.g. grammaticality
- Items = clusters of minimally different conditions
- Conditions = levels of factors
- Trial = ordered event in experiment
- Lists = Sets of trials to be shown to participants

TERMS

1a) I know what who ordered.

1b) I know which drink which customer ordered.

- Factor = Superiority
- Conditions = a & b
- Item = {a, b} = 1
- Trial = e.g.1a appears as the 54 sentence

DESIGN TYPES

- Between subject
- Within subject

BETWEEN SUBJECTS

- Participants are grouped according to condition
  - ex. How does rate of compensation affect judgments?
    - Group 1 = $.01/correct answer
    - Group 2 = $.25/correct answer

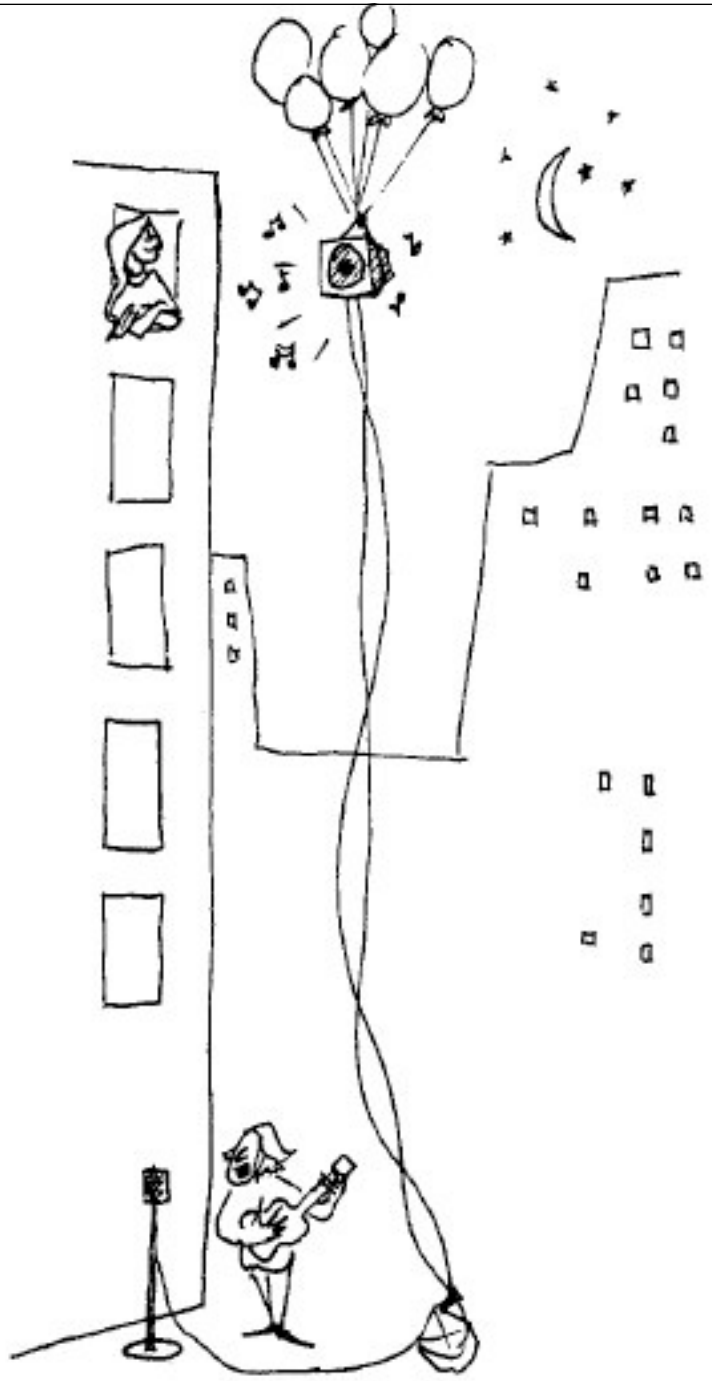
BETWEEN SUBJECT DESIGNS

Bransford & Johnson (1982)

If the balloons popped, the sound wouldn't be able carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying, since most buildings tend to be well-insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course, the fellow could shout but the human voice is not loud enough to carry that far.

Participants asked to rate for comprehensibility and later asked to recall what they had read

BETWEEN SUBJECT DESIGNS



Bransford & Johnson (1982)

If the balloons popped, the sound wouldn't be able carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying, since most buildings tend to be well-insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course, the fellow could shout but the human voice is not loud enough to carry that far.

Participants asked to rate for comprehensibility and later asked to recall what they had read

WITHIN SUBJECTS

- The same subject is used in different experimental conditions; each subject is in every group
- Subject should be exposed to an equal # of each condition

BETWEEN OR WITHIN?

- Advantages of within-subjects design
  - All groups are equal on every factor
  - Lower # of subjects typically required
  - Greater sensitivity to treatment effects

BETWEEN OR WITHIN?

- Disadvantages of within-subjects design
  - Participants may notice manipulations more easily
  - Intermixed conditions can have unanticipated effects on each other

BETWEEN OR WITHIN?

- Long-lasting effects of conditions
  - e.g. syntactic priming
    - How does a syntactic prime (e.g. an NP PP sentence) affect how a participant produces a subsequent dative sentence?
    - Hypothesis: A syntactic prime of the form NP PP or NP NP will bias the speaker towards producing a similar form
      - `Prime: gave the bottle to him – gave him the bottle`

BETWEEN OR WITHIN?

- Long-lasting effects of conditions
  - Participants' output (NP PP / NP NP construction) may influence the output on the next item
  - NP PP response could bias next response to be NP PP, even when the prime is NP NP
  - This could obliterate signs of priming in the NP NP condition, which might show up in a between-subjects design

FACTORIAL DESIGNS

- Some of the most interesting hypotheses are tested when we look at how multiple factors interact

FACTORIAL DESIGNS

- What did who read? = [bare bare]
- Which book did who read? = [complex bare]
- What did which student read? = [bare complex]
- Which book did which student read? = [complex complex]

FACTORIAL DESIGN

- Two factors with 2 factor/treatment levels
  - 2 x 2 design

FACTORIAL DESIGN

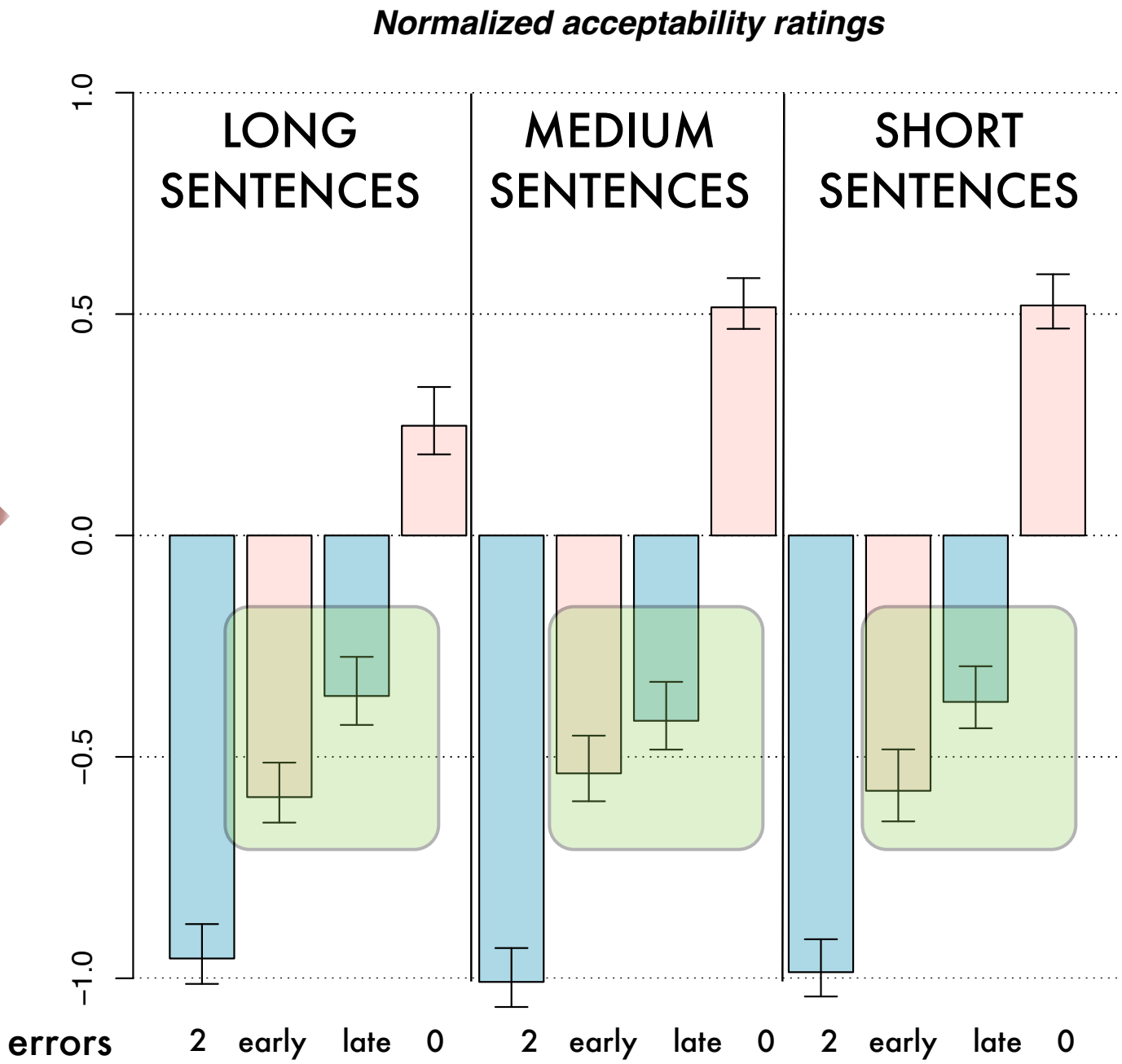
- A 2 x 2 design with 24 items means that each participant will see 6 versions of each condition in a within-subject design

FACTORIAL DESIGN

- Note that increasing the # of factors increases the chances for 1 treatment level to be 'off'
  - For example, in a 3x 2 x 2 study, if one factor patterns in an unpredicted way, the entire dataset can be hard to interpret

- People who always **play/playing** video games are slightly less likely to have **enacted/enacting** violence.
- People who always **play/playing** violent video games are actually slightly less likely to have **enacted/enacting** violence.
- People who always **play/playing** violent video games are actually slightly less likely than their otherwise similar peers to have **enacted/enacting** violence.

FACTORIAL DESIGN (3 X 2 X 2)



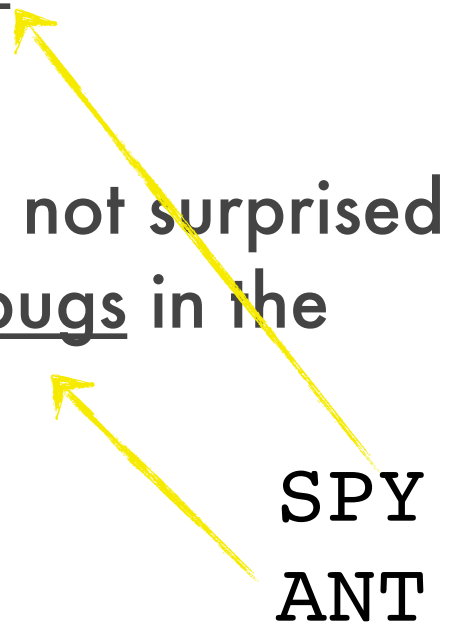
- Think through the possible results. Are they interpretable? Is it possible to be right? Is it possible to be wrong?

SWINNEY (1979)

- H_0 : All possible meanings of an ambiguous word are activated initially
- H_1 : Only contextually consistent meanings of ambiguous words are activated initially

HYPOTHESIS TESTING

- Swinney (1979)
  - BIASED: The man was not surprised when he found several spiders, roaches, and other bugs in the apartment
  - NEUTRAL: The man was not surprised when he found several bugs in the apartment



HYPOTHESIS TESTING

- Possible findings:
  - RTs faster to ANT vs. SPY in neutral & biased condition
  - RTs not statistically different
  - RTs faster to ANT vs. SPY only in biased condition

HYPOTHESIS TESTING

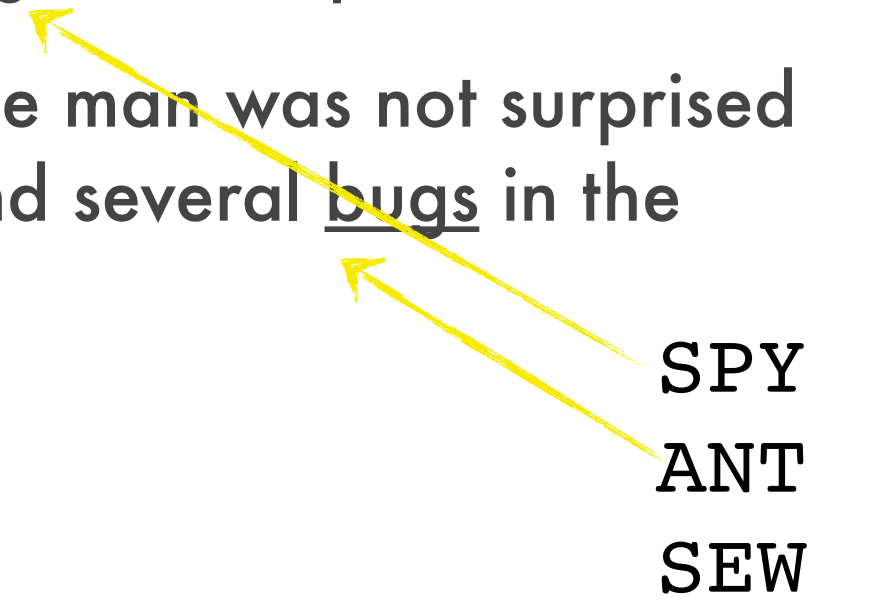
- Possible findings:
  - RTs faster to ANT vs. SPY in neutral & biased condition = H_0 not rejected
  - RTs not statistically different = H_0 not rejected
  - RTs faster to ANT vs. SPY only in biased condition = H_0 not rejected!

HYPOTHESIS TESTING

- None of these possibilities allows for the null hypothesis to be rejected!
  - Why?
    - Design lacks a control so far
    - We need to know whether SPY / ANT has been primed relative to a baseline condition

HYPOTHESIS TESTING

- Swinney (1979)
  - BIASED: The man was not surprised when he found several spiders, roaches, and other bugs in the apartment
  - NEUTRAL: The man was not surprised when he found several bugs in the apartment



HYPOTHESIS TESTING

- Equivalent priming after BUGS for both SPY & ANT compared to SEW
- No priming after 3-syllable interval for SPY in BIASED condition

HYPOTHESIS TESTING

- Moral of the story
  - Think through the possible results and determine if they will be interpretable
  - Consider using a baseline or control condition

End of Part 1