

Announcements

- Applied for this course: 45 students vs. our capability: 30 students
 - Limitation: manpower to give detailed feedback to the assignments (7 submissions) and **ideal group size (3–4 students/group)**
- **Compromise: Off-loading learning responsibility to you!**
 - Groups of 6–7 students \Rightarrow Communication overload and other problems
 - You are responsible to organize your groups
 - Issues: Load-balancing, communication, sharing documents, drop-outs
 - Strategies: Divide & Conquer, Mirroring
- **Do:** Find yourself a group for the first assignment, download the group registration form from our website, fill in the details and email Chat by tomorrow 17:00
 - You will be unregistered from the class if you cannot find a group
 - You are prohibited to stay in the same group for assignment 1–3
 - You may freely choose your group for the mini project



Last Tuesday in Current Topics...

- Three approaches to HCI research
- Three steps in the empirical science approach
- Three strategies in the planned observation



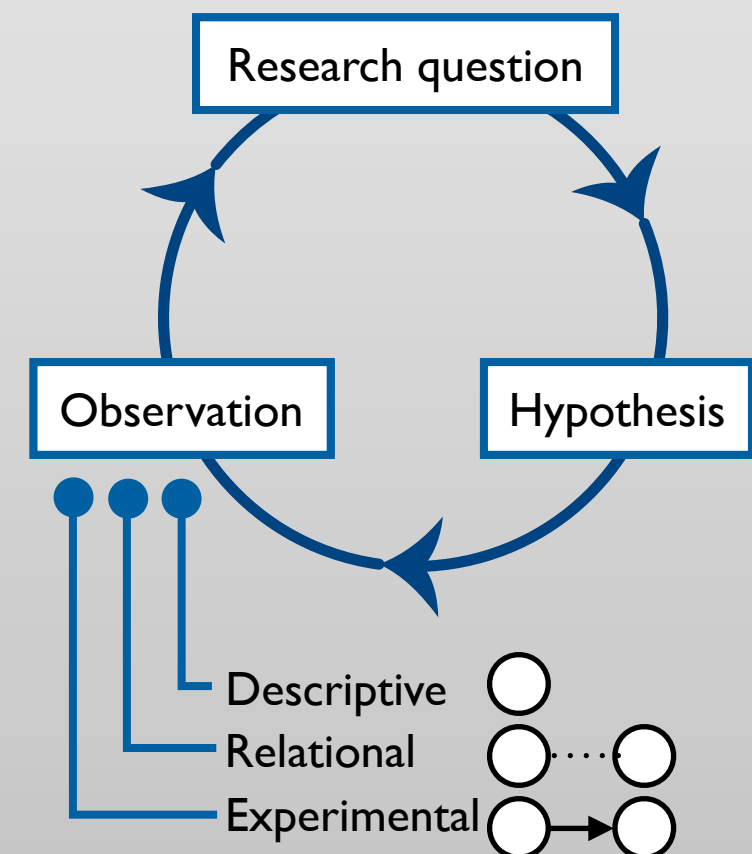
Empirical science



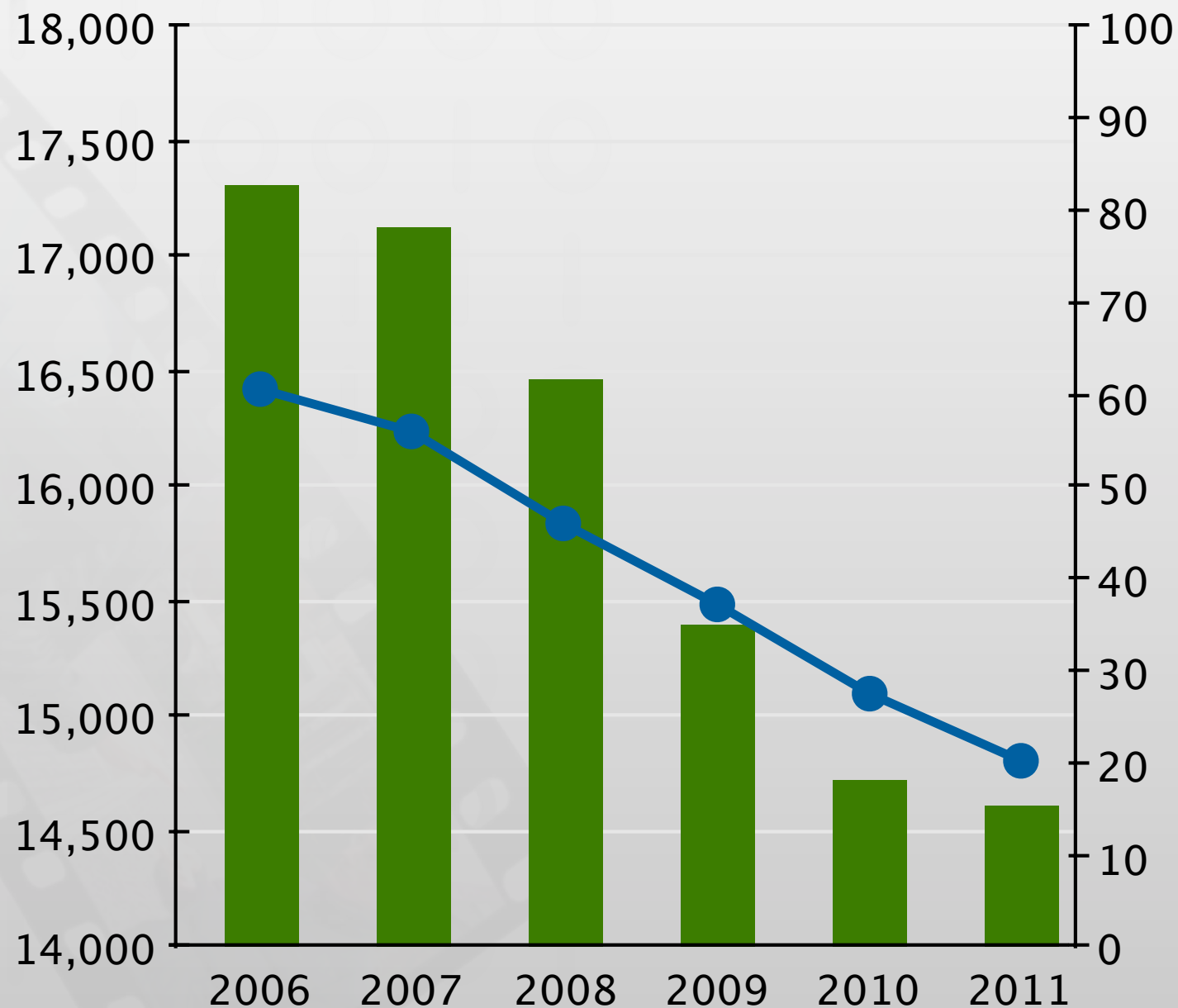
Ethnography



Engineering and design



Correlation Does Not Imply Causation

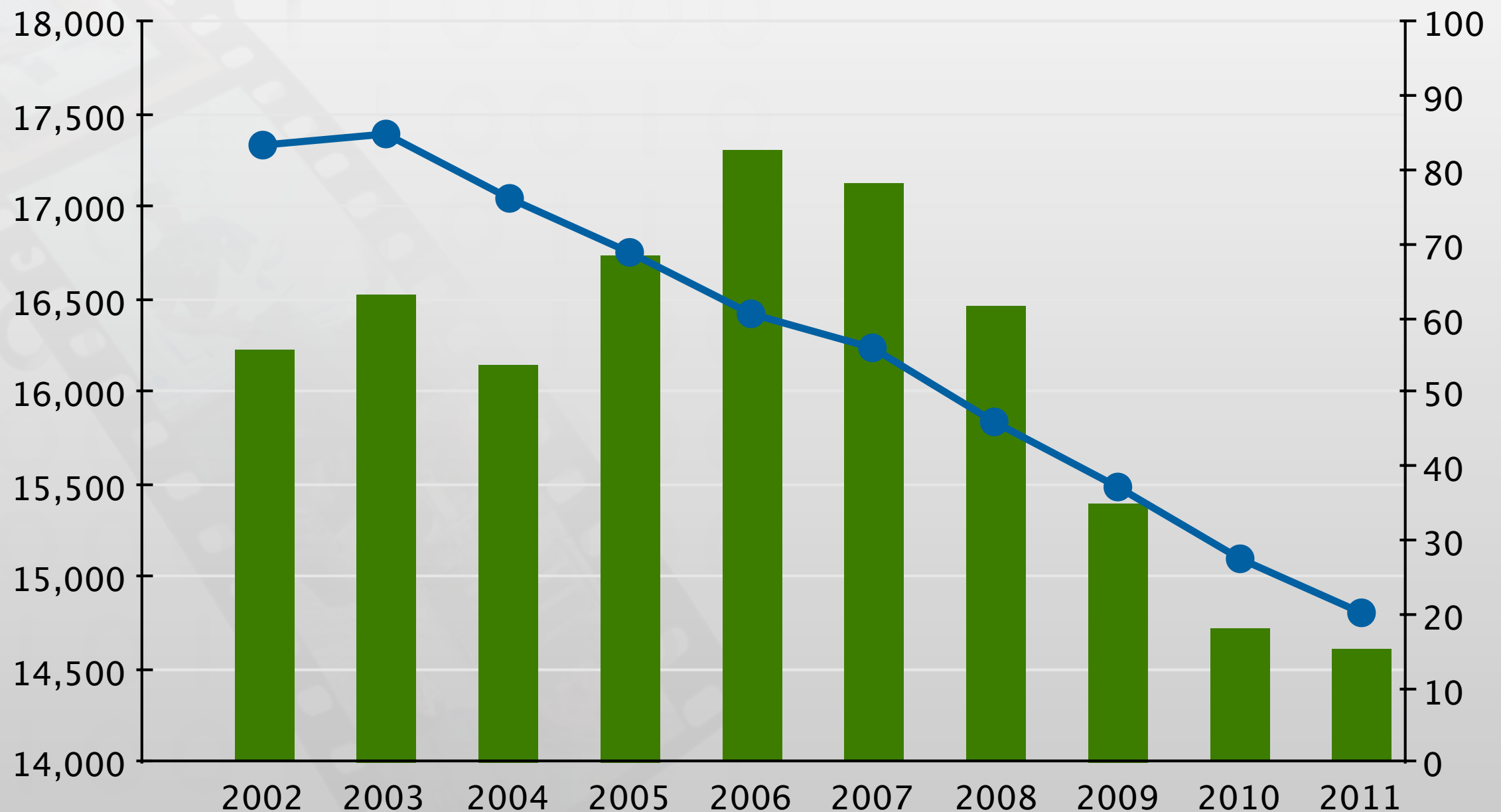


■ Internet Explorer Market Share ● Murders in the US

Adapted from a tweet of @altonncf with data from FBI and W3Schools



Correlation Does Not Imply Causation



Internet Explorer Market Share Murders in the US

Adapted from a tweet of @altonncf with data from FBI and W3Schools



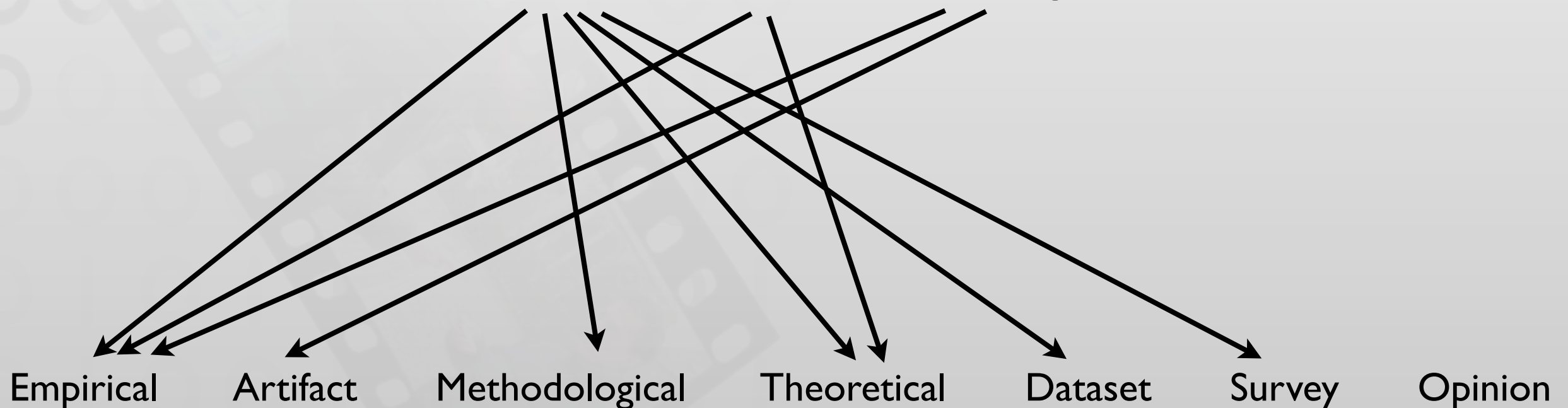
Research Approaches vs. Contribution Types



Empirical
science

Ethnography

Engineering
and design



Seven Research Contribution types

[Wobbrock, 2014]



In-Class Exercise:

Contributions and Benefits

Vulture: A Mid-Air Word-Gesture Keyboard

Markussen et al., CHI 2014

“Word-gesture keyboards enable fast text entry by letting users draw the shape of a word on the input surface. Such keyboards have been used extensively for touch devices, but not in mid-air, even though their fluent gestural input seems well suited for this modality. We present Vulture, a word-gesture keyboard for mid-air operation. Vulture adapts touch based word-gesture algorithms to work in mid-air, projects users’ movement onto the display, and uses pinch as a word delimiter. A first 10-session study suggests text-entry rates of 20.6 Words Per Minute (WPM) and finds hand-movement speed to be the primary predictor of WPM. A second study shows that with training on a few phrases, participants do 28.1 WPM, 59% of the text-entry rate of direct touch input. Participants’ recall of trained gestures in mid-air was low, suggesting that visual feedback is important but also limits performance. Based on data from the studies, we discuss improvements to Vulture and some alternative designs for mid-air text entry.”



Vulture: A Mid-Air Word-Gesture Keyboard

Available at:

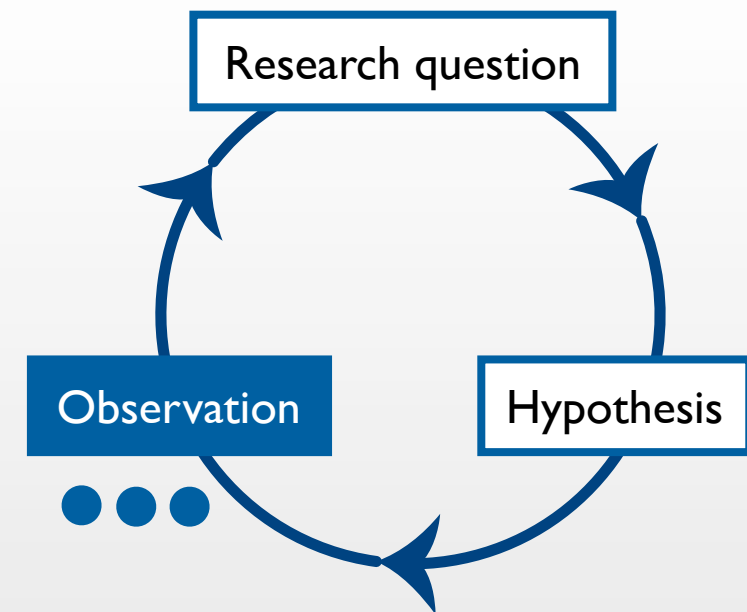
https://www.youtube.com/watch?v=7CKJ6B_dFhc

In-Class Exercise: Contributions and Benefits

- Contributions & Benefits:
 - “Presents an **empirical evaluation** of the potential for **Word-Gesture Keyboards (WGKs)** in **mid-air text entry** and compares how performance compares to **touch based WGKs**.” [Markussen et al., CHI 2014]



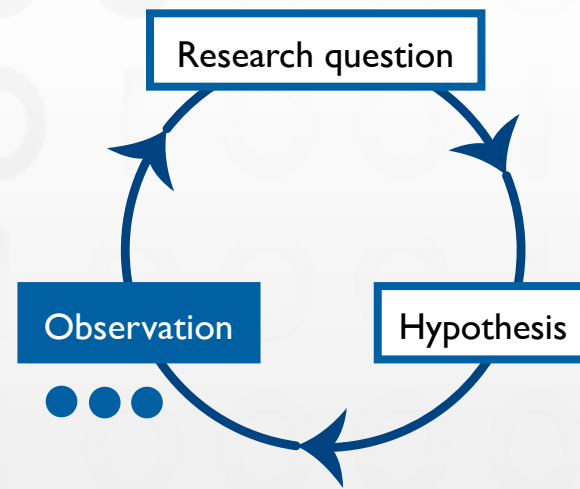
Experimental Research



- Purpose: To infer cause-and-effect relationship
- Controlling **independent variable**
- Observe the change in the **dependent variables**
- In-class exercise: recall the following experimental designs
 - Between-group vs. within-group
 - Benefits and drawbacks
- More details in next lecture

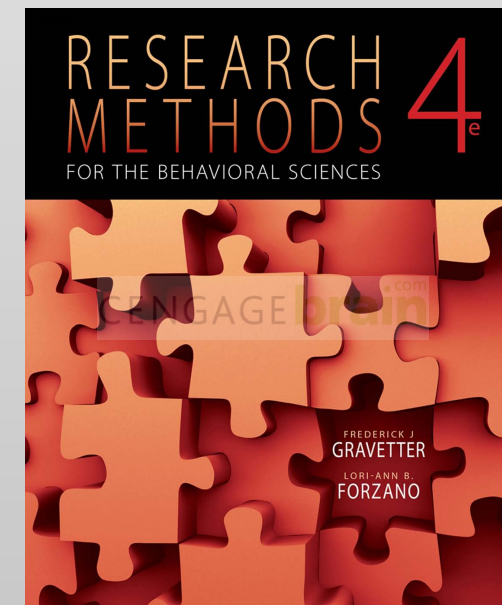
From the last lecture





Experimental Research in HCI

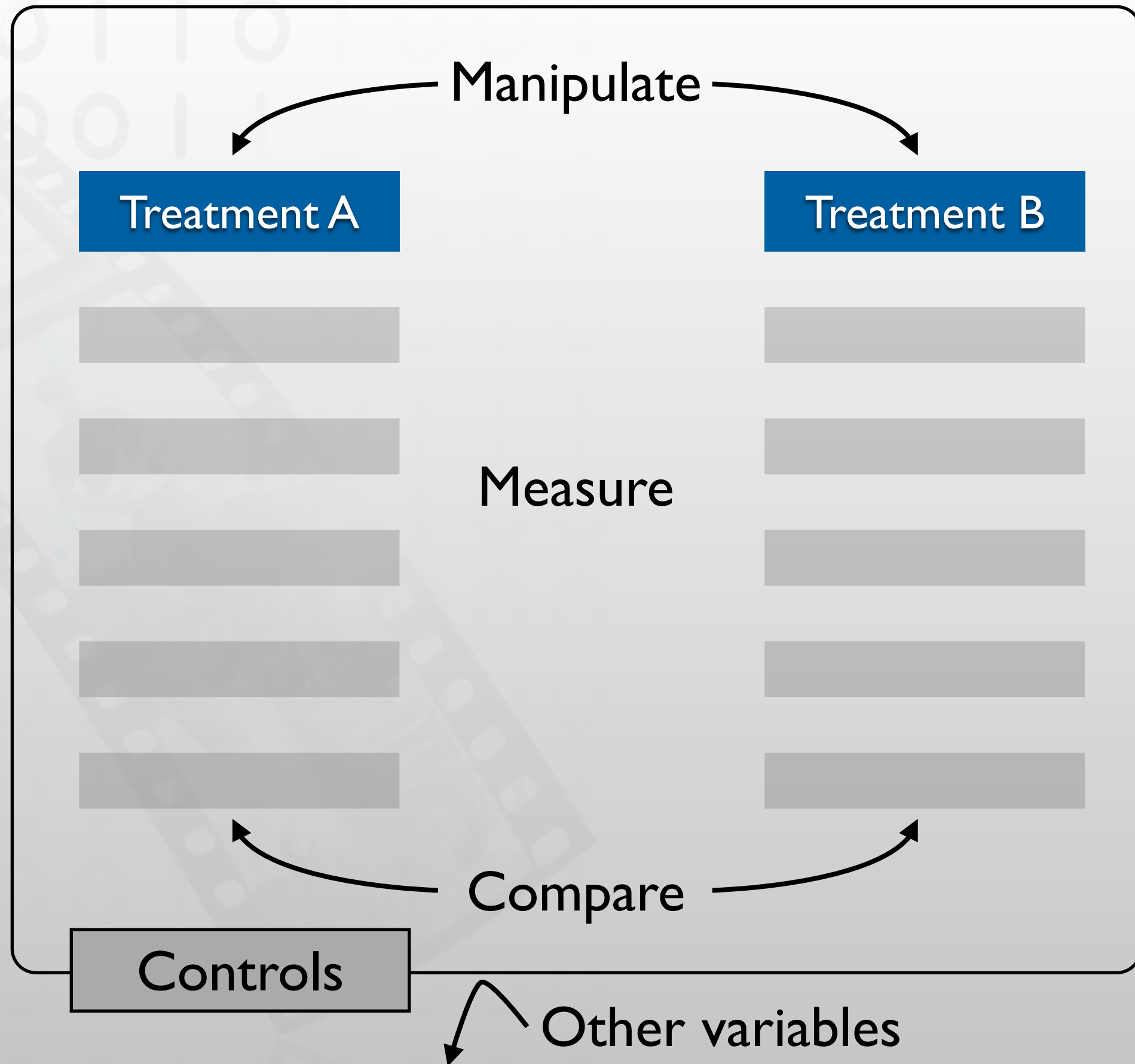
Illustrated through Text Entry Research



Further reading:

Research Methods for the Behavioral Sciences (Gravetter and Forzano, 2012)





Adapted from (Gravetter and Forzano, 2012)



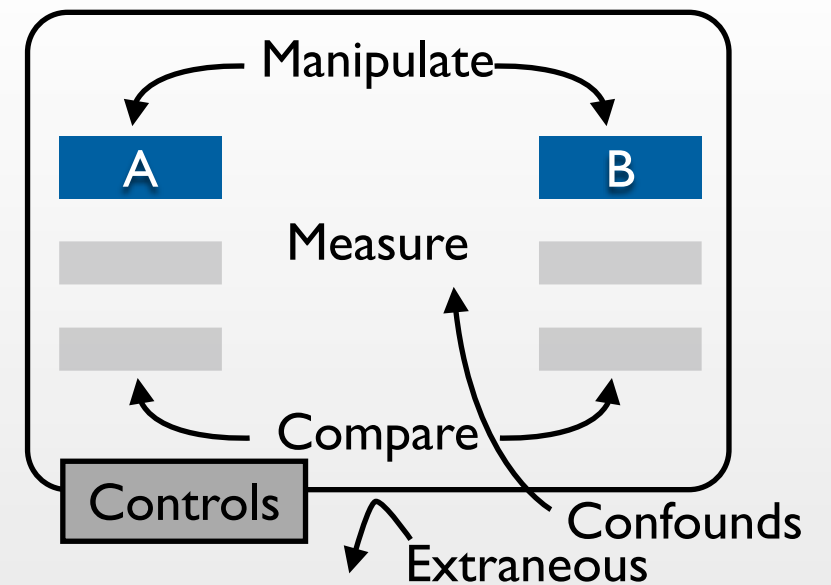
Basic Elements of Experimental Study

- **Manipulation:** Changing the value of the independent variable to create treatment conditions
- **Measurement:** Measure the value of the dependent variable in each treatment condition
- **Comparison:** The score of one treatment condition is compared with another. Consistent differences between treatments \Rightarrow evidence of causality
- **Control:** Other variables are controlled to be sure that they do not influence the two variables being examined

Definitions from (Gravetter and Forzano, 2012)



Variables



- **Independent variable** is manipulated by the researcher
- **Dependent variable** is observed for changes to assess the effect of the independent variable
- All other variables: **extraneous variables**
- A **confounding variable** is an extraneous variable that changes systematically along with IV and DVs \Rightarrow alternative explanation of the relationship between the two variables



Scales of Measurement

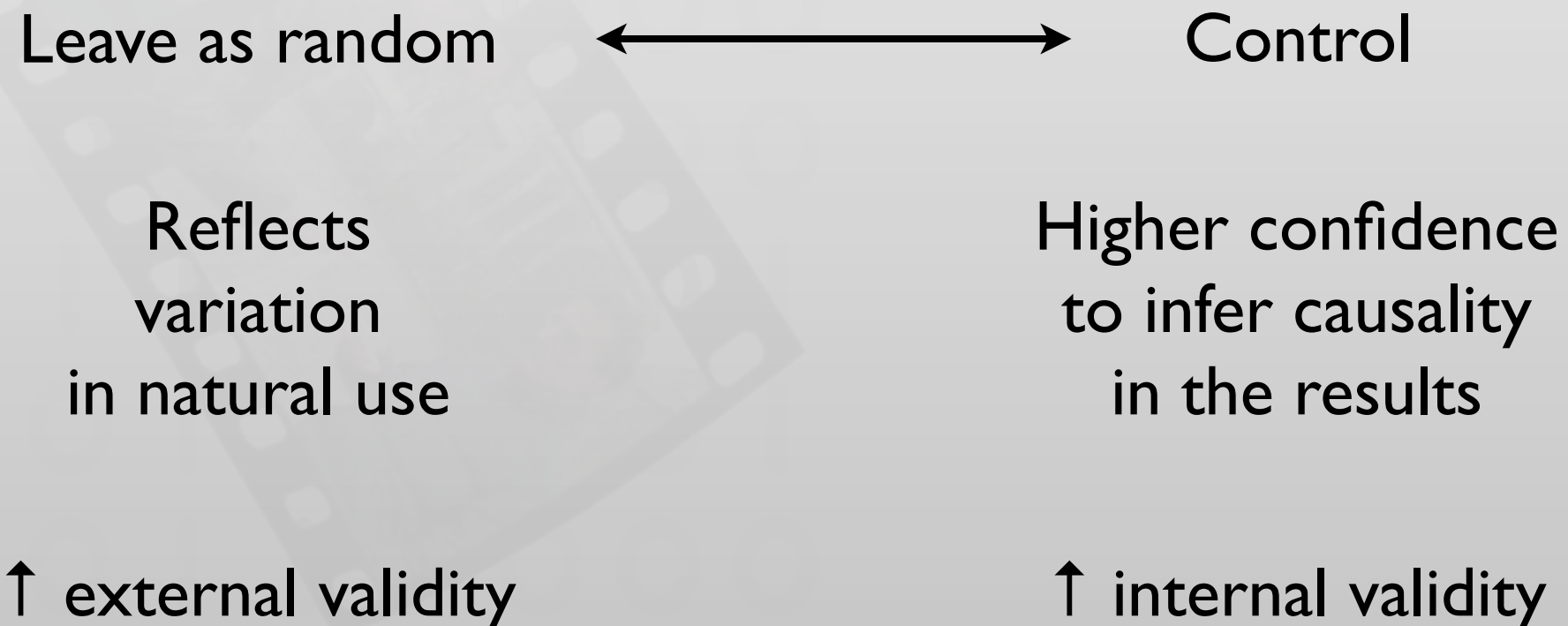
- **Nominal scale:** discrete, qualitative, categorical differences, ignoring the order
 - E.g., input techniques: mouse vs. touchscreen (IV), whether the user made an error or not (DV)
- **Ordinal scale:** sequentially ranked categories, ignoring magnitude of differences
 - E.g., size of keyboard buttons (IV), Likert (5-point) scale answers* (DV)
- **Interval scale:** sequentially organized categories, all categories have the same size (possible to determine relative distances)
- **Ratio scale:** interval scale that zero represents complete absence (possible to determine absolute distances)
 - E.g., Task completion time in seconds (DV), error rate in percent (DV)

* Can be treated as ordinal (strictly according to the definition) or interval (empirically verified over 50 years to be OK)

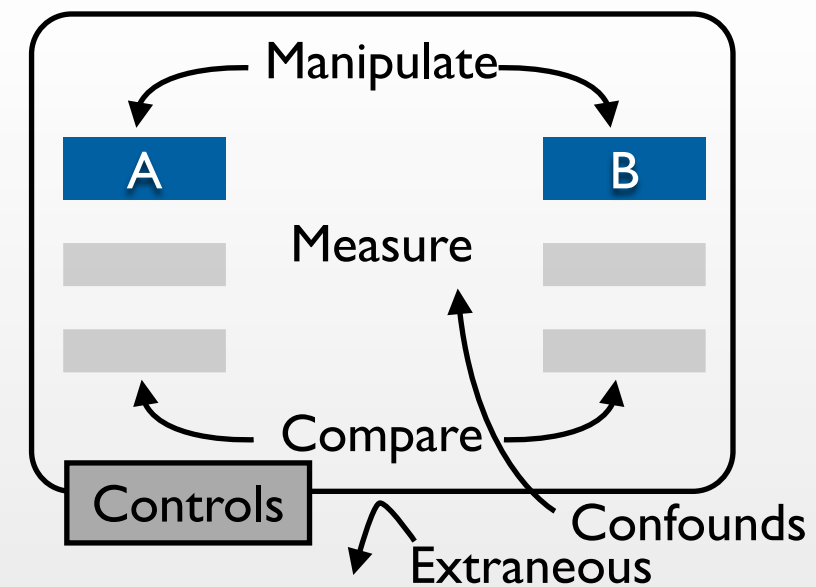


Dealing Extraneous Variables

- Include them as IVs \Rightarrow too many experimental conditions!



Validity



- A study has **internal validity** if it produces a single, unambiguous explanation for the relationship between two variables
 - Threats: e.g., confounding variables, experimenter bias, learning effect, **Hawthorne effect** (being observed causes the changes)
- **External validity** refers to the extent to which we can generalize the results to people, settings, times, measures, and characteristics other than those used in that study
 - Threats: e.g., generalizing across participants, multiple IVs interference
- Always a trade-off, strike an appropriate balance depending on the goal of your research

Definitions from (Gravetter and Forzano, 2012)

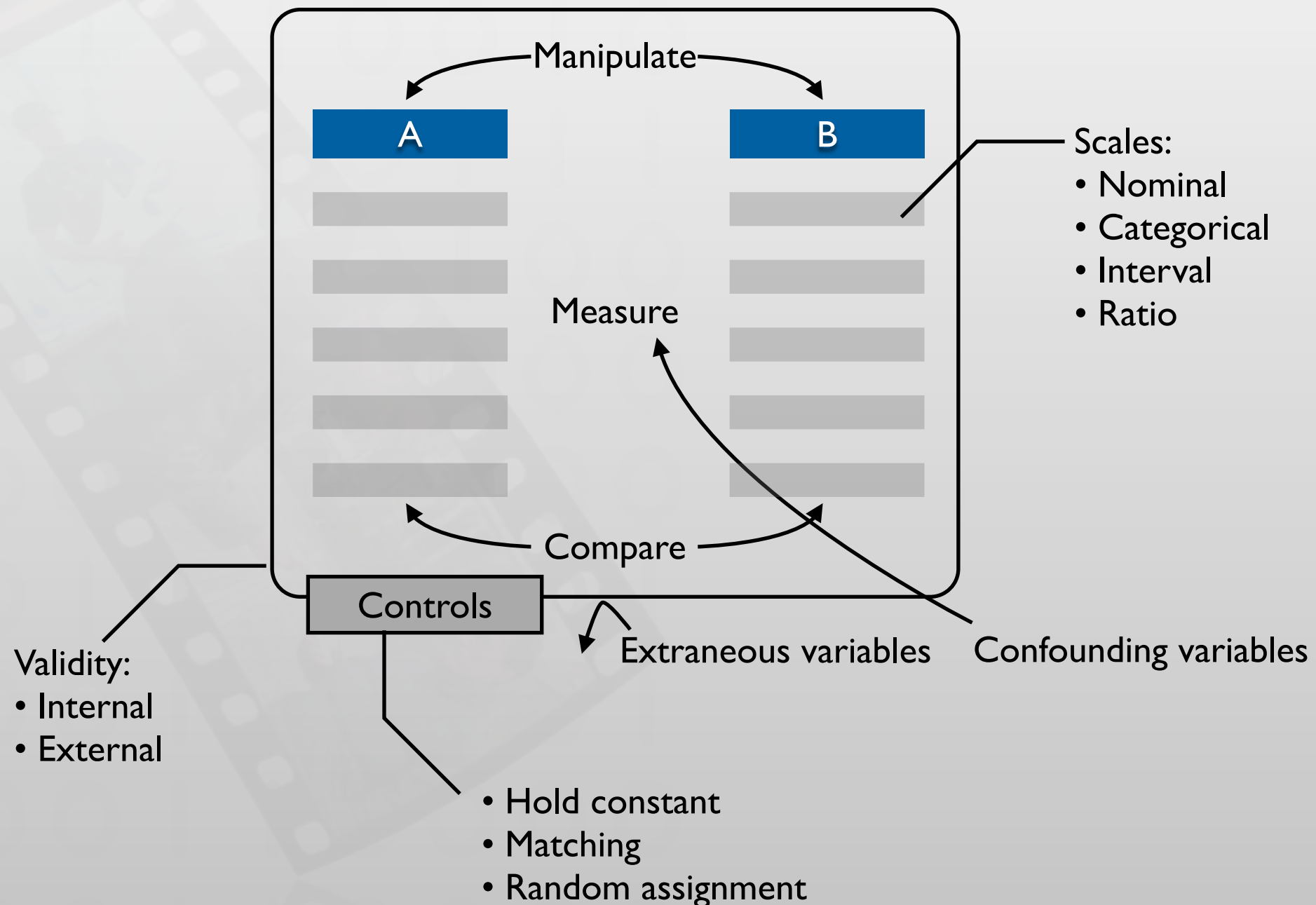


Controlling Extraneous Variables

- **Hold constant**, e.g., selecting participants in the same gender/age
- **Matching** the same number of participants with the same extraneous variable
 - E.g., gender, age, or level of expertise
- **Random assignment** of participants to treatment conditions
 - Other random assignment, e.g., time slot

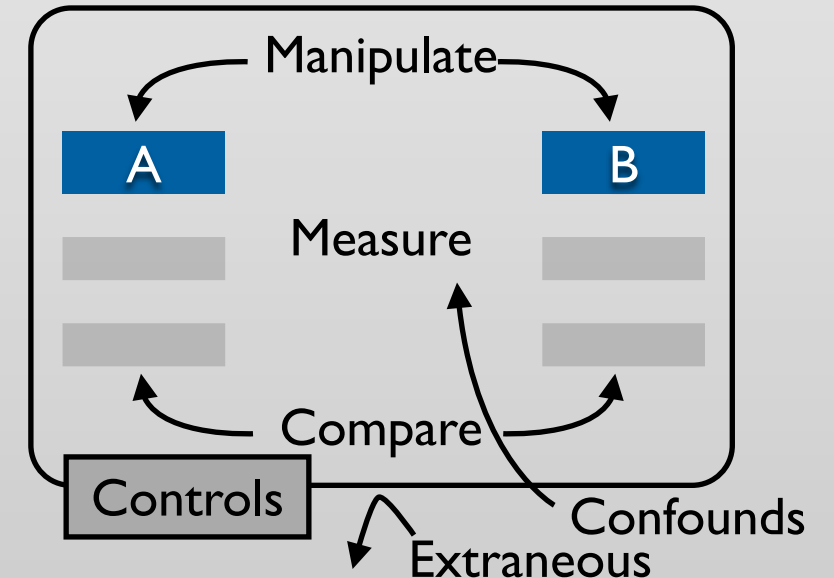


Basic Elements of Experimental Study




Example: Text Entry Research

- You have designed a new keyboard layout, and you want to know how good it is
- Strategy: compare it with existing techniques
- Basic research questions
 - How fast is it?
 - How accurate is it?
 - How satisfied the users are?
- In-class exercise: Identify
 - Independent variables
 - Dependent variables
 - Extraneous variables and potential confounding variables



Dependent Variables in Text Entry Experiments

- Speed
 - Accuracy
 - Qualitative feedback
 - Comfort
 - Device impressions
 - Report as anecdotes or quotes
 - In-class exercise:
Give an **operational definition** of each variable, and indicate on which **scale** it is measured
- 
- A diagram illustrating a trade-off between Speed and Accuracy. A blue-outlined box labeled "Trade-off" has two curved arrows pointing from it to the words "Speed" and "Accuracy" in the list above.



Speed Measures: Words per Minute

$$\text{WPM} = \frac{|T| - 1}{S} \times 60 \times \frac{1}{5}$$

$|T|$ Length of the transcribed string

– 1 Timing begins after the first character was pressed

S Duration in seconds

$\frac{1}{5}$ Estimated length of a word: 5 characters including spaces (Yamada, 1980)

+ Easiest measure, you just need a watch

– Disregards errors in the final text

- Alternative: insist on the user correcting all errors, but may lead to user frustration

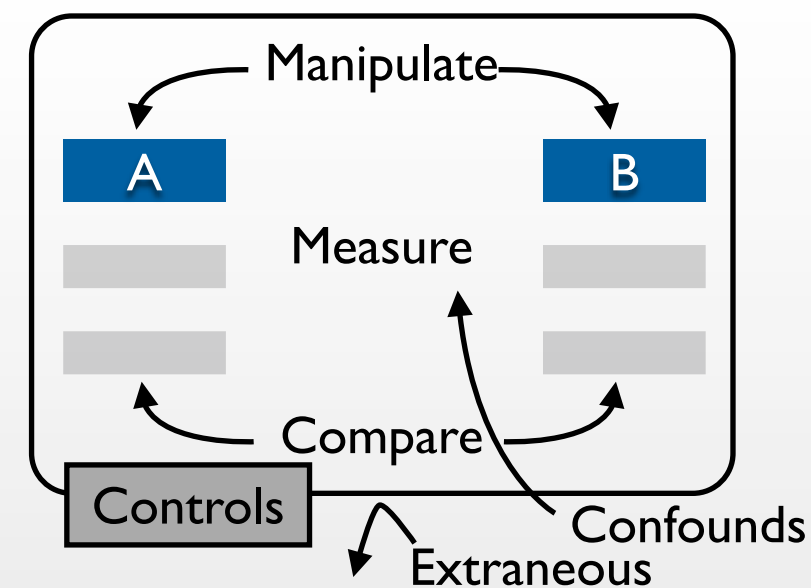
– Disregards the process of entering

- E.g., it doesn't matter how many times you pressed the backspace key.



Text Entry Tasks

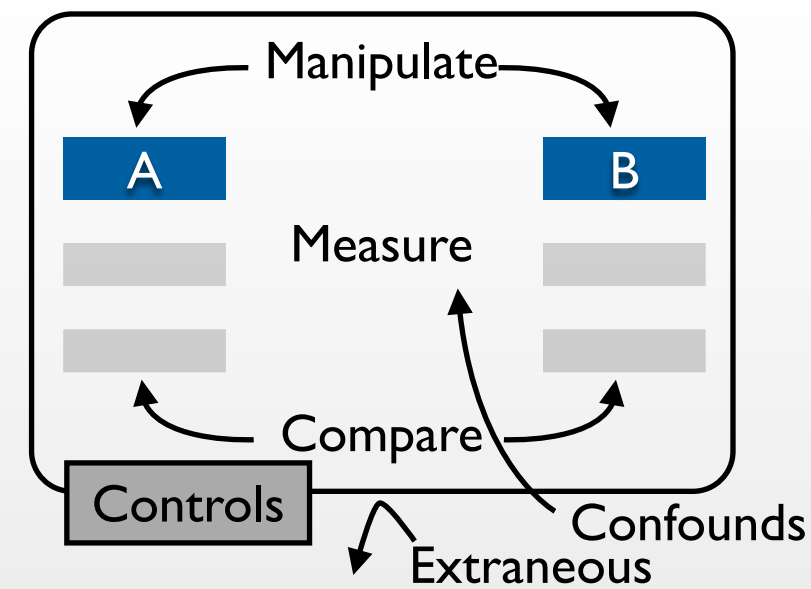
- **Composition:** user create his own text
 - More realistic
 - Users may take inconsistent durations to think about what to write
 - Error identification is difficult
- **Transcription:** copy text
 - Exclude behaviors that may compromise the measures, e.g., pondering what to write
 - Allows identifying error because the content is known
 - Can control the distribution of letters and words
- **Read and memorize** a short sentence before entering
 - Reduce participants' tendency to switch between the displayed text and the entry text field
 - Faster typing but the overall experiment takes longer due to the memorizing [Kristensson & Vertanen, IUI'12]



there will be some fog tonight

there w_

Standard Dataset for Transcription Task



- MacKenzie and Soukoreff (CHI 2003)
- 500 English phrases in moderate length, easy to remember, and representative of the target language (in term of letter frequency correlation)
- Ignore case and enter all characters in lowercase.

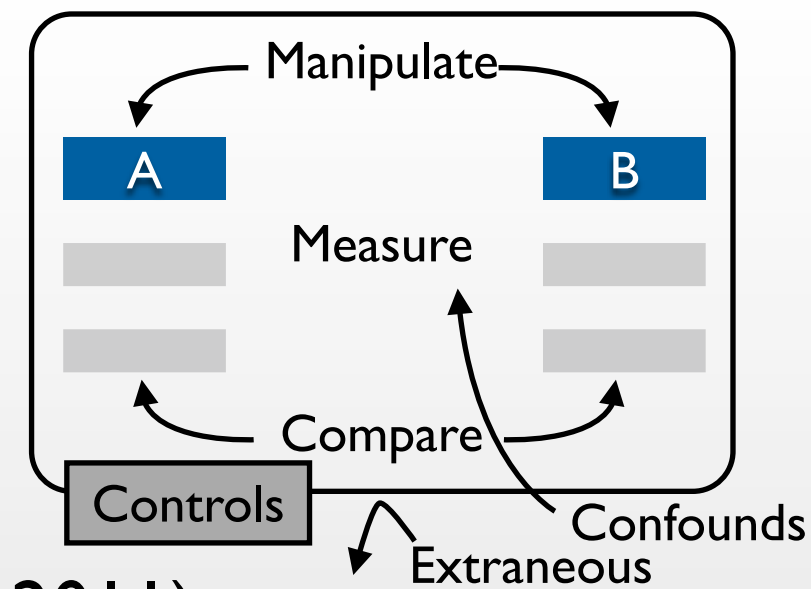
+ Allows replication

- Examples:

there will be some fog tonight
round robin scheduling
time to go shopping
frequently asked questions



Standard Dataset for Transcription Task



- EnronMobile: Vertanen & Kristensson (MobileHCI 2011)
- 200 sentences extracted from real-world mobile phone text entry (BlackBerry QWERTY), tested for memorability and representative character distribution of mobile texting

+ Better external validity for mobile phone text entry studies

- Examples:

Mackenzie & Soukoreff

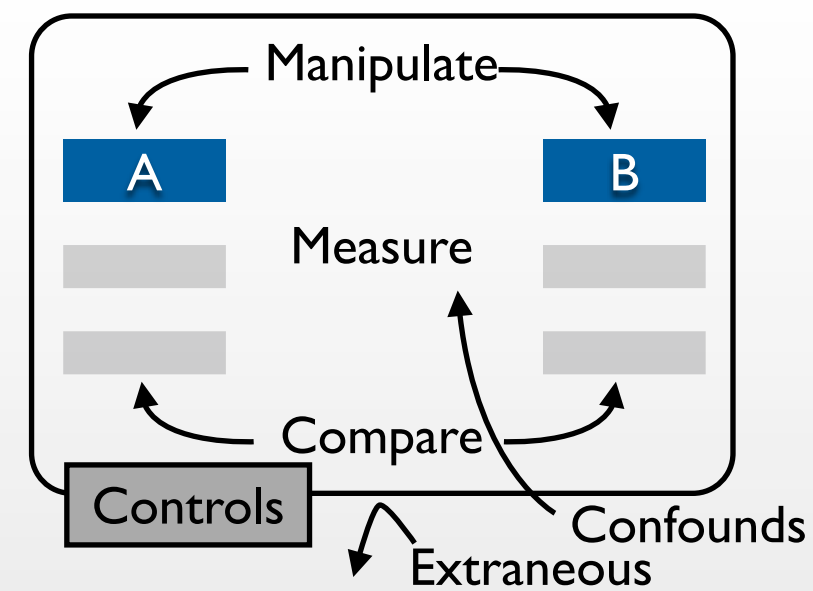
there will be some fog tonight
round robin scheduling
time to go shopping
frequently asked questions

EnronMobile

Thanks, I will look at it tonight.
Interesting, are you around for a late lunch?
Are you going to join us for lunch?
Thanks for the surprise



Text Composition Task



- Problem:
 - Users may take inconsistent durations to think about what to write
 - Error identification is difficult
- Vertanen and Kristensson (TOCHI 2014) characterizes and fine-tune text composition task with four experiments with Amazon Mechanical Turks
- Composition task variants:
 - Copy, reply, situational composition, free composition, aiding communication
- Instructions variants
 - E.g., “Say the intended message before typing” or “Do not use slang”
- Results: Composition tasks take longer and have more edits



Text Composition Task

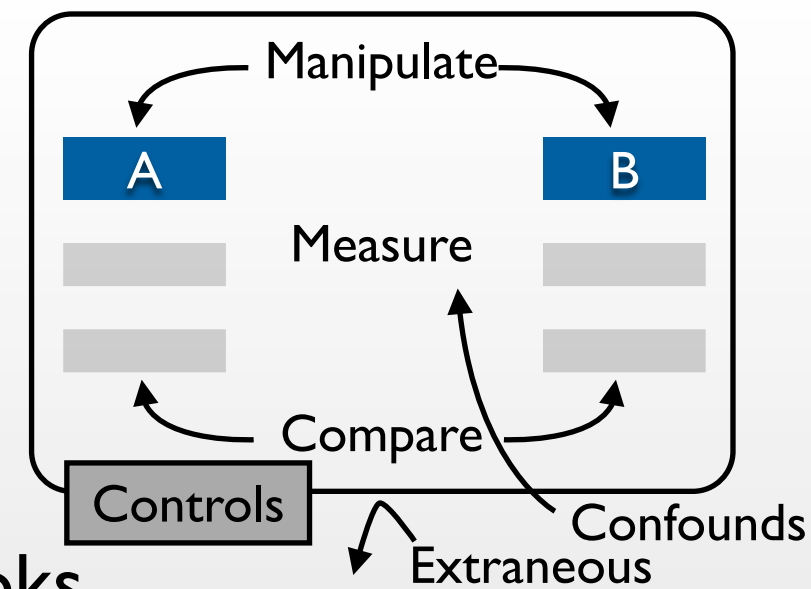
- Task description is adequate to control the quality
 - “Imagine you are **using a mobile device and need to write a message**. We want you to invent and type in a fictitious (but plausible) message. Use your imagination. If you are struggling for ideas, think about things you often write about using your own mobile device

Please write **complete sentences** with **good grammar and spelling**. Do NOT use texting **abbreviations or slang**.”

- Error identification: Use median score from multiple judges or crowdsourcing



Real-world Text Entry



- Evans and Wobbrock, CHI 2012
- Observe keyboards and mouse with low-level hooks
- Segmenting trials by heuristics
 - Segment on Enter key, end-of-sentence punctuations, and pauses
 - At least 24 characters

Close to average length of the MacKenzie & Soukoreff phrase set

- Use online search engine for distinguishing edits from errors and to calculate uncorrected error rate
- Found correlation between the WPM and error measures extracted from this method and the lab study

+ Better external validity



Real-world Text Entry

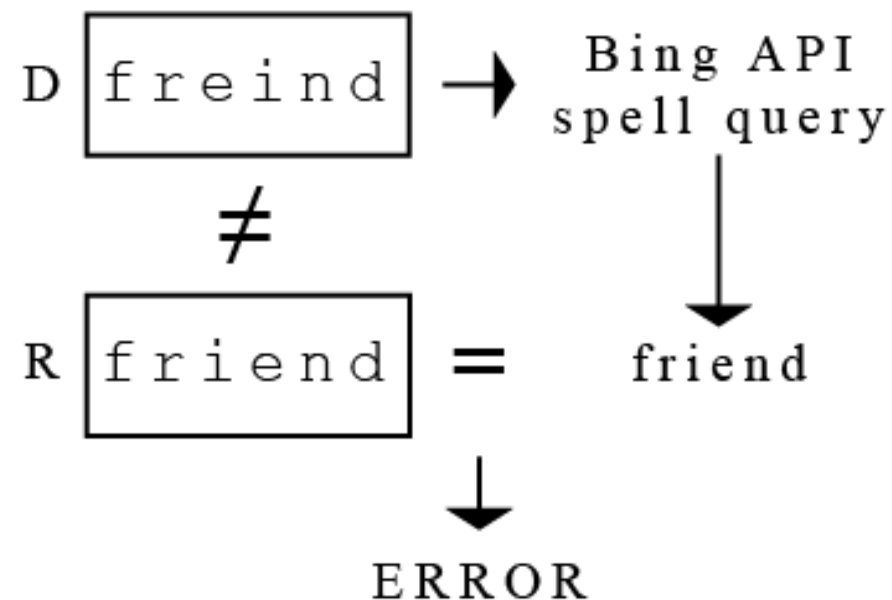
(a)

My best freind
 <<<<
 iend

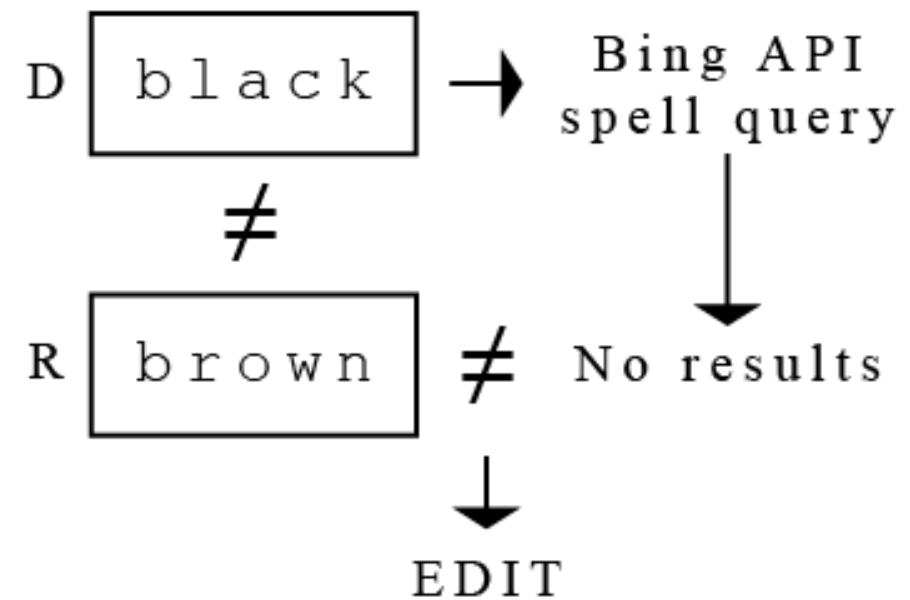
(c)

The quick black fox
 <<<<<<<<
 rown fox

(b)



(d)



[Evans and Wobbrock, CHI 2012]



Basic Experimental Designs

From DISI

- Between-subjects design

- Each subject only does one variant of the experiment
- There are at least 2 groups to isolate effect of manipulation:

Treatment group and control group

+ No practice effects across variants

Good for tasks that are simple and involve limited cognitive processes,
e.g., tapping, dragging, or visual search

— But: requires more users

- Within-subjects design

- Each subject does all variants of the experiment

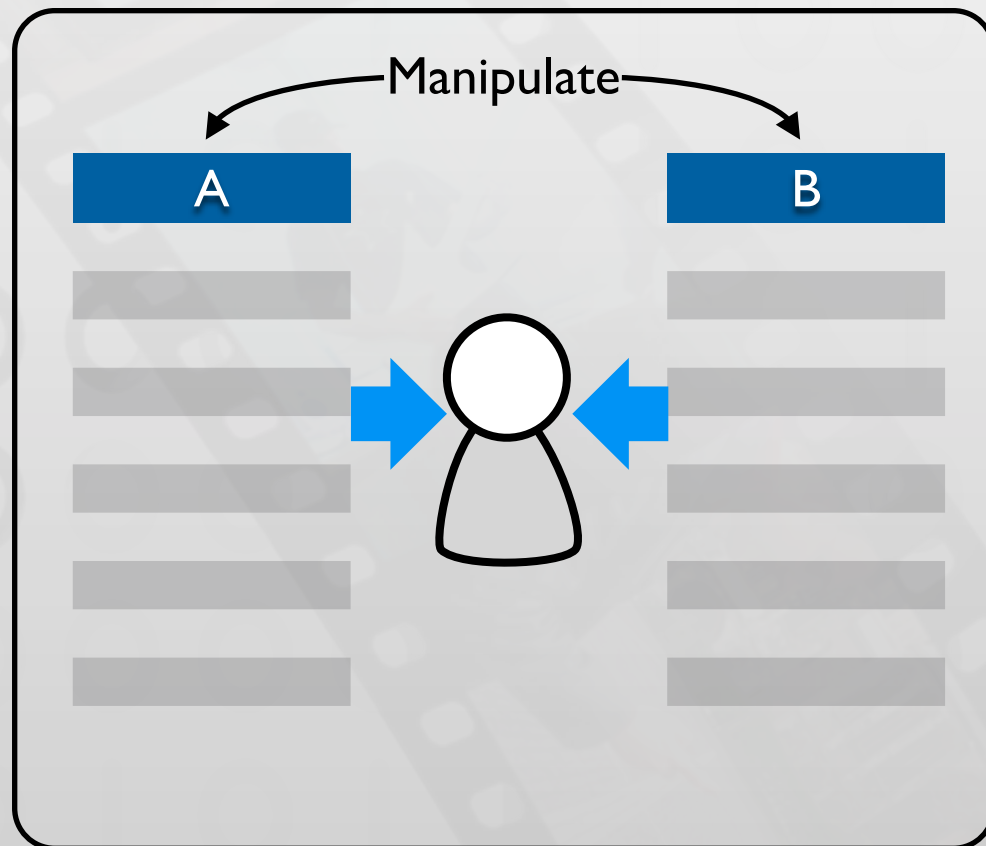
+ Fewer users required, individual differences canceled out

Good for complex tasks, e.g., typing, reading, composition, problem solving

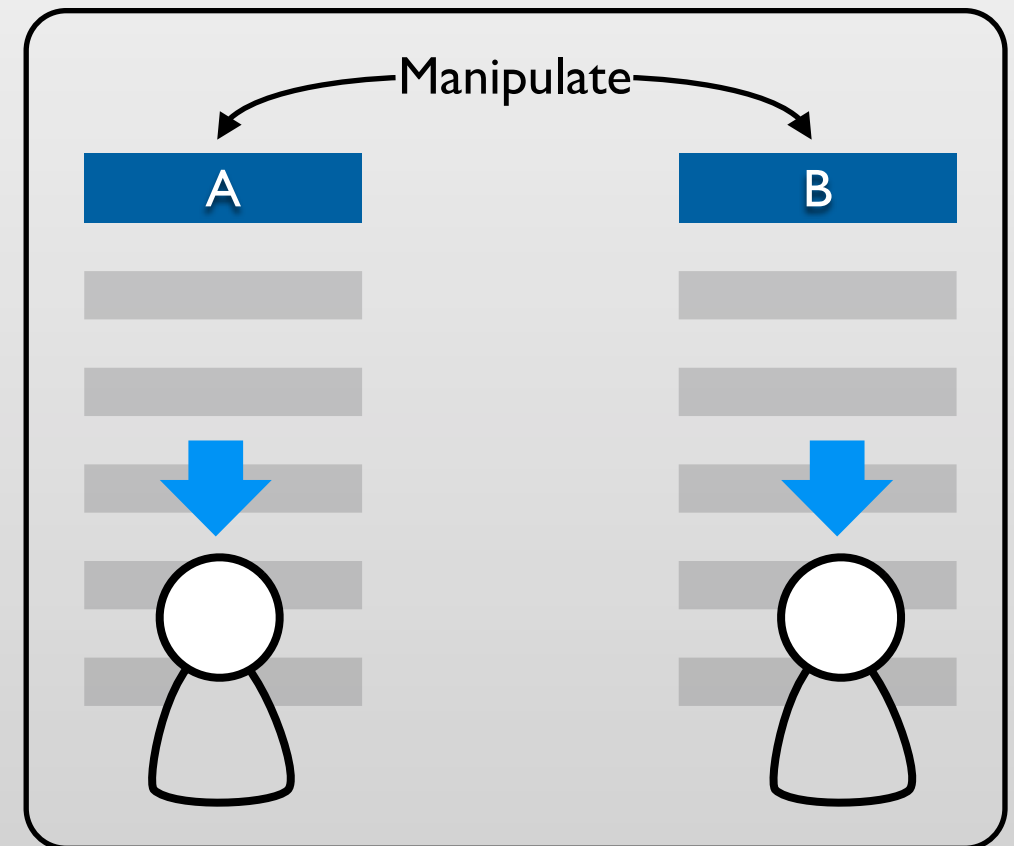
— But: practice effects may occur



Basic Experimental Designs

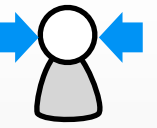


Within-subjects design



Between-subjects design

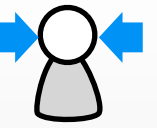




Order Effects

- Within-subjects design
- The behavior may be influenced by experience that occurred earlier in the sequence
- **Carryover effects:** changes caused by the lingering aftereffects of an earlier treatment condition.
 - E.g., Testing the first condition causes users finger to hurt, degrading their performance in the second condition
- **Progressive error:** changes that are related to general experience in the study but unrelated to specific treatments
 - Practice effects and fatigue
 - E.g., The experiment overall takes too long





Counterbalancing

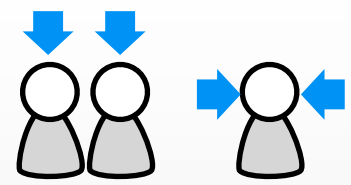
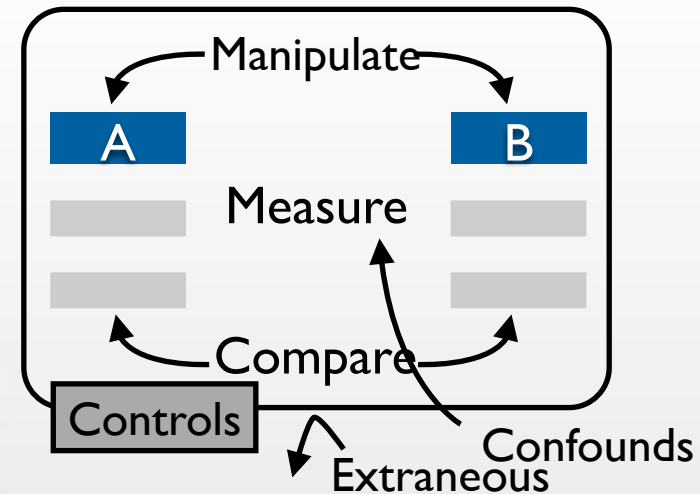
From DISI

- Use every possible order of treatments with an equal number of individual participants
- Latin Square
 - Each condition appears at each ordinal position
 - Each condition precedes and follows each condition one time
 - Example: six treatments: A, B, C, D, E, F

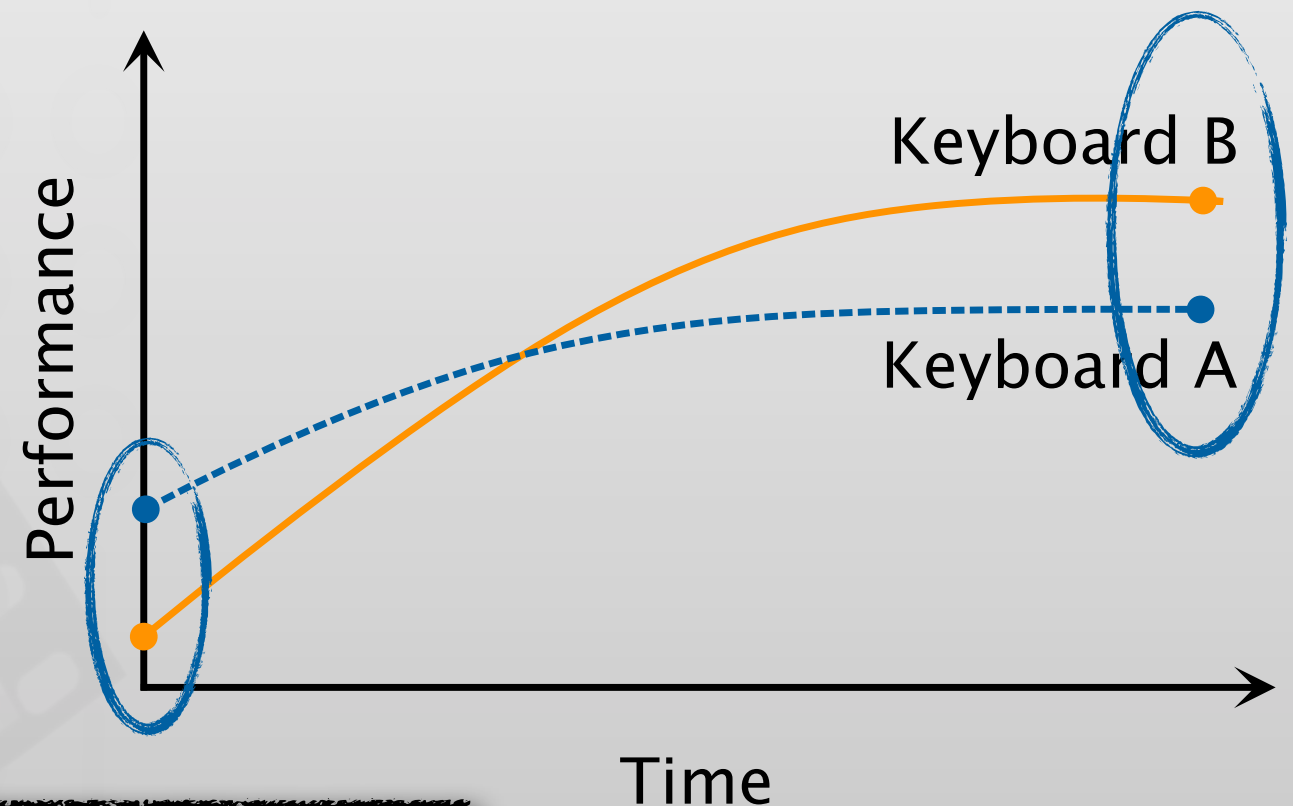
| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | A | B | F | C | E | D |
| 2 | B | C | A | D | F | E |
| 3 | C | D | B | E | A | F |
| 4 | D | E | C | F | B | A |
| 5 | E | F | D | A | C | B |
| 6 | F | A | E | B | D | C |



Learning Curve



- **Learning curve:** relationship between experience (or time) and performance
- Rapid raise at the beginning follow by a plateau
- In general, start measuring when the learning effect is gone!



Skilled use

Immediate usability

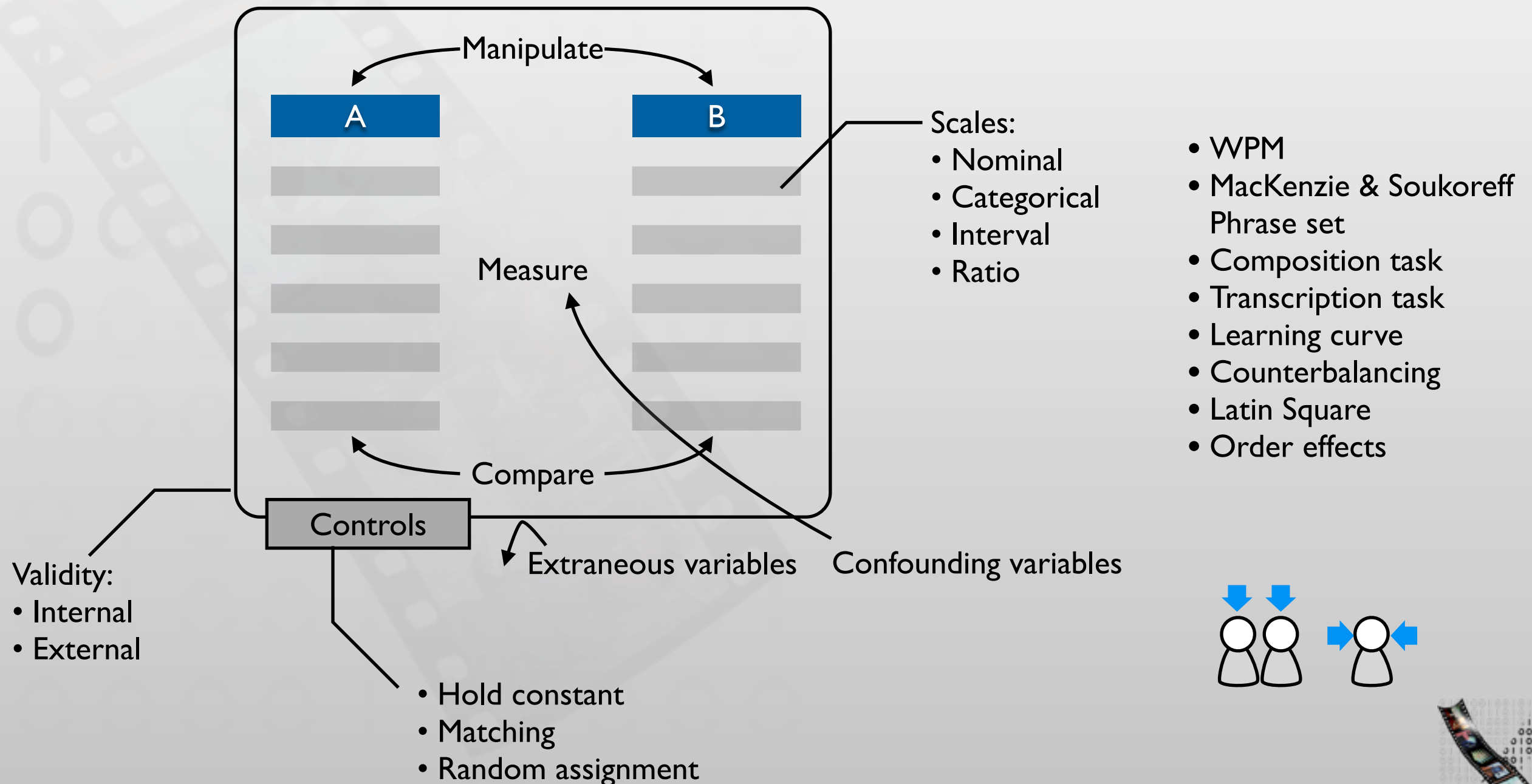


Experimental Design in Text Entry Research

- Usually preferred: **within-group design**
 - Minimizes confounding effects from the behavioral differences between participants
- Sometimes, we need a **between-groups design**
 - E.g., when testing whether a keyboard favors users with right-handedness over those with left-handedness
 - When there are interferences between conditions, e.g., different keyboard layouts on the same hardware




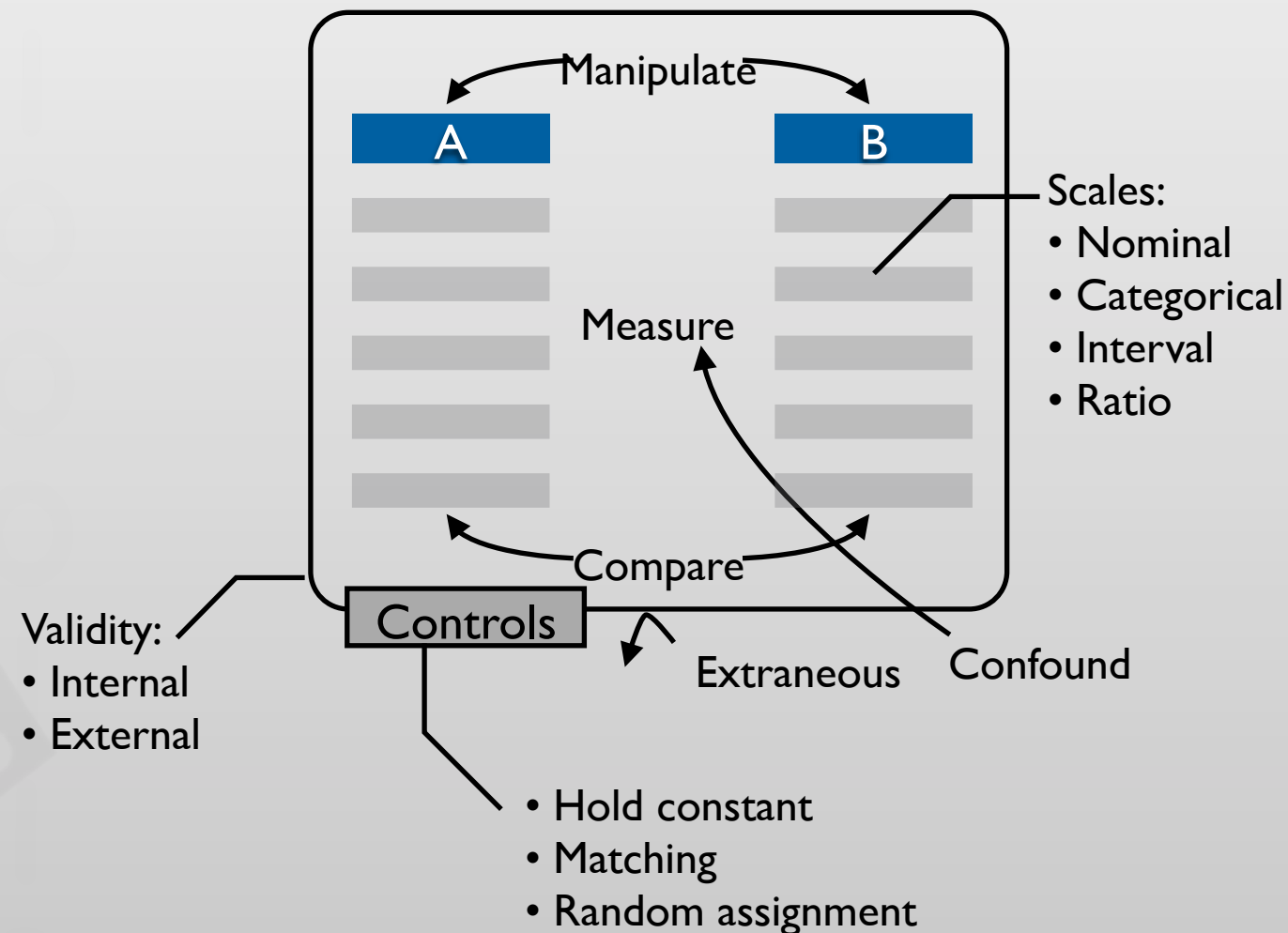
Basic Elements of Experimental Study in Text Entry Studies



Reverse-Engineer

An Experimental Study

- Gestures and widgets: performance in text editing on multi-touch capable mobile devices
- Fuccella et al., CHI '13 
- Contributions & Benefits
 - “We present the **design** and **evaluation** of a gestural text editing technique for touchscreens. Gestures drawn on the soft keyboard are often faster than conventional editing techniques.”



What You Need To Do Now

- Read this paper this week
 - [Evaluation of Text Entry Techniques](#) — MacKenzie, 2007
- Optional reading
 - [Complementing Text Entry Evaluations with a Composition Task](#) — Vertanen and Kristensson, TOCHI 2014
 - [Measures of Text Entry Performance](#) — Wobbrock, 2007
- First assignment is out tomorrow!
 - Identifying contributions and benefits of research articles
 - Find yourself a group for the first assignment, download the group registration form from our website, fill in the details and email Chat by tomorrow 17:00
 - Submission deadline for peer review: April 23

