

Last Tuesday in Current Topics...

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



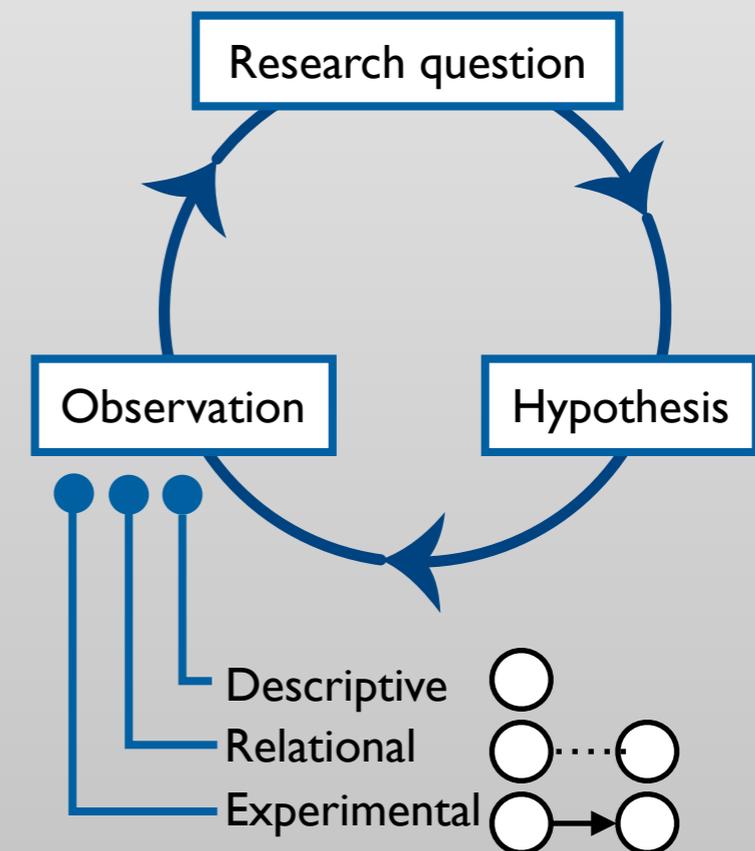
Empirical science



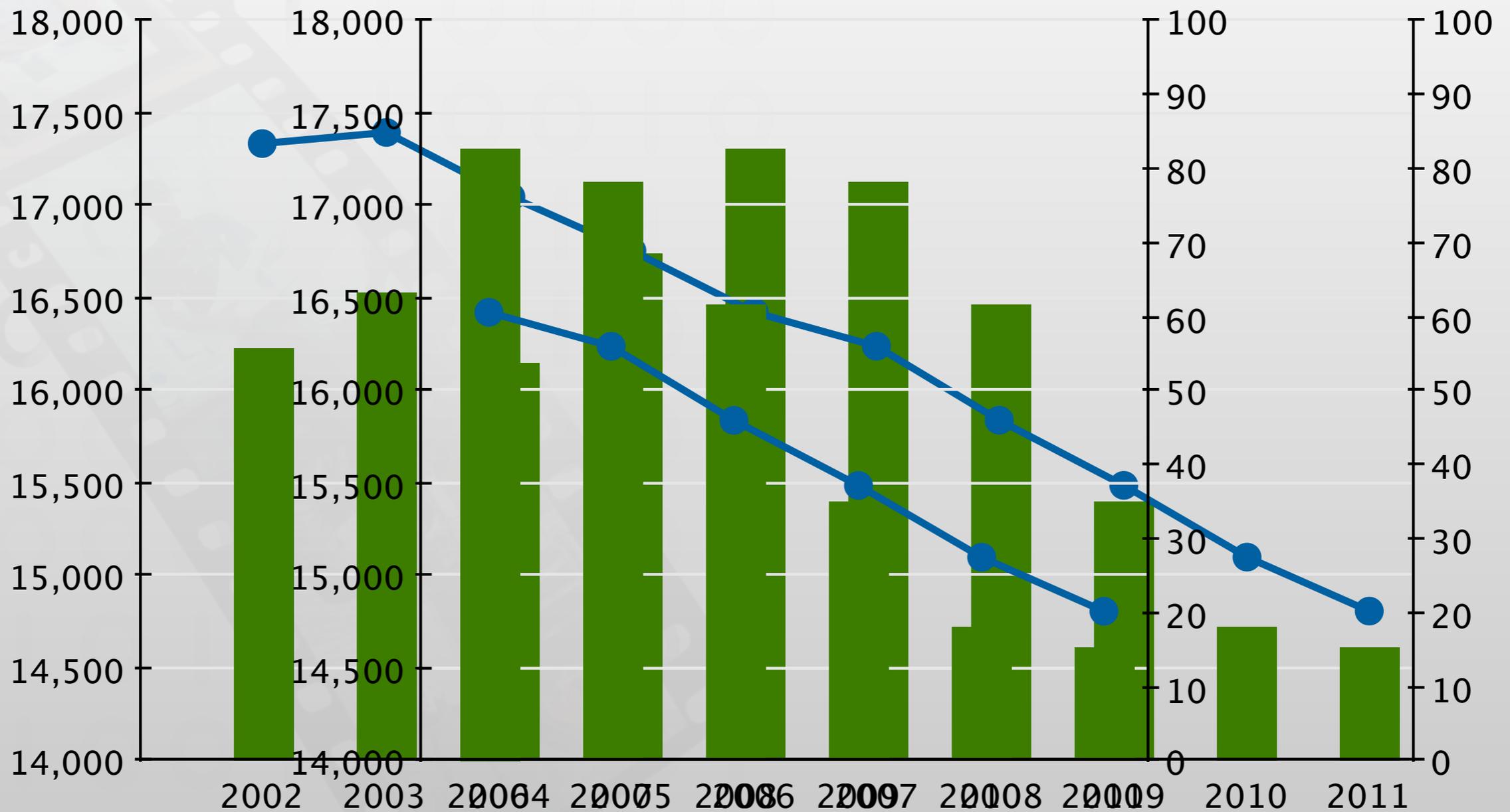
Ethnography



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



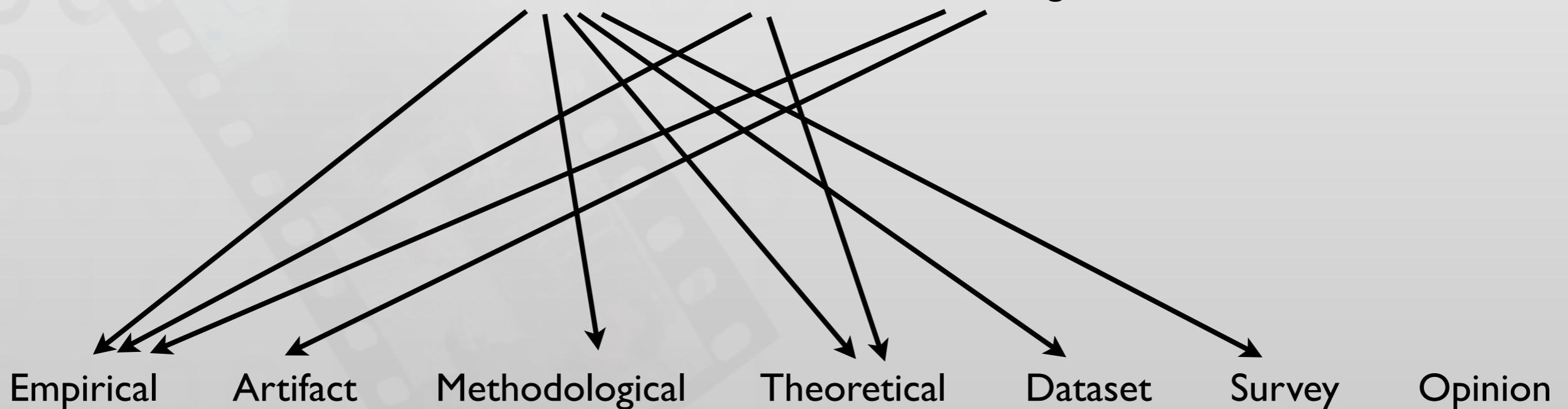
Research Approaches vs. Contribution Types



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

Anders Markussen Mikkel R. Jakobsen Kasper Hornbæk

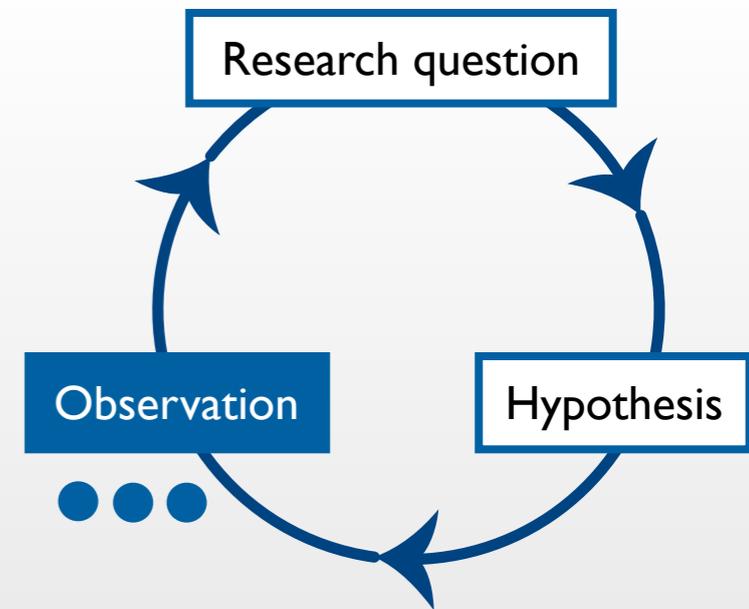
Department of Computer Science
University of Copenhagen

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]



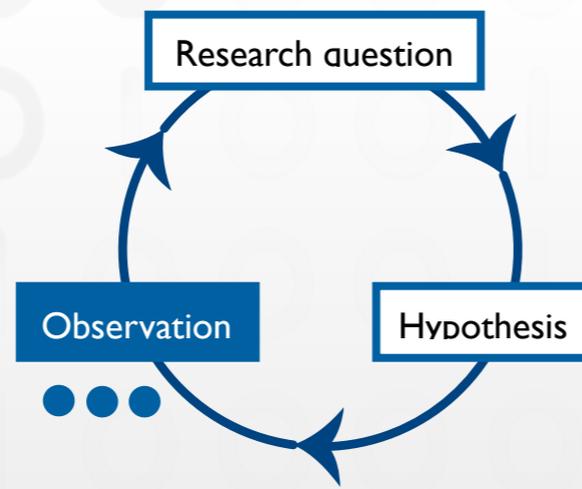
Experimental Research



- 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

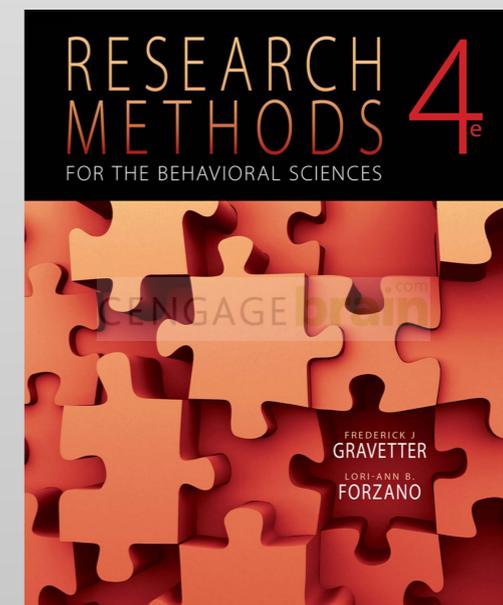
From the last lecture





Experimental Research in HCI

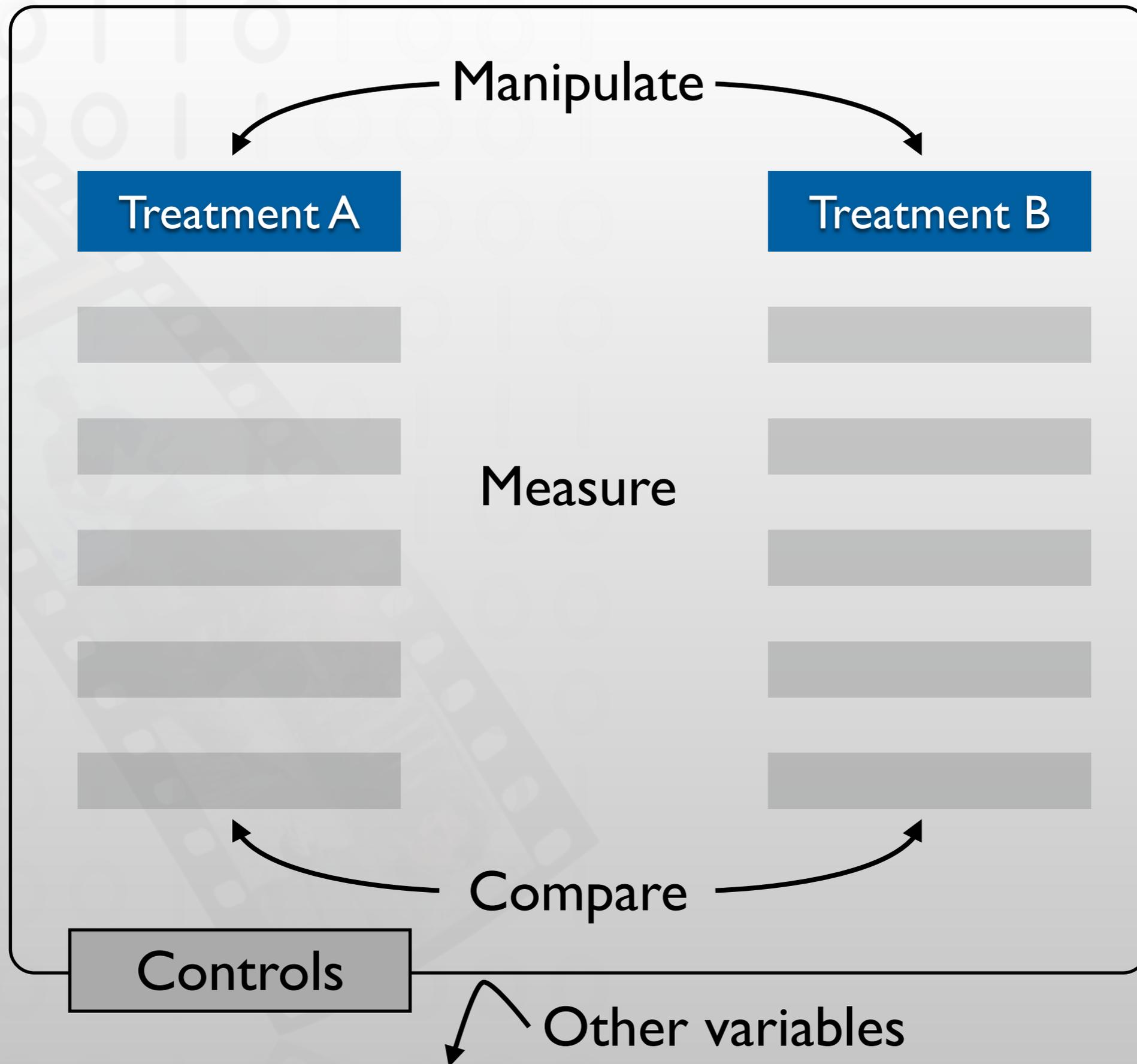
Illustrated through Text Entry Research



Further reading:

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





Adapted from (Gravetter and Forzano, 2012)



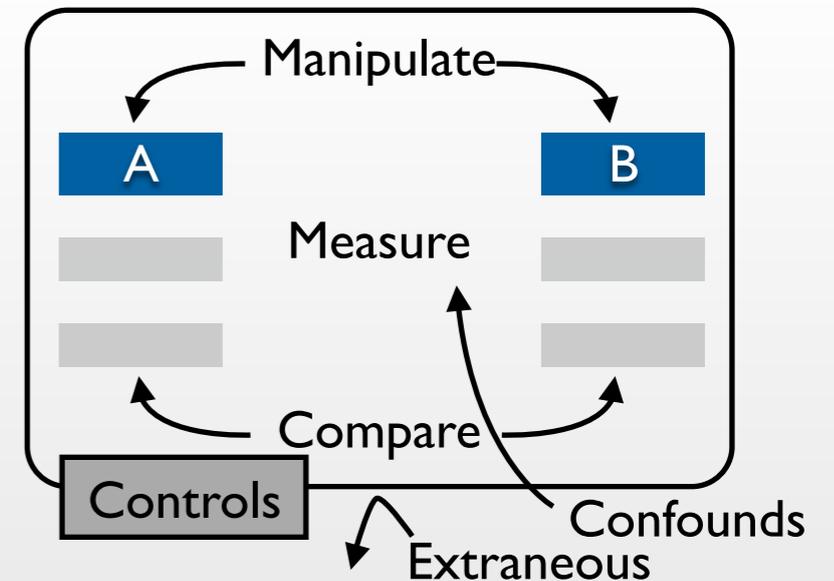
Basic Elements of Experimental Study

- **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!

Leave as random



Control

Reflects
variation
in natural use

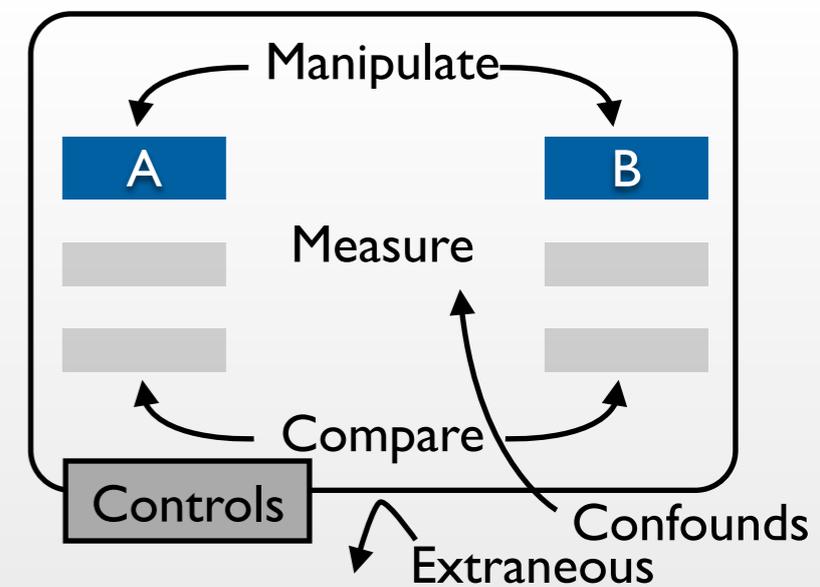
↑ external validity

Higher confidence
to infer causality
in the results

↑ internal validity



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)

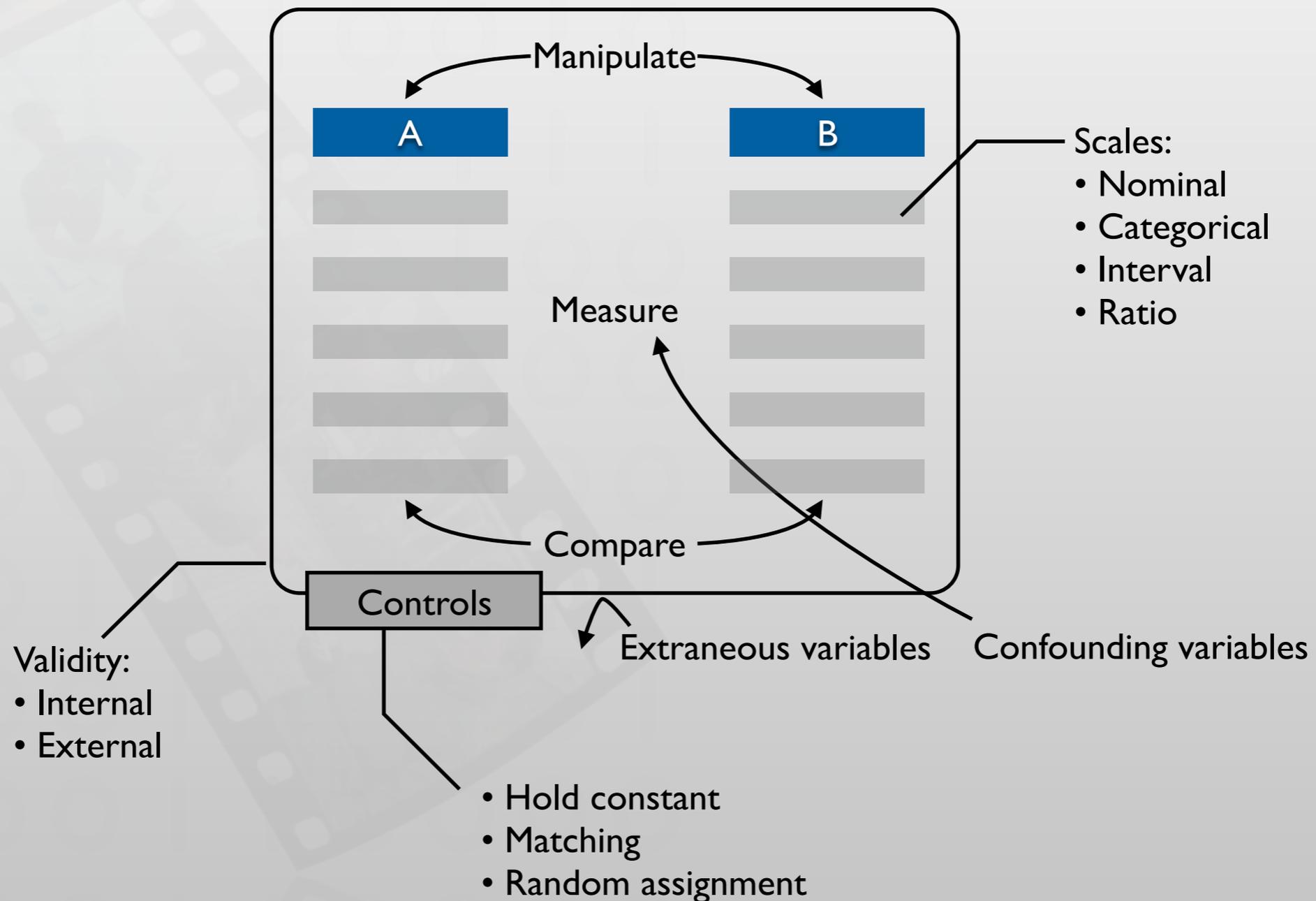


Controlling Extraneous Variables

- **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

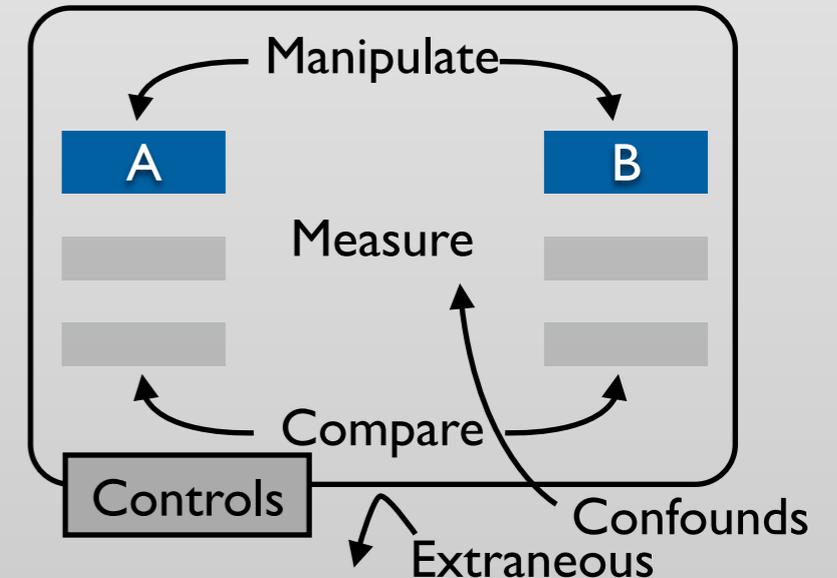


Basic Elements of Experimental Study



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
 - **Operational definition:** an exact description of what the variables are and how they are measured in your study.
 - In-class exercise: Give an **operational definition** of each variable, and indicate on which **scale** it is measured
- 



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

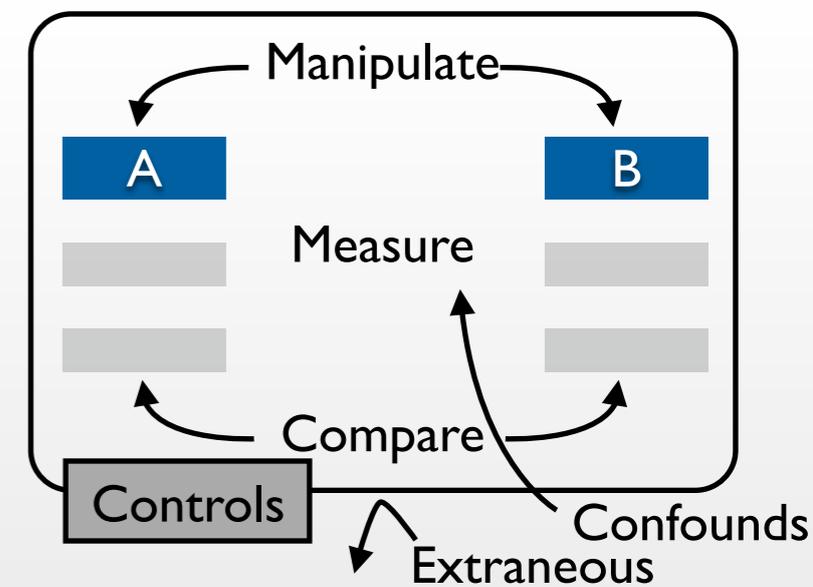
– Disregards the process of entering

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



Text Entry Tasks

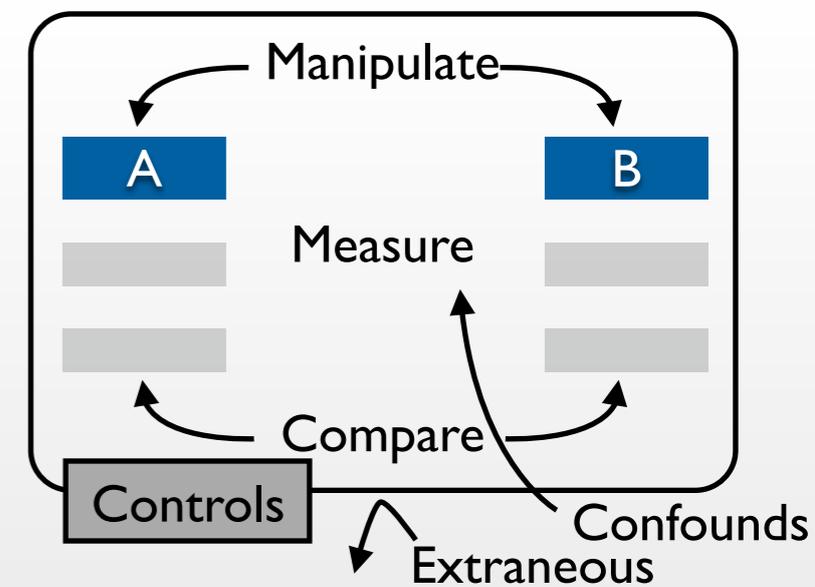
- **Composition:** user create his own text
 - More realistic
 - Users may take inconsistent durations to think about what to write
 - Error identification is difficult
- **Transcription:** copy text
 - Exclude behaviors that may compromise the measures, e.g., pondering what to write
 - Allows identifying 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.

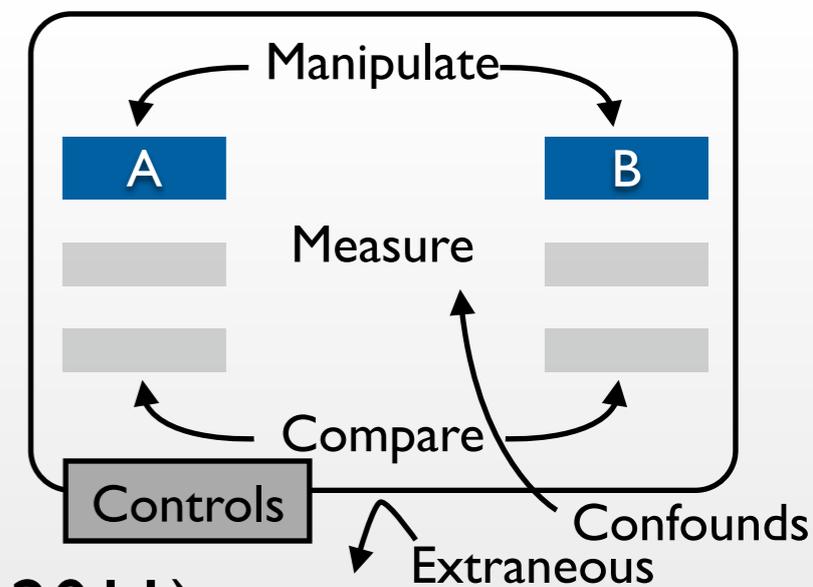
+ Allows replication

- 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

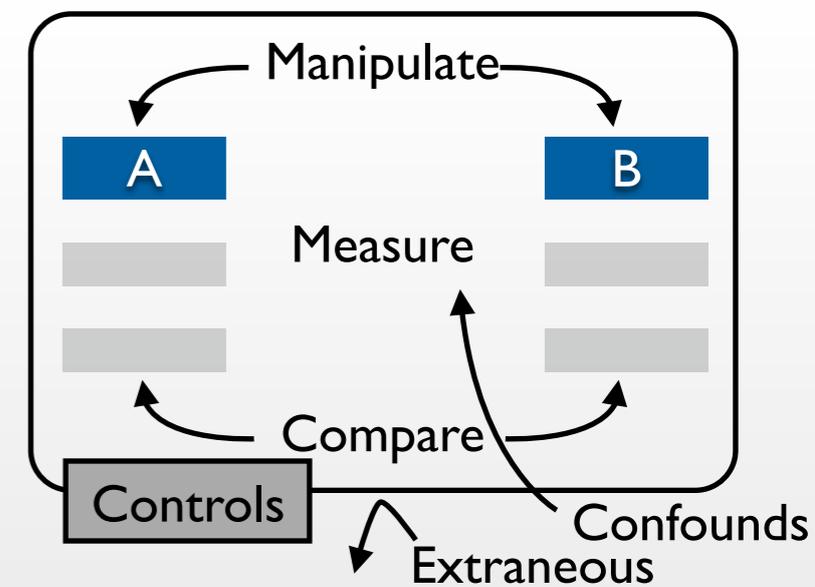
there will be some fog tonight
round robin scheduling
time to go shopping
frequently asked questions

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



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

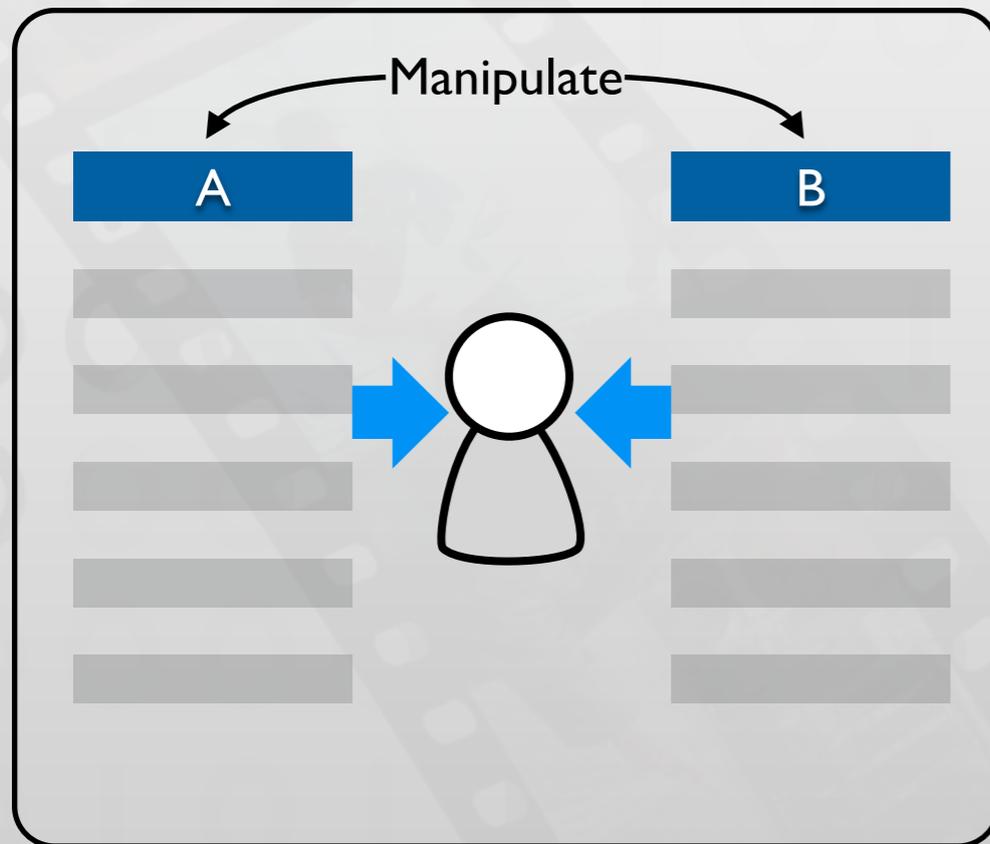
- + Fewer users required, individual differences canceled out

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

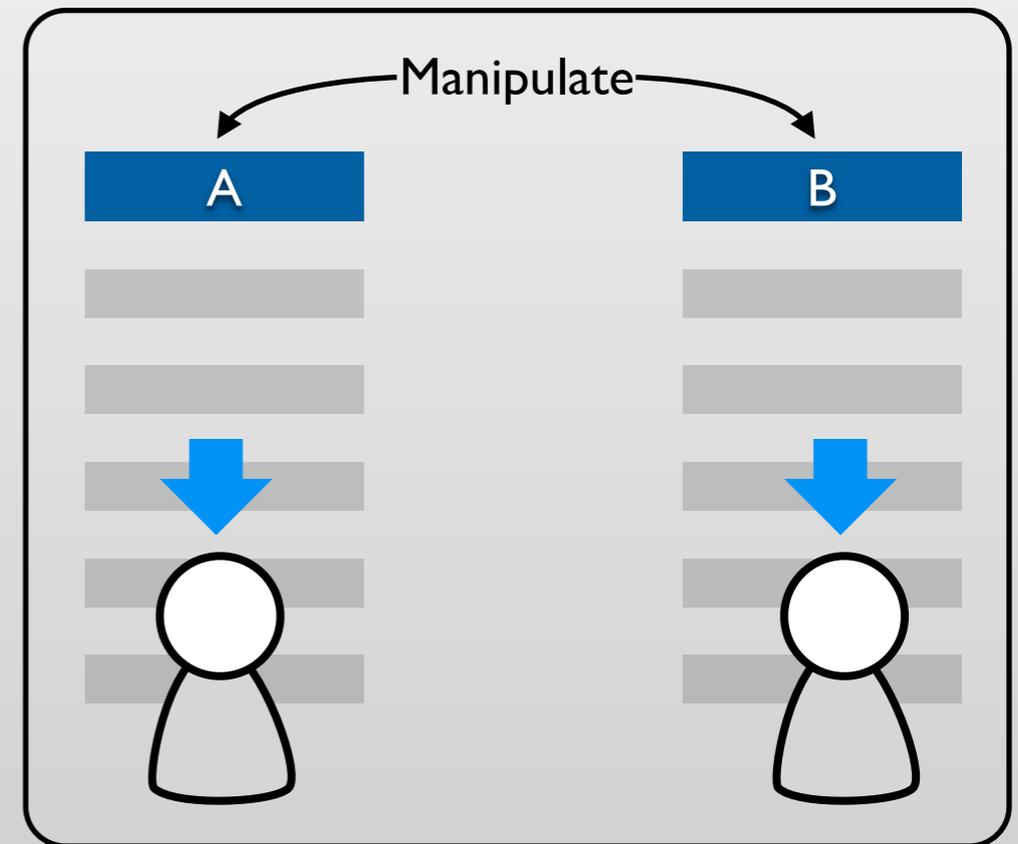
- But: practice effects may occur



Basic Experimental Designs

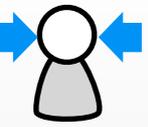


Within-subjects design



Between-subjects design

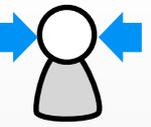




Order Effects

- 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

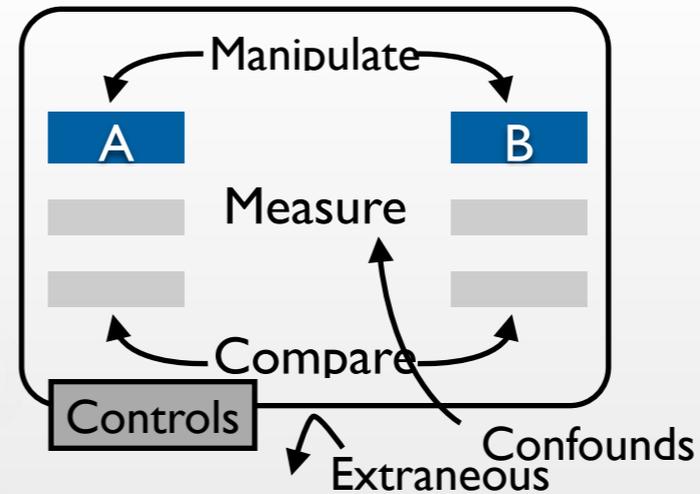
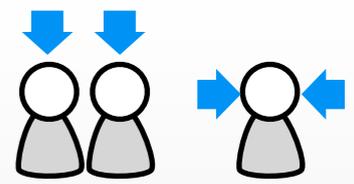
From DISI

- 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

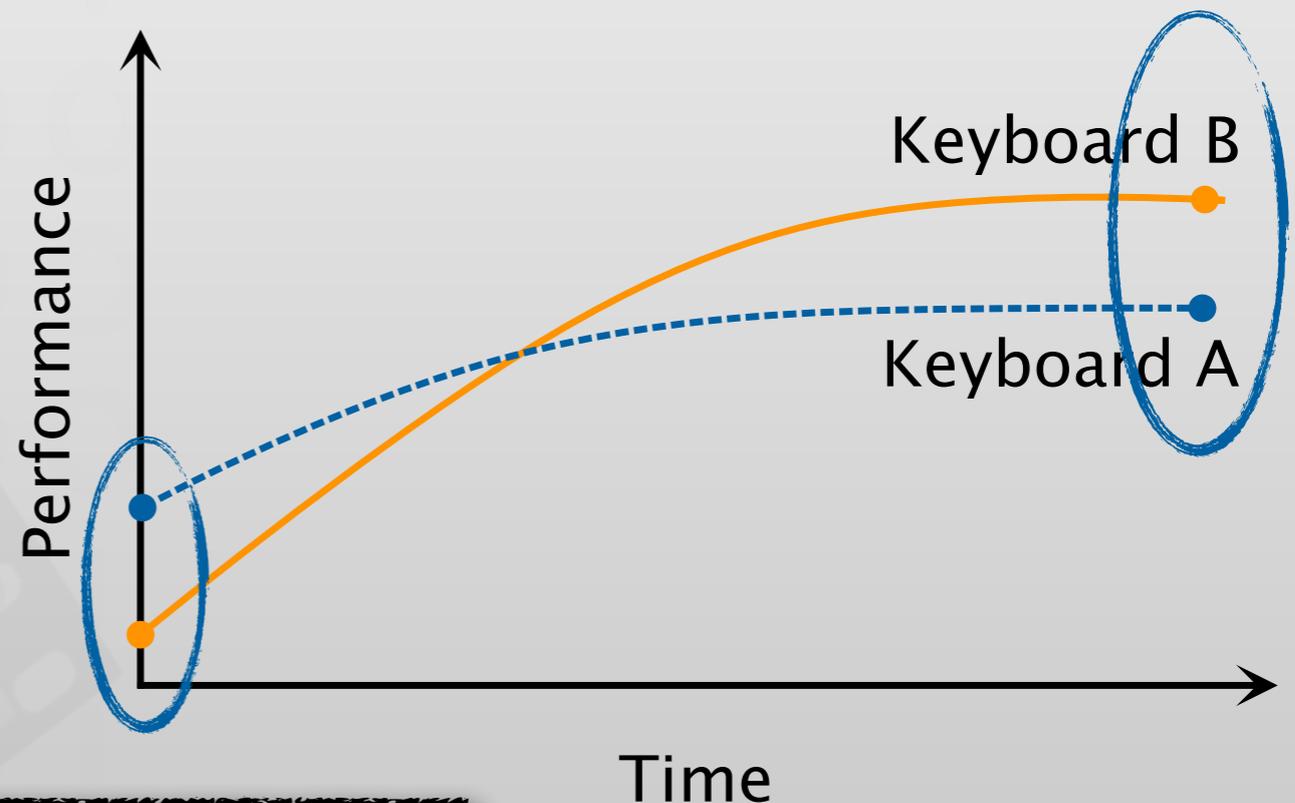


Learning Curve



- **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

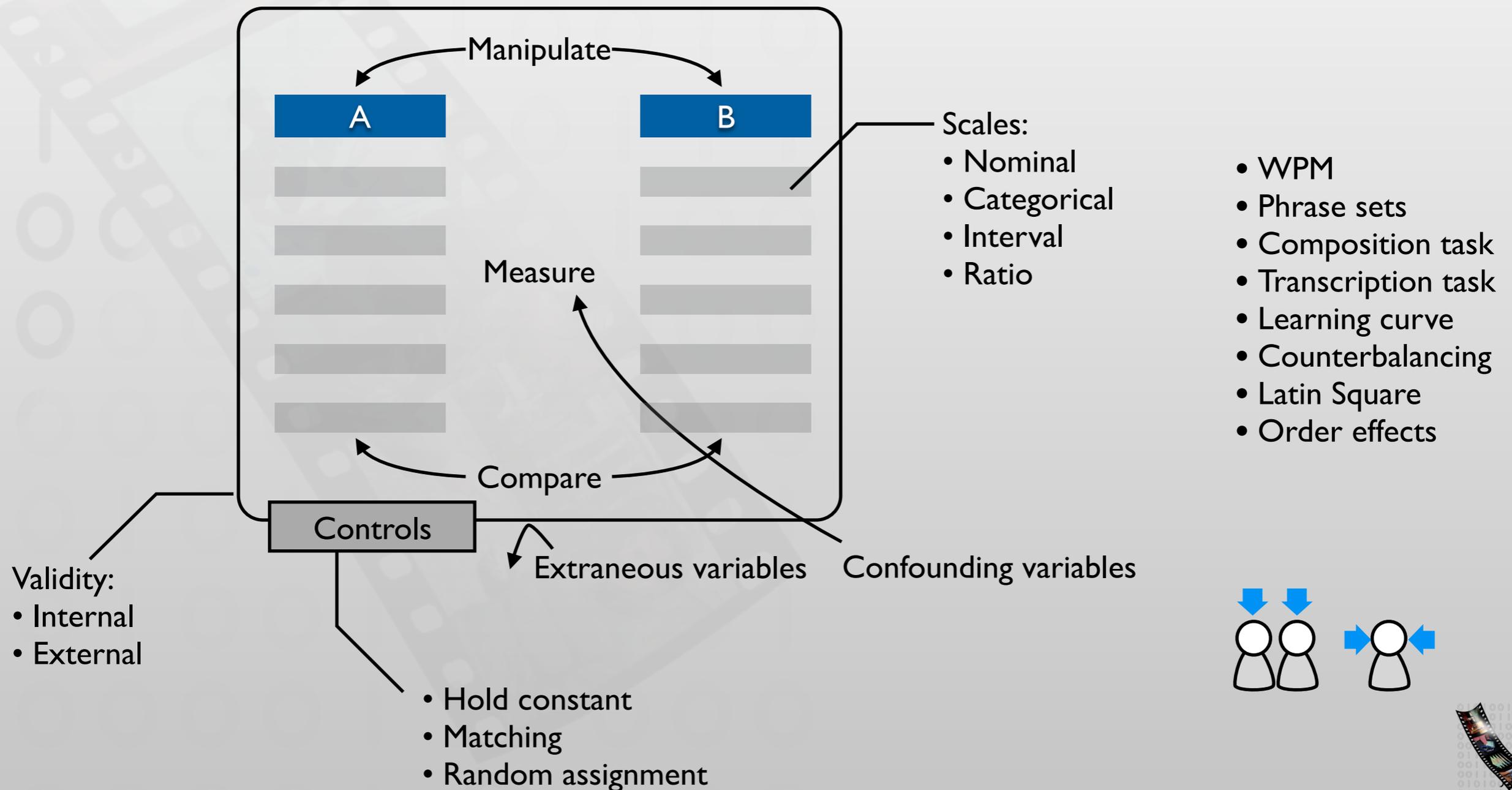


Experimental Design in Text Entry Research

- 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



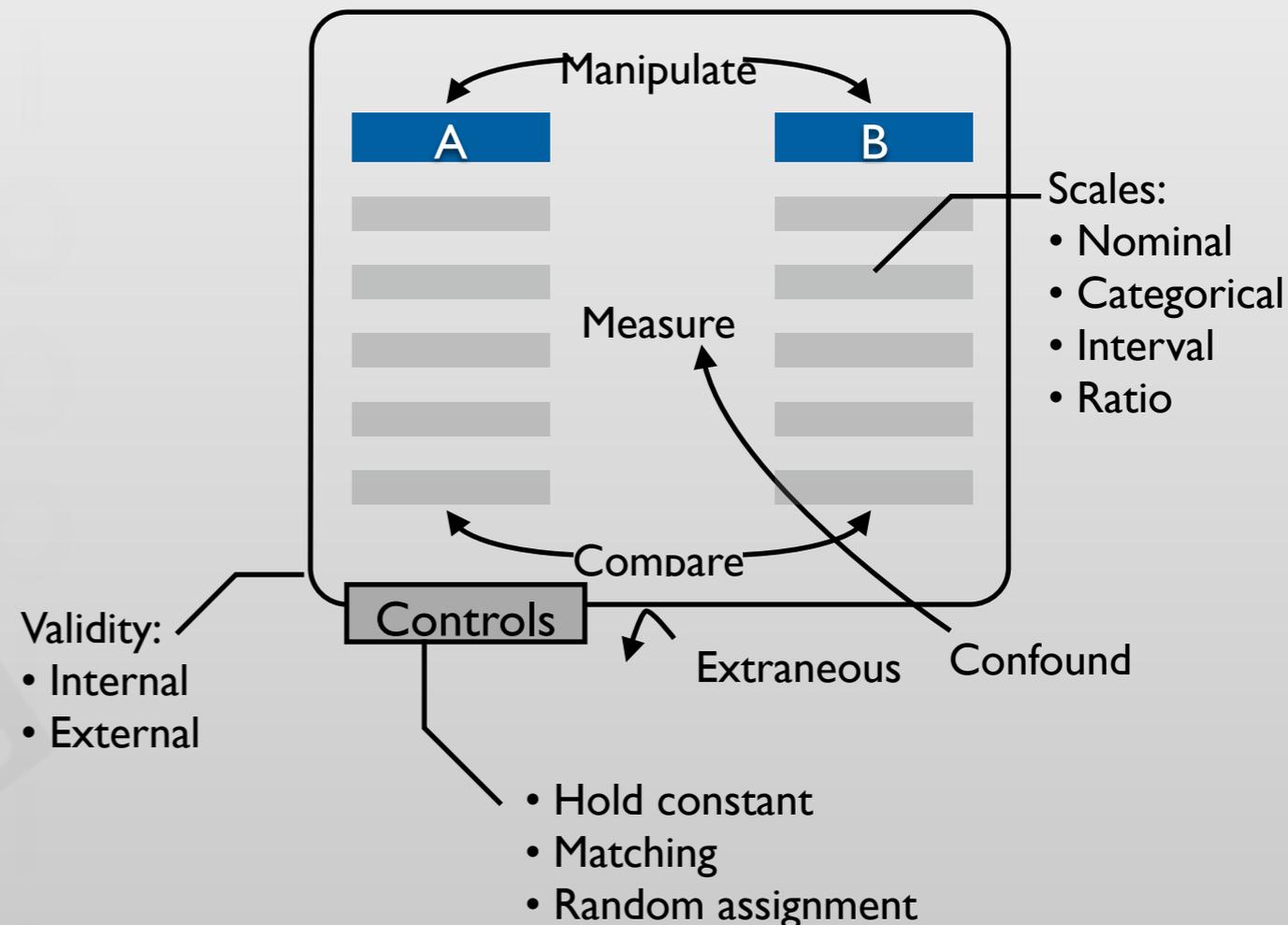
Basic Elements of Experimental Study in Text Entry Studies



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 (all links on our public website)
 - [Evaluation of Text Entry Techniques](#) — MacKenzie, 2007
- Practice reverse engineering
 - [Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild](#) — Kristensson, 2015
- Optional reading
 - [Complementing Text Entry Evaluations with a Composition Task](#) — Vertanen and Kristensson, TOCHI 2014
 - [Measures of Text Entry Performance](#) — Wobbrock, 2007

CHI 2015: <http://confer.csail.mit.edu/chi2015/papers>



S02 Referenced Literature 1/2

- Anders Markussen, Mikkel R Jakobsen, and Kasper Hornbaek. Vulture: A Mid-Air Word-Gesture Keyboard. In (To appear) CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, ON, Canada, 2014. ACM.
- Wobbrock, Jacob O. "Measures of Text Entry Performance." In "Text Entry Systems: Mobility, Accessibility, Universality" (2007): 75-101.
- Gravetter, Frederick J., and Lori-Ann B. Forzano. Research methods for the behavioral sciences. Wadsworth Publishing Company, 2011.
- I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03). ACM, New York, NY, USA, 754-755. DOI=10.1145/765891.765971 <http://doi.acm.org/10.1145/765891.765971>
- Keith Vertanen and Per Ola Kristensson. 2011. A versatile dataset for text entry evaluations based on genuine mobile emails. In Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11). ACM, New York, NY, USA, 295-298. DOI=10.1145/2037373.2037418 <http://doi.acm.org/10.1145/2037373.2037418>



S02 Referenced Literature 2/2

- Per Ola Kristensson and Keith Vertanen. 2012. Performance comparisons of phrase sets and presentation styles for text entry evaluations. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (IUI '12). ACM, New York, NY, USA, 29-32. DOI=10.1145/2166966.2166972 <http://doi.acm.org/10.1145/2166966.2166972>
- Keith Vertanen and Per Ola Kristensson. 2014. Complementing text entry evaluations with a composition task. ACM Trans. Comput.-Hum. Interact. 21, 2, Article 8 (February 2014), 33 pages. DOI=10.1145/2555691 <http://doi.acm.org/10.1145/2555691>
- Vittorio Fucella, Poika Isokoski, and Benoit Martin. 2013. Gestures and widgets: performance in text editing on multi-touch capable mobile devices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13). ACM, New York, NY, USA, 2785-2794. DOI=10.1145/2470654.2481385 <http://doi.acm.org/10.1145/2470654.2481385>
- MacKenzie, I. Scott, and K. Tanaka-Ishii. "Evaluation of text entry techniques." In "Text entry systems: Mobility, accessibility, universality" (2007): 75-101.
- Wobbrock, Jacob O. Measures of text entry performance. In "Text entry systems: Mobility, accessibility, universality" (2007): 47

