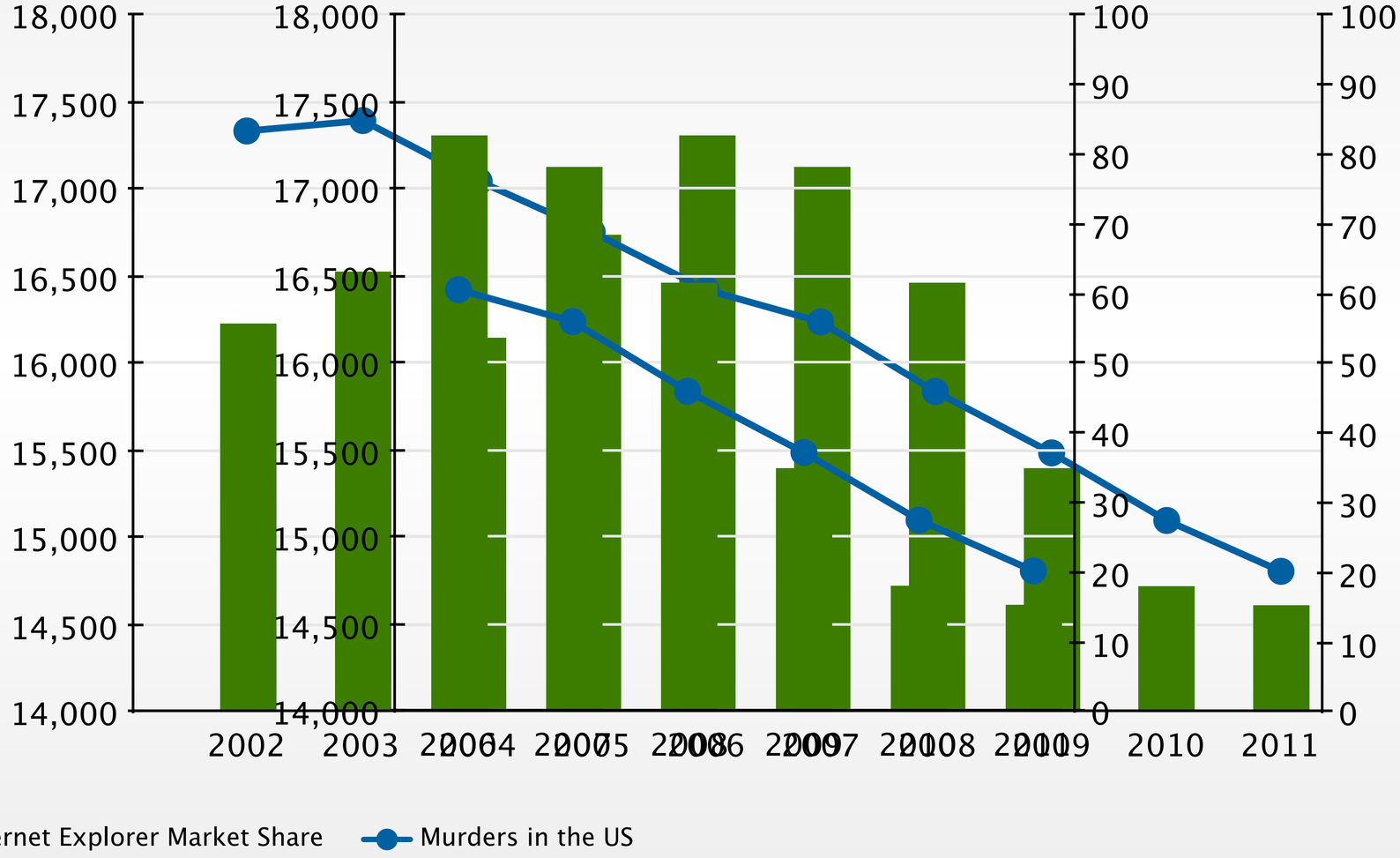
WARNING

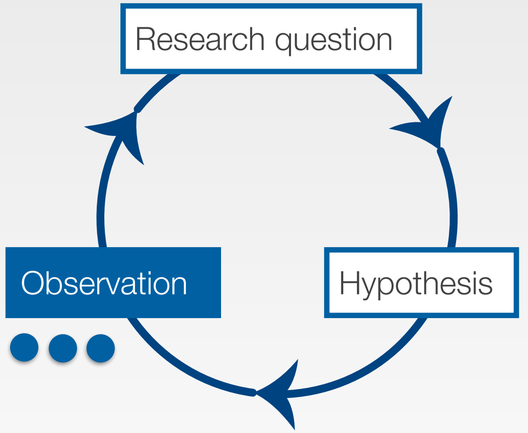
Correlation Does Not Imply Causation



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

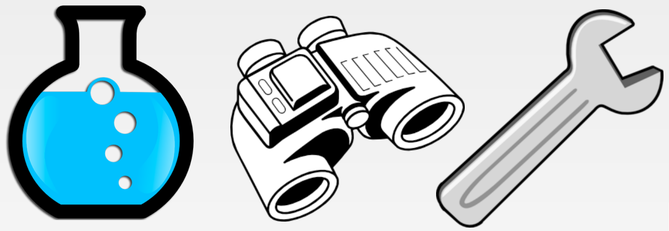
From Correlation to Causation: More about Experimental Research

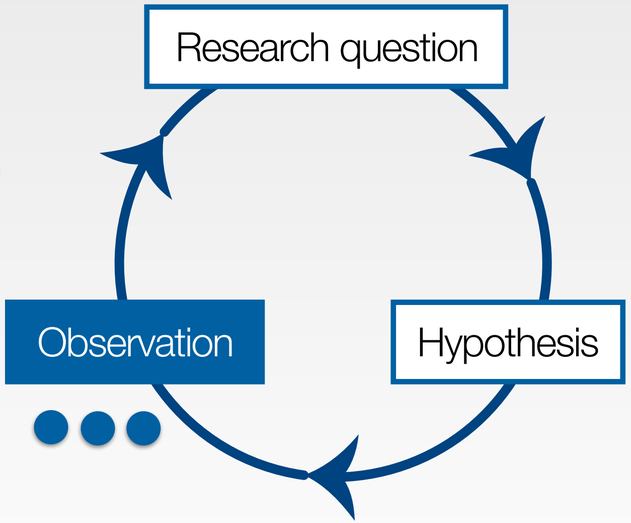
Experimental Research



- 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

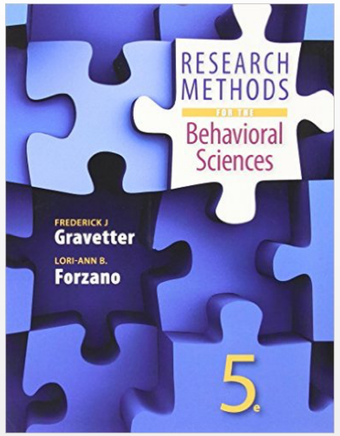
From the last lecture





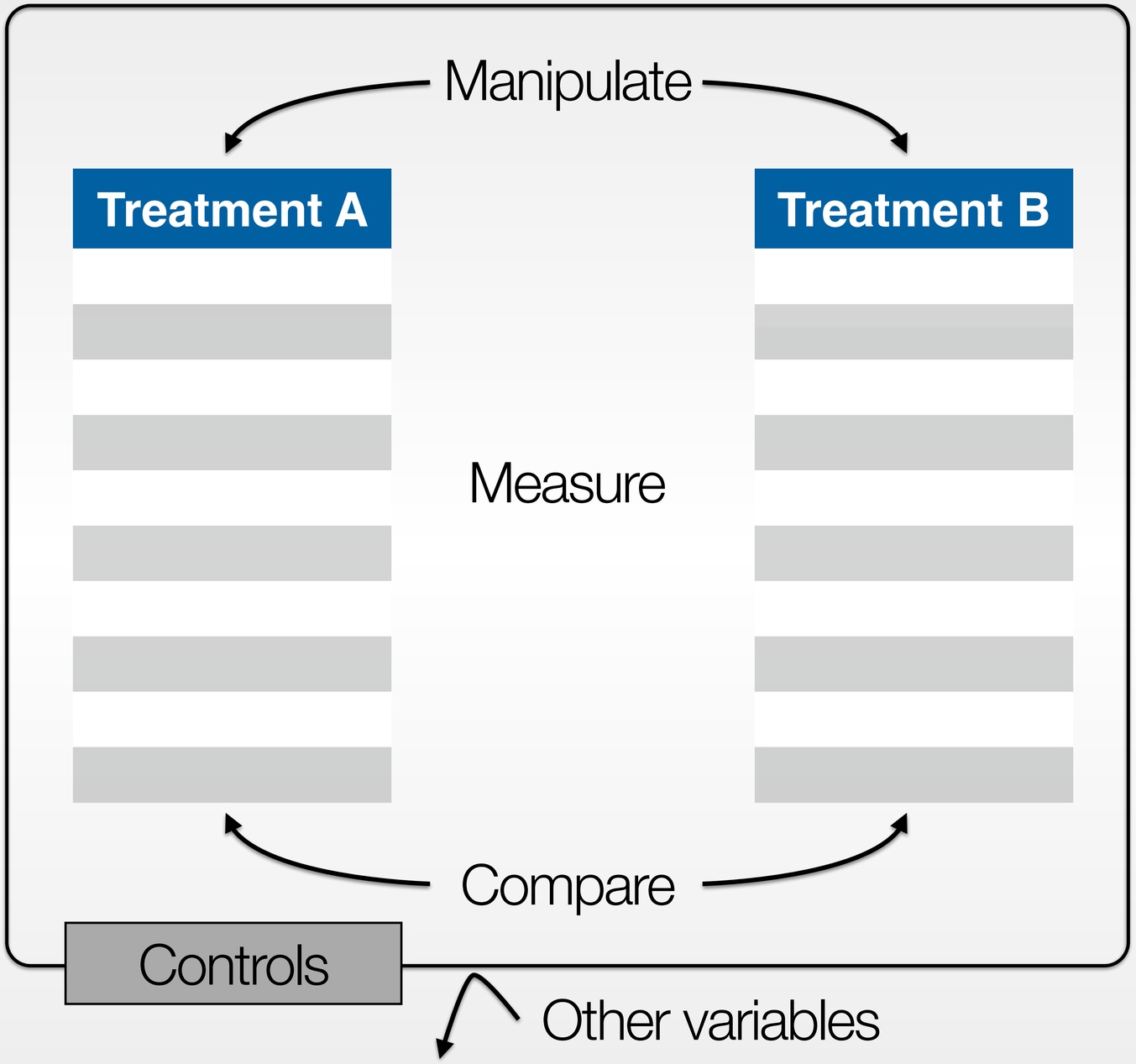
Experimental Research in HCI

Illustrated through Text Entry Research



Further reading:

Research Methods for the Behavioral Sciences (Gravetter and Forzano, 2015)



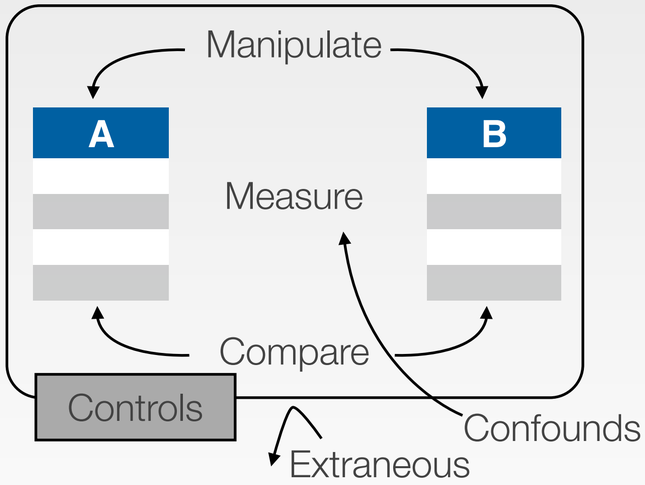
Adapted from Gravetter and Forzano

Basic Elements of Experimental Studies

- 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 ⇒ 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

Variables



- Independent variables are manipulated by the researcher
- Dependent variables are observed for changes to assess the effect of the independent variables
- All other variables: extraneous variables
- A confounding variable is an extraneous variable that changes systematically along with IV and DVs ⇒ 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 in which zero represents complete absence (possible to determine absolute distances)
  - E.g., Task completion time in seconds (DV), error rate in percent (DV)

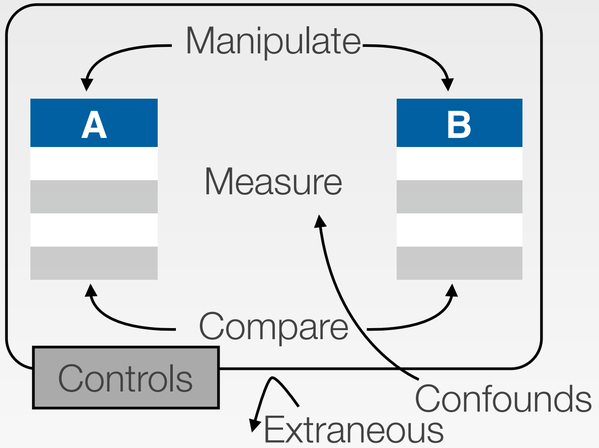
* Can be treated as ordinal (strictly according to the definition) or interval (empirically verified over 50 years to be OK)

Dealing with Extraneous Variables

- Include them as IVs ⇒ too many experimental conditions!

Leave as random	⟷	Control
Reflects variation in natural use		Higher confidence to infer causality in the results
↑ external validity		↑ internal validity

Validity



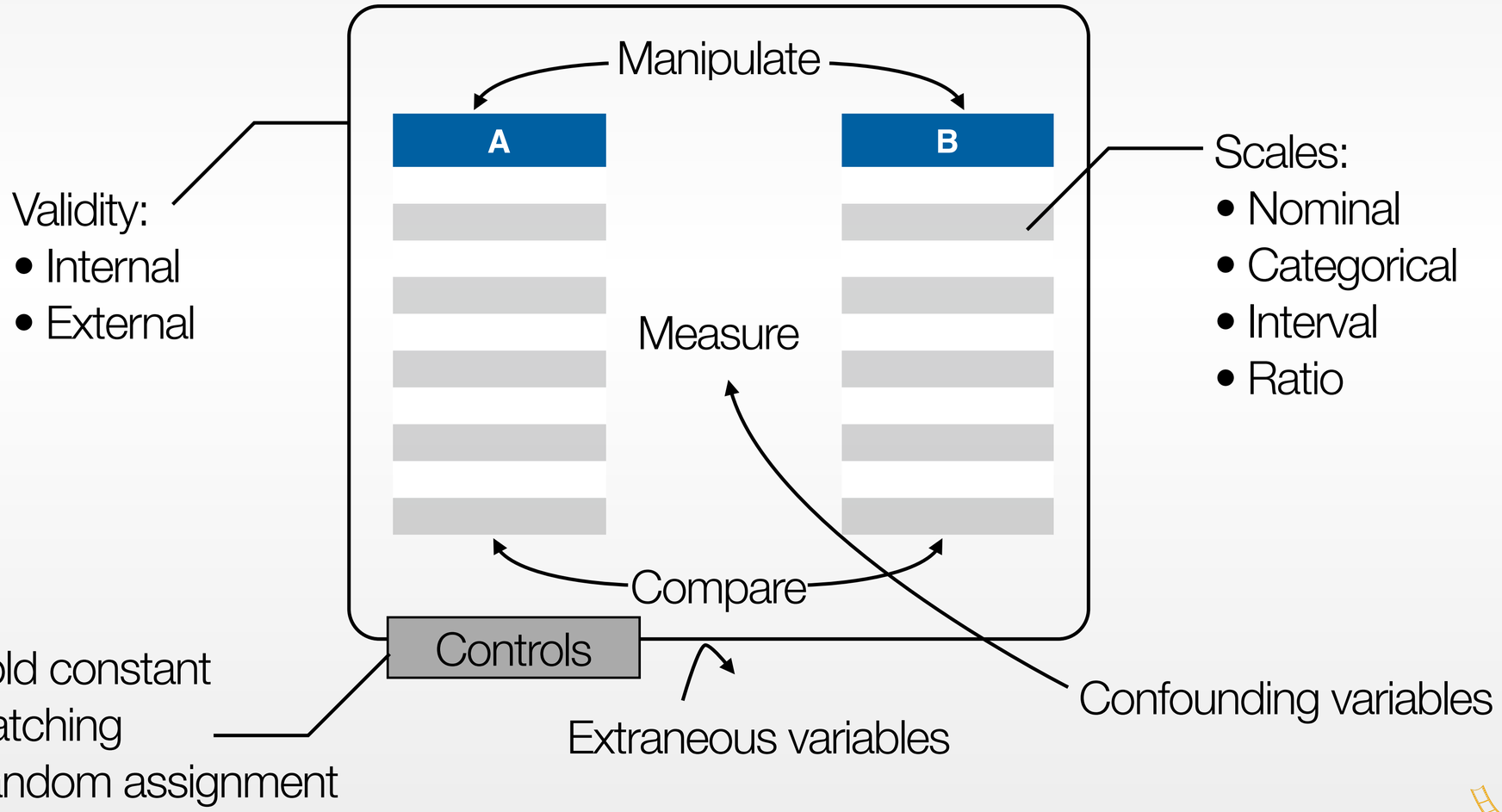
- A study has internal validity if it produces a single, unambiguous explanation for the relationship between two variables
  - Threats: e.g., confounding variables, experimenter bias, learning effect, Hawthorne effect (being observed causes the changes)
- External validity refers to the extent to which 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

Controlling Extraneous Variables

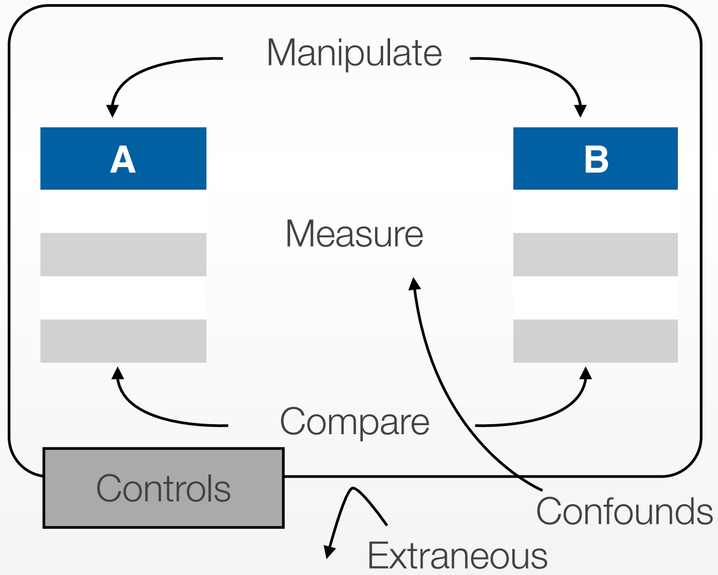
- Hold constant, e.g., selecting participants in the same gender/age
- Matching the same number of participants with the same extraneous variable
  - E.g., gender, age, or level of expertise
- Random assignment of participants to treatment conditions
  - Other random assignment, e.g., time slot

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 are users?
- 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

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

$\lvert T \rvert$	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

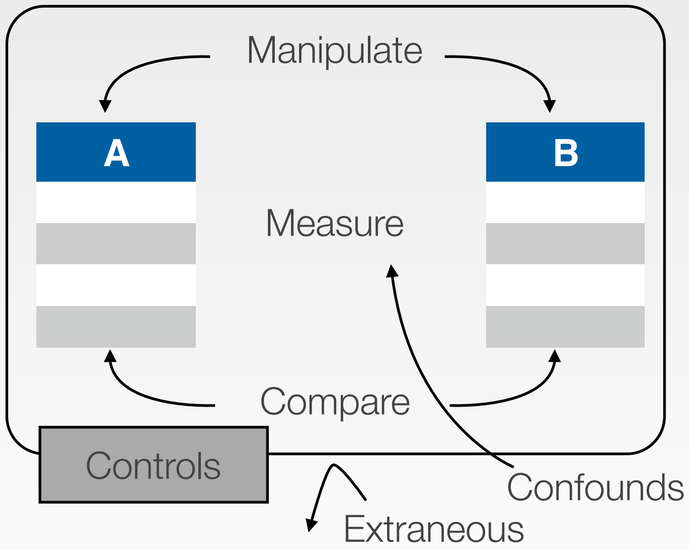
\– Disregards errors in the final text

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

\– Disregards the process of entering

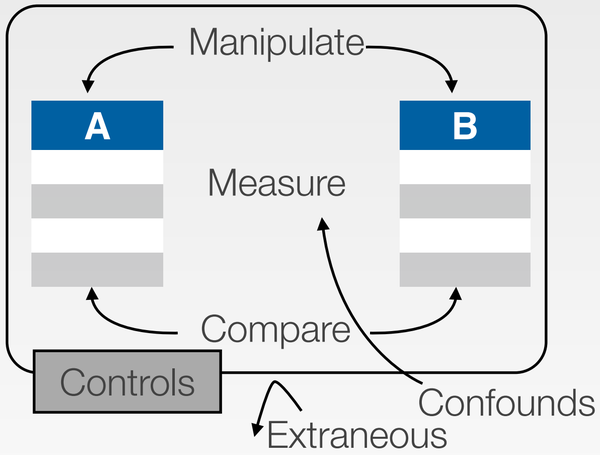
- E.g., it doesn't matter how many times you pressed the backspace key.

Text Entry Tasks



- Composition: users create their own text
  - More realistic
  - Users may take inconsistent durations to think about what to write
  - Error identification is difficult
- Transcription: users copy a given text
  - Excludes behaviors that may compromise measurements, e.g., pondering what to write
  - Allows identifying errors, because the content is known
  - Allows controlling the distribution of letters and words

Text Entry Tasks

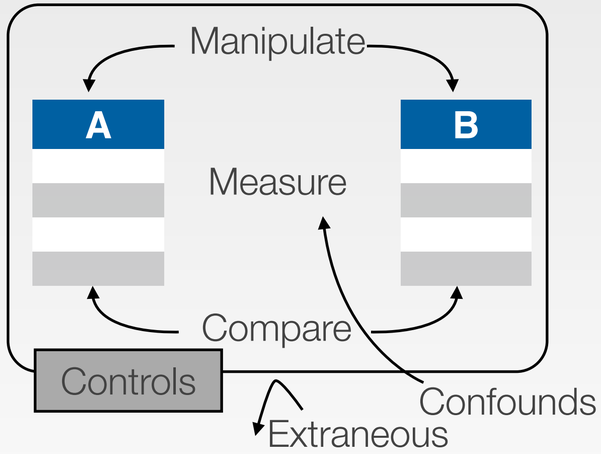


- Read and memorize a short sentence before entering
  - Reduce participants' tendency to switch between the displayed text and the entry text field
  - Faster typing but the overall experiment takes longer due to the memorizing [Kristensson & Vertanen, IUI'12]

there will be some fog tonight

there w_

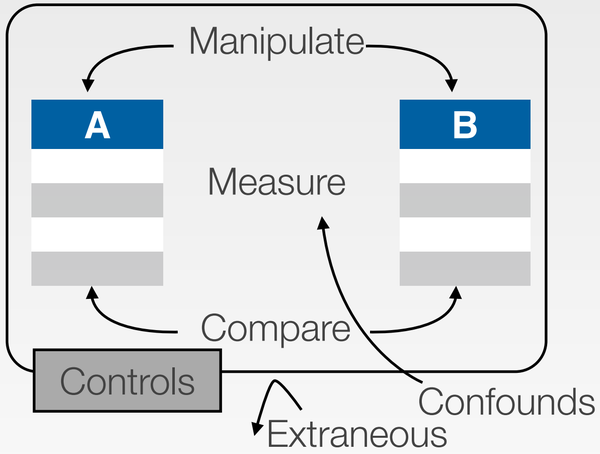
Standard Dataset for Transcription Task



- MacKenzie and Soukoreff (CHI 2003)
- 500 English phrases of moderate length, easy to remember, representative of the English language (in terms 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:

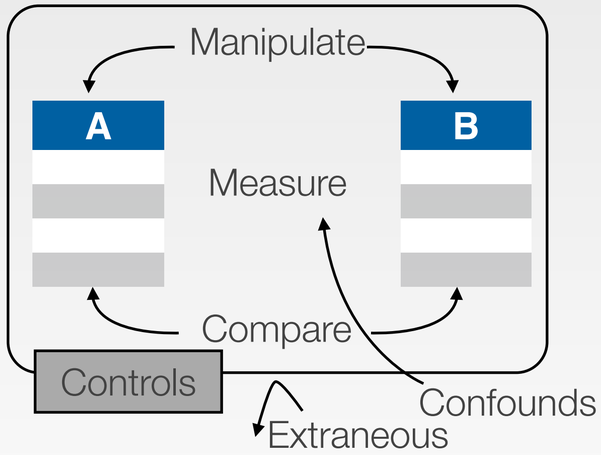
MacKenzie & Soukoreff

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

EnronMobile

```
Thanks, I will look at it tonight.
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) characterize and fine-tune text composition tasks 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

- Ensure task description is adequate, to control the quality. Example:
  - “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 DIS1

- Between-groups 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-groups 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