Photo credit: Dave DiCello

17-803 Empirical Methods

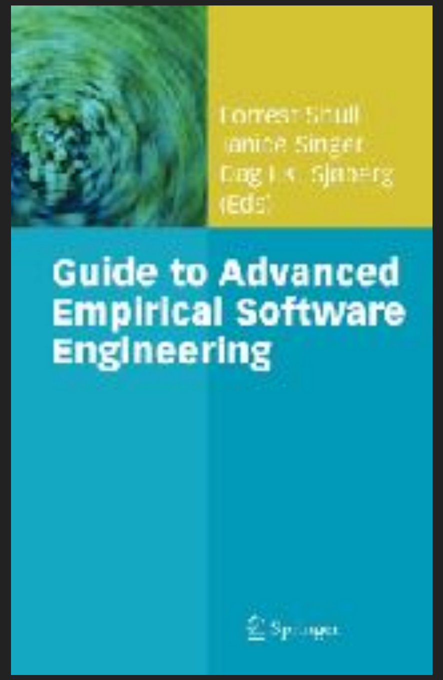
Bogdan Vasilescu, S3D

Designing Experiments (I)

Thursday, February 29, 2024

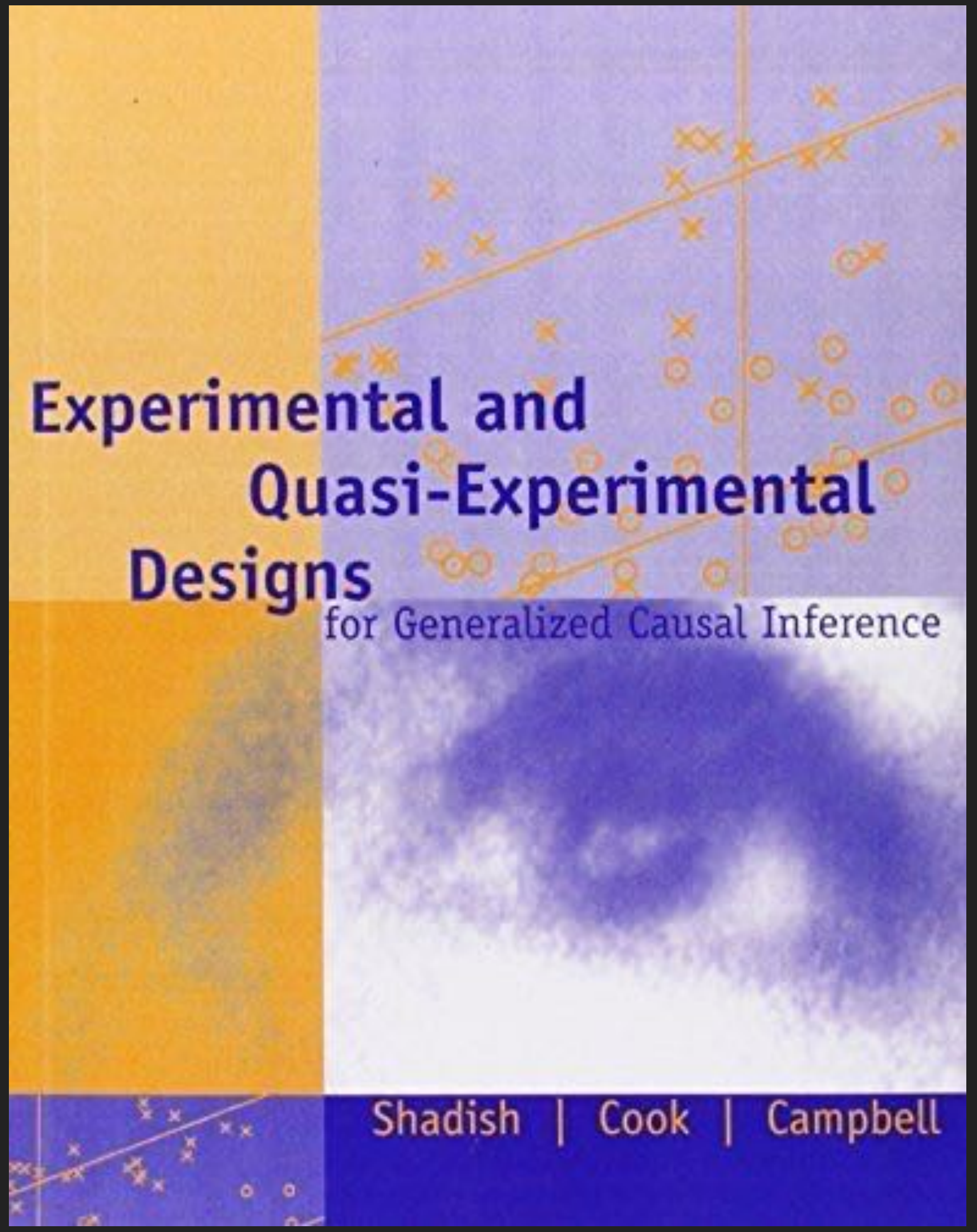
Readings

Ch 10 (Analysis and interpretation)

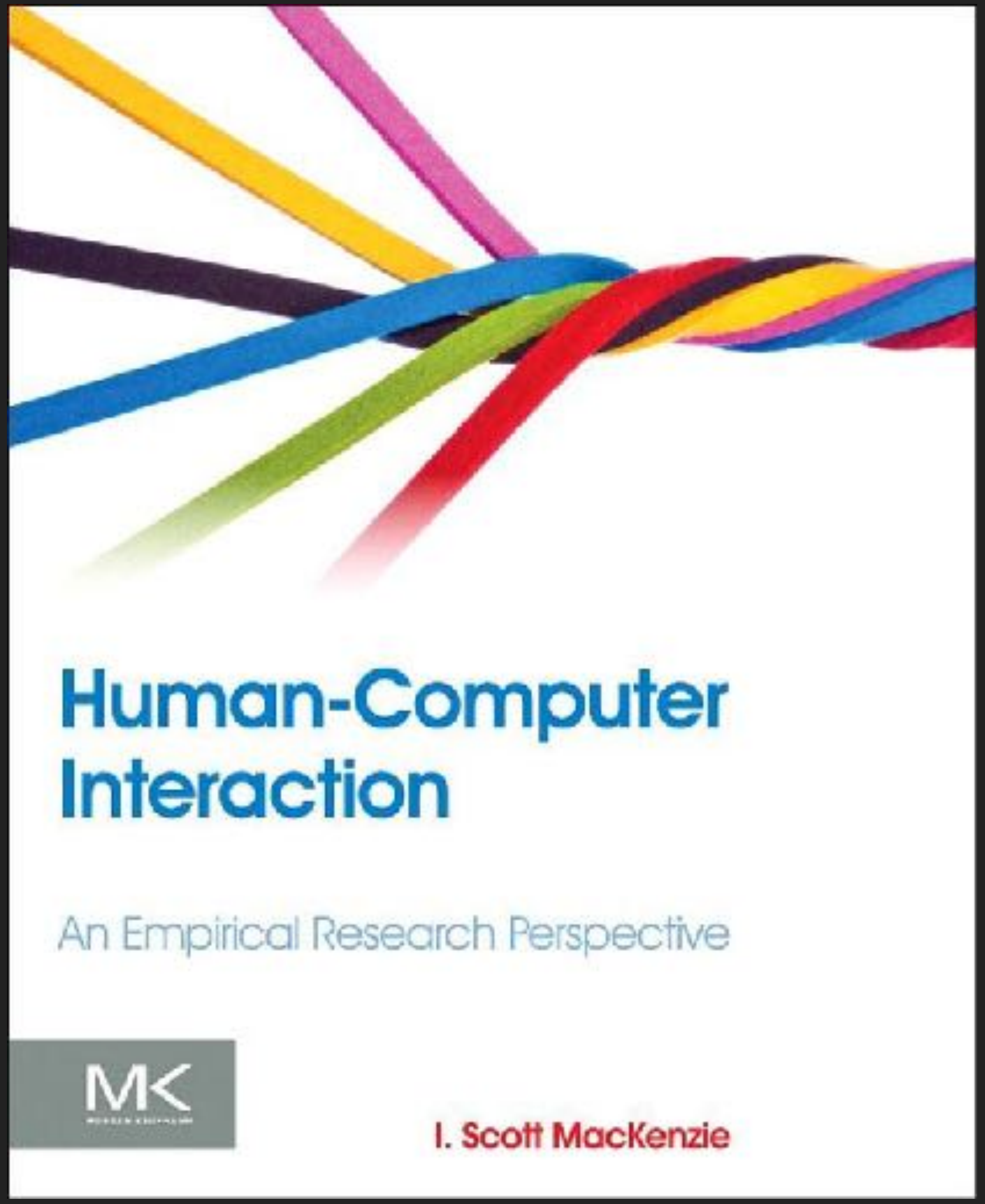


Ch 6 (Statistical methods and measurement)

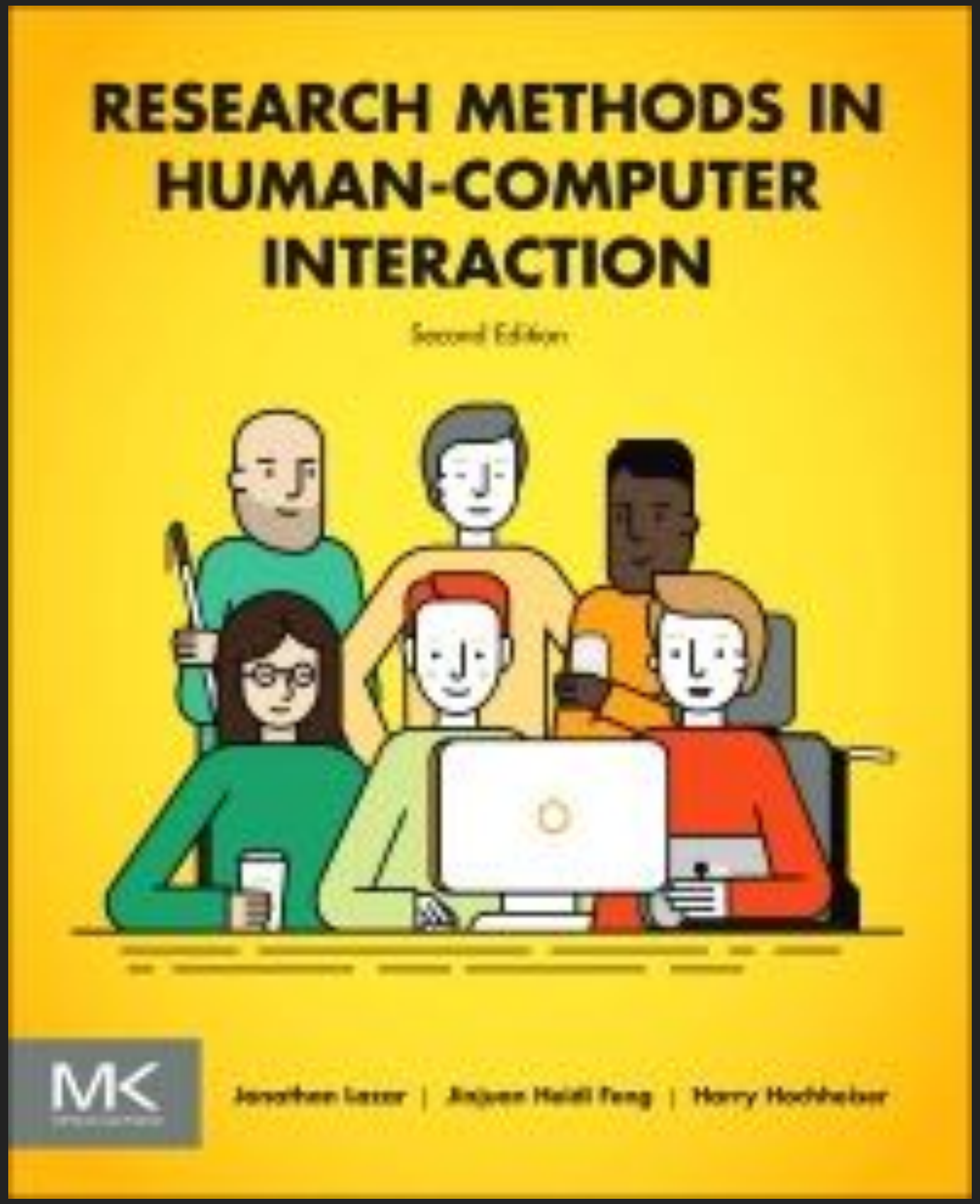
Ch 5 (Effect sizes and power analysis)
Ch 13 (Fair statistical communication)
Ch 14 (Improving statistical practice)



Ch 1 (Experiments and causality)
Ch 2 & 3 (Validity)
Ch 8 (Randomized experiments)



Ch 5 (Designing HCI Exp.)
Ch 6 (Hypothesis testing)



Ch 3 (Experimental design)
Ch 4 (Statistical analysis)

Credits

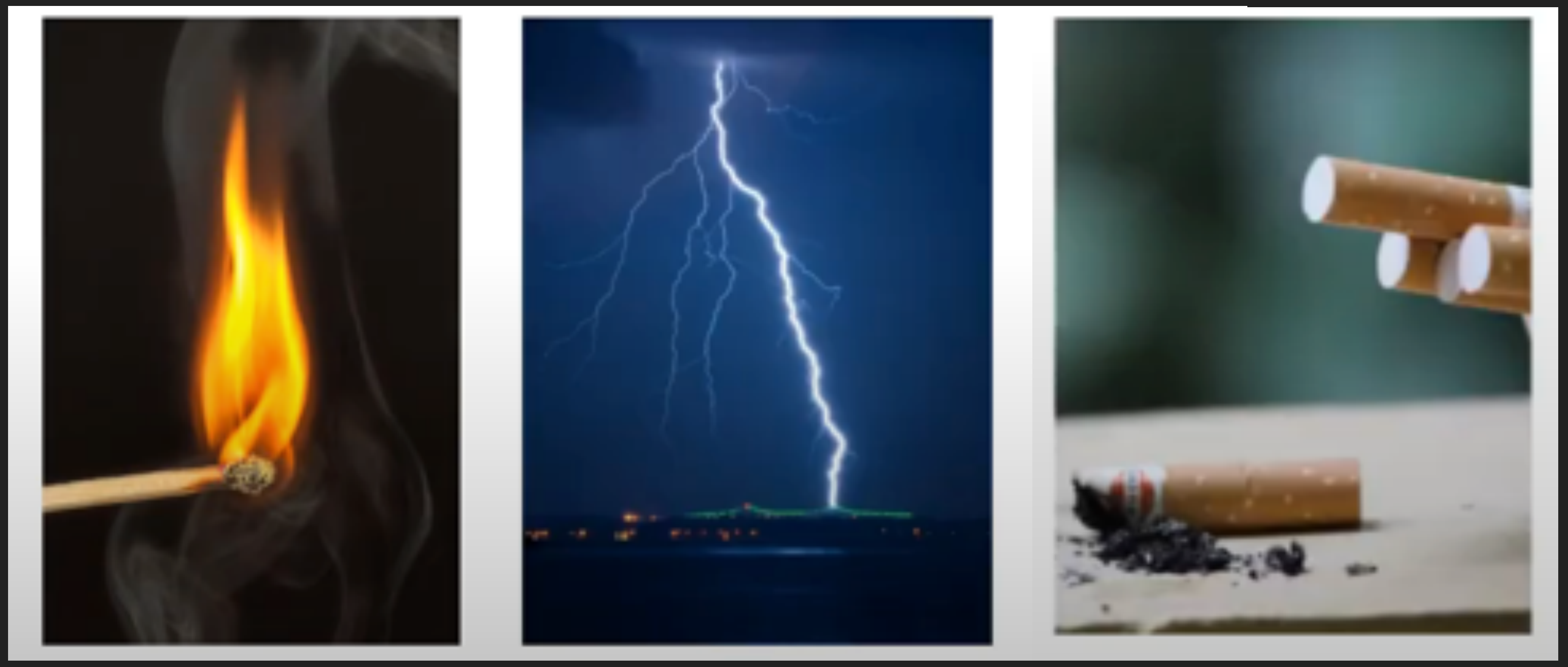
- Graphics:
  - Dave DiCello photography (cover)
- Content:
  - Chapters from Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Publishing
    - Ch1: Experiments and generalized causal inference
    - Ch2: Statistical conclusion validity and internal validity
    - Ch3: Construct validity and external validity
    - Ch8: Randomized experiments
  - Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media.
  - Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (2007). Statistics.

Causal relationships

Cause

- *inus condition* – "insufficient but nonredundant part of an unnecessary but sufficient condition"

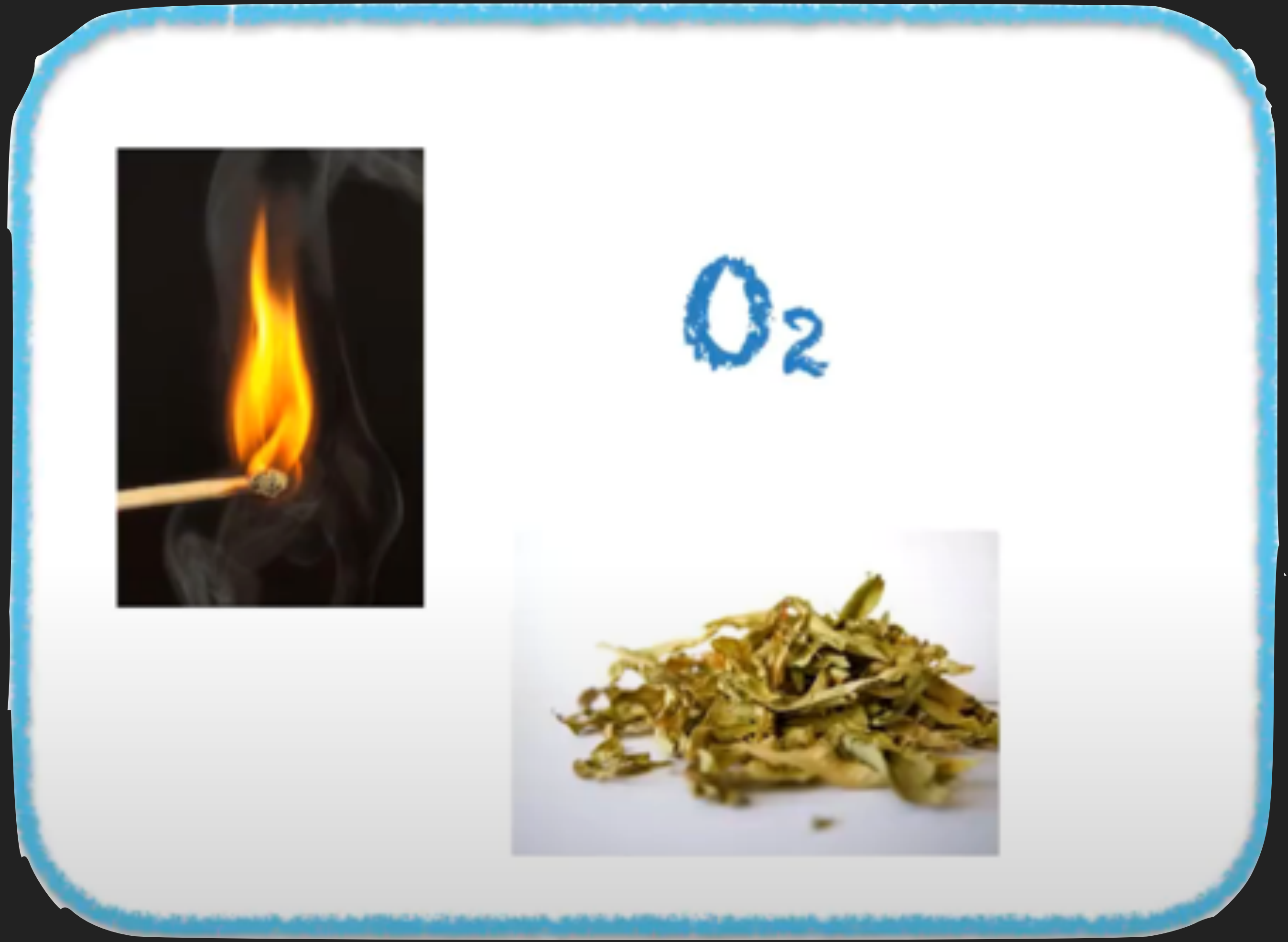
- Example: match to start a forest fire
- Fires can start even without matches
  - → Match is not a necessary condition
- Matches don't always start forest fires (e.g., not on long enough, rainy weather)
  - → Match is not a sufficient condition



Cause

- *inus condition* – "insufficient but nonredundant part of an unnecessary but sufficient condition"

- Match is part of a bigger constellation of conditions without which a fire would not result
  - Insufficient: needs oxygen, dry leaves, etc
  - Nonredundant: needs to add something unique besides oxygen, dry leaves, etc



Effect

- Counterfactual: what would have happened to these subjects had the cause not been present?
  - What did happen when people received a treatment, vs
  - What would have happened to those same people if they simultaneously had not received the treatment ("counterfactual", i.e., contrary to fact)
  - Effect is distance between the two

- Can't observe, must infer / approximate.

Experimental design:

- Creating a high-quality but necessarily imperfect source of counterfactual inference
- Understanding how this source differs from the treatment condition

Ingredients for Establishing a Causal Relationship?

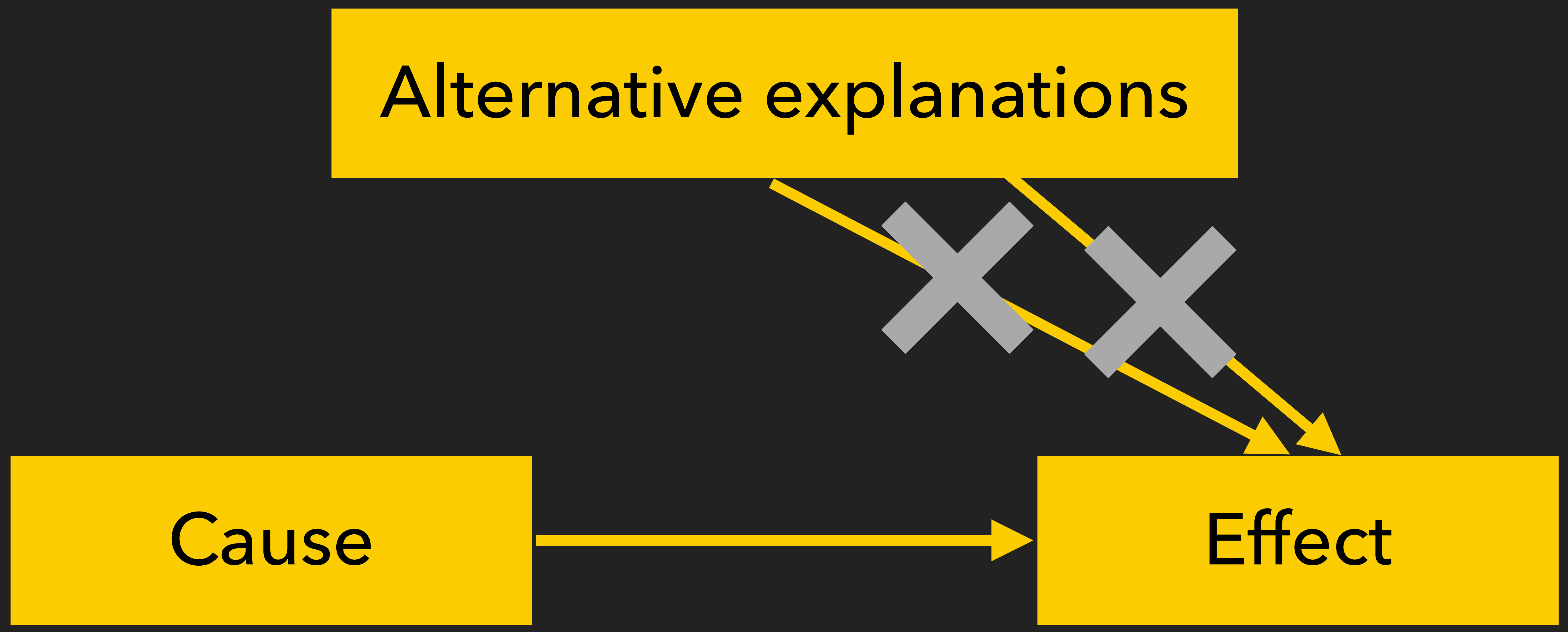
Ingredients for Establishing a Causal Relationship

The cause preceded the effect

The cause was related to the effect

We can find no plausible alternative explanation for the effect other than the cause

Ingredients for Establishing a Causal Relationship



Alternative explanations

Cause

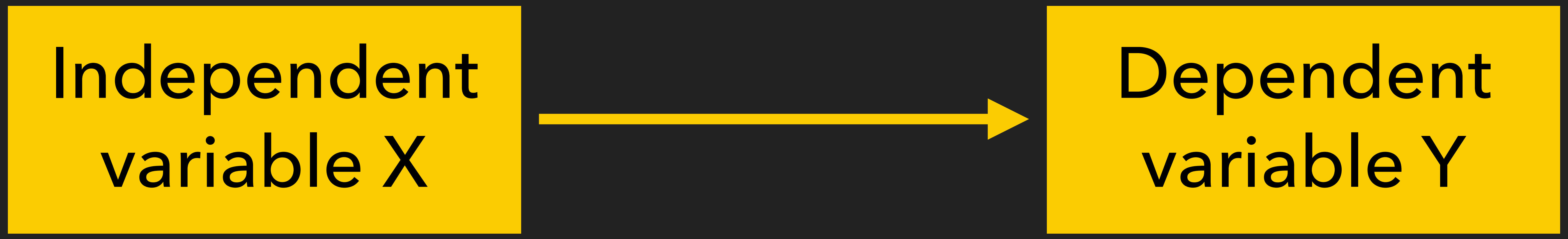
Effect

Note how this mirror what happens in experiments.

No other scientific method regularly matches the characteristics of causal relationships so well.

Aside: Mediators & Moderators

Mediators and Moderators

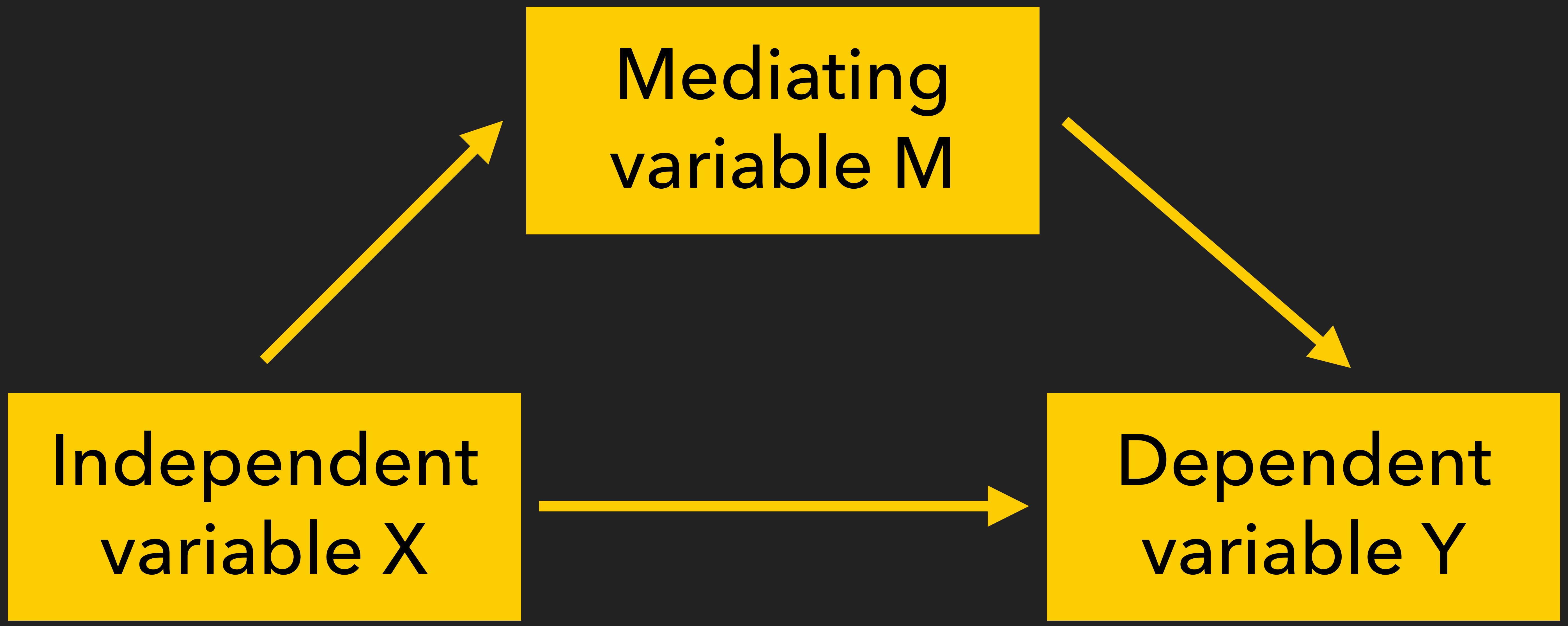


Independent variable X

Dependent variable Y

Mediators and Moderators

Links in the explanatory chain: Mediator

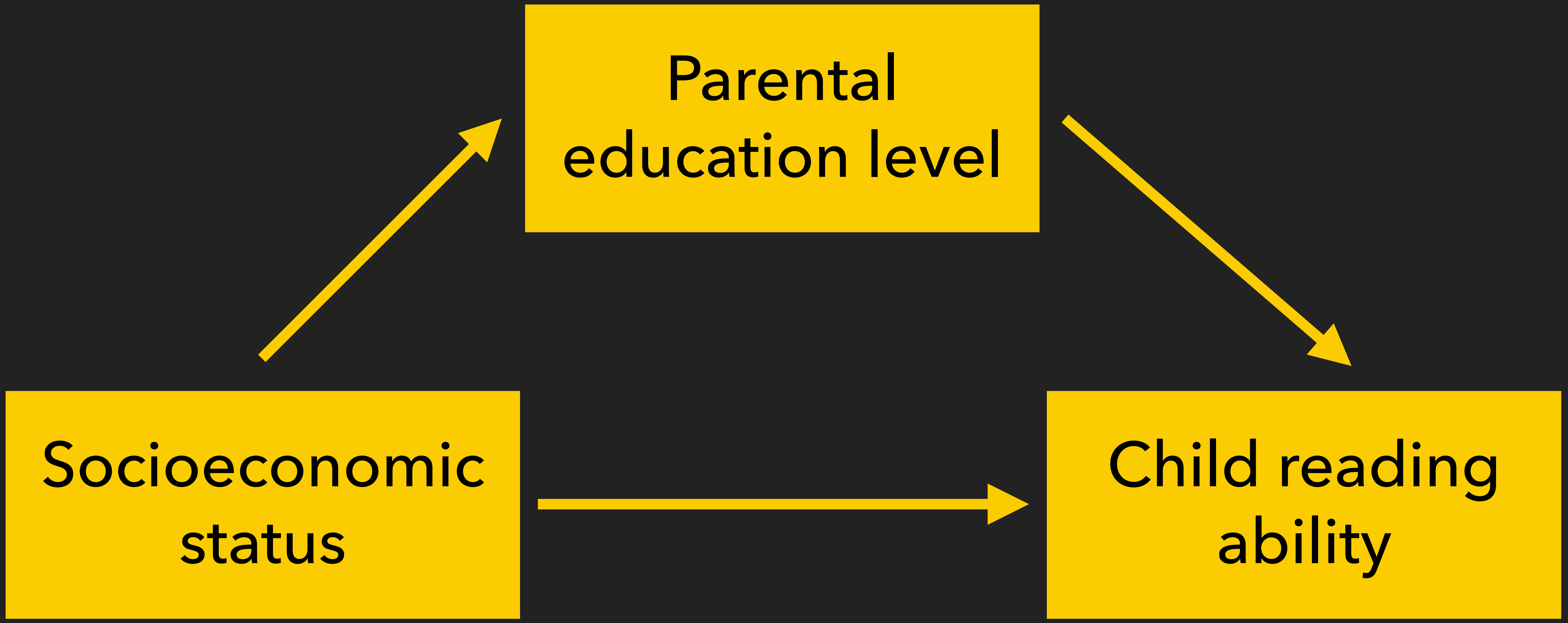


Mediating variable M

Independent variable X

Dependent variable Y

Mediators and Moderators

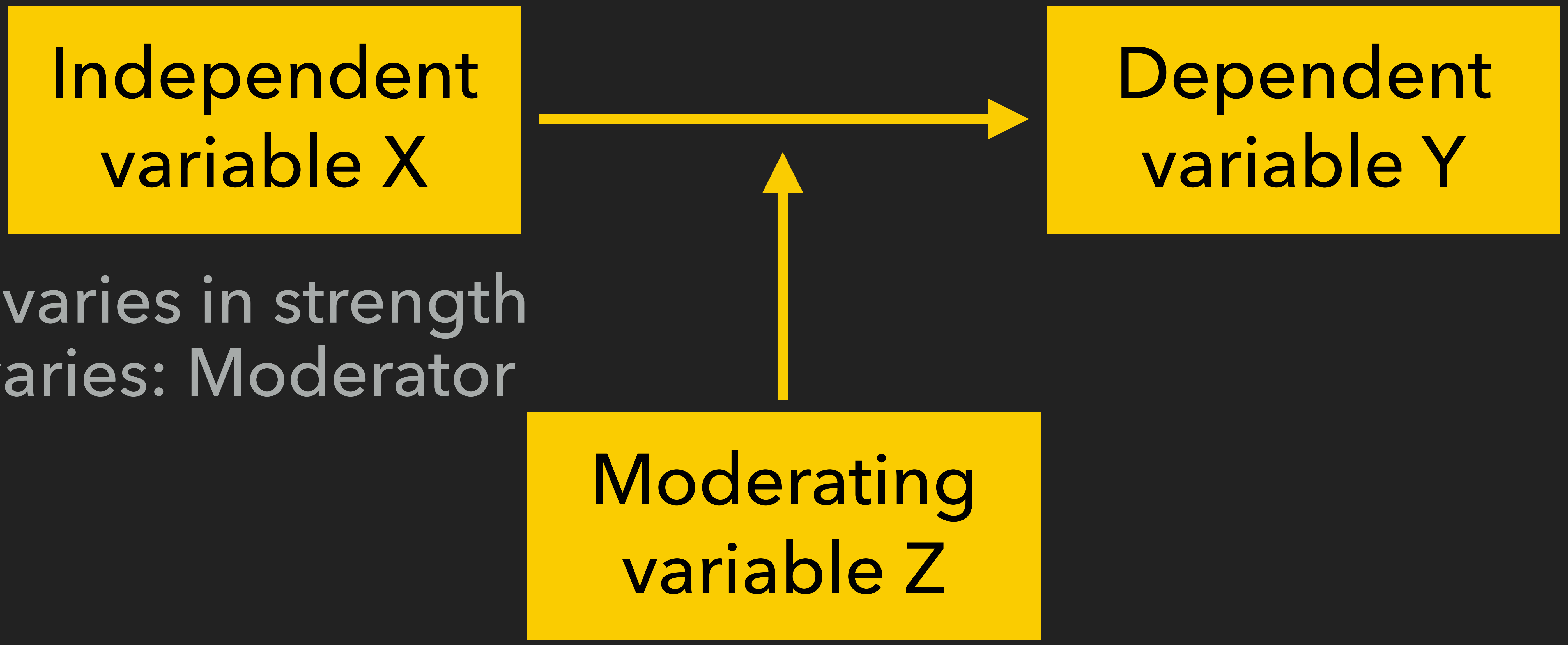


Parental education level

Socioeconomic status

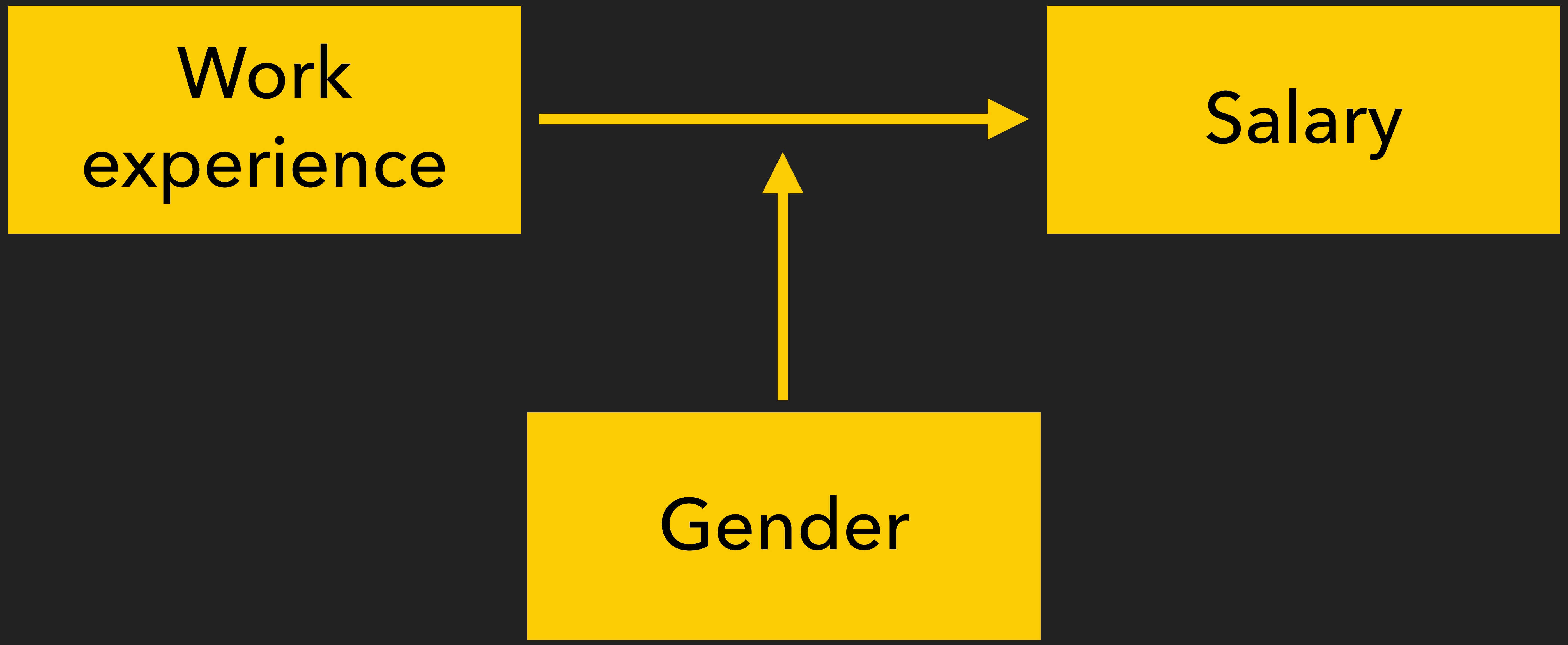
Child reading ability

Mediators and Moderators



Causal relationship varies in strength (or direction) as Z varies: Moderator

Mediators and Moderators

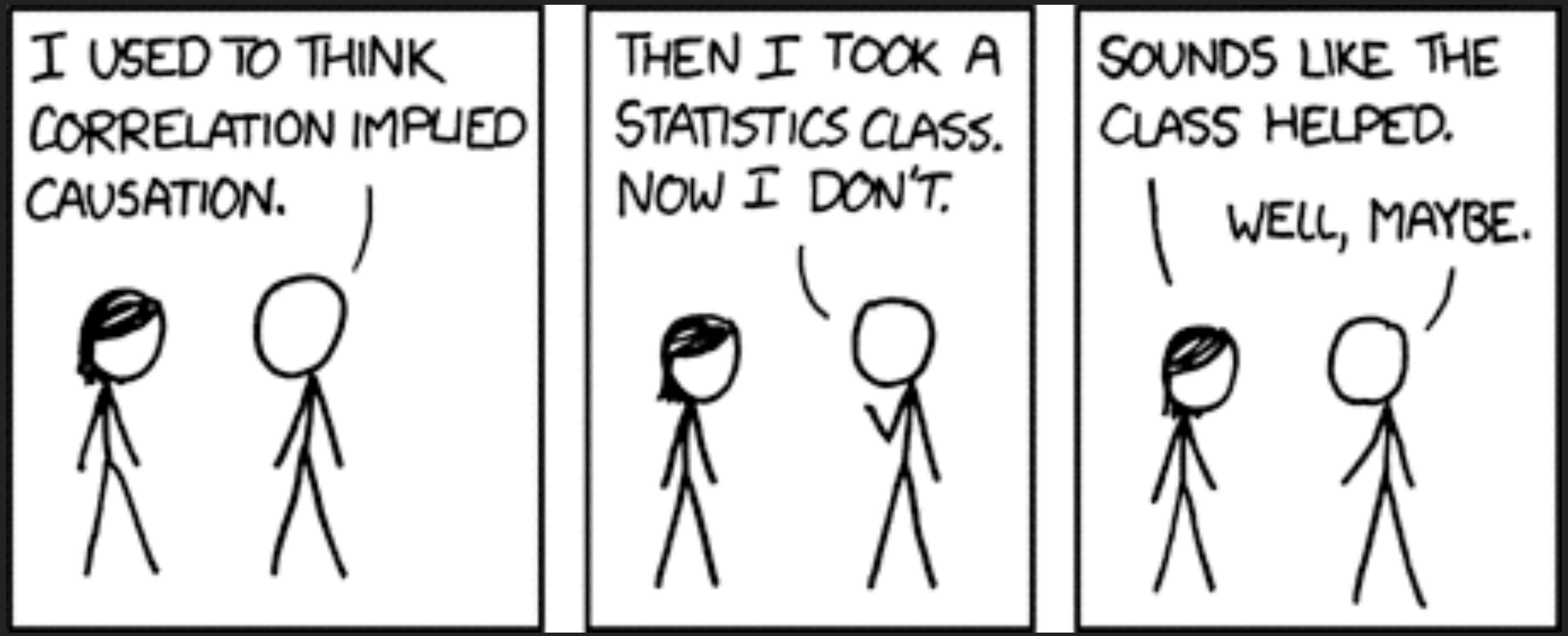


Aside: Correlation is not enough!

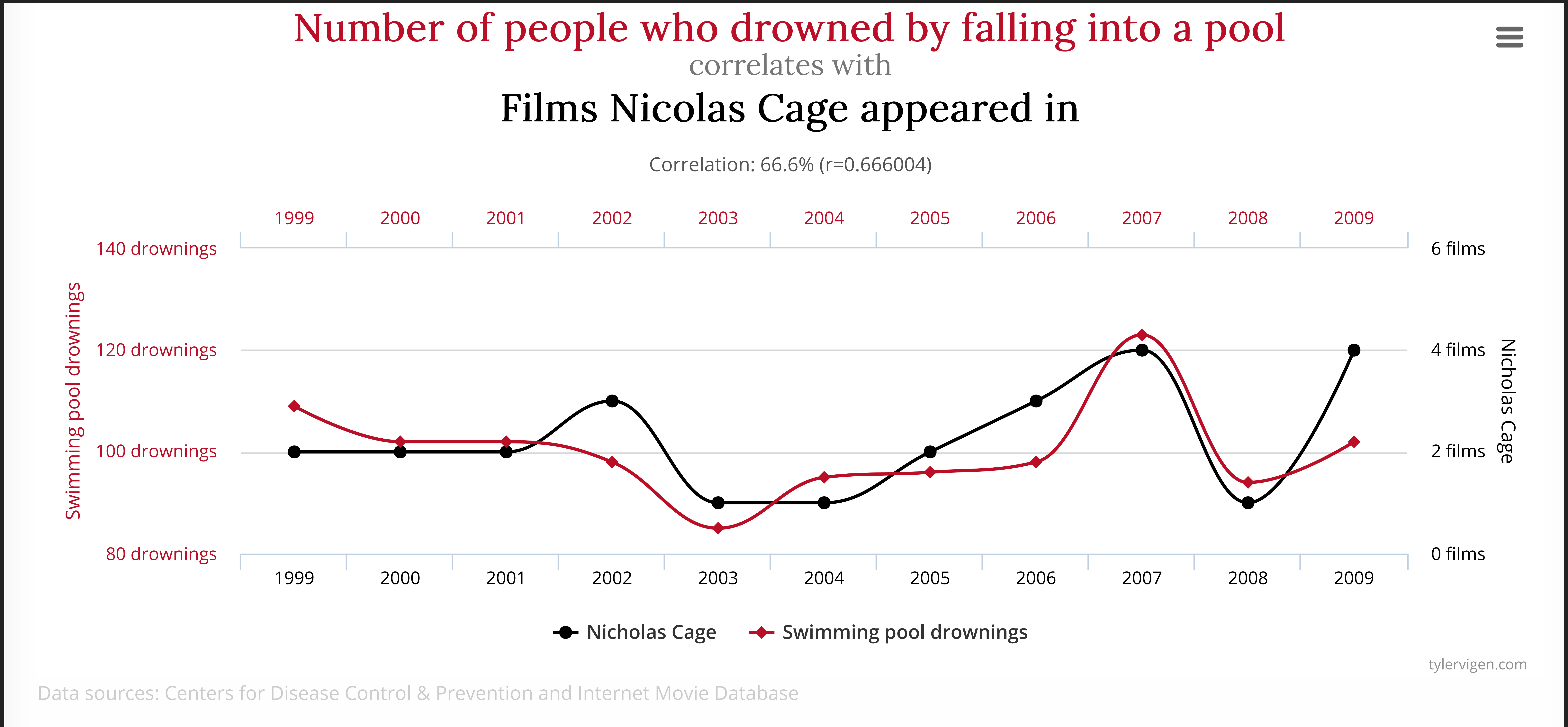
Correlation Does Not Prove Causation

- Which variable came first?
- Are there alternative explanations for the presumed effect?

- Example: income ~ education or education ~ income?
  - Confounding variables: intelligence, family socioeconomic status (causes both high education and high income), …


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

http://www.tylervigen.com/spurious-correlations



Number of people who drowned by falling into a pool

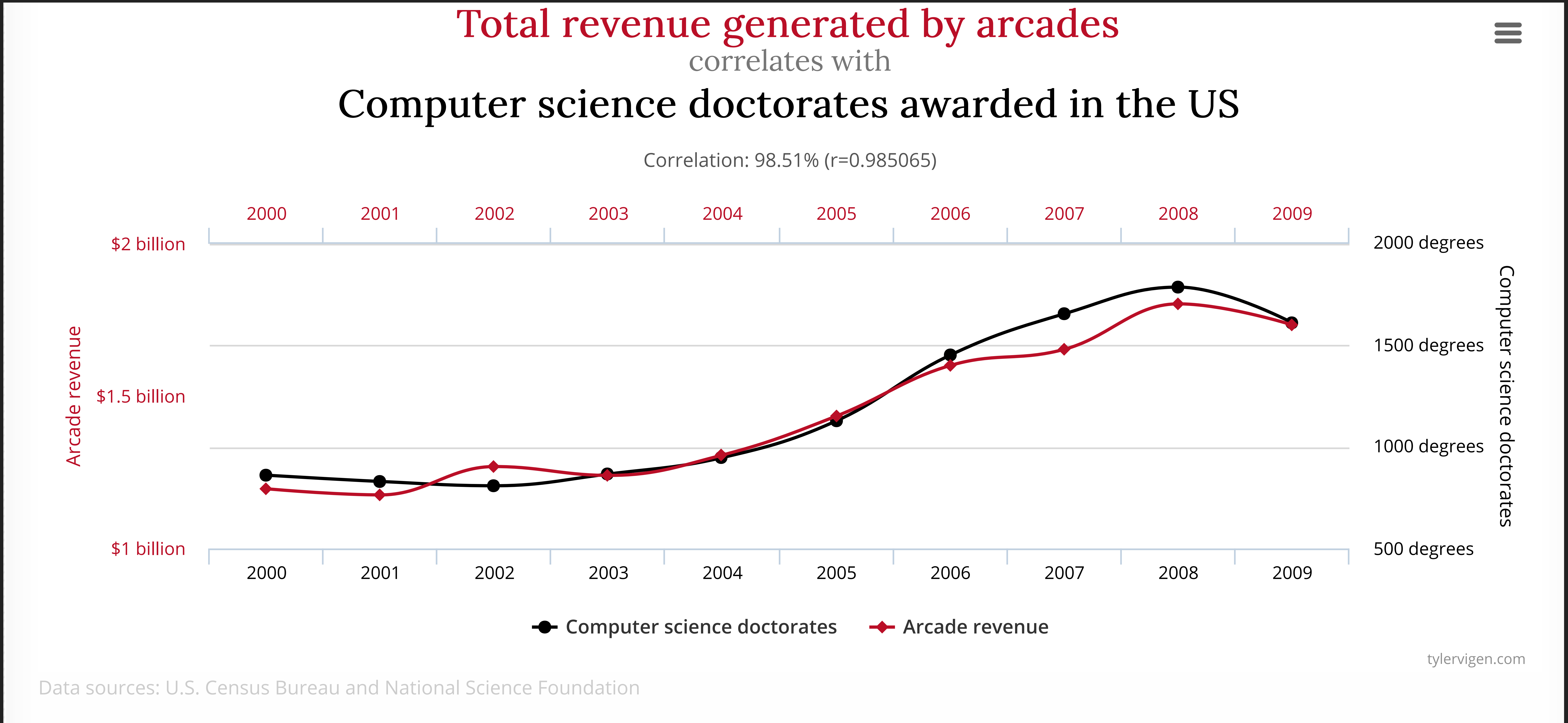
correlates with

Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)

Data sources: Centers for Disease Control & Prevention and Internet Movie Database

http://www.tylervigen.com/spurious-correlations


Total revenue generated by arcades

correlates with

Computer science doctorates awarded in the US

Correlation: 98.51% (r=0.985065)

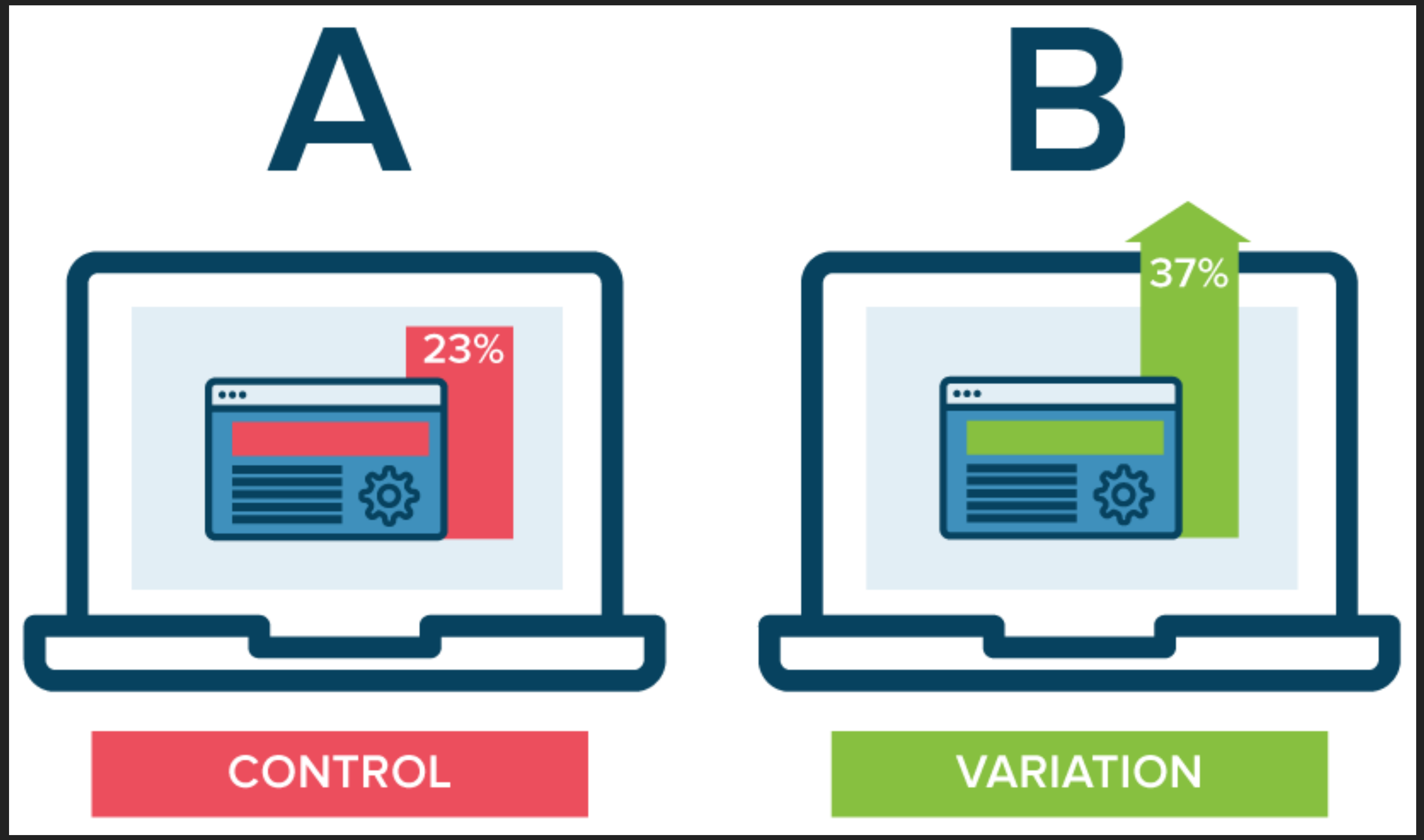
Data sources: U.S. Census Bureau and National Science Foundation

Experiments: Summary Pros and Cons

Advantages and Disadvantages of Experiments

- Disadvantages of experiments:
  - Conditions may be unrealistic
  - Tell nothing about how and why effects occurred
  - Cannot deal with cases when we first observe effect and need to look for causes

"O.K., let's slowly lower in the grant money."



Advantages and Disadvantages of Experiments

- Disadvantages of experiments:
  - Conditions may be unrealistic
  - Tell nothing about how and why effects occurred
  - Cannot deal with cases when we first observe effect and need to look for causes

- Unique advantage:
  - Causal description: describe consequences attributable to deliberately varying a treatment
  - (But not causal explanation / mechanisms)

The vocabulary of experiments

The Vocabulary of Experiments

Experiment

A study in which an intervention is deliberately introduced to observe its effects

Randomized Experiment

An experiment in which units are assigned to receive the treatment or an alternative condition by a random process

Quasi-Experiment

An experiment in which units are not assigned to conditions randomly

Natural Experiment

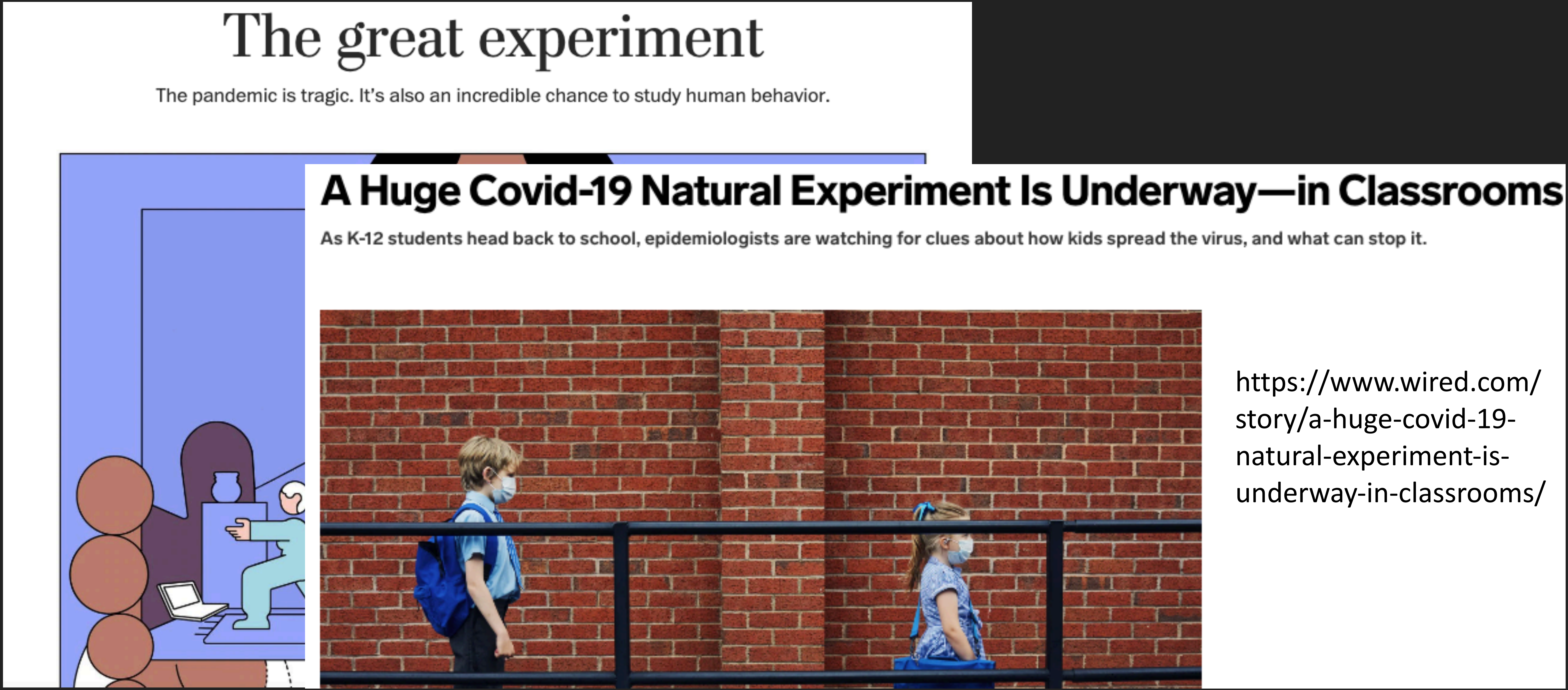
The cause usually can't be manipulated.

A study that contrasts a naturally occurring event such as an earthquake with a comparison condition

Correlational Study

Aka "observational study."

A study that simply observes the size and direction of a relationship among variables



The great experiment

The pandemic is tragic. It’s also an incredible chance to study human behavior.

A Huge Covid-19 Natural Experiment Is Underway—in Classrooms

As K-12 students head back to school, epidemiologists are watching for clues about how kids spread the virus, and what can stop it.

https://www.wired.com/story/a-huge-covid-19-natural-experiment-is-underway-in-classrooms/

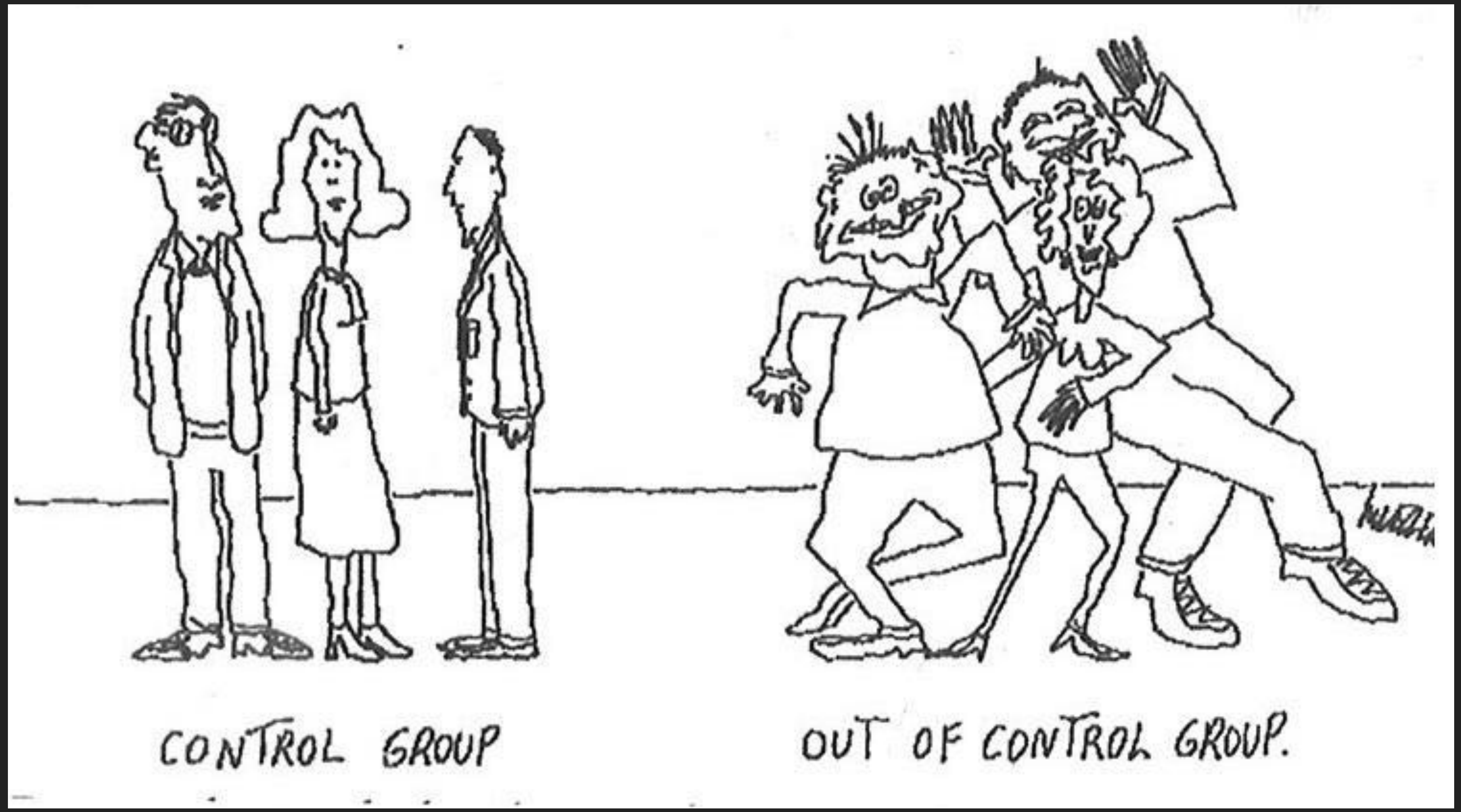
https://www.washingtonpost.com/outlook/2020/09/10/coronavirus-research-experiment-behavior

Randomized Experiment (Sometimes “True Experiment”)

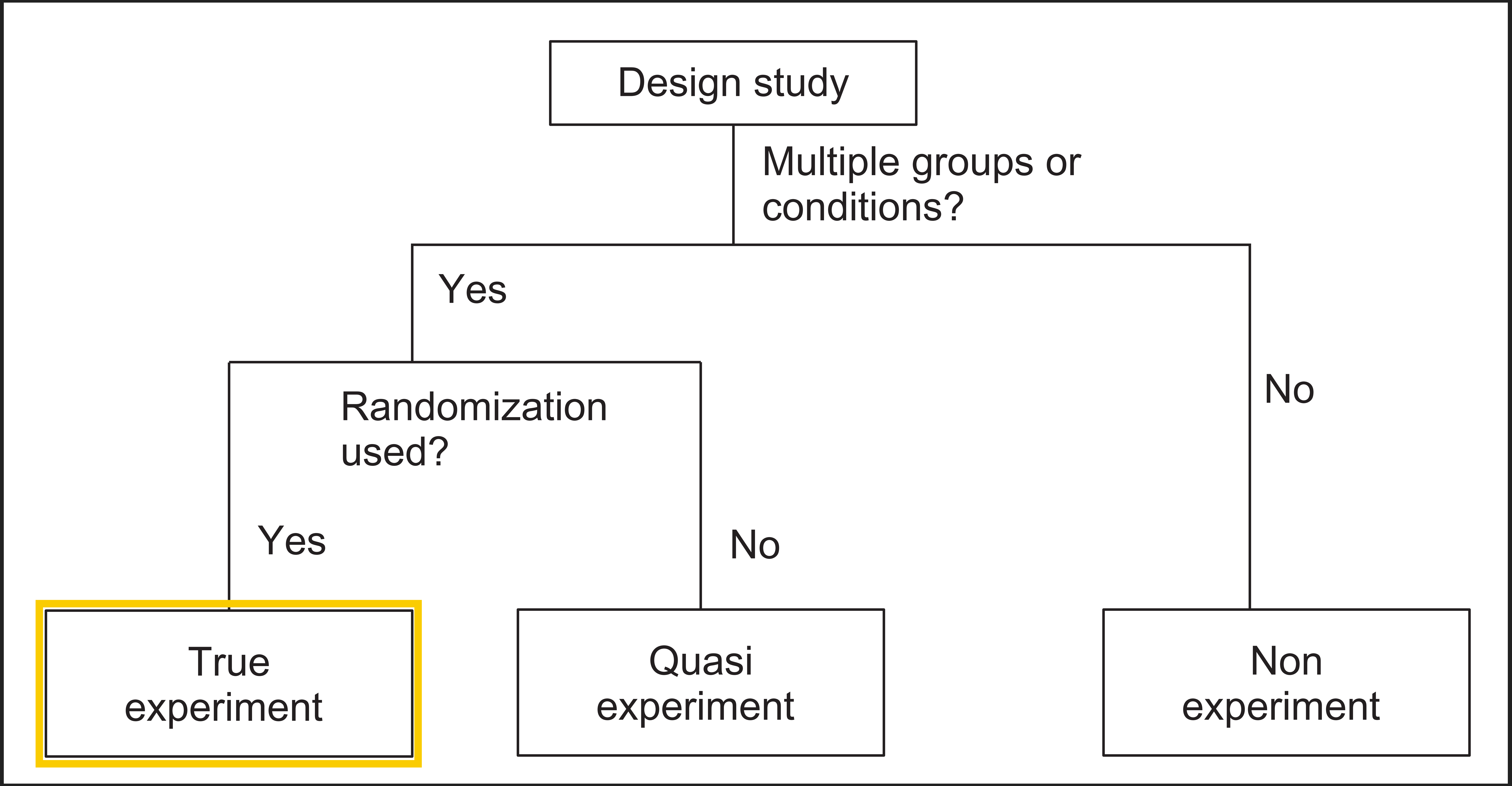
- Various treatments being contrasted (including no treatment at all) are assigned to experimental units by chance.

- Resulting 2+ groups of units are probabilistically similar to each other on the average.

- Outcome differences are likely due to treatment.



Are You Really Doing an "Experiment"?

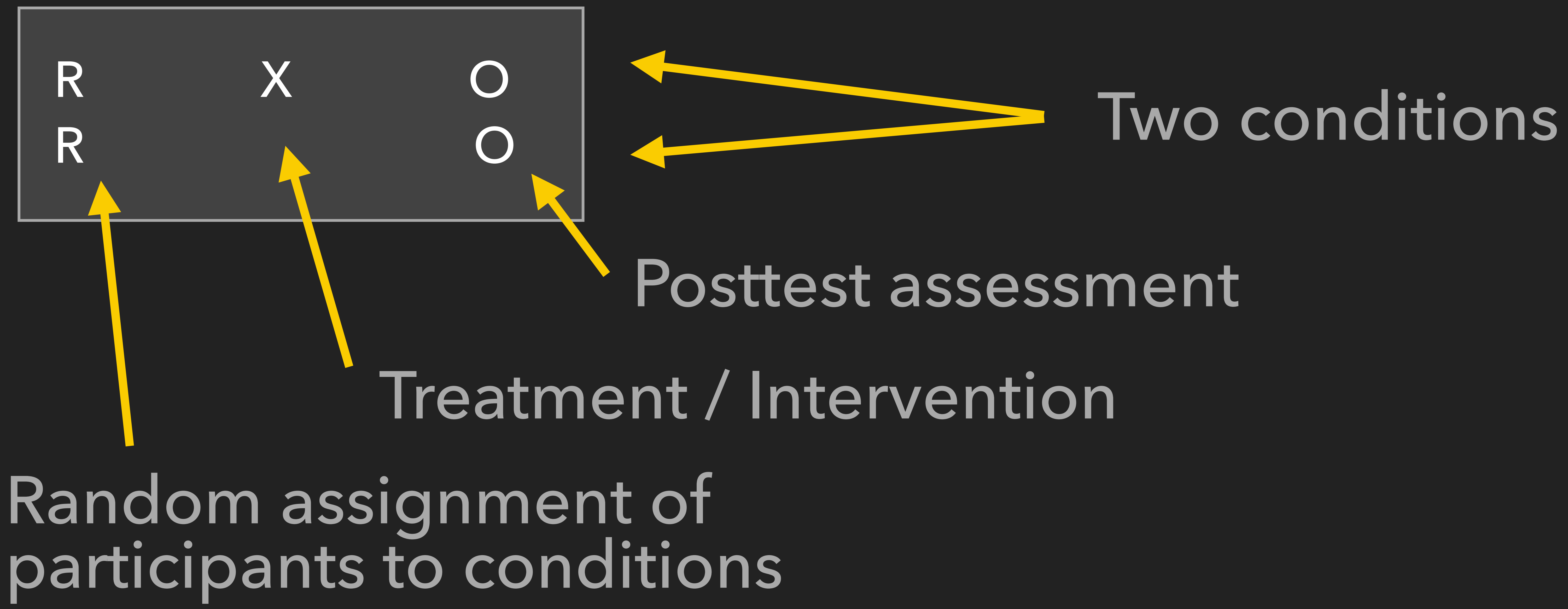


Some designs used with random assignment

Basic X vs C

R	X	O
R		O

Basic X vs C



R X O
R O

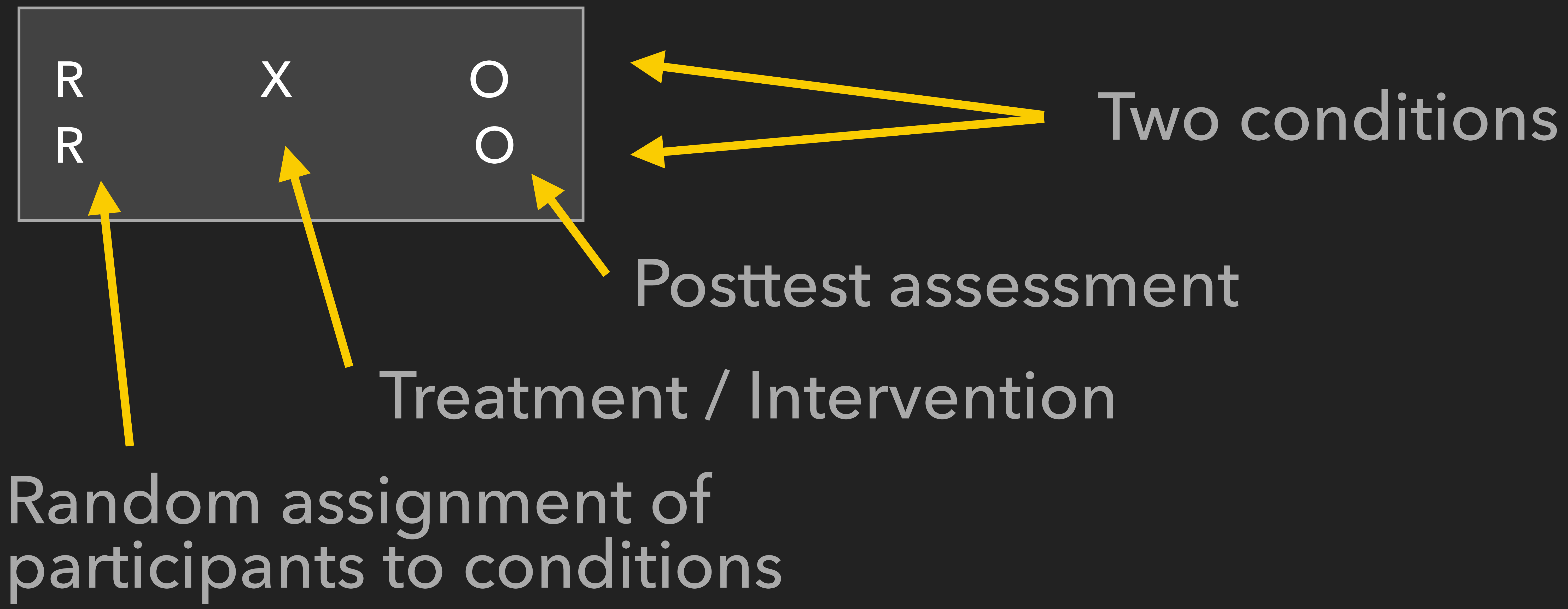
Two conditions

Posttest assessment

Treatment / Intervention

Random assignment of participants to conditions

Basic X vs C



R	X	O
R		O

Two conditions

Posttest assessment

Treatment / Intervention

Random assignment of participants to conditions

- **Limitation:**

Can't separate active ingredients in treatment from the experience of being treated

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

- Innovative treatment vs gold standard
- Limitation:
  - If no effect, can't distinguish if both treatments were equally effective or equally ineffective

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

- Innovative treatment vs gold standard vs control

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

- Common **limitation**: Lack of pretest
  - Especially if attrition
  - But not always undesirable
    - E.g., unwanted sensitization effect from pretest, physically impossible to collect, constant (all alive)

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

Pretest-posttest

R	O	X	O
R	O		O

Alternative Xs with pretest

R	O	X_A	O
R	O	X_B	O

- Some extra statistical analysis advantages, besides robustness to attrition.

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

Pretest-posttest

R	O	X	O
R	O		O

Alternative Xs with pretest

R	O	X_A	O
R	O	X_B	O

Factorial

R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

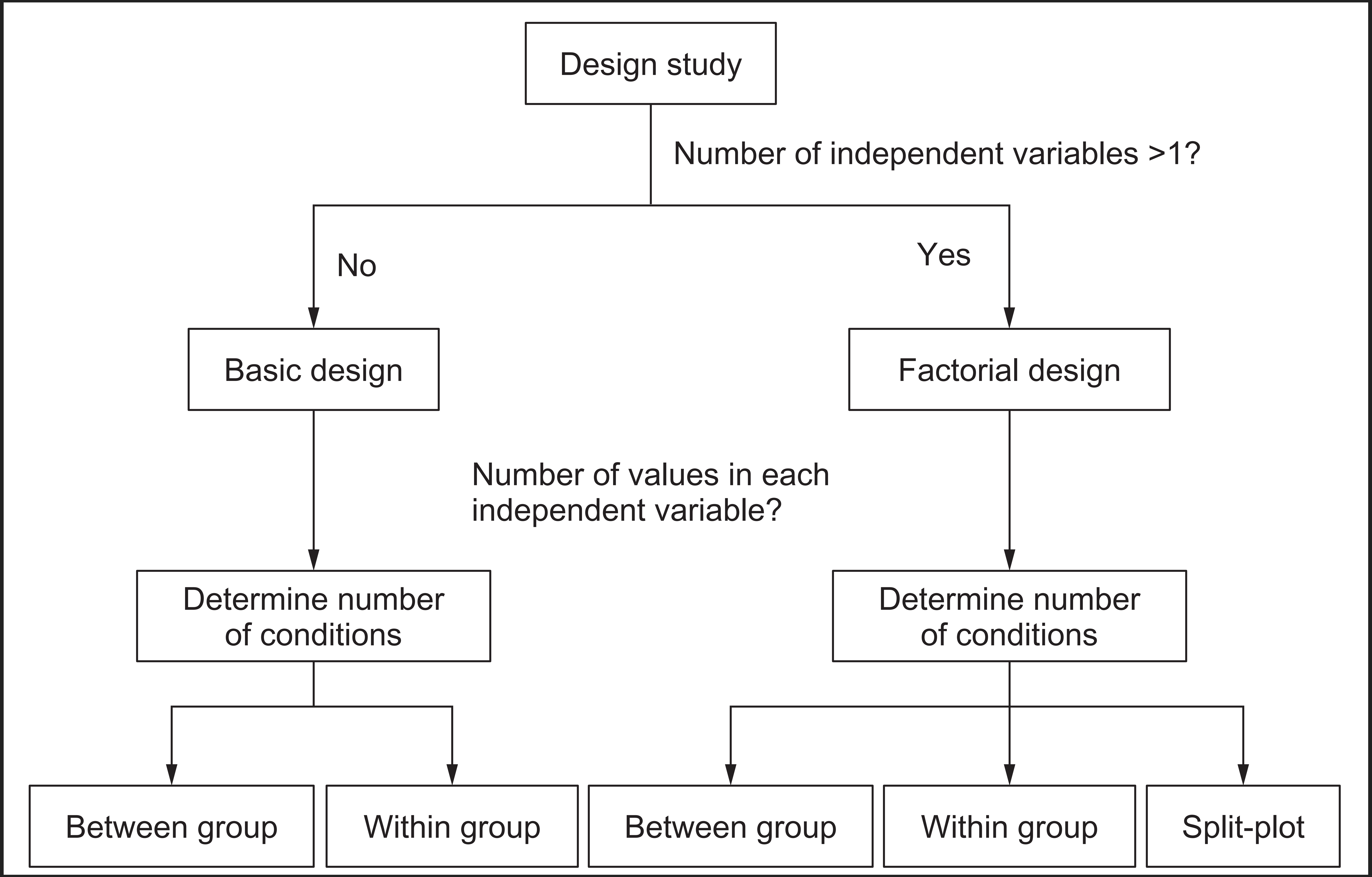
Longitudinal

R	O … O	X	O … O
R	O … O		O … O

Crossover

R	O	X_A	O	X_B	O
R	O	X_B	O	X_A	O

Another way to think about designs


Design study

Number of independent variables >1?

No

Yes

Basic design

Factorial design

Number of values in each independent variable?

Determine number of conditions

Determine number of conditions

Between group

Within group

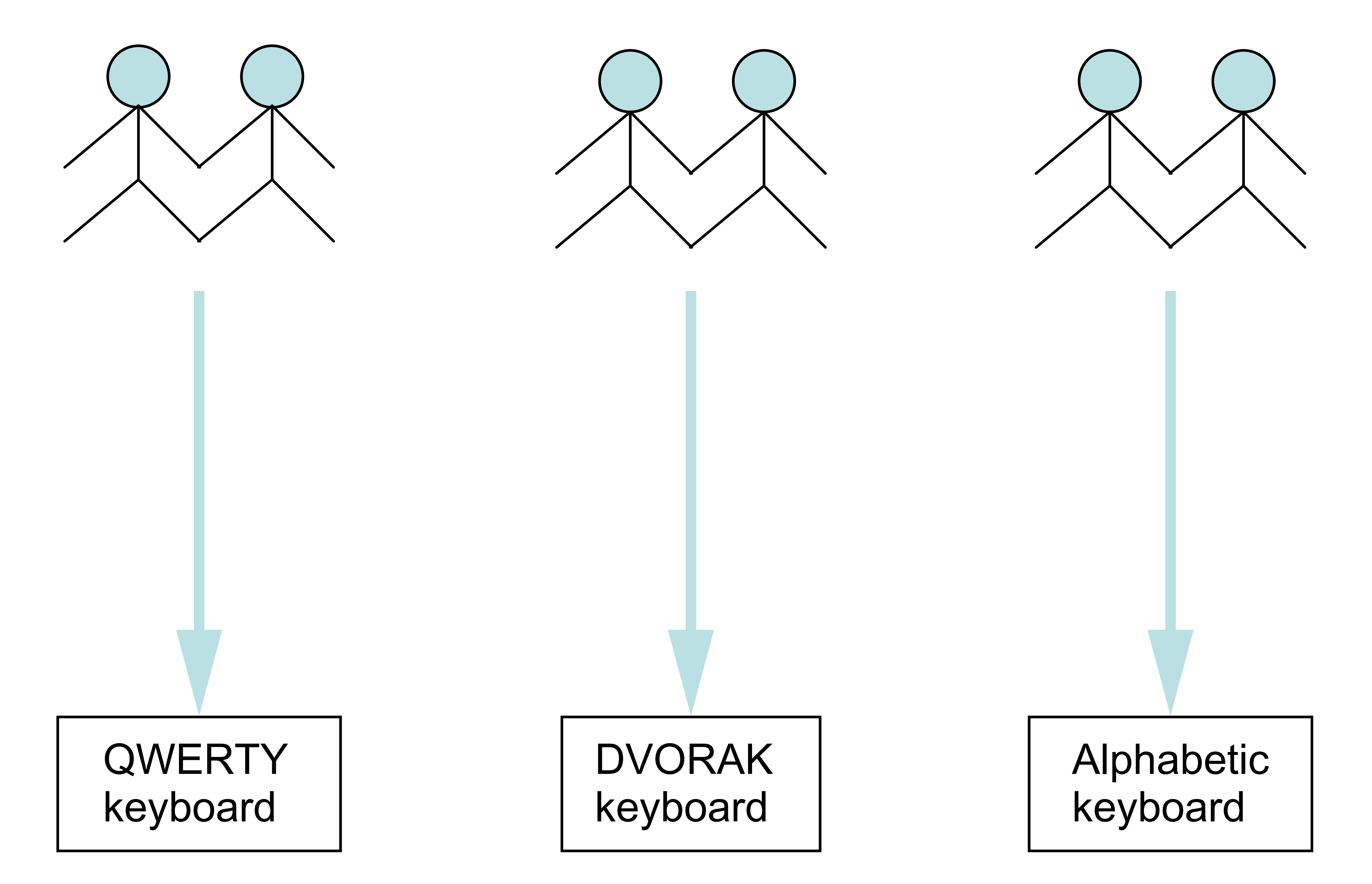
Between group

Within group

Split-plot

Between-Group Design

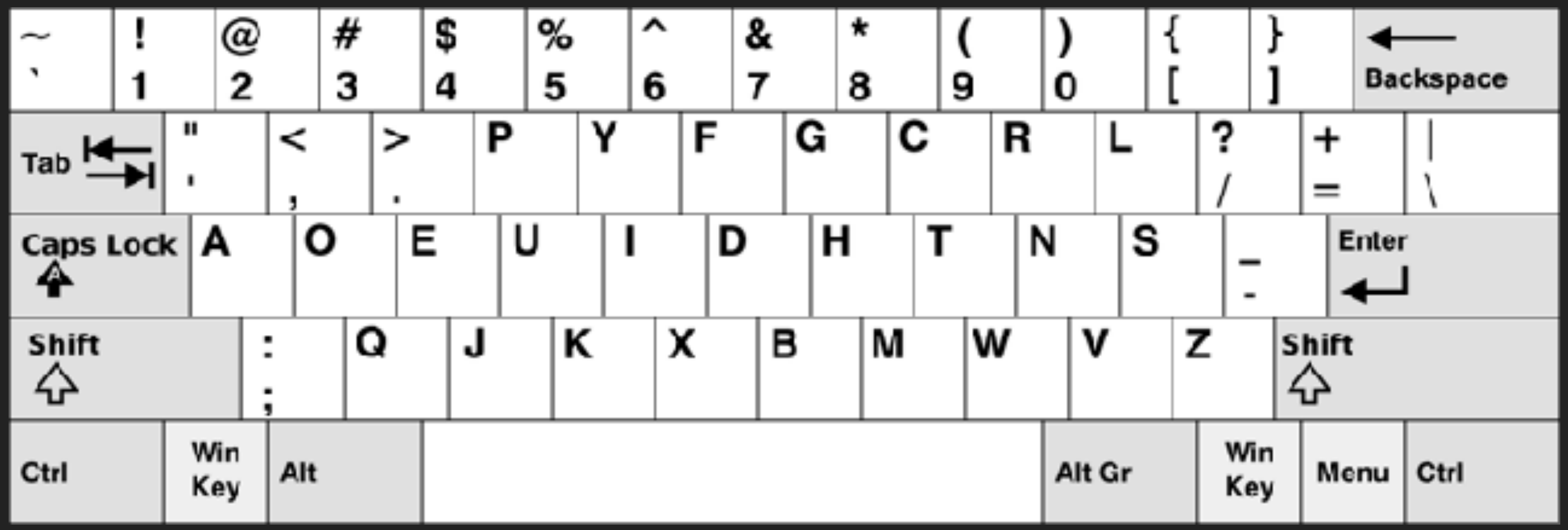
- Aka “between-subject design.”
- Each participant is only exposed to one experimental condition.
- E.g., if the task is to type a 500-word doc, each participant types one doc using one of the keyboards.



QWERTY keyboard

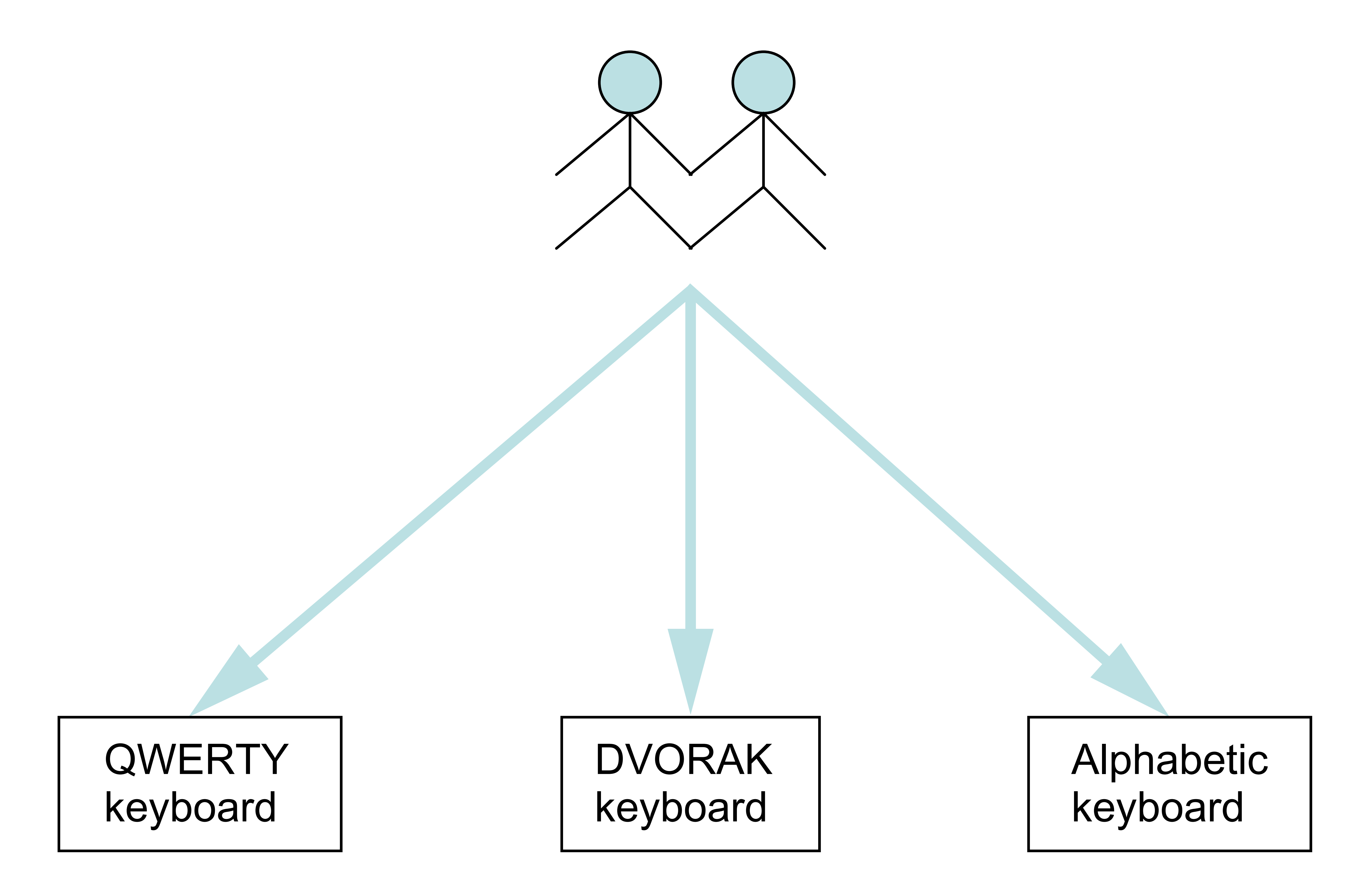
DVORAK keyboard

Alphabetic keyboard



Within-Group Design

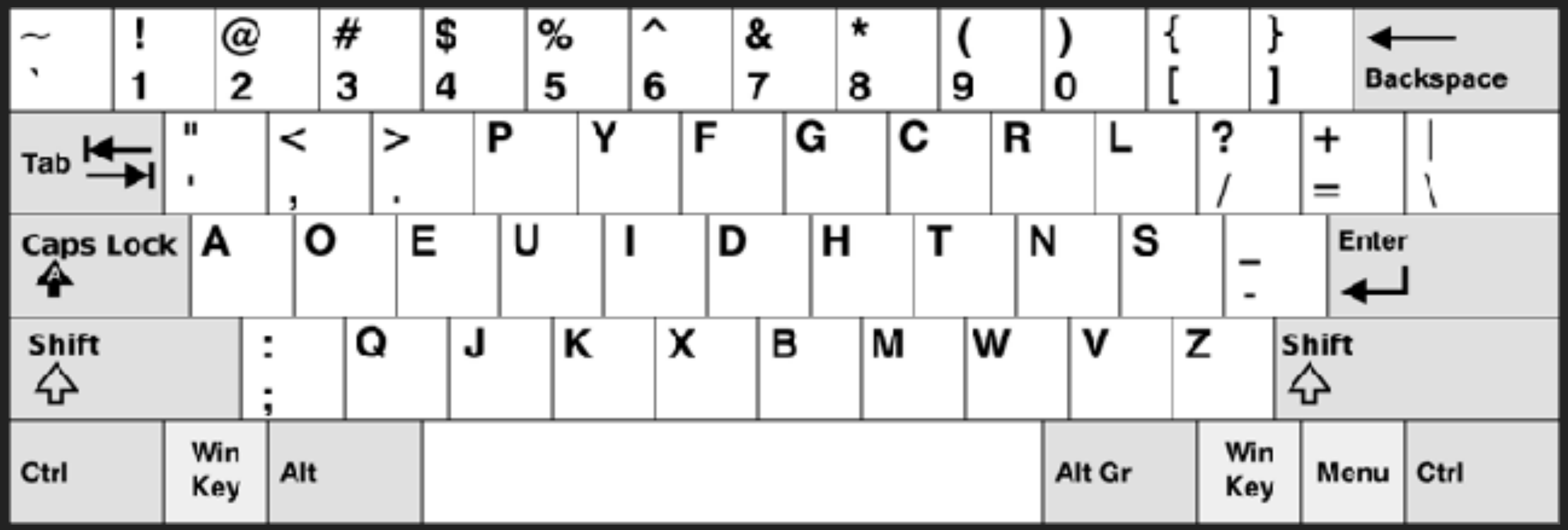
- Aka “within-subject design.”
- Each participant is exposed to multiple experimental conditions.
- E.g., each participant types three docs, using each of the three keyboards for one doc.



QWERTY keyboard

DVORAK keyboard

Alphabetic keyboard


~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 { [}] Backspace
Tab " ' < , > . P Y F G C R L ? / + = | \
Caps Lock A O E U I D H T N S _ - Enter
Shift : ; Q J K X B M W V Z Shift
Ctrl Win Key Alt Alt Gr Win Key Menu Ctrl

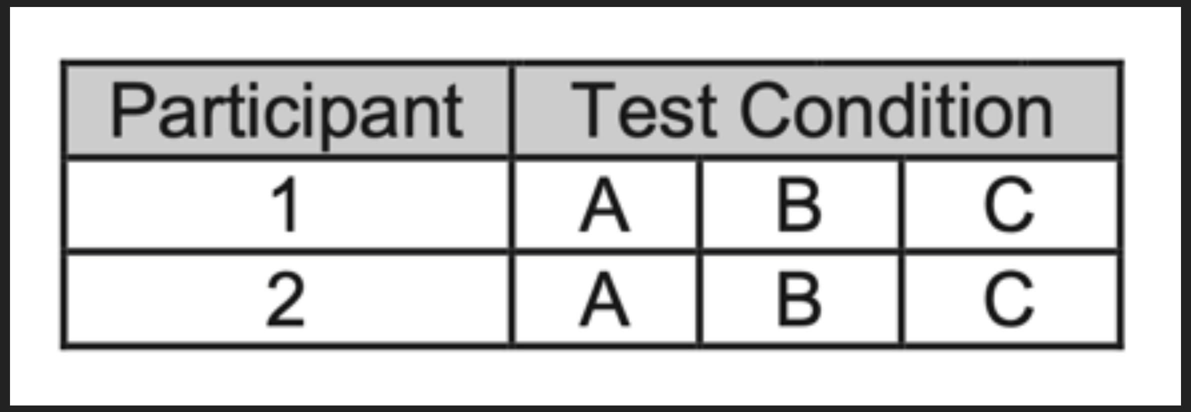
Between-Subjects vs Within-Subjects?

Between-Subjects vs Within-Subjects Considerations

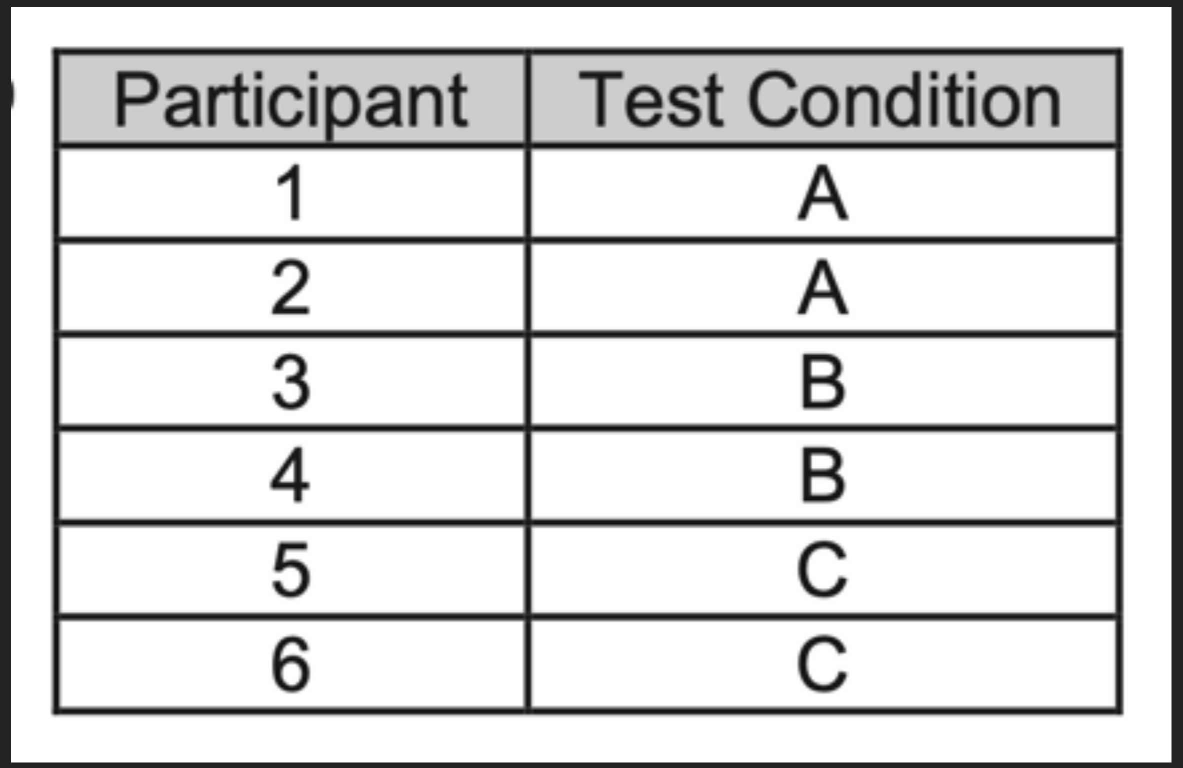
- Order effects
    - Learning - favors conditions completed toward the end of the experiment
    - Fatigue - negatively impacts on the performance of conditions completed toward the end of the experiment
- Win: Between-subjects
    - No learning effects.
        - Any participant is only exposed to one condition
    - Takes less time to complete.
        - Confounding factors such as fatigue and frustration can be more effectively controlled.

Between-Subjects vs Within-Subjects Considerations

- Impacts from individual differences can obscure effect
- Win: Within-subjects
  - Requires a much smaller sample size
    - We are comparing the performances of the same participants under different conditions.
    - Therefore, the impact of individual differences is effectively isolated.
- But, sometimes it's totally impossible
  - e.g., "There is no difference in the time required to implement a web server in Python between novice developers and experienced developers."



Participant	Test Condition		
1	A	B	C
2	A	B	C



Participant	Test Condition
1	A
2	A
3	B
4	B
5	C
6	C

Comparison of Between-Group and Within-Group Designs

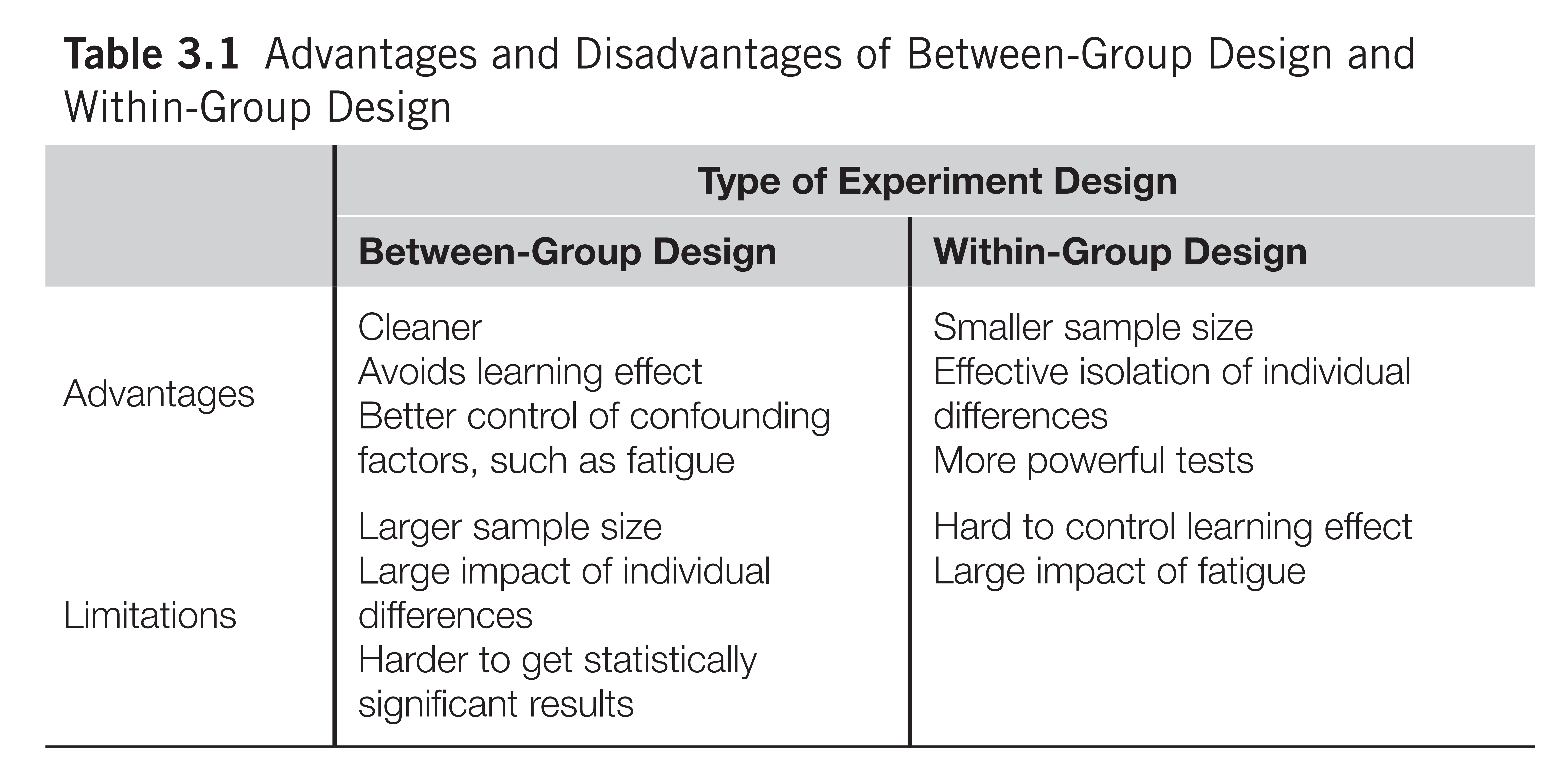


Table 3.1 Advantages and Disadvantages of Between-Group Design and Within-Group Design

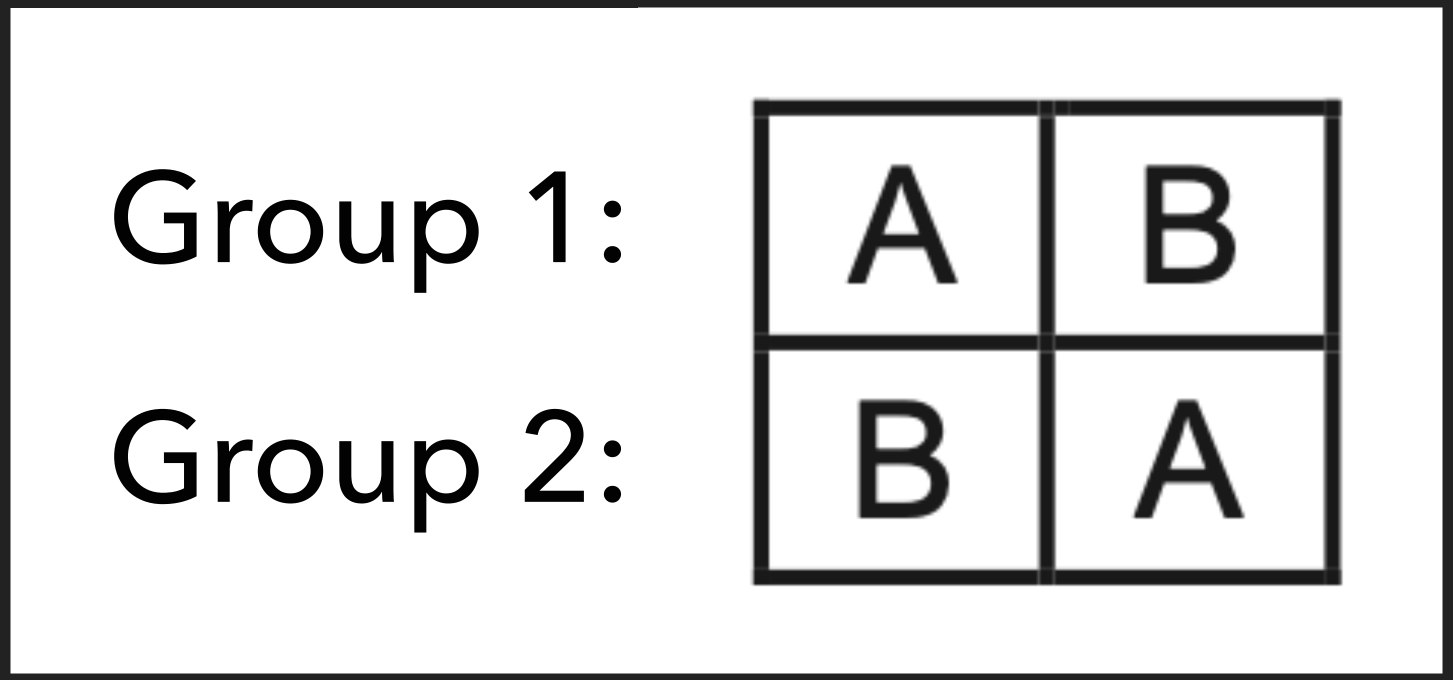
	Type of Experiment Design	
	Between-Group Design	**Within-Group Design**
Advantages	Cleaner Avoids learning effect Better control of confounding factors, such as fatigue	Smaller sample size Effective isolation of individual differences More powerful tests
Limitations	Larger sample size Large impact of individual differences Harder to get statistically significant results	Hard to control learning effect Large impact of fatigue

Order effects, counterbalancing, and latin squares

The most common method of compensating for an order effect is to divide participants into groups and administer the conditions in a different order for each group. The compensatory ordering of test conditions to offset practice effects is called counterbalancing.

Example

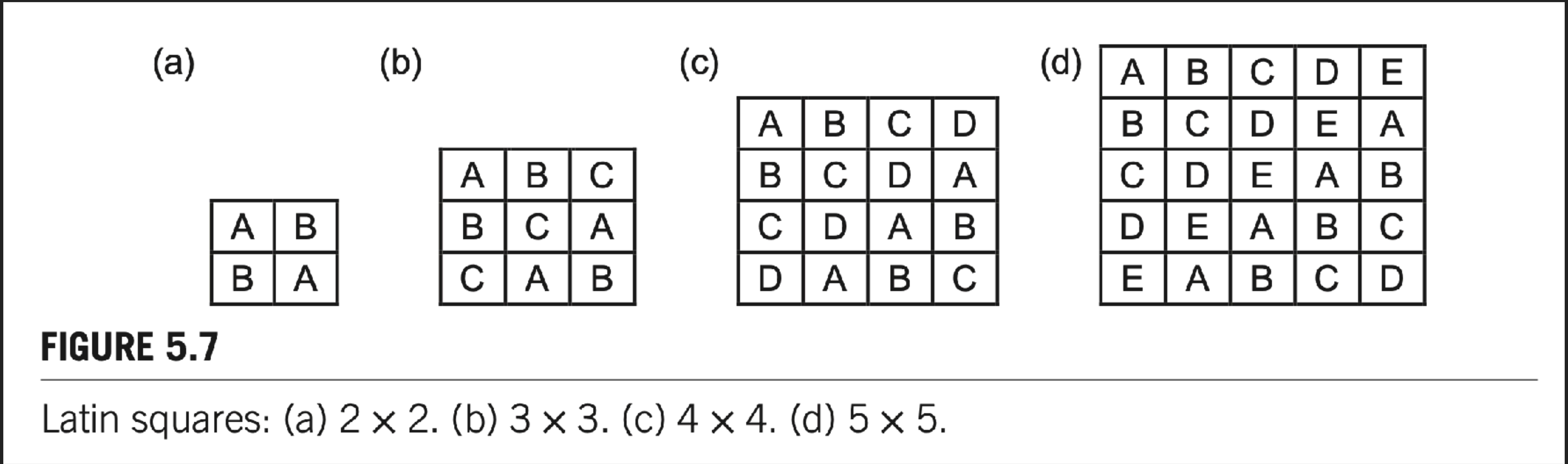
- In the simplest case of a factor with two levels, say, A and B, participants are divided into two groups.
- If there are 12 participants overall, then Group 1 has 6 participants and Group 2 has 6 participants.
- Group 1 is tested first on condition A, then on condition B. Group 2 is given the test conditions in the reverse order.



Group 1:	A	B
Group 2:	B	A

2 x 2 Latin square

Latin Squares: (a) 2 × 2. (b) 3 × 3. (c) 4 × 4. (d) 5 × 5



(a)

A	B
B	A

(b)

A	B	C
B	C	A
C	A	B

(c)

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

(d)

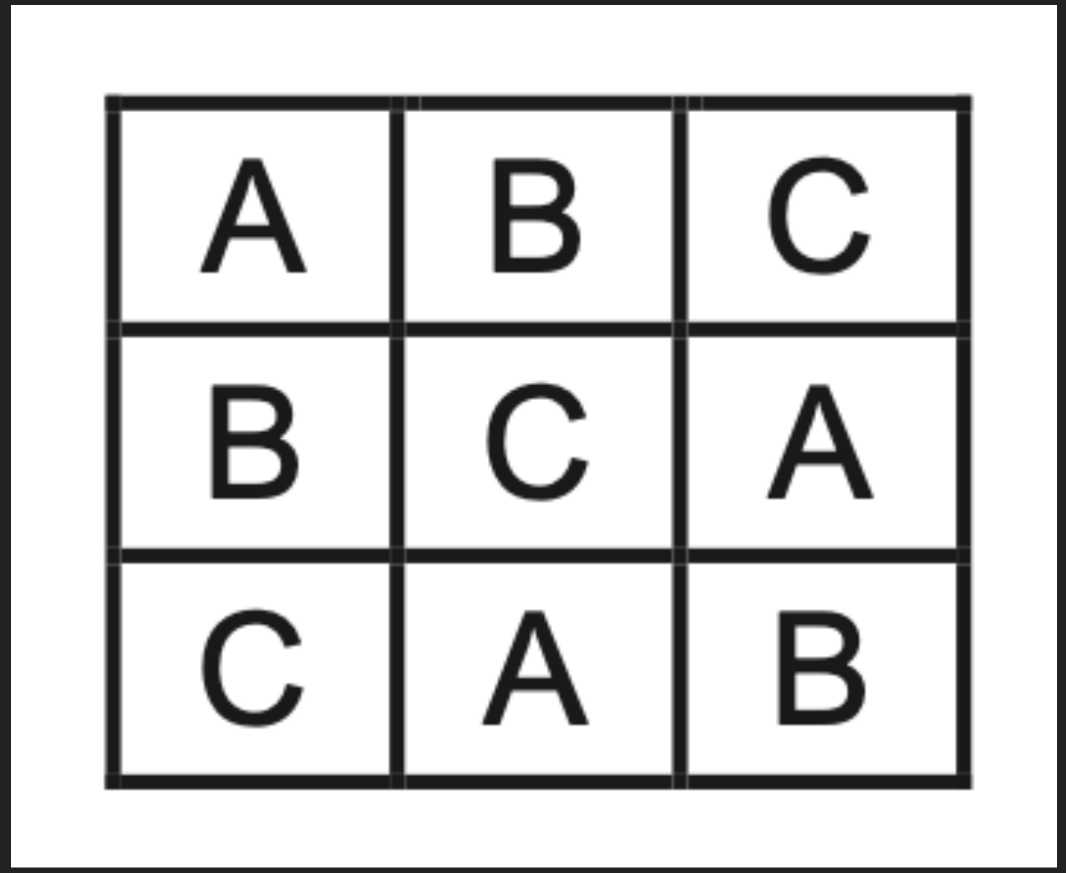
A	B	C	D	E
B	C	D	E	A
C	D	E	A	B
D	E	A	B	C
E	A	B	C	D

FIGURE 5.7

Latin squares: (a) 2 × 2. (b) 3 × 3. (c) 4 × 4. (d) 5 × 5.

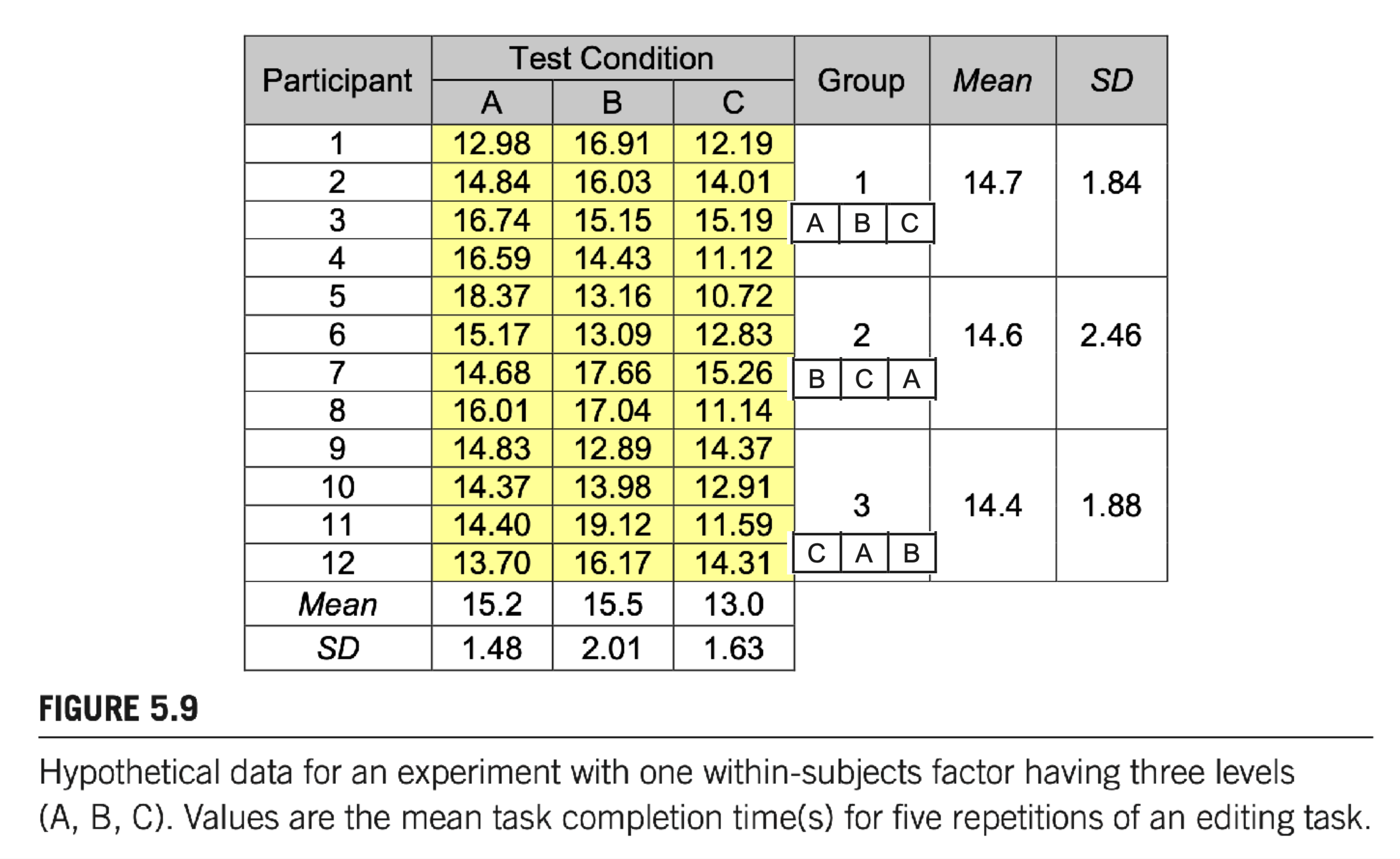
Example

- An experimenter seeks to determine if three editing methods (A, B, C) differ in the time required for common editing tasks.
  - Method A: arrow keys, backspace, type
  - Method B: search and replace dialog
  - Method C: point and double click with the mouse, type
- Twelve participants are recruited. To counterbalance for learning effects, participants are divided into three groups with the tasks administered according to a Latin square.
- Each participant does the task five times with one editing method, then again with the second editing method, then again with the third.



A	B	C
B	C	A
C	A	B

Example (continued)



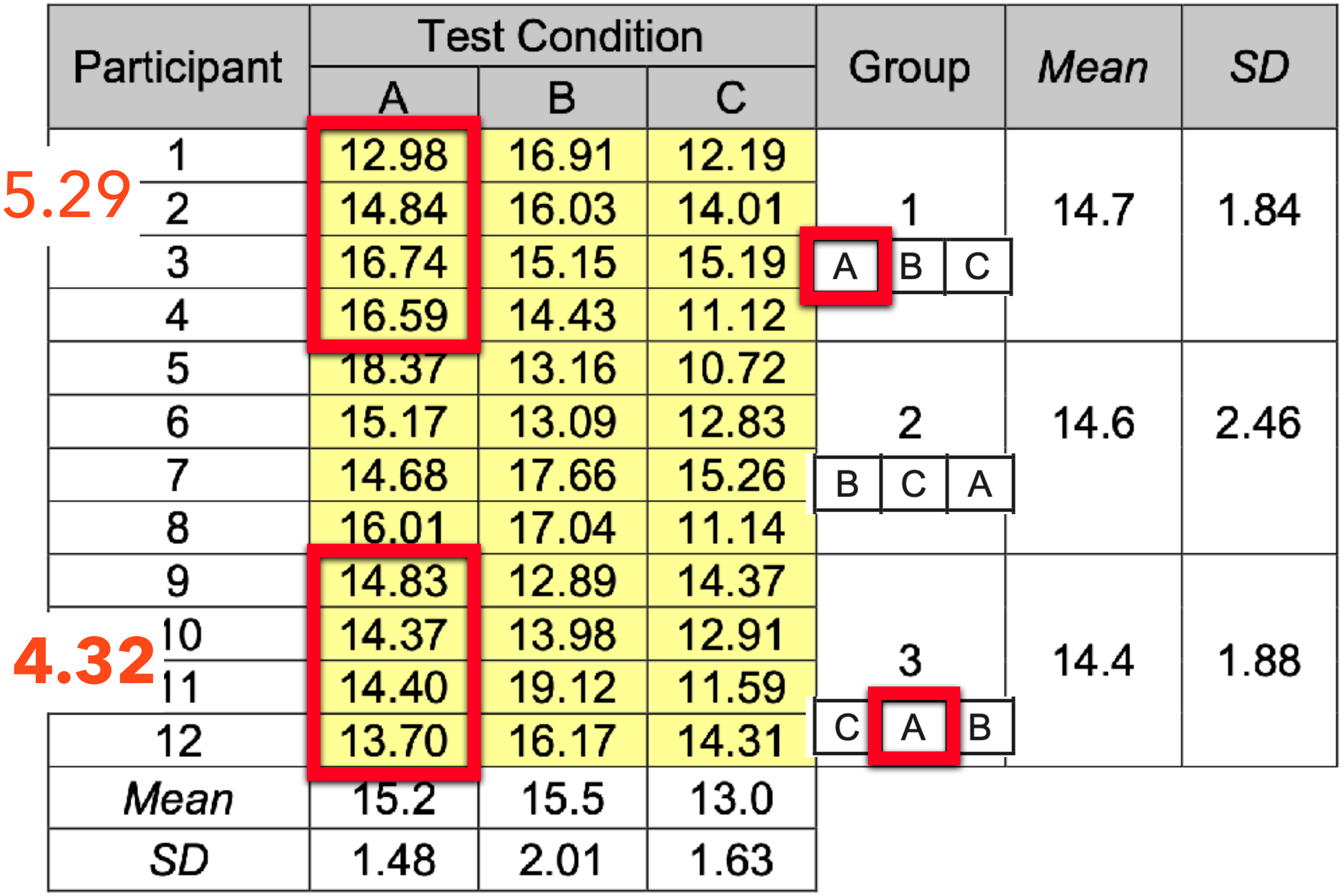
Participant	Test Condition			Group	*Mean*	*SD*
	A	B	C			
1	12.98	16.91	12.19	1 A B C	14.7	1.84
2	14.84	16.03	14.01			
3	16.74	15.15	15.19			
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2 B C A	14.6	2.46
6	15.17	13.09	12.83			
7	14.68	17.66	15.26			
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3 C A B	14.4	1.88
10	14.37	13.98	12.91			
11	14.40	19.12	11.59			
12	13.70	16.17	14.31			
Mean	15.2	15.5	13.0			
SD	1.48	2.01	1.63			

FIGURE 5.9

Hypothetical data for an experiment with one within-subjects factor having three levels (A, B, C). Values are the mean task completion time(s) for five repetitions of an editing task.

Example (continued)

Learning?



Mean = 15.29

Mean = 14.32

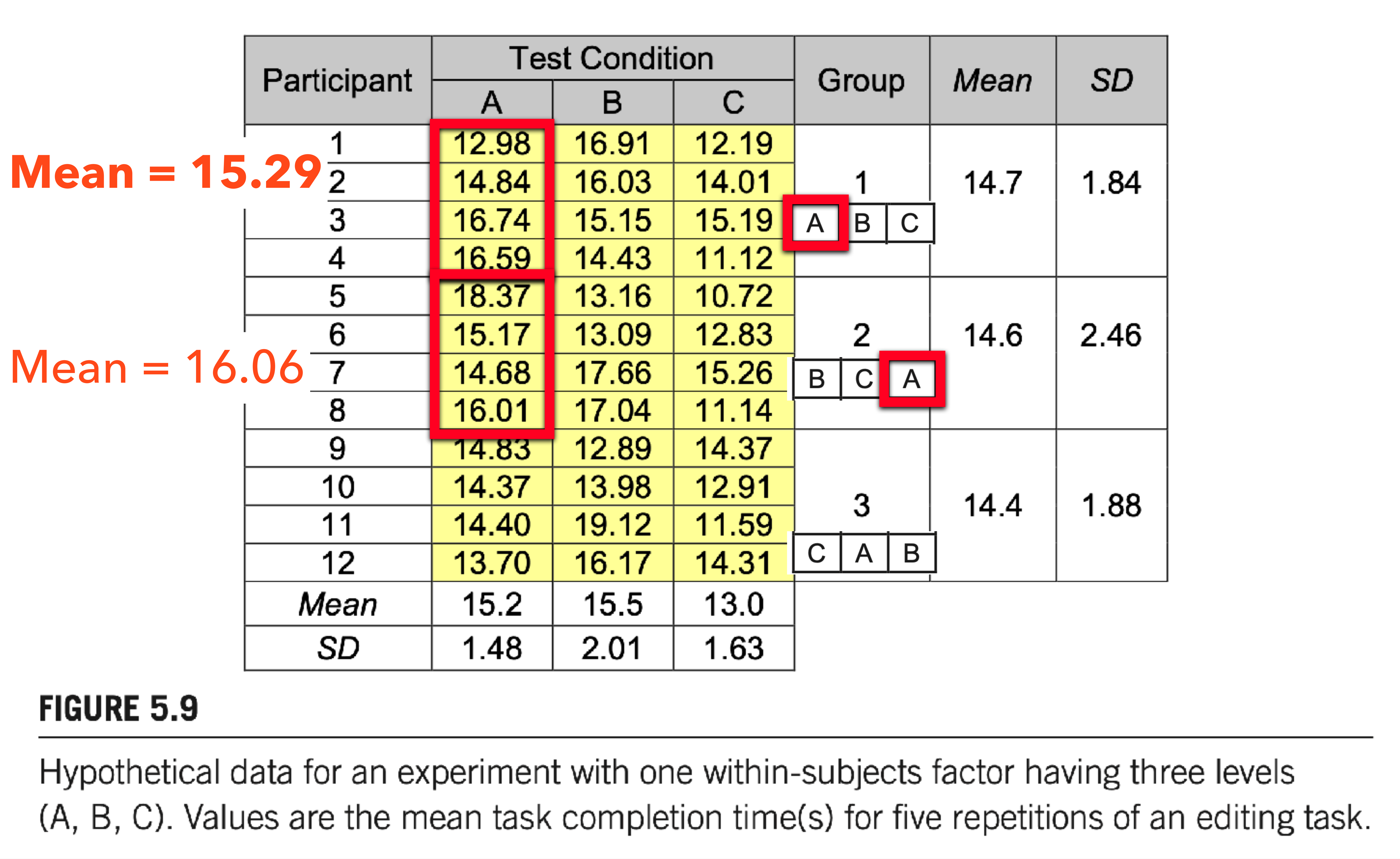
Participant	Test Condition			Group	*Mean*	*SD*
	A	B	C			
1	12.98	16.91	12.19	1 A B C	14.7	1.84
2	14.84	16.03	14.01			
3	16.74	15.15	15.19			
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2 B C A	14.6	2.46
6	15.17	13.09	12.83			
7	14.68	17.66	15.26			
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3 C A B	14.4	1.88
10	14.37	13.98	12.91			
11	14.40	19.12	11.59			
12	13.70	16.17	14.31			
Mean	15.2	15.5	13.0			
SD	1.48	2.01	1.63			

FIGURE 5.9

Hypothetical data for an experiment with one within-subjects factor having three levels (A, B, C). Values are the mean task completion time(s) for five repetitions of an editing task.

Example (continued)

Fatigue?



Mean = 15.29

Mean = 16.06

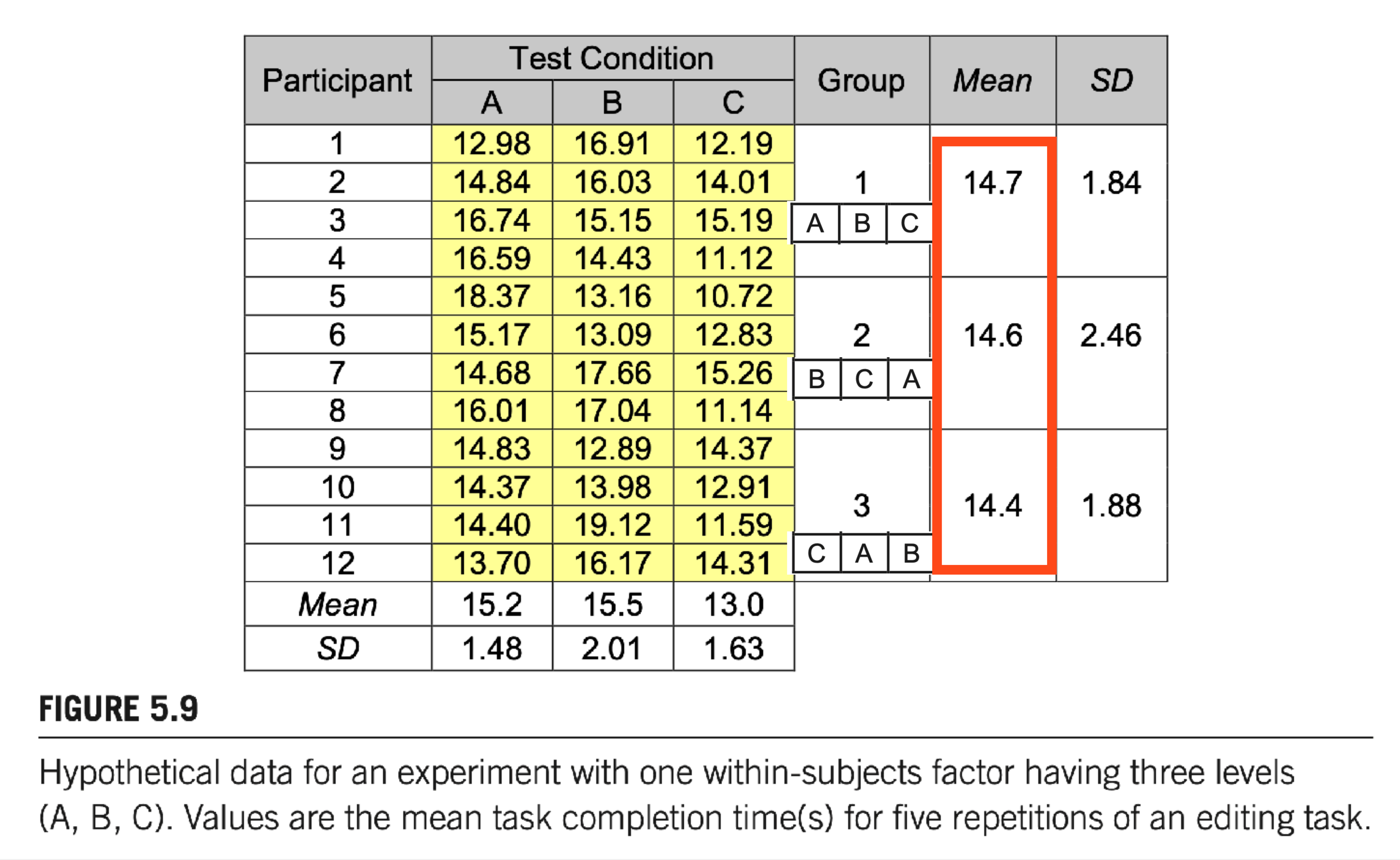
Participant	Test Condition			Group	*Mean*	*SD*
	A	B	C			
1	12.98	16.91	12.19	1 A B C	14.7	1.84
2	14.84	16.03	14.01			
3	16.74	15.15	15.19			
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2 B C A	14.6	2.46
6	15.17	13.09	12.83			
7	14.68	17.66	15.26			
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3 C A B	14.4	1.88
10	14.37	13.98	12.91			
11	14.40	19.12	11.59			
12	13.70	16.17	14.31			
Mean	15.2	15.5	13.0			
SD	1.48	2.01	1.63			

FIGURE 5.9

Hypothetical data for an experiment with one within-subjects factor having three levels (A, B, C). Values are the mean task completion time(s) for five repetitions of an editing task.

Example (continued)

Counterbalancing worked!



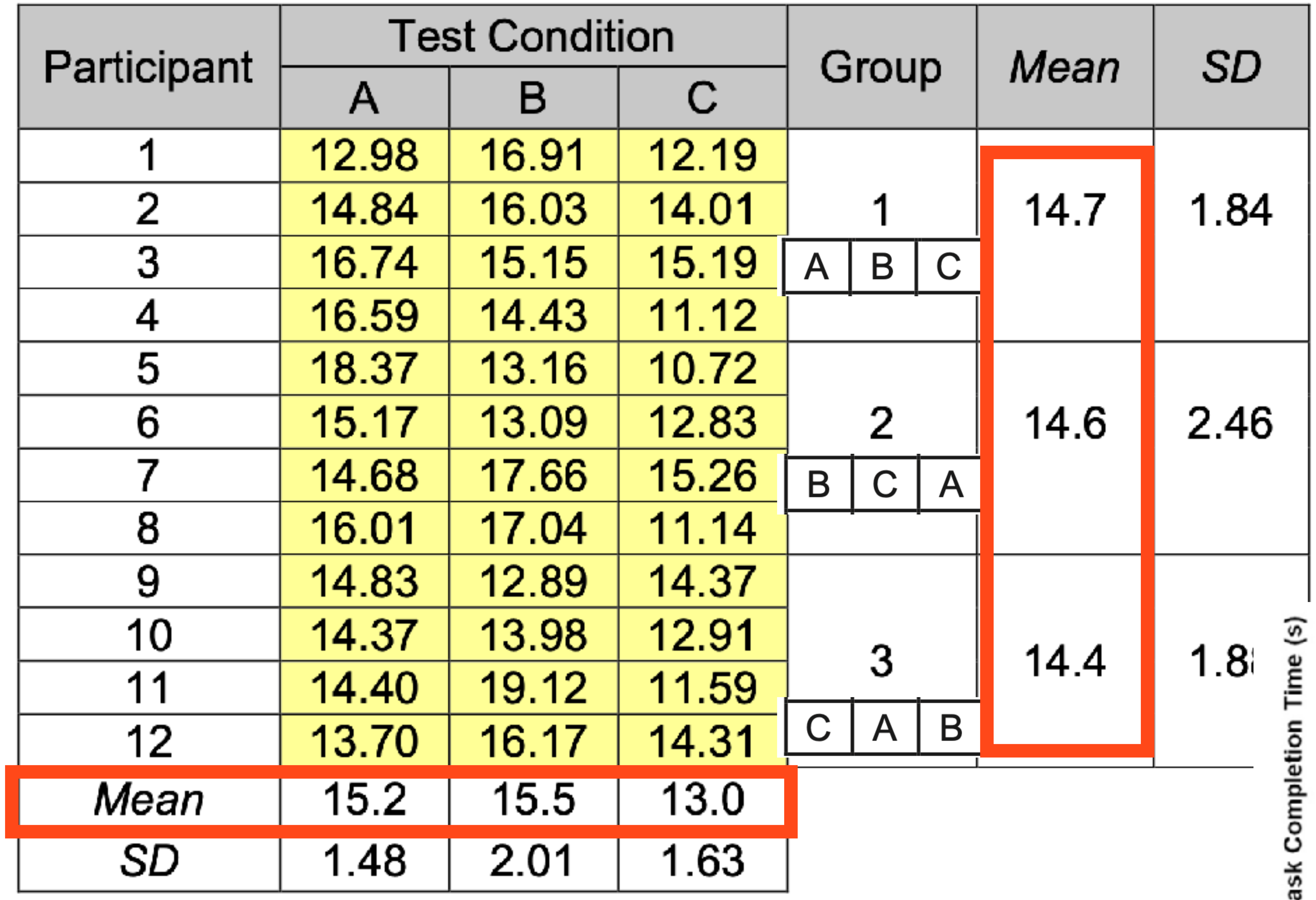
Participant	Test Condition			Group	*Mean*	*SD*
	A	B	C			
1	12.98	16.91	12.19	1 A | B | C	14.7	1.84
2	14.84	16.03	14.01			
3	16.74	15.15	15.19			
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2 B | C | A	14.6	2.46
6	15.17	13.09	12.83			
7	14.68	17.66	15.26			
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3 C | A | B	14.4	1.88
10	14.37	13.98	12.91			
11	14.40	19.12	11.59			
12	13.70	16.17	14.31			
Mean	15.2	15.5	13.0			
SD	1.48	2.01	1.63			

FIGURE 5.9

Hypothetical data for an experiment with one within-subjects factor having three levels (A, B, C). Values are the mean task completion time(s) for five repetitions of an editing task.

Example (continued)

Counterbalancing worked!



Participant	Test Condition			Group	*Mean*	*SD*
	A	B	C			
1	12.98	16.91	12.19	1 A | B | C	14.7	1.84
2	14.84	16.03	14.01			
3	16.74	15.15	15.19			
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2 B | C | A	14.6	2.46
6	15.17	13.09	12.83			
7	14.68	17.66	15.26			
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3 C | A | B	14.4	1.8
10	14.37	13.98	12.91			
11	14.40	19.12	11.59			
12	13.70	16.17	14.31			
Mean	15.2	15.5	13.0			
SD	1.48	2.01	1.63			

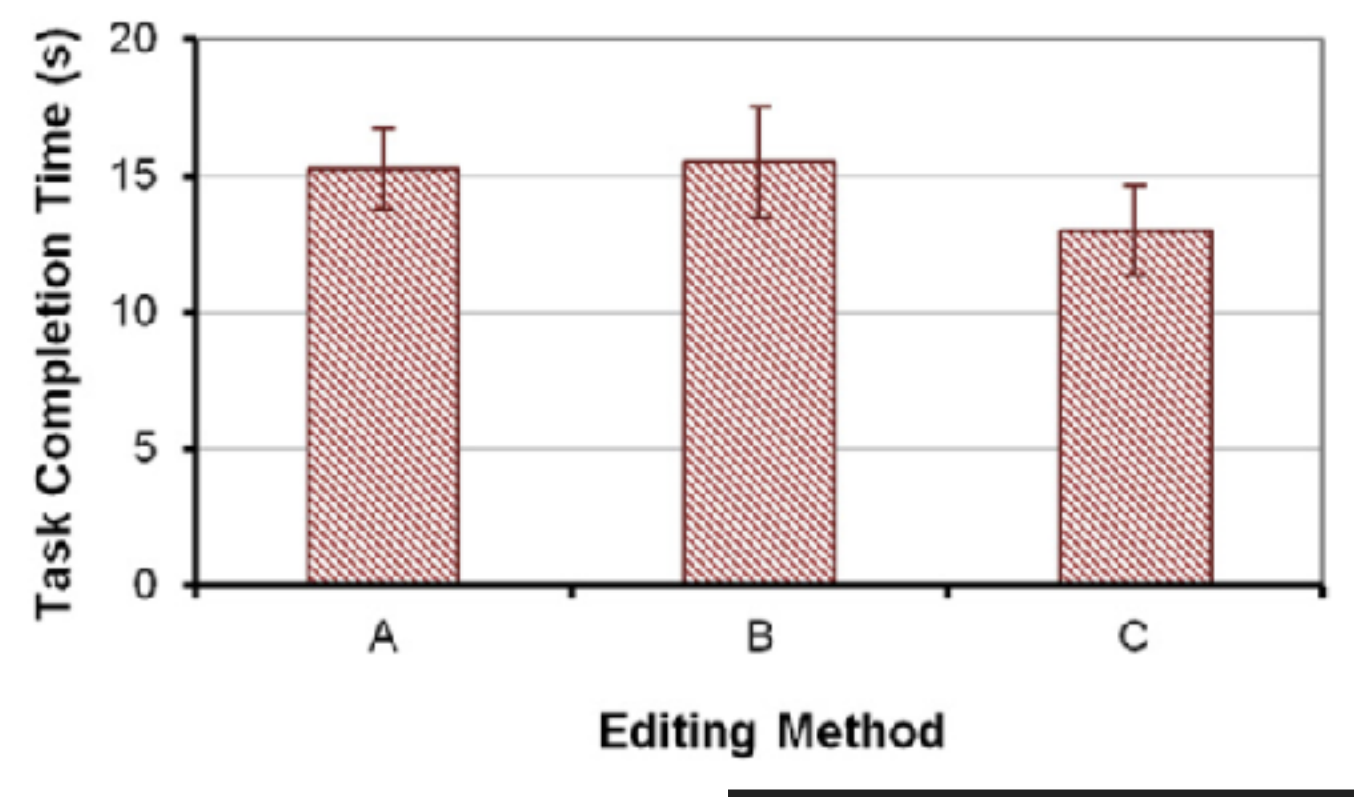
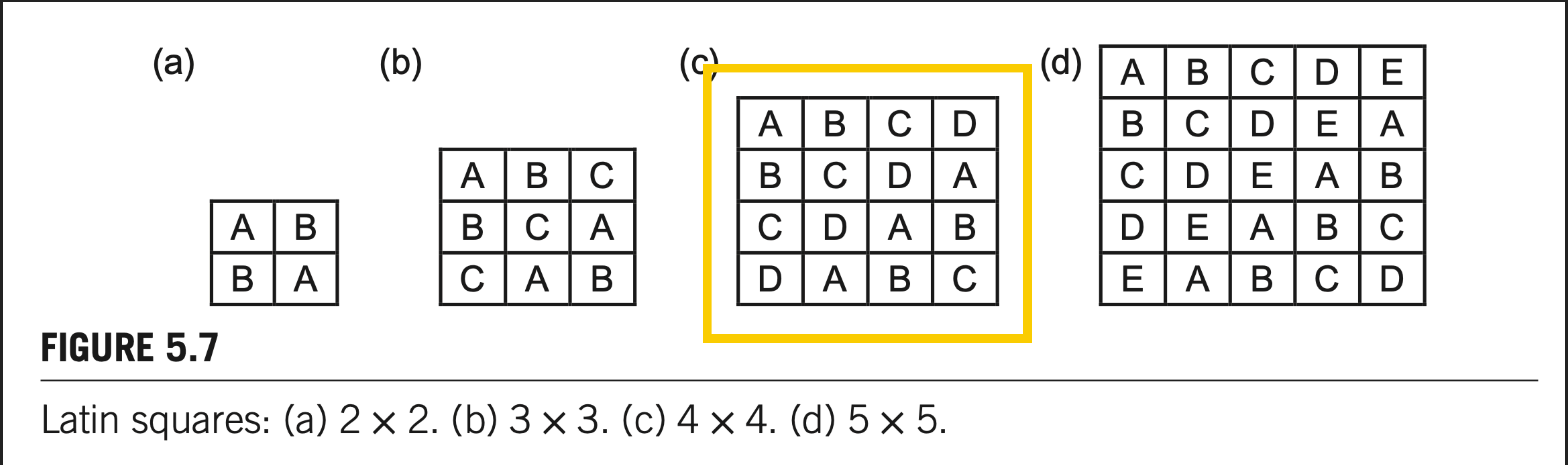


FIGURE 5.9

Hypothetical data for an experiment with one within-subjects factor having three levels (A, B, C). Values are the mean task completion time(s) for five repetitions of an editing task.

Latin Squares: (a) 2 × 2. (b) 3 × 3. (c) 4 × 4. (d) 5 × 5



(a)

A	B
B	A

(b)

A	B	C
B	C	A
C	A	B

(c)

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

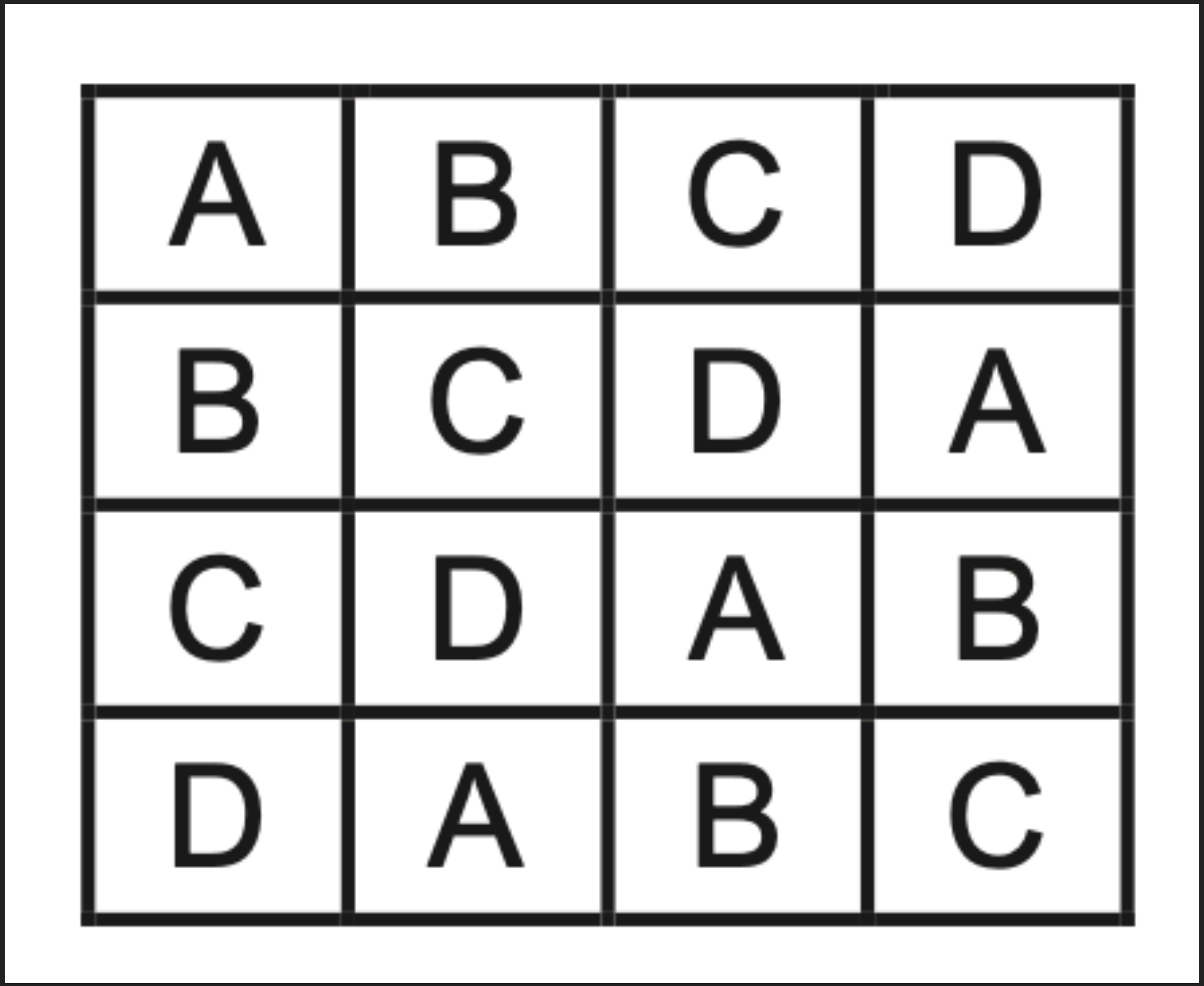
(d)

A	B	C	D	E
B	C	D	E	A
C	D	E	A	B
D	E	A	B	C
E	A	B	C	D

FIGURE 5.7

Latin squares: (a) 2 × 2. (b) 3 × 3. (c) 4 × 4. (d) 5 × 5.

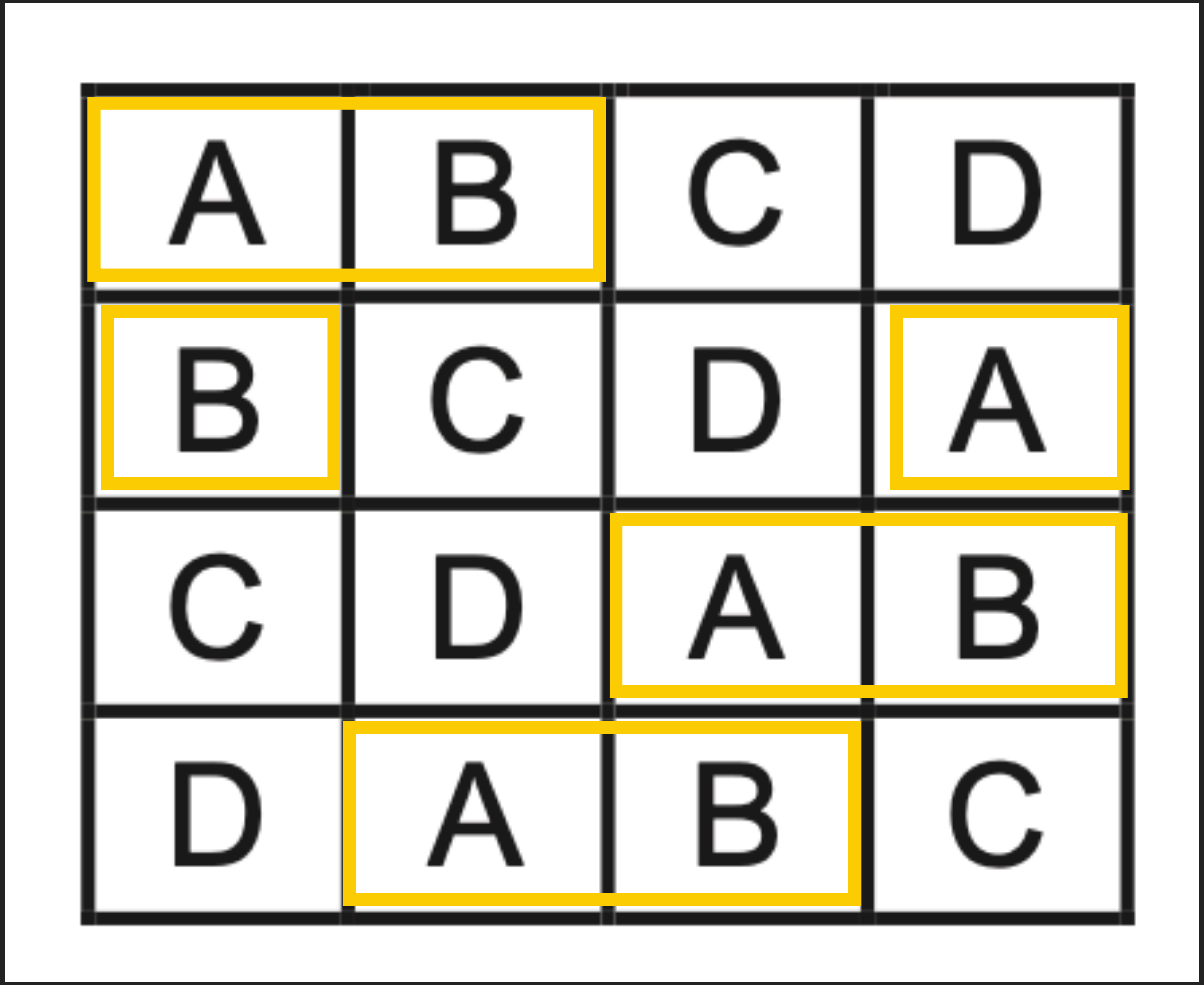
What's wrong with this?



A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

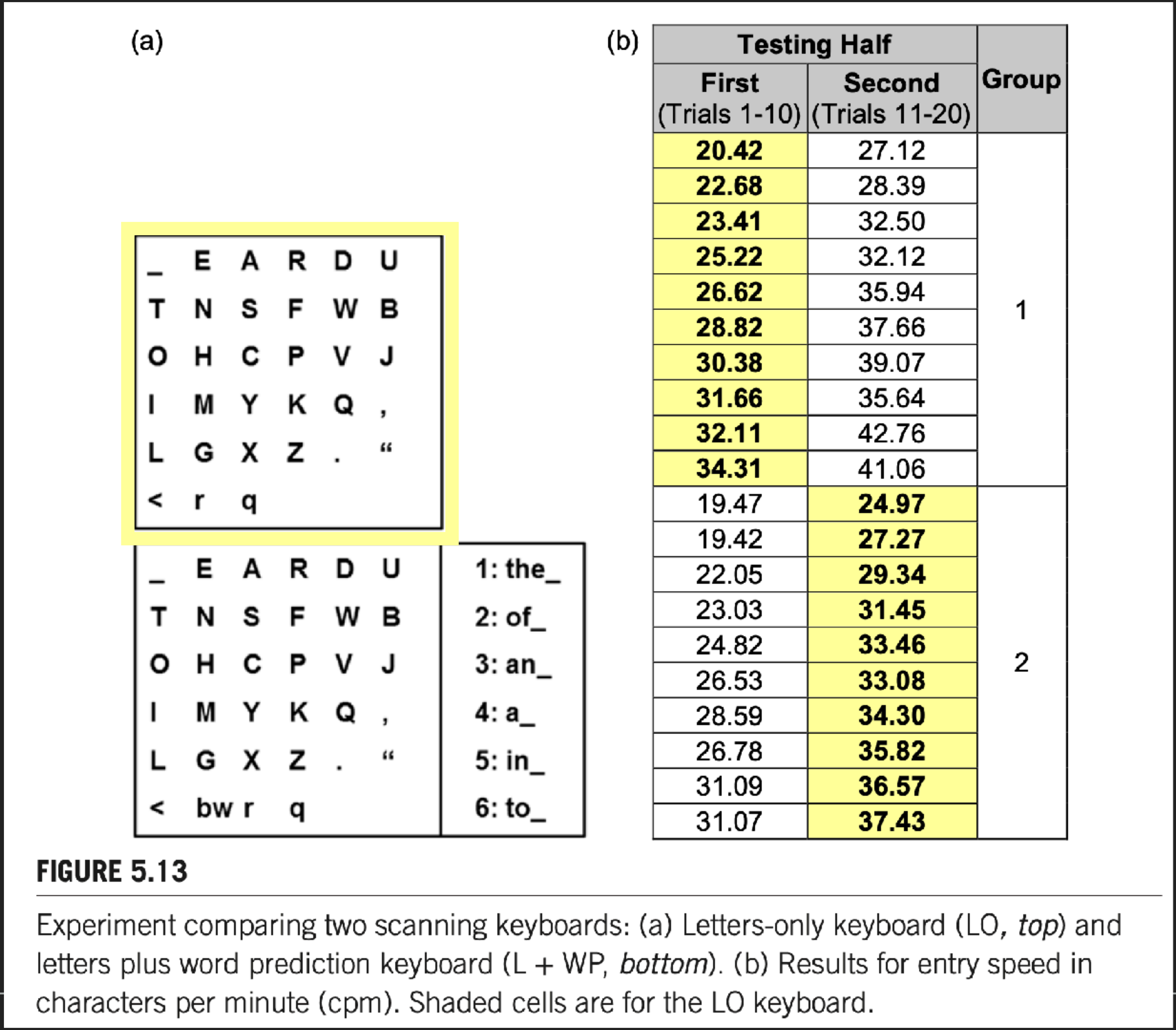
A deficiency in Latin squares of order 3 and higher is that conditions precede and follow other conditions an unequal number of times.

If present, an A-B sequence effect is not fully compensated for.



A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Experiment Comparing Two Scanning Keyboards



(a)

_	E	A	R	D	U
T	N	S	F	W	B
O	H	C	P	V	J
I	M	Y	K	Q	,
L	G	X	Z	.	“
<	r	q			

_	E	A	R	D	U	1: the_
T	N	S	F	W	B	2: of_
O	H	C	P	V	J	3: an_
I	M	Y	K	Q	,	4: a_
L	G	X	Z	.	“	5: in_
<	bw	r	q			6: to_

(b)

Testing Half		Group
First (Trials 1-10)	Second (Trials 11-20)	
20.42	27.12	1
22.68	28.39	
23.41	32.50	
25.22	32.12	
26.62	35.94	
28.82	37.66	
30.38	39.07	
31.66	35.64	
32.11	42.76	
34.31	41.06	
19.47	**24.97**	2
19.42	**27.27**	
22.05	**29.34**	
23.03	**31.45**	
24.82	**33.46**	
26.53	**33.08**	
28.59	**34.30**	
26.78	**35.82**	
31.09	**36.57**	
31.07	**37.43**	

FIGURE 5.13

Experiment comparing two scanning keyboards: (a) Letters-only keyboard (LO, *top*) and letters plus word prediction keyboard (L + WP, *bottom*). (b) Results for entry speed in characters per minute (cpm). Shaded cells are for the LO keyboard.

Example (continued)

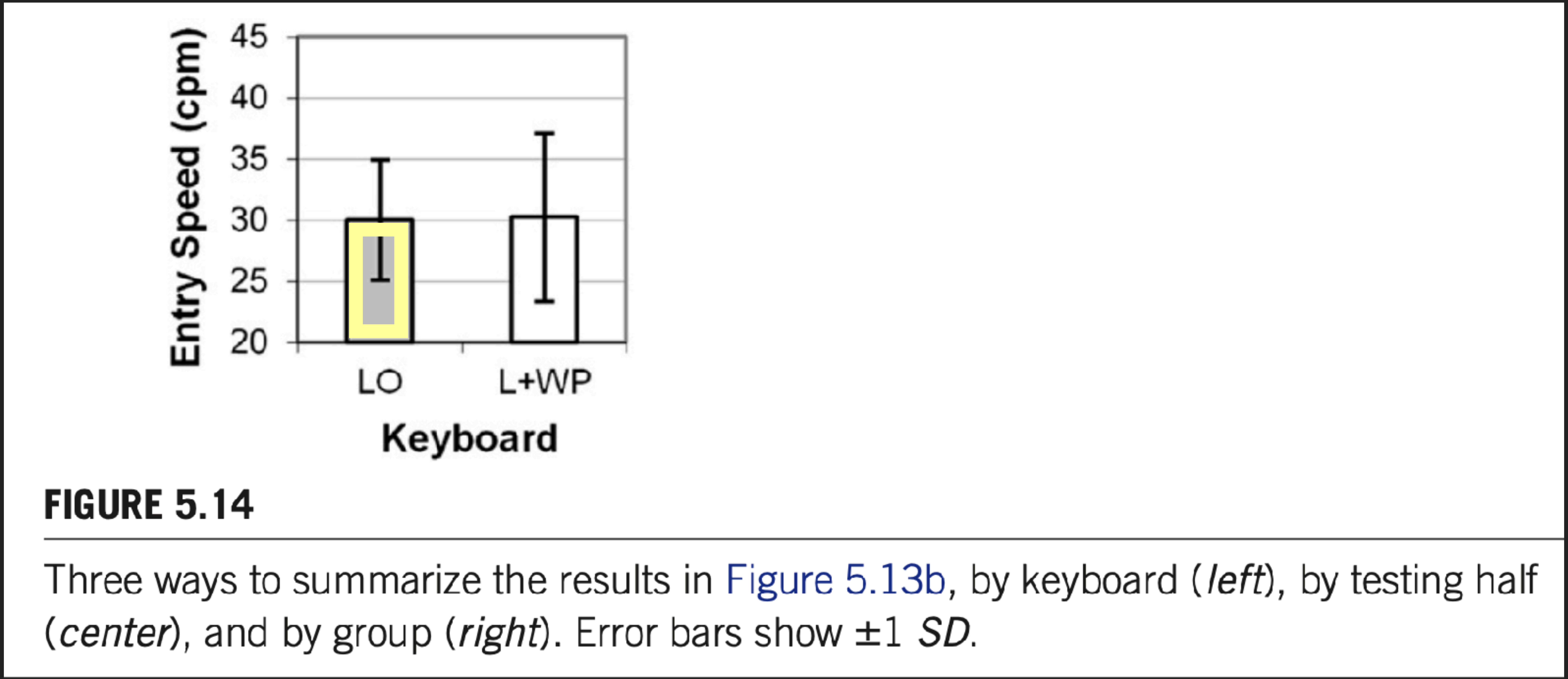


FIGURE 5.14

Three ways to summarize the results in Figure 5.13b, by keyboard (*left*), by testing half (*center*), and by group (*right*). Error bars show ±1 *SD*.

Example (continued)

Learning effect

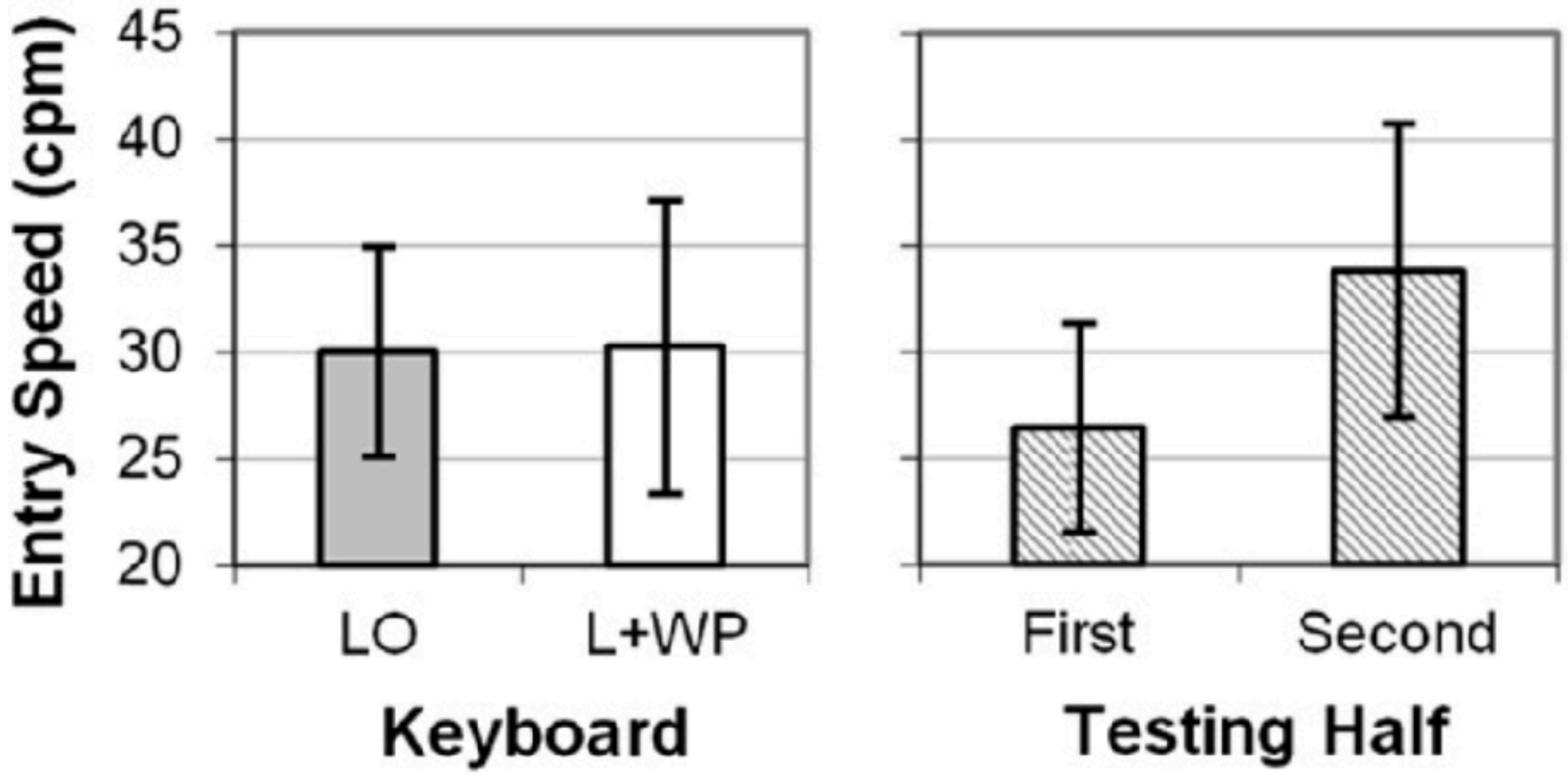


FIGURE 5.14

Three ways to summarize the results in Figure 5.13b, by keyboard (*left*), by testing half (*center*), and by group (*right*). Error bars show ±1 *SD*.

Example (continued)

Learning effect

Asymmetric skill transfer!

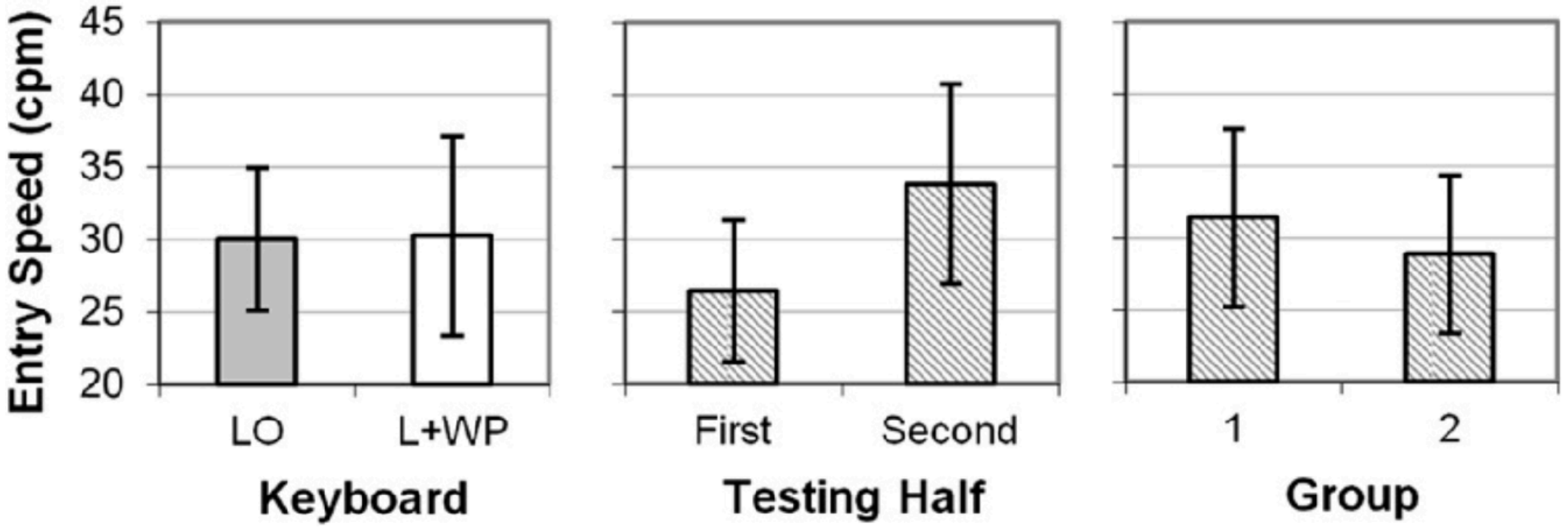


FIGURE 5.14

Three ways to summarize the results in Figure 5.13b, by keyboard (*left*), by testing half (*center*), and by group (*right*). Error bars show ±1 *SD*.

Counterbalancing only works if the order effects are the same or similar.

Example (continued)

Learning: Both groups improved, at comparable rates

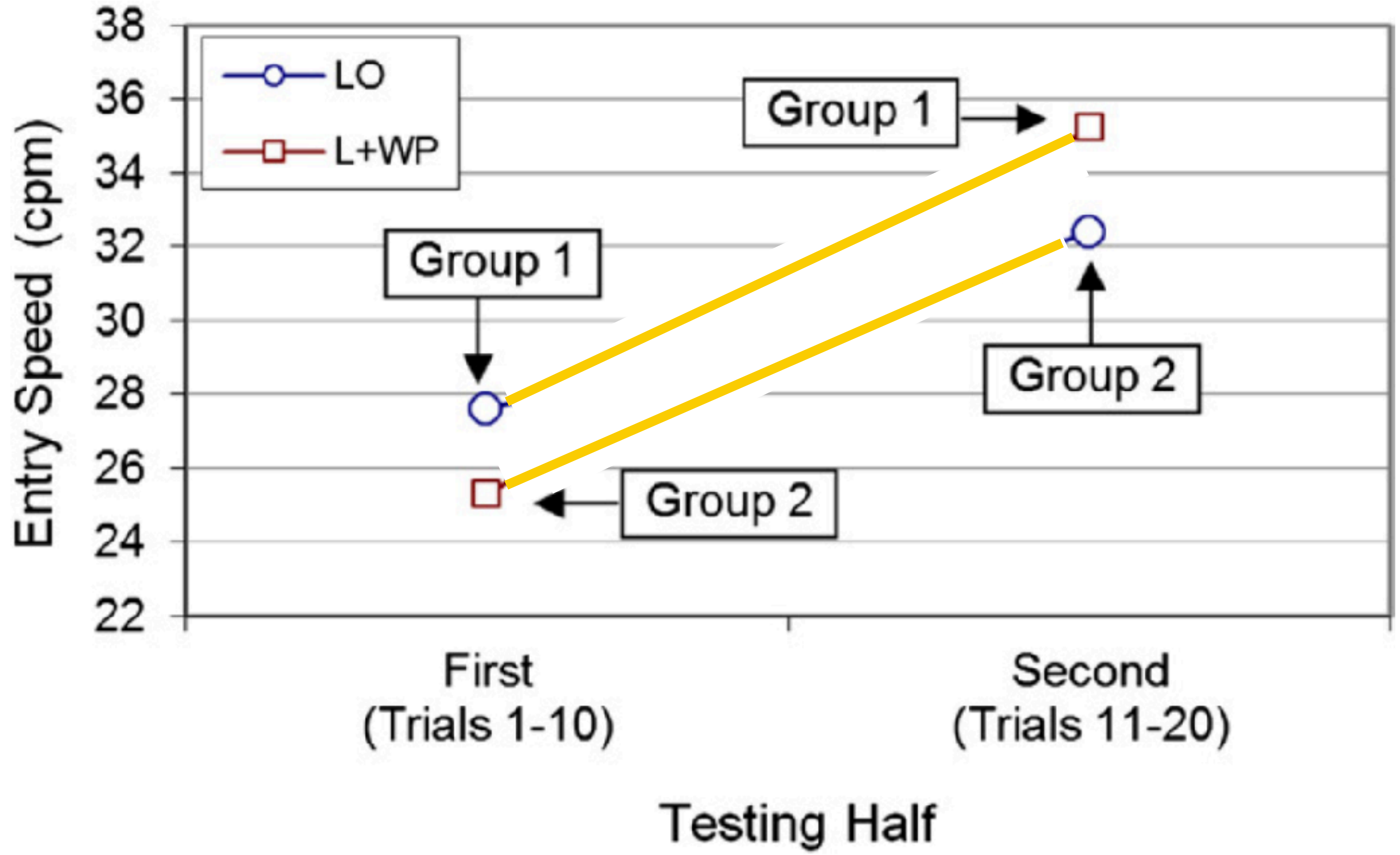


FIGURE 5.15

Demonstration of asymmetric skill transfer. The chart uses the data in Figure 5.13b.

Example (continued)

Harder to start with the more complex keyboard

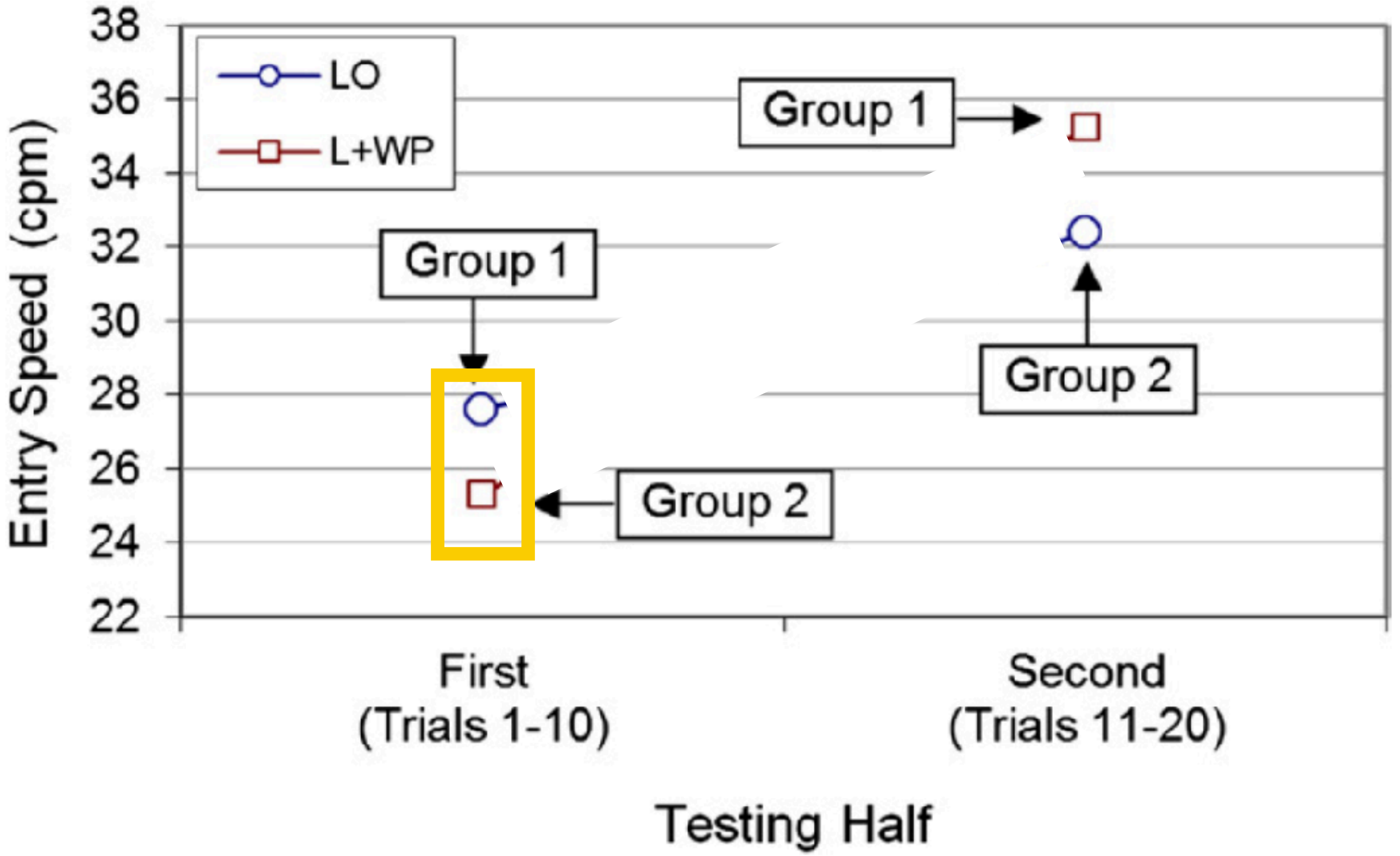


FIGURE 5.15

Demonstration of asymmetric skill transfer. The chart uses the data in Figure 5.13b.

Example (continued)

But: higher efficiency eventually with the more complex keyboard

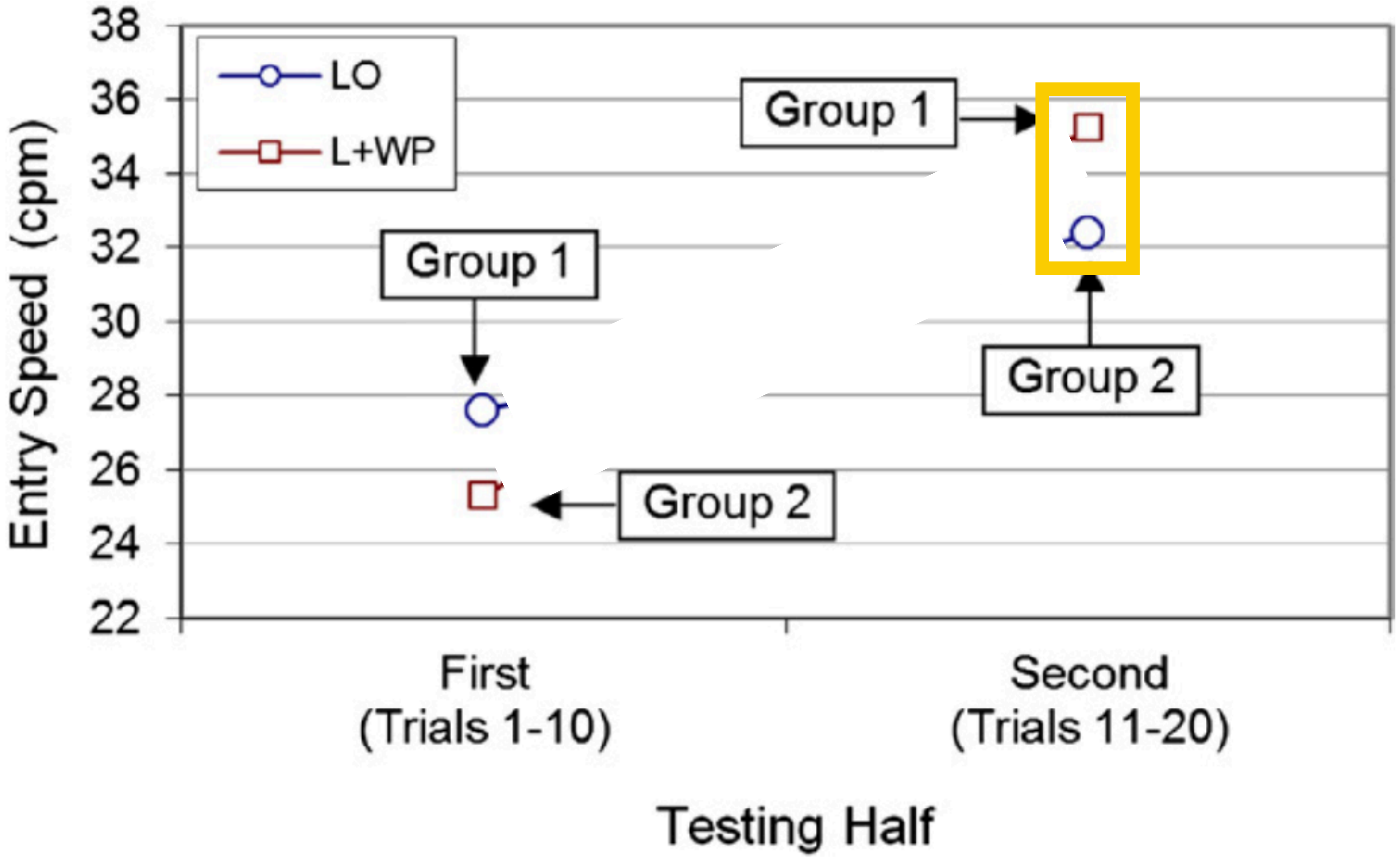


FIGURE 5.15

Demonstration of asymmetric skill transfer. The chart uses the data in Figure 5.13b.

Example (continued)

Asymmetric skill transfer!

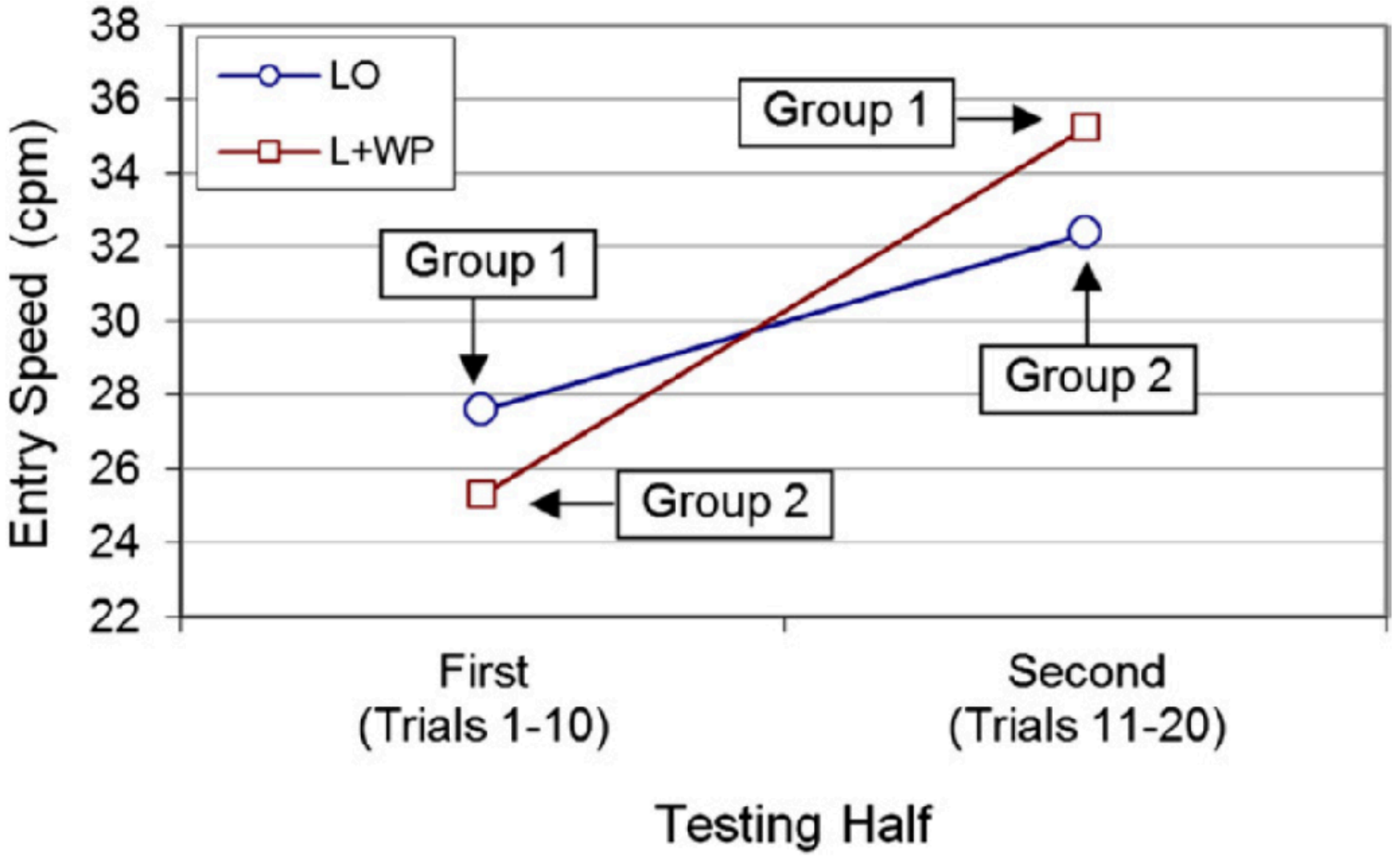


FIGURE 5.15

Demonstration of asymmetric skill transfer. The chart uses the data in Figure 5.13b.

Investigating more than one independent variable

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

Pretest-posttest

R	O	X	O
R	O		O

Alternative Xs with pretest

R	O	X_A	O
R	O	X_B	O

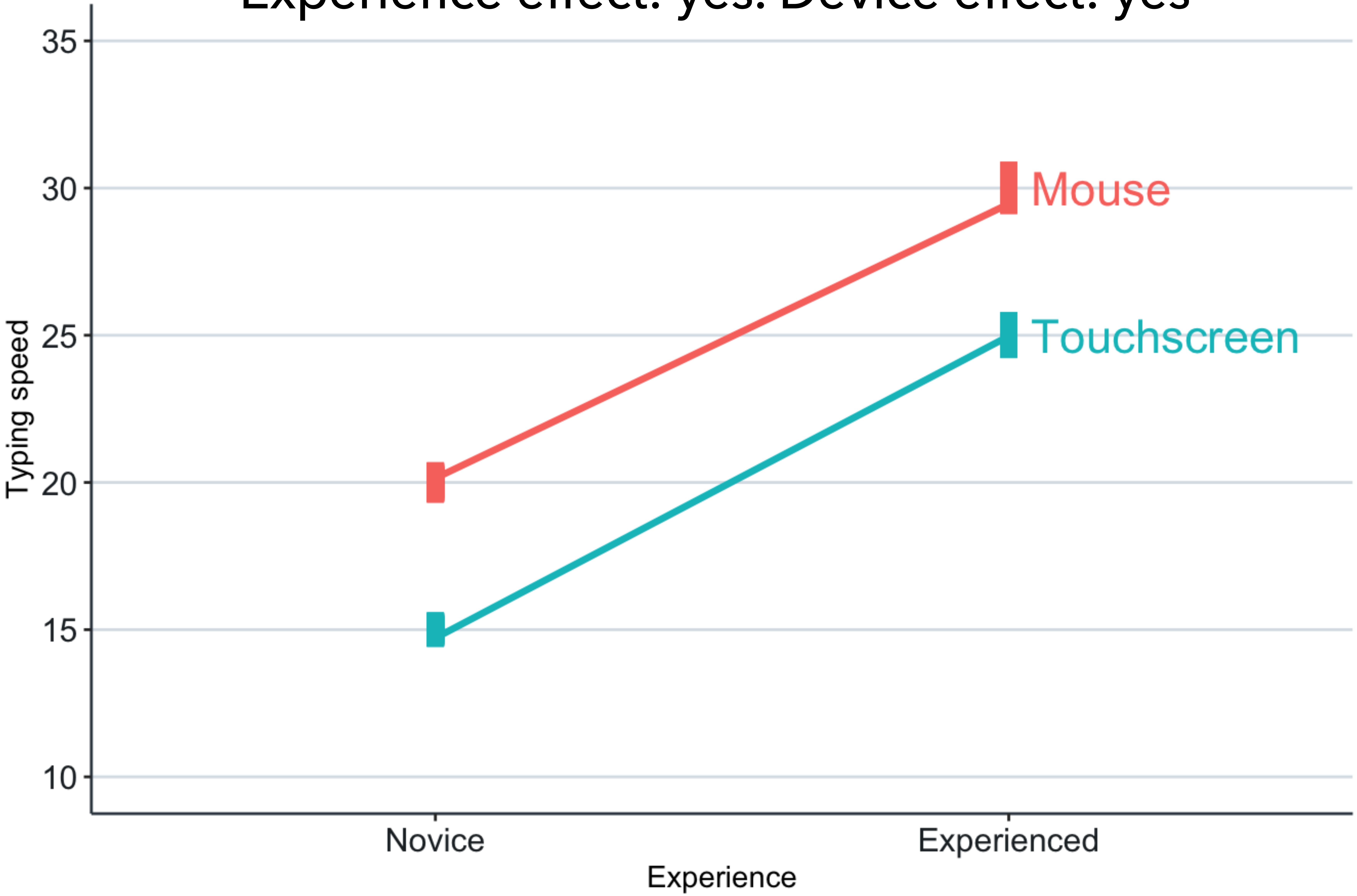
Factorial

R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

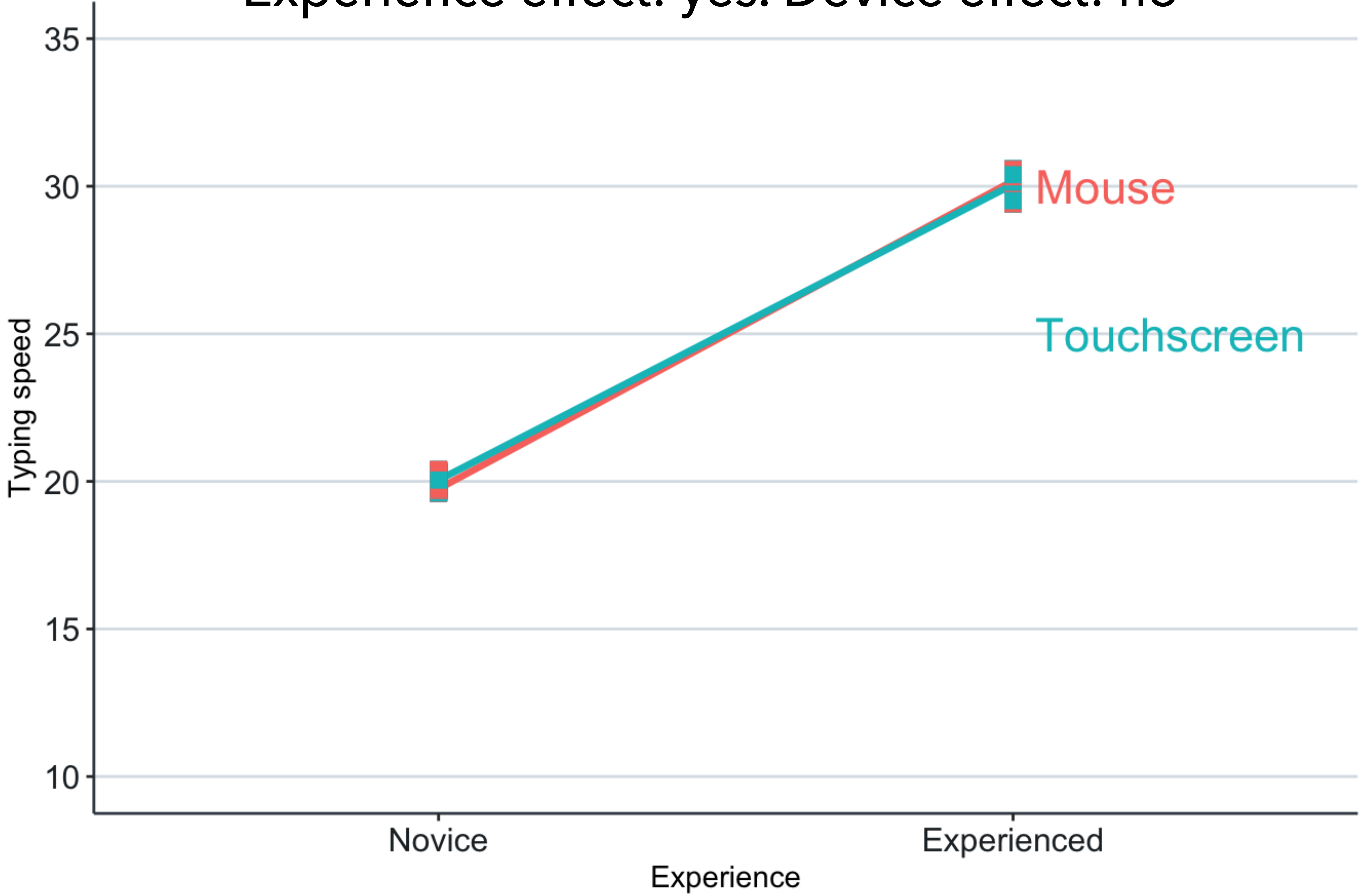
- Three major advantages:
    - They often require fewer units.
    - They allow testing combinations of treatments more easily.
    - They allow testing interactions.

Example: Typing speed = f(Experience, Device)

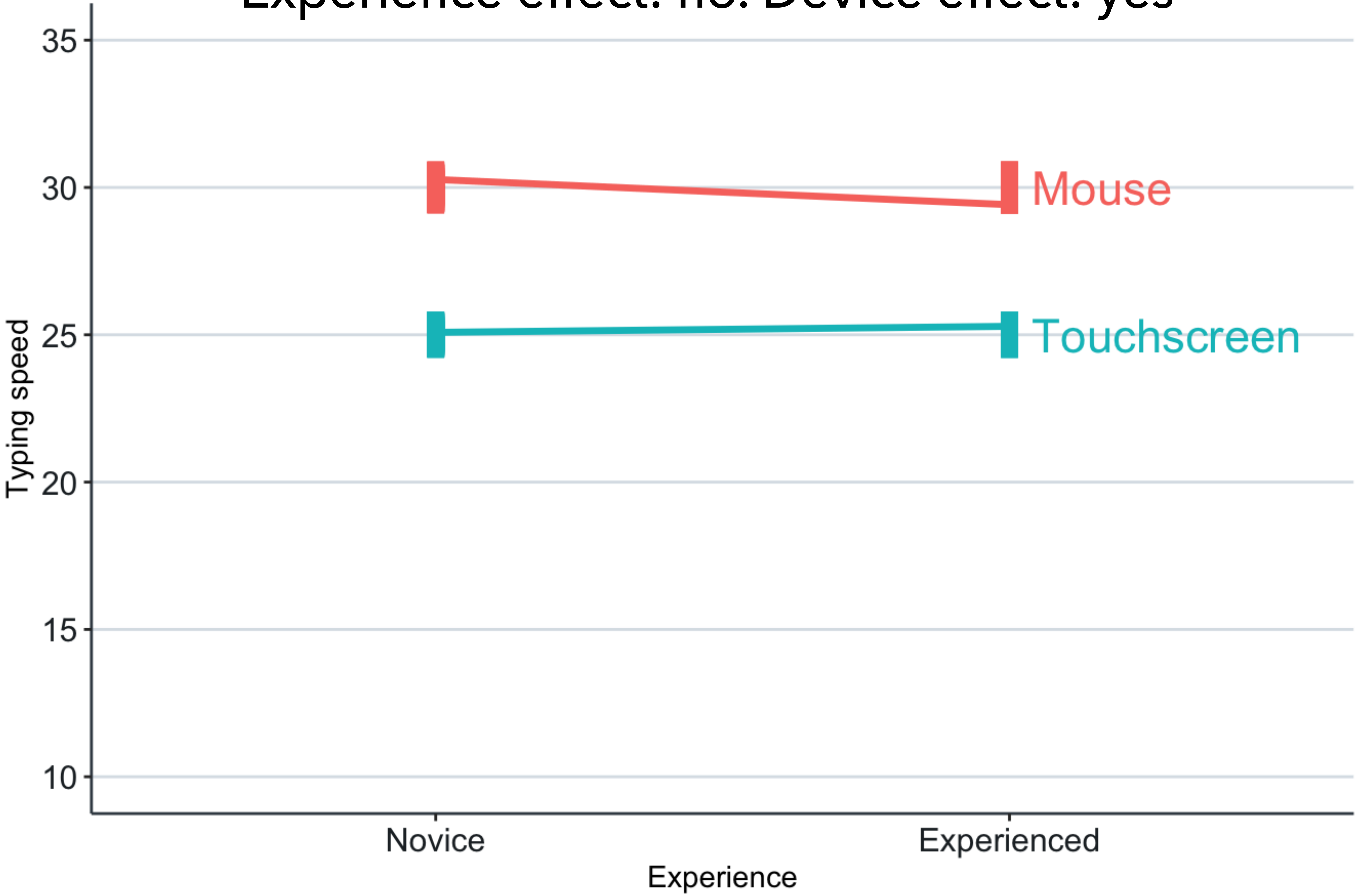
Experience effect: yes. Device effect: yes



Experience effect: yes. Device effect: no

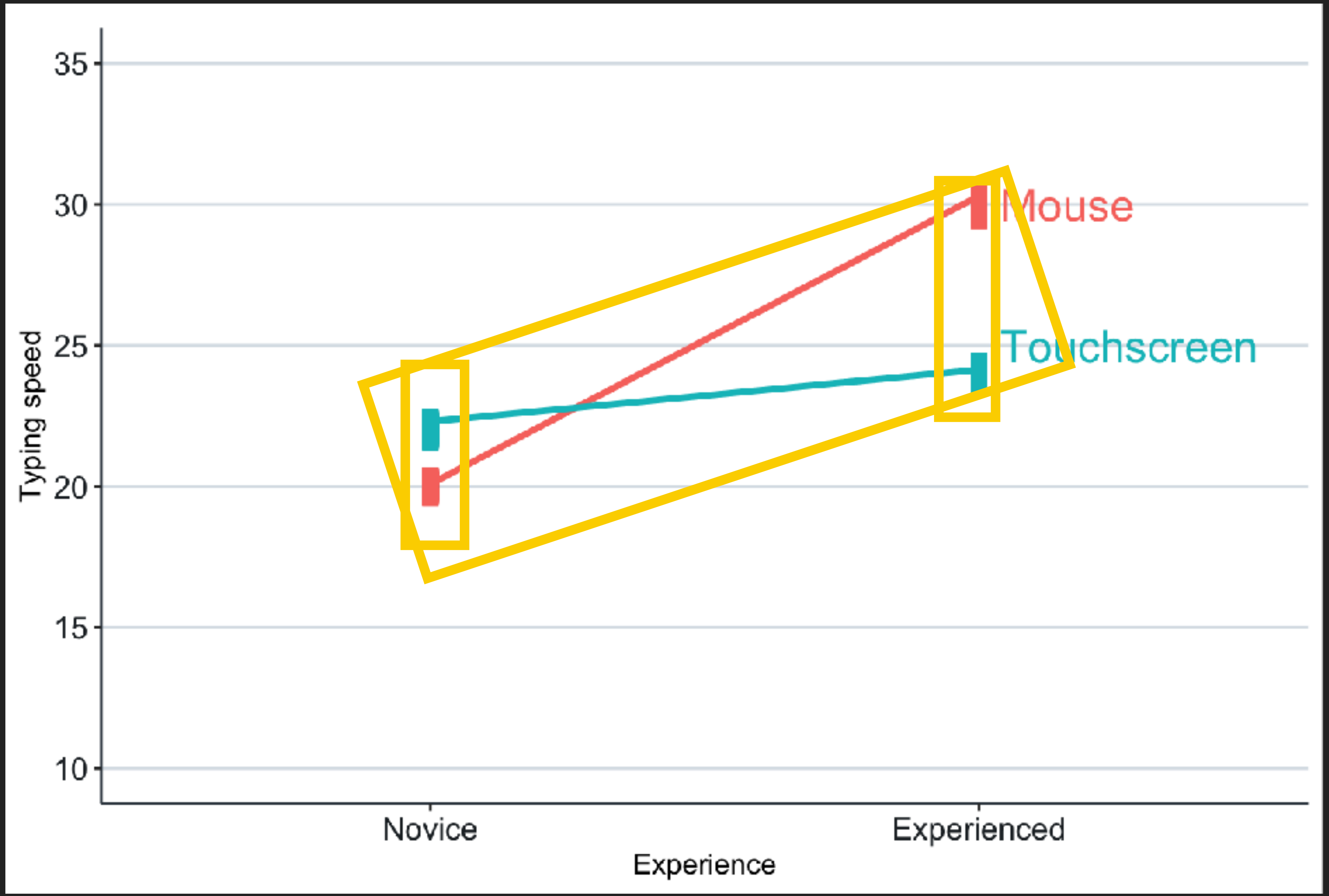


Experience effect: no. Device effect: yes



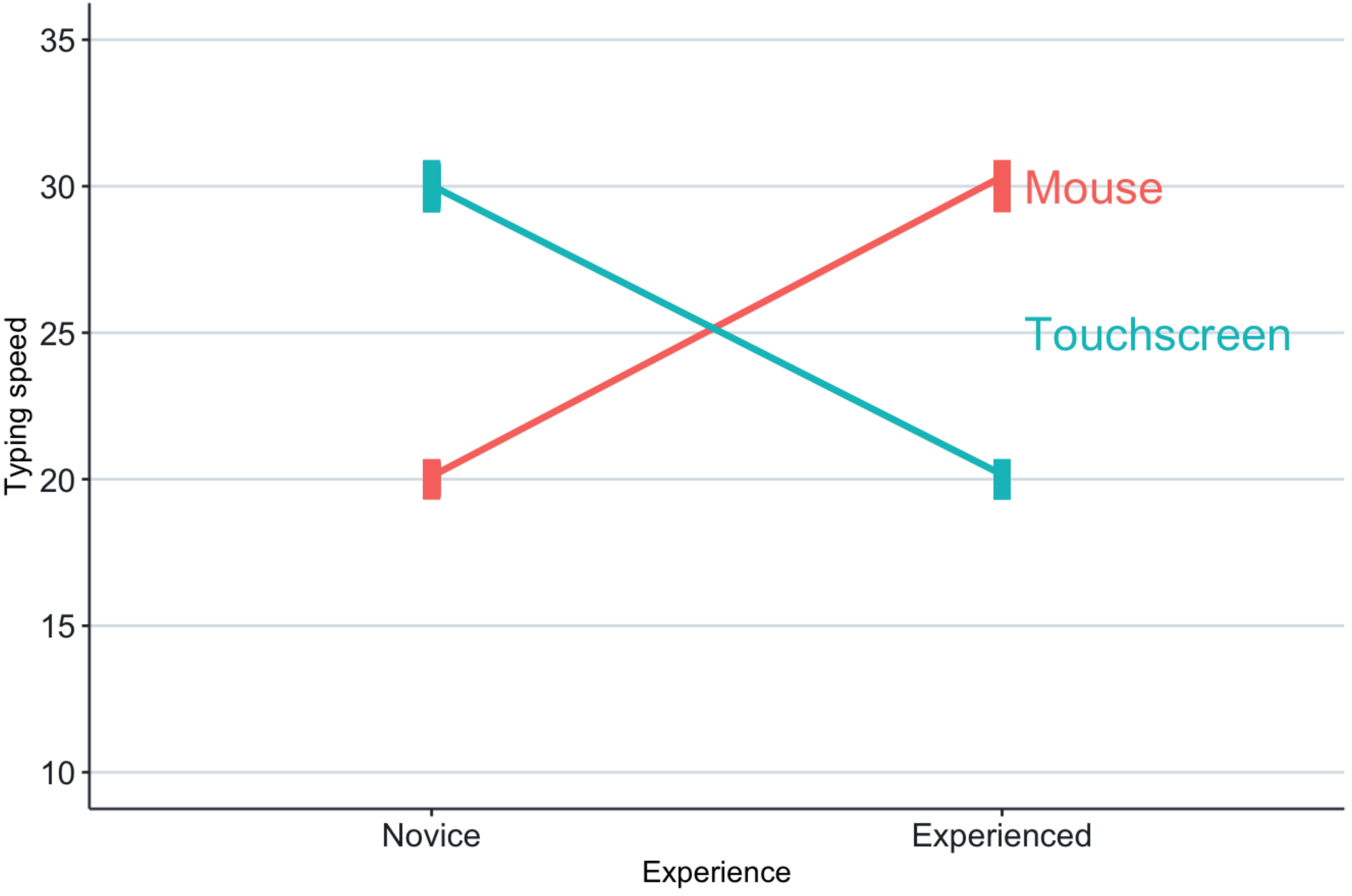
Example of Interaction Effects

- Novice users can select targets faster with a touchscreen than with a mouse.
- Experienced users can select targets faster with a mouse than with a touchscreen.
- The target selection speeds for both the mouse and the touchscreen increase as the user gains more experience with the device.
- However, the increase in speed is much larger for the mouse than for the touchscreen.



😱

Experience effect: no. Device effect: no. Interaction: yes



Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

Pretest-posttest

R	O	X	O
R	O		O

Alternative Xs with pretest

R	O	X_A	O
R	O	X_B	O

Factorial

R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

Longitudinal

R	O … O	X	O … O
R	O … O		O … O

- Examine how effects change over time

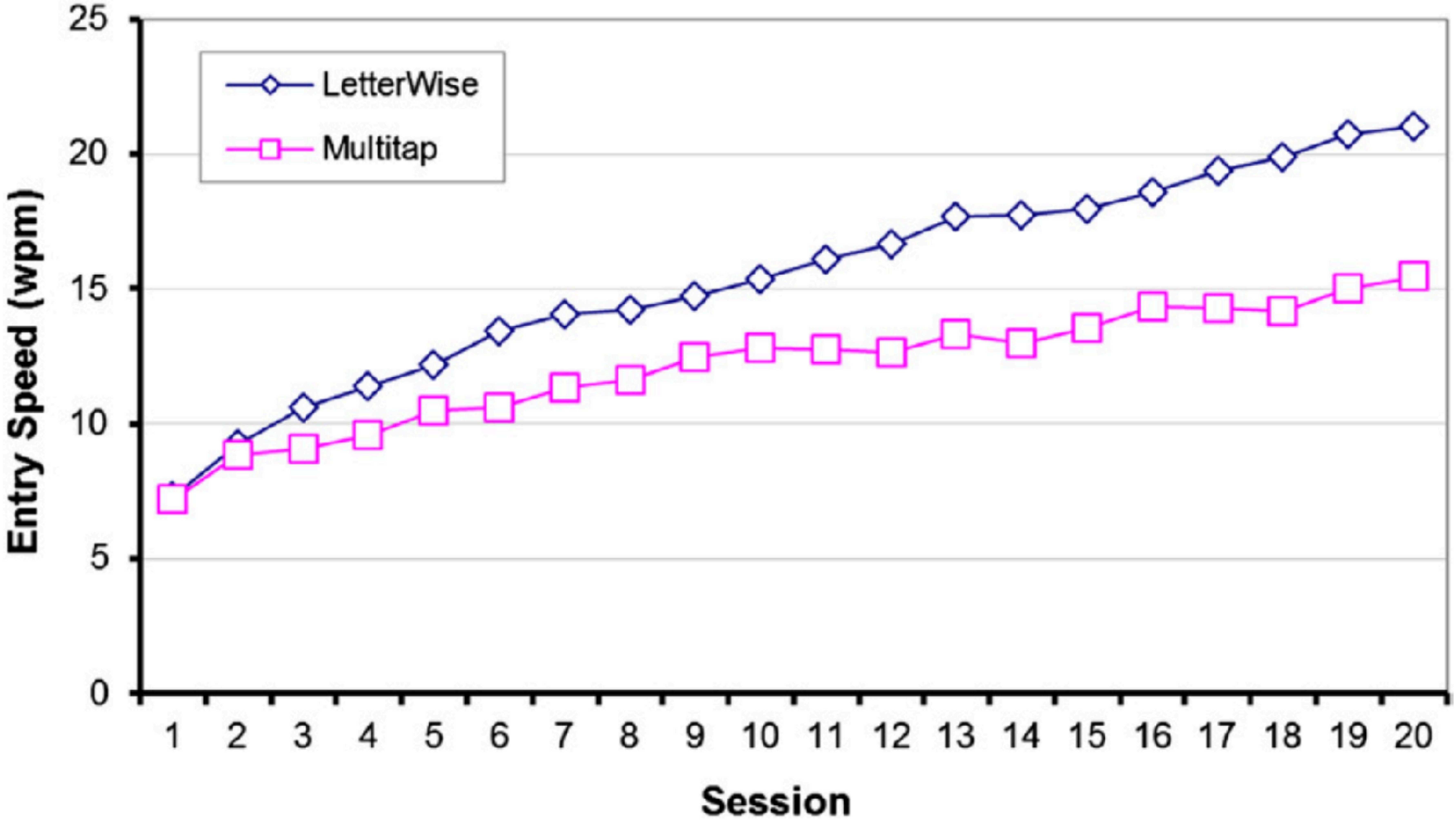


FIGURE 5.16

Example of a longitudinal study. Two text entry methods were tested and compared over 20 sessions of input. Each session involved about 30 minutes of text entry.

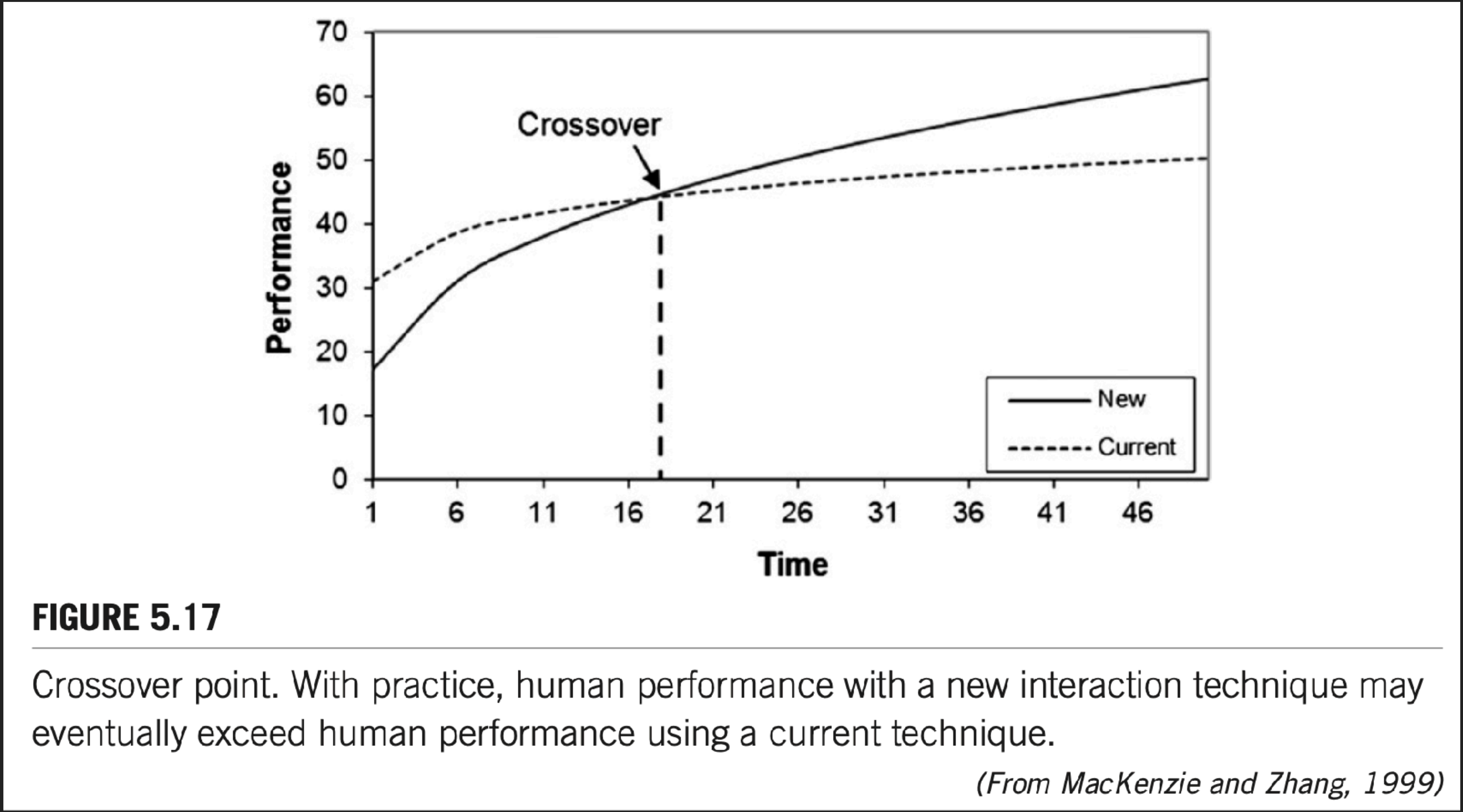


FIGURE 5.17

Crossover point. With practice, human performance with a new interaction technique may eventually exceed human performance using a current technique.

(From MacKenzie and Zhang, 1999)

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

Pretest-posttest

R	O	X	O
R	O		O

Alternative Xs with pretest

R	O	X_A	O
R	O	X_B	O

Factorial

R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

Crossover

R	O	X_A	O	X_B	O
R	O	X_B	O	X_A	O

- Used to counterbalance and assess order effects with multiple treatments

. . . to be continued