Photo credit: Dave DiCello

17-803 Empirical Methods

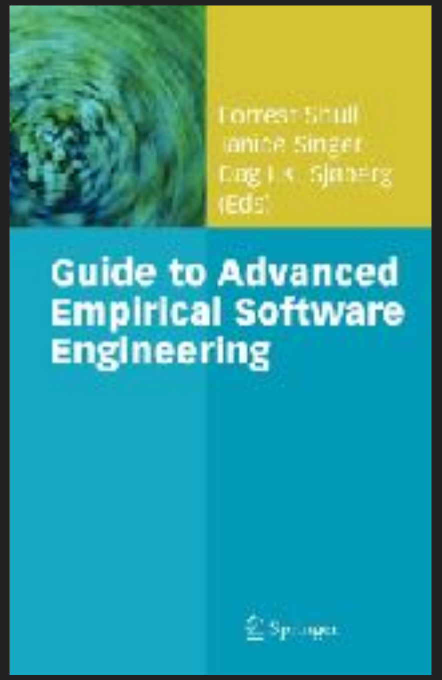
Bogdan Vasilescu, S3D

Designing Experiments (II)

Thursday, October 13, 2022

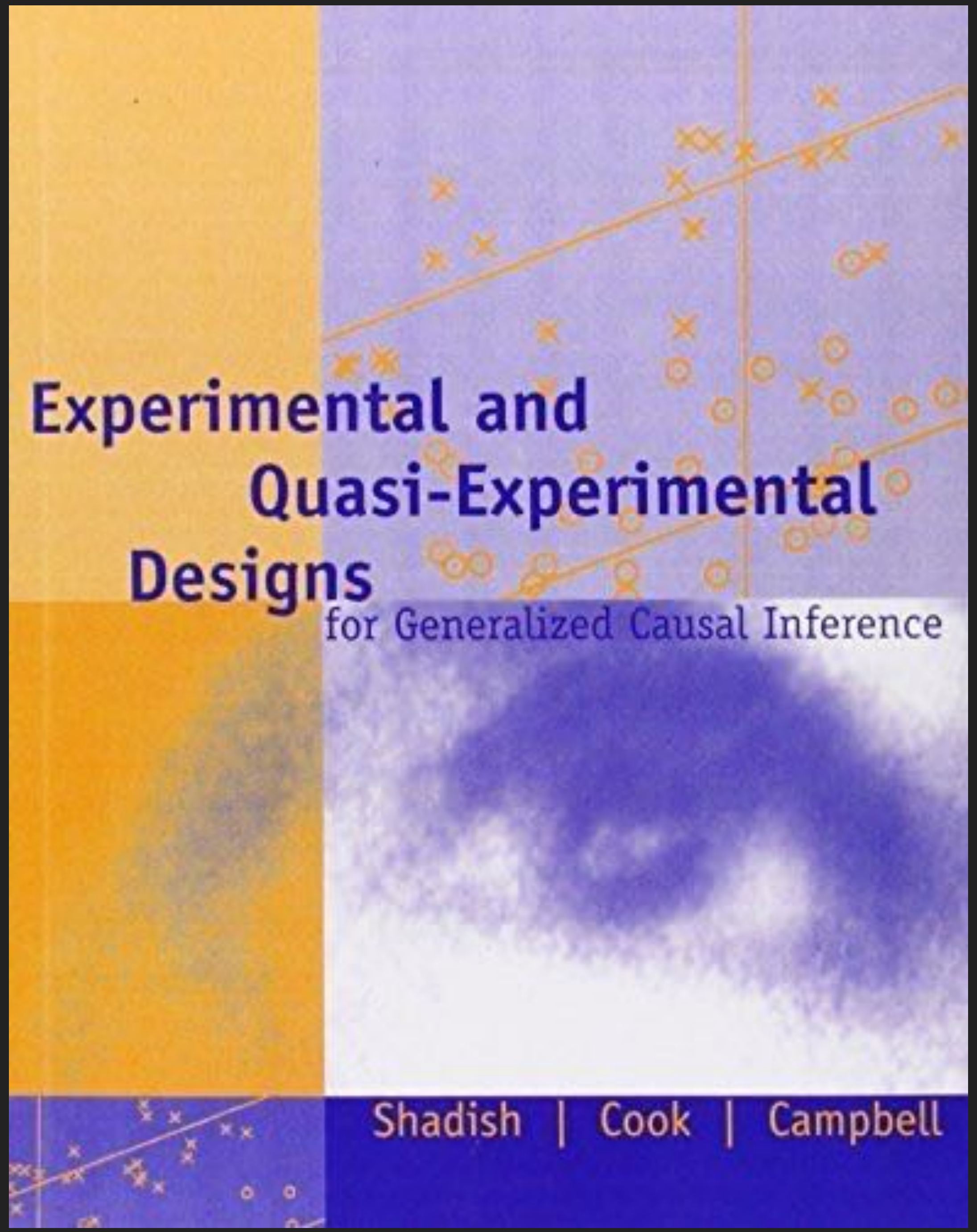
Readings

Ch 10 (Analysis and interpretation)

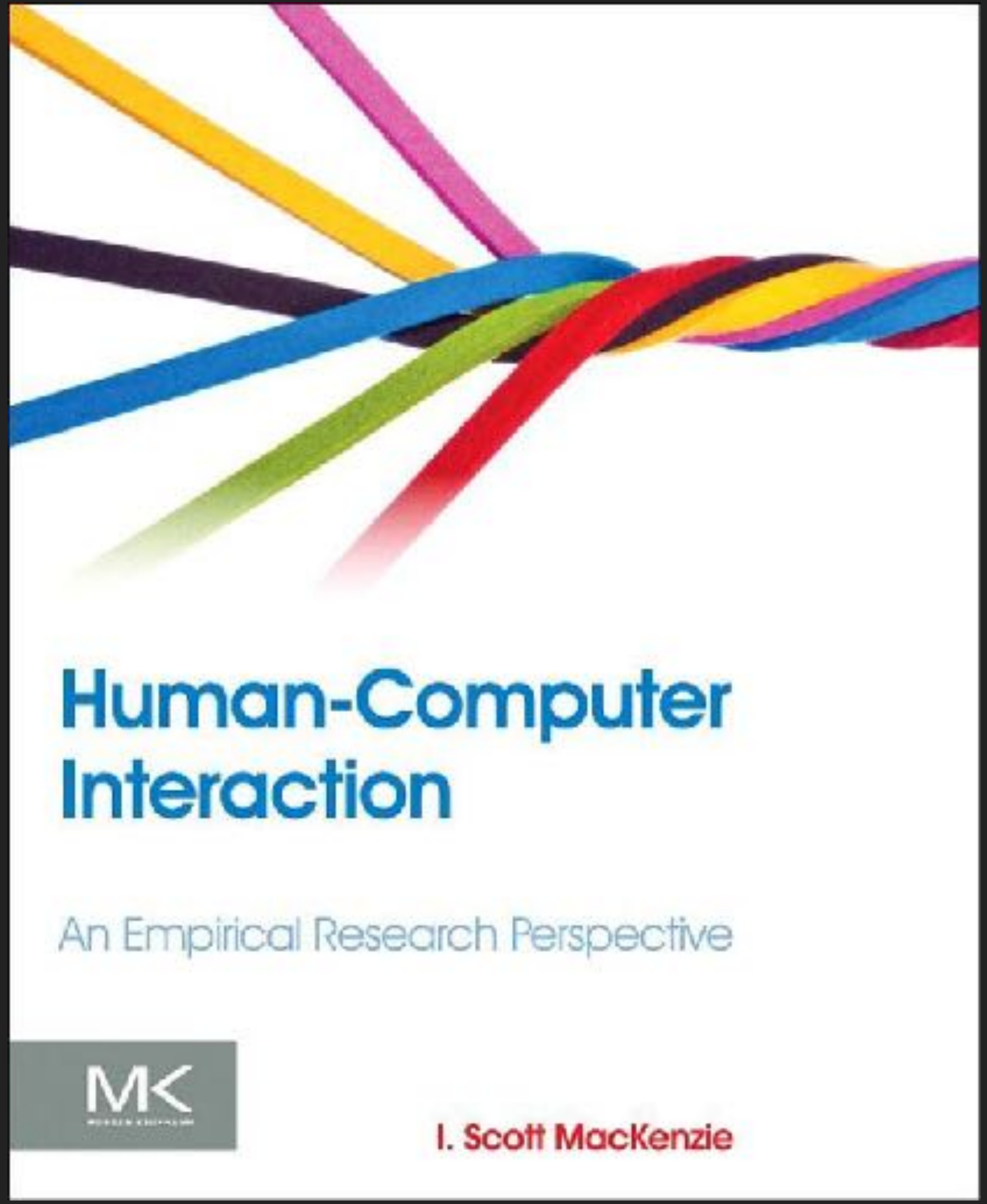


Ch 6 (Statistical methods and measurement)

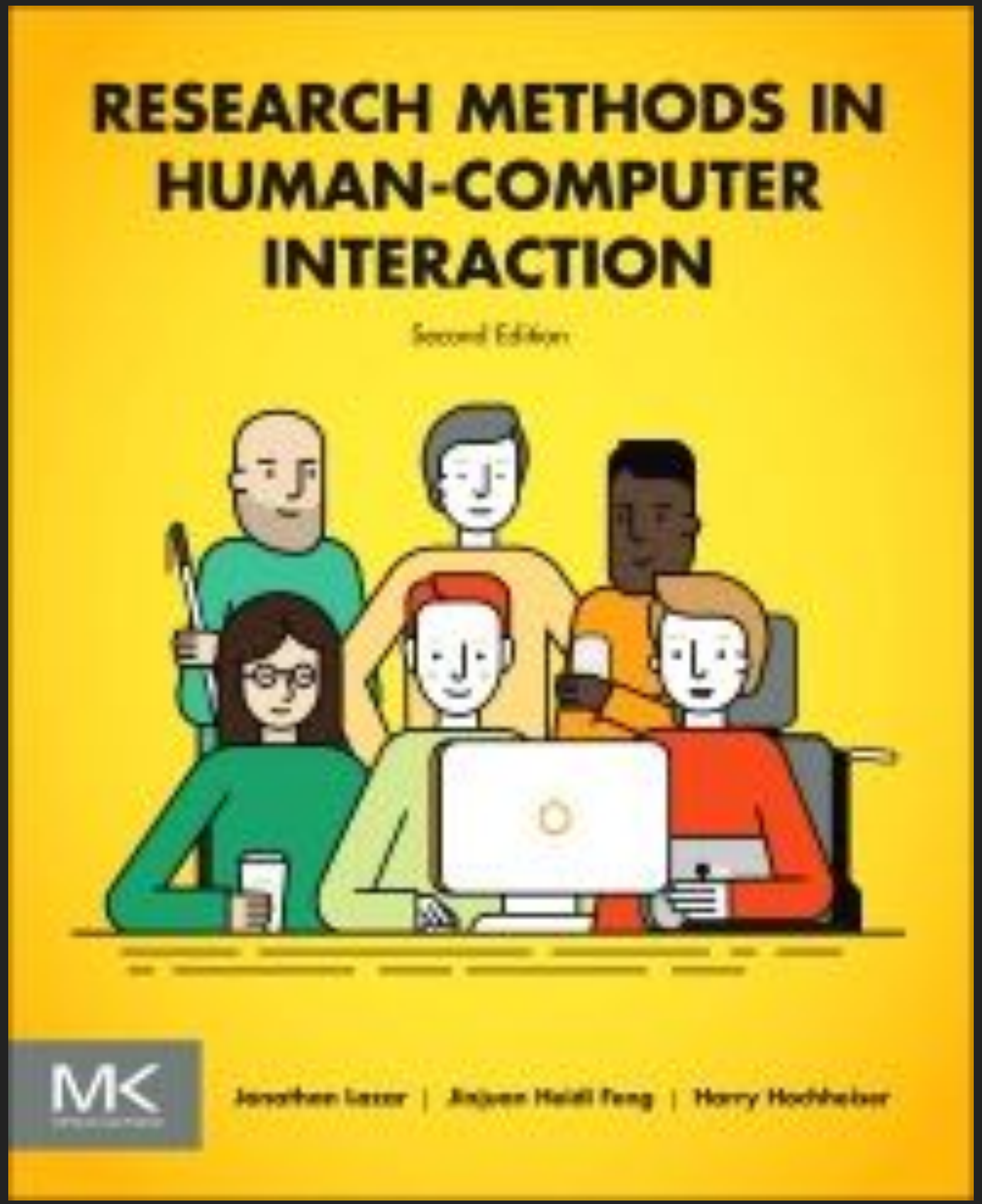
Ch 5 (Effect sizes and power analysis)
Ch 13 (Fair statistical communication)
Ch 14 (Improving statistical practice)



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



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



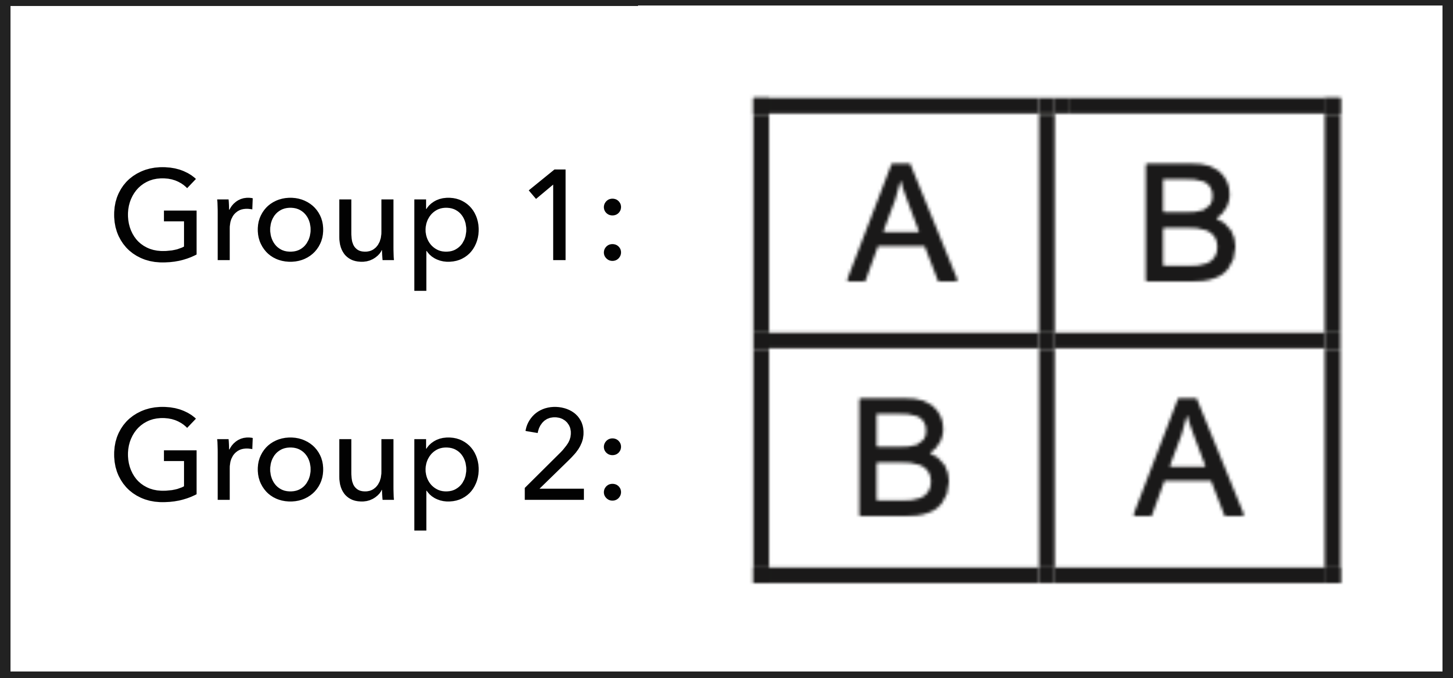
Ch 3 (Experimental design)
Ch 4 (Statistical analysis)

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.



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	14.7	1.84
2	14.84	16.03	14.01	A B C		
3	16.74	15.15	15.19			
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2	14.6	2.46
6	15.17	13.09	12.83	B C A		
7	14.68	17.66	15.26			
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3	14.4	1.88
10	14.37	13.98	12.91	C A B		
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	14.7	1.84
2	14.84	16.03	14.01			
3	16.74	15.15	15.19	A B C		
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2	14.6	2.46
6	15.17	13.09	12.83			
7	14.68	17.66	15.26	B C A		
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3	14.4	1.88
10	14.37	13.98	12.91			
11	14.40	19.12	11.59	C A B		
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 A	B	C	Group	*Mean*	*SD*
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 A	Test Condition B	Test Condition C	Group	*Mean*	*SD*
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 A	Test Condition B	Test Condition C	Group	*Mean*	*SD*
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			

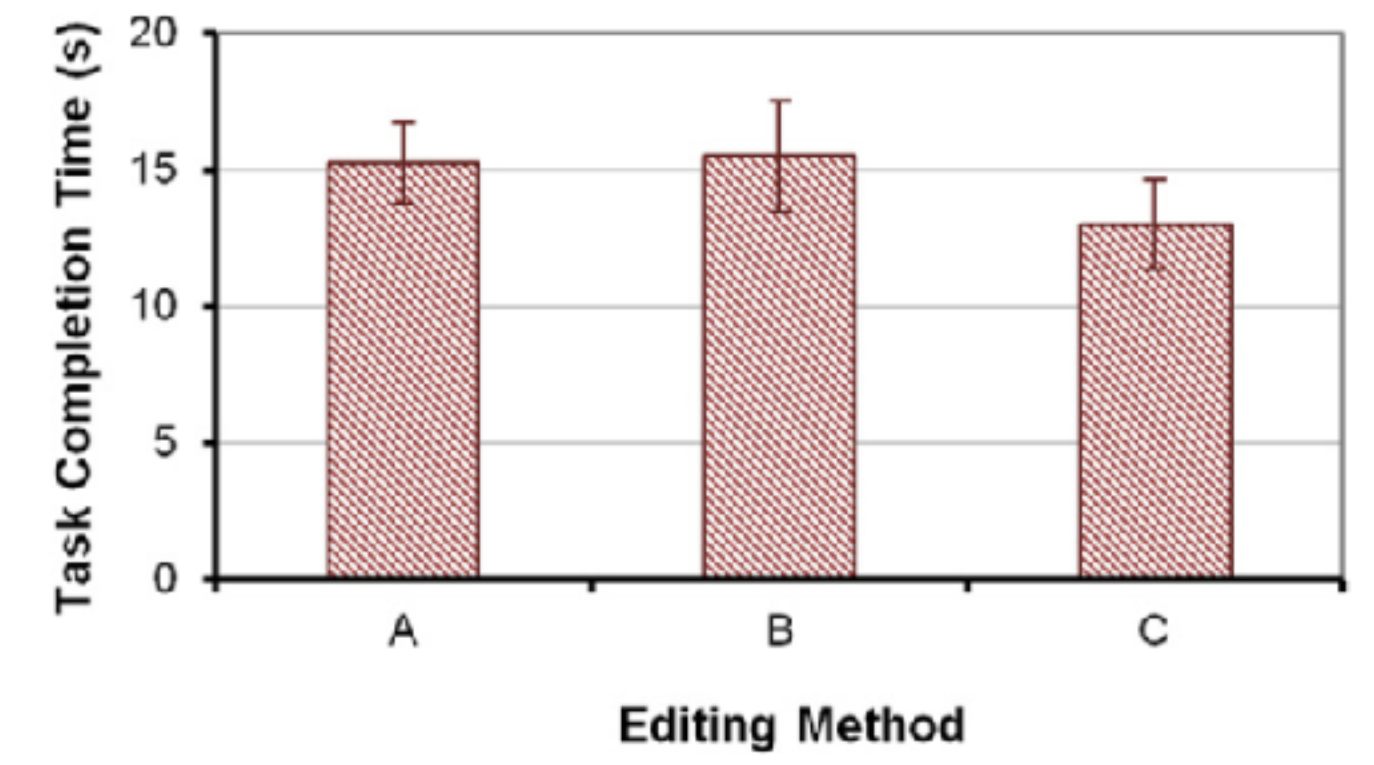


FIGURE 5.9

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

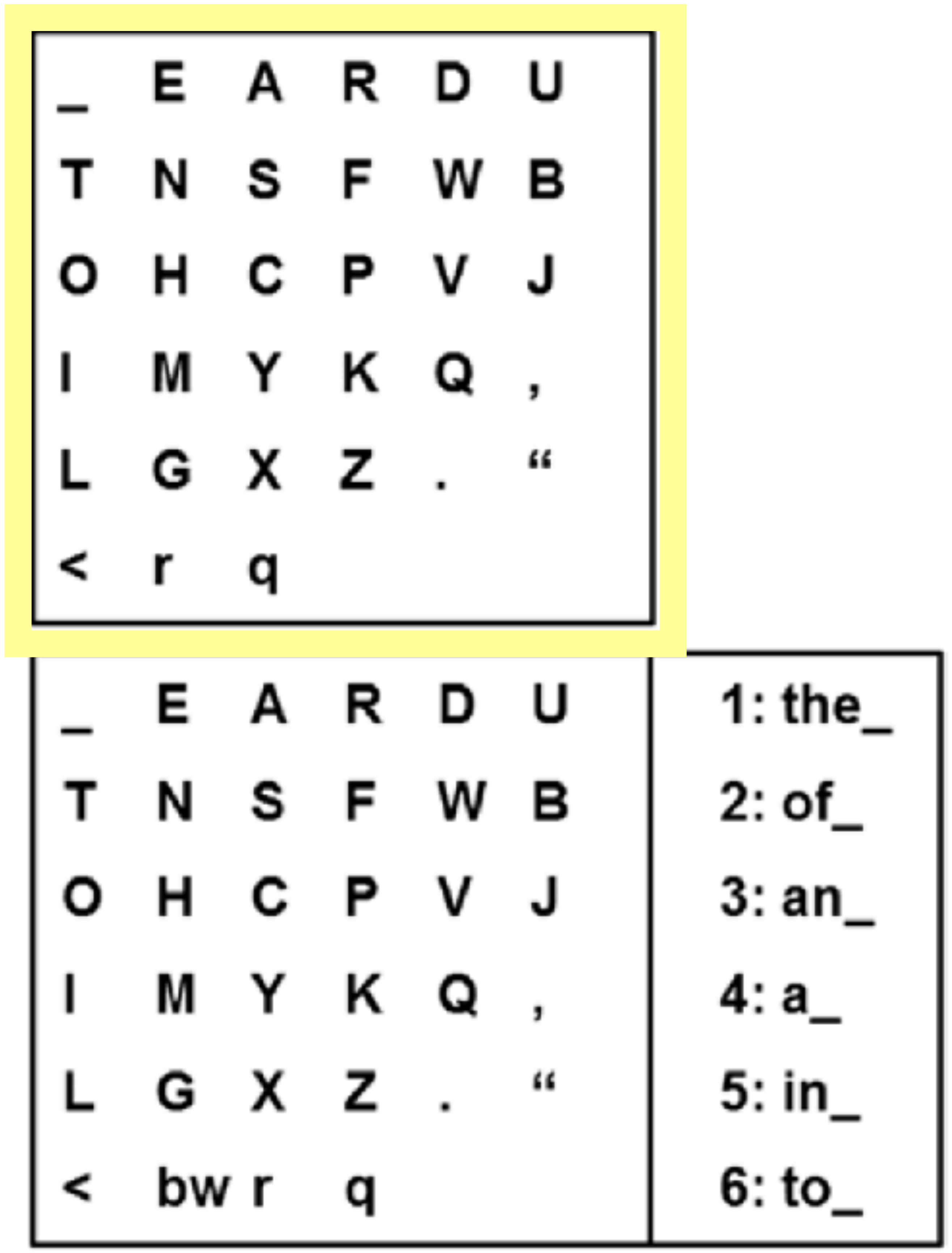
A deficiency in Latin squares of order 3 and higher is that conditions precede and follow other conditions an unequal number of times.

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)



(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)

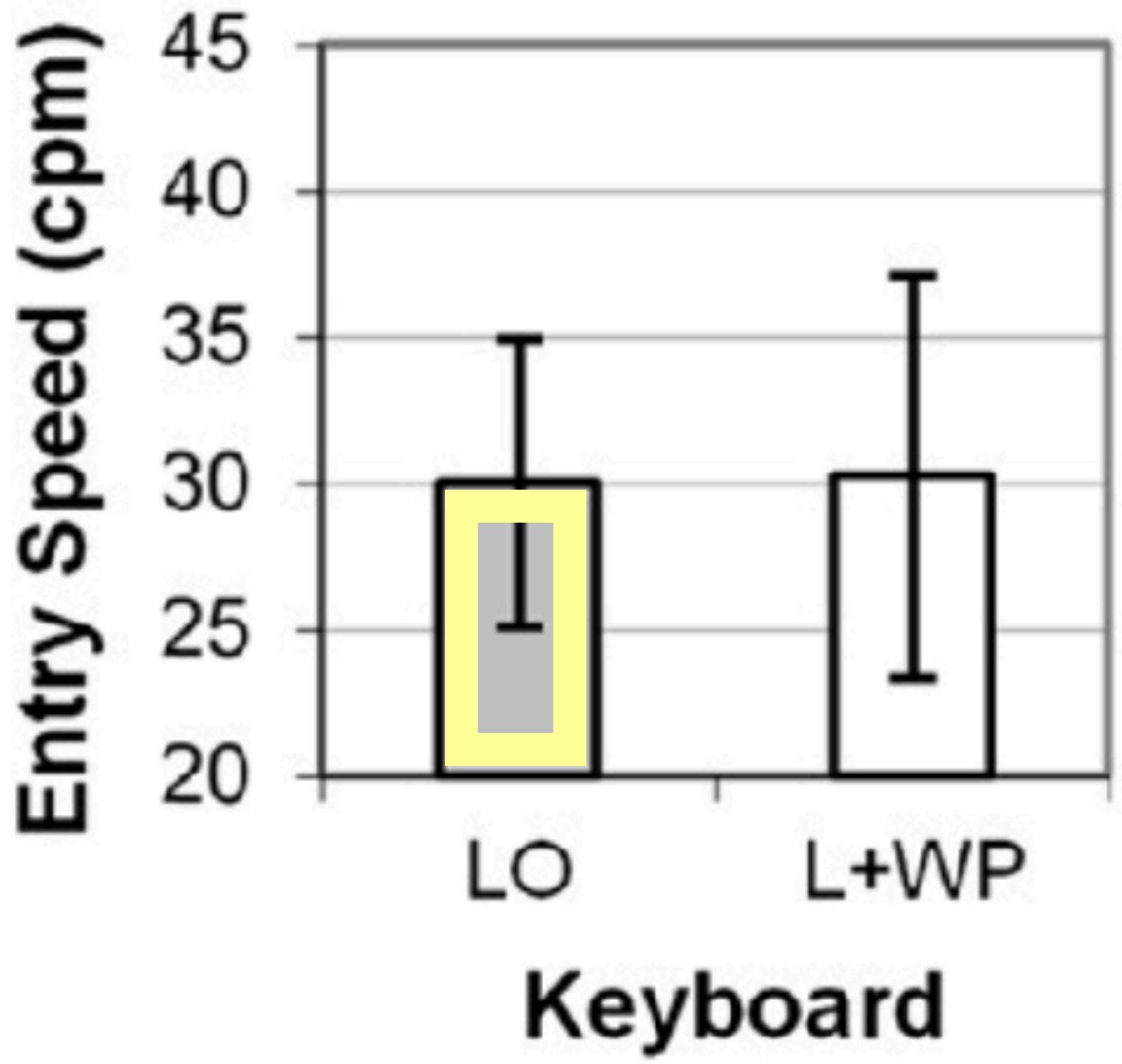


FIGURE 5.14

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

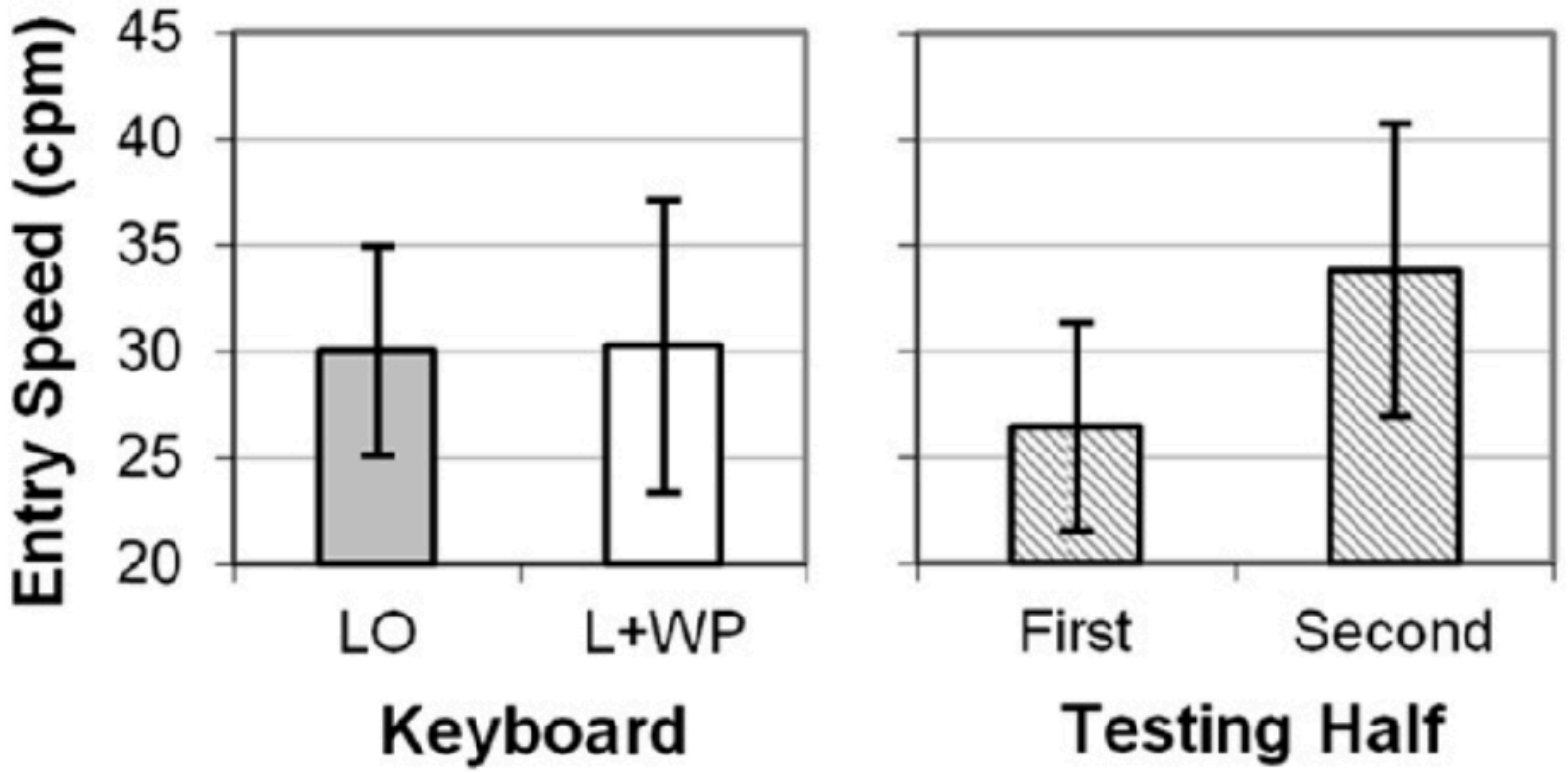


FIGURE 5.14

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!

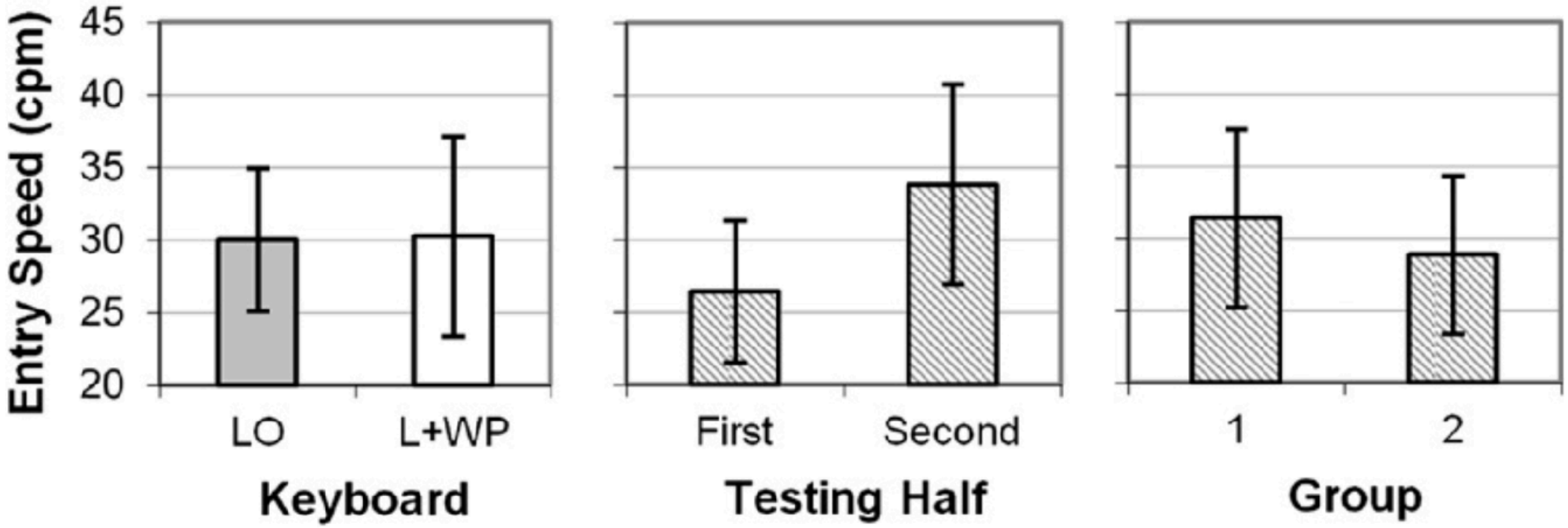


FIGURE 5.14

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

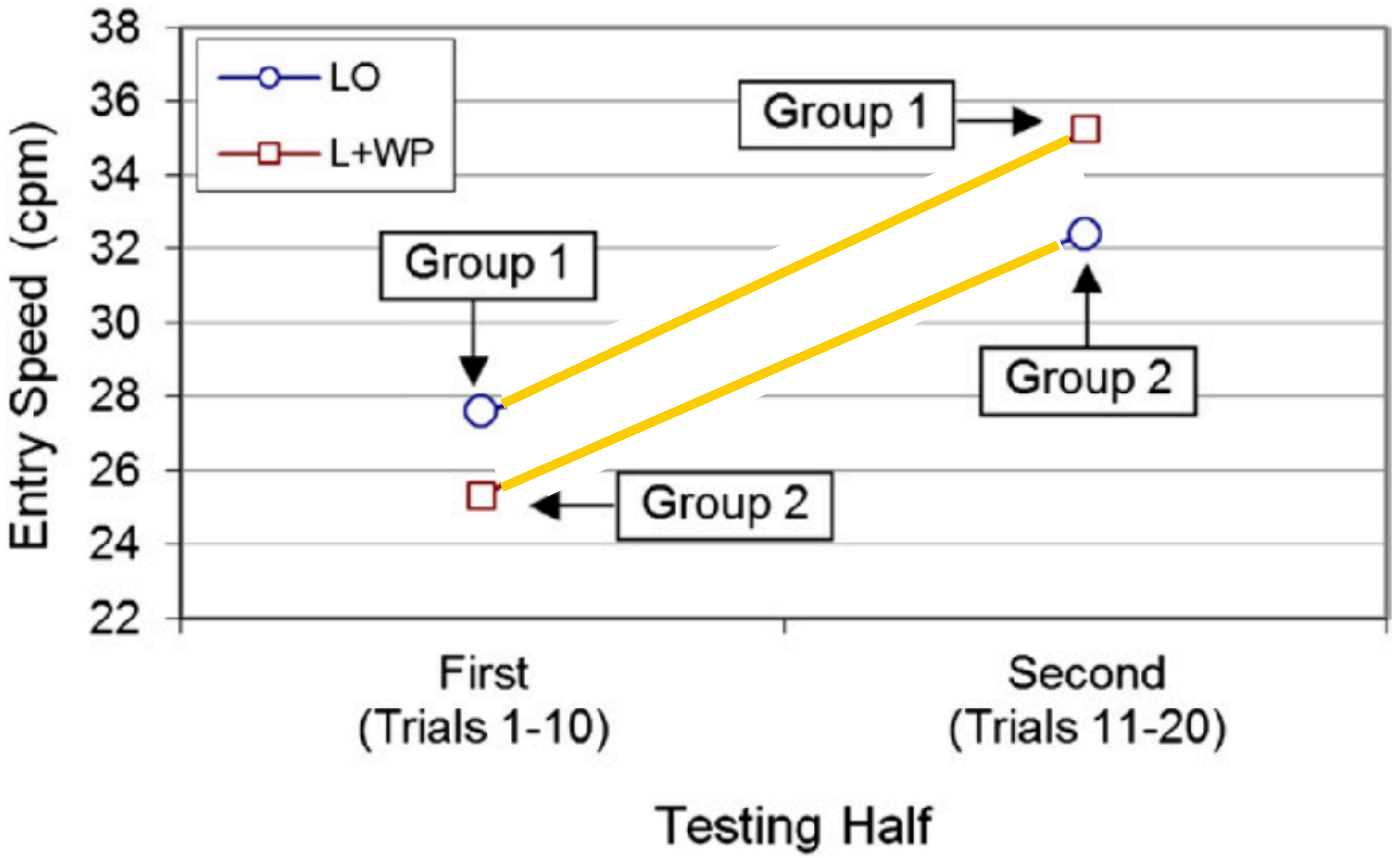


FIGURE 5.15

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

Example (continued)

Harder to start with the more complex keyboard

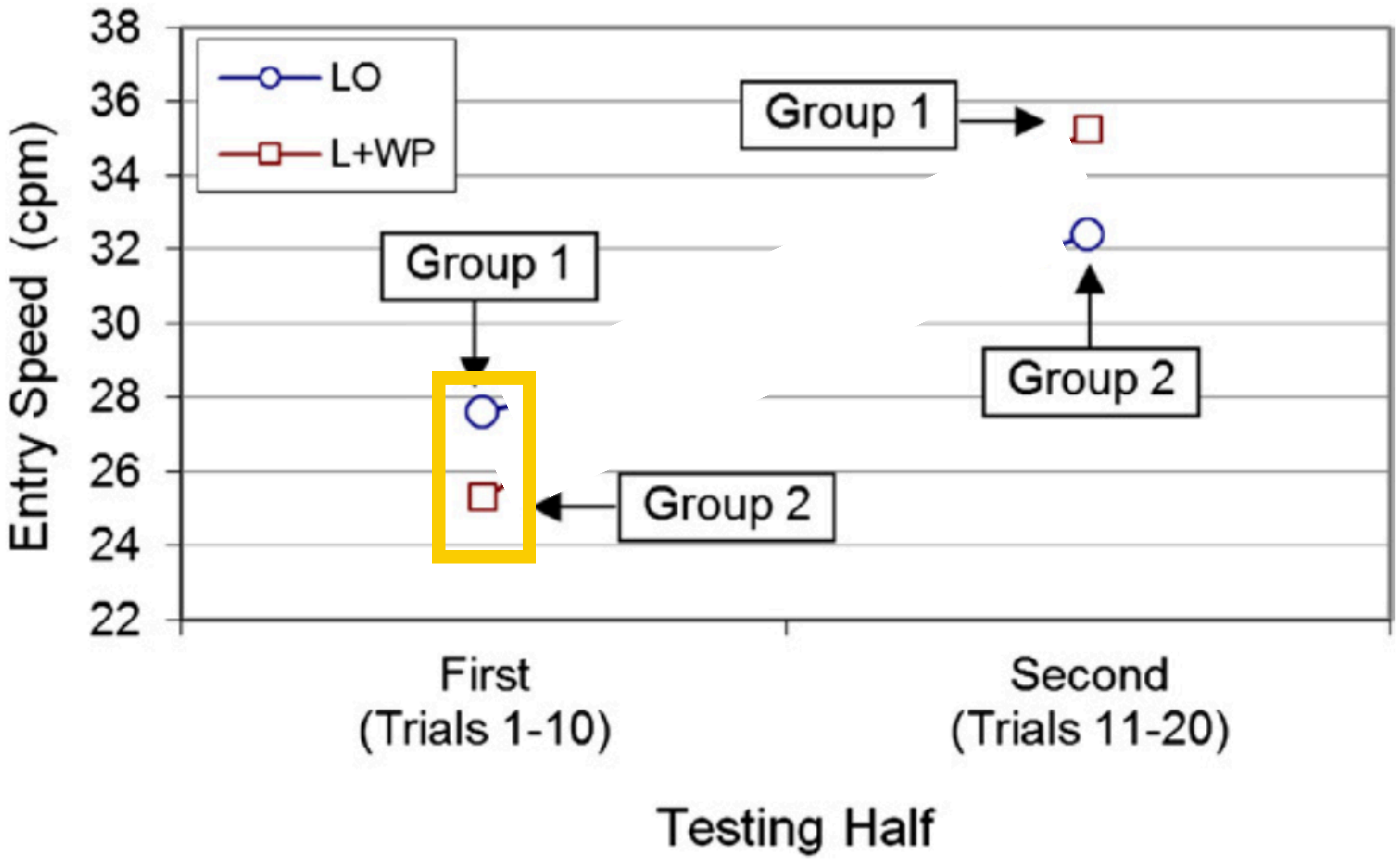


FIGURE 5.15

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

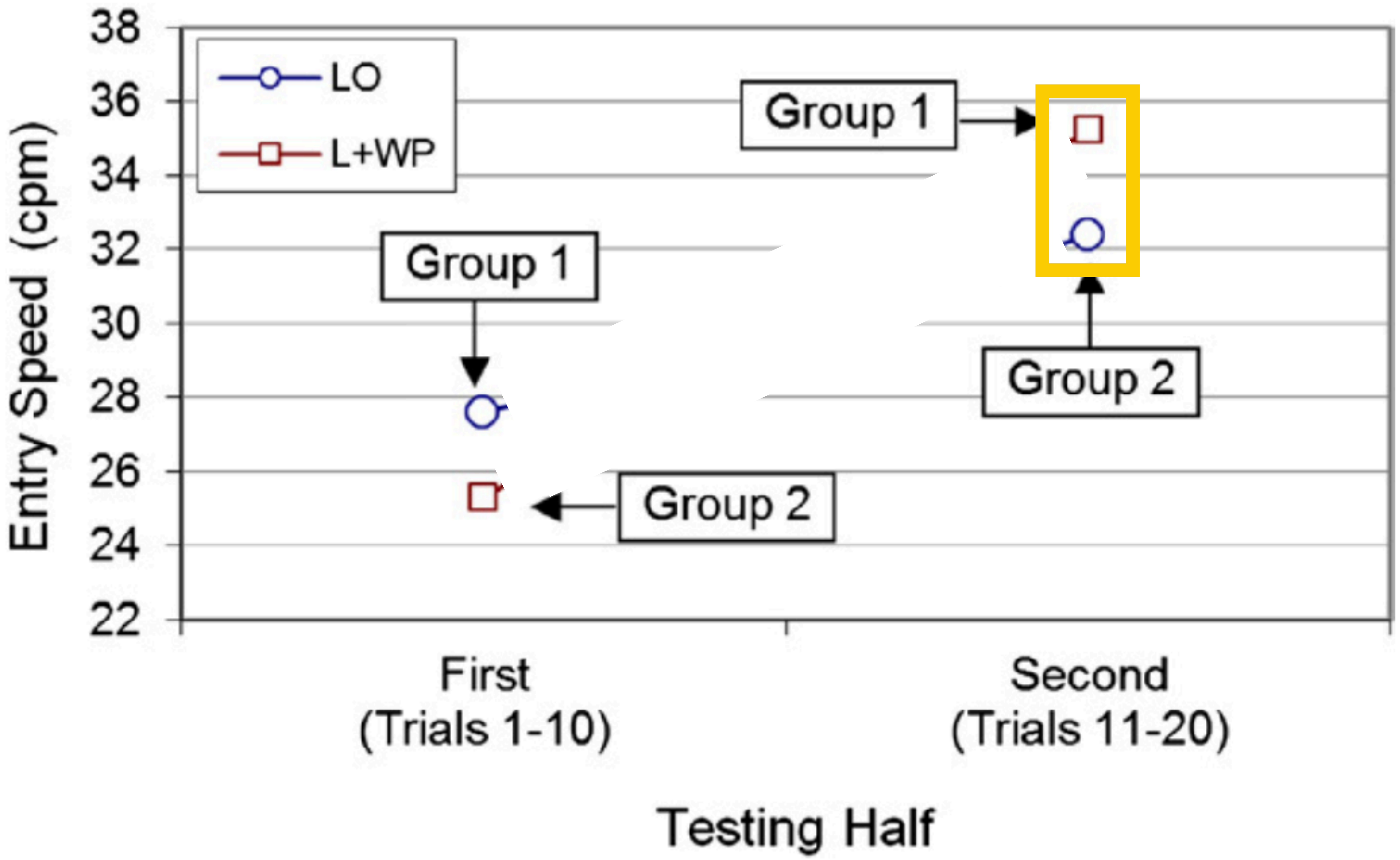


FIGURE 5.15

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

Example (continued)

Asymmetric skill transfer!

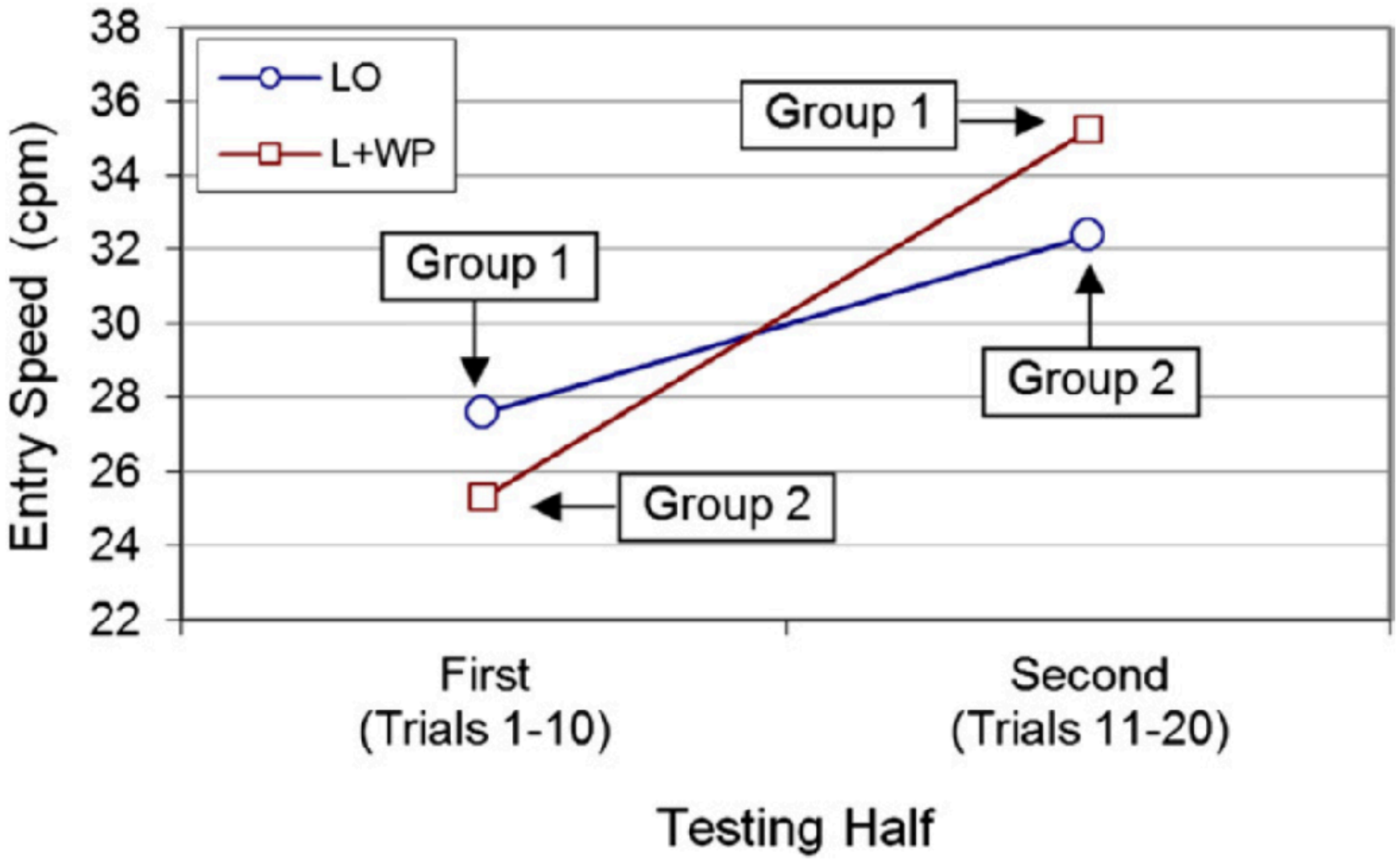


FIGURE 5.15

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

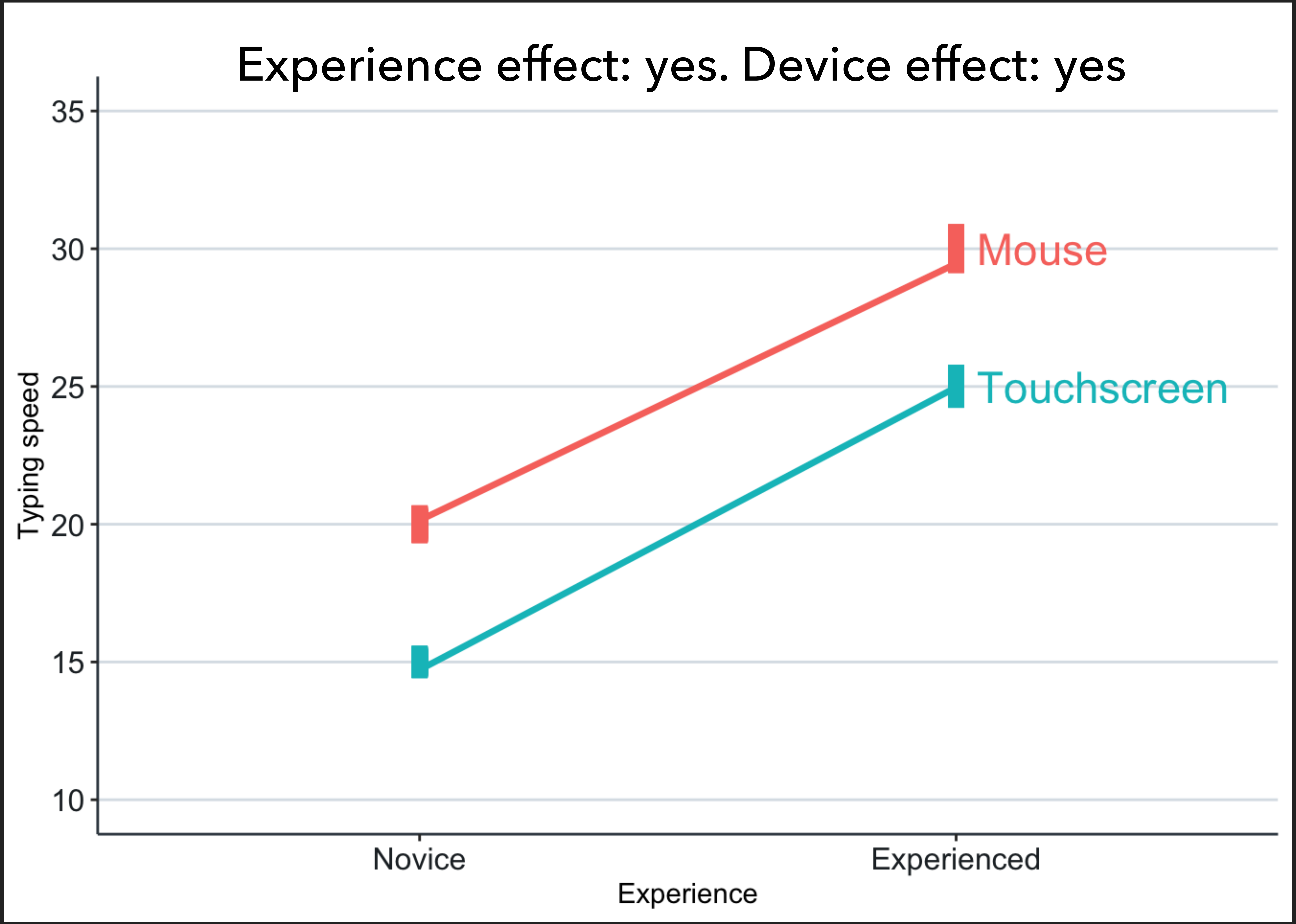
R	O	X_A	O
R	O	X_B	O

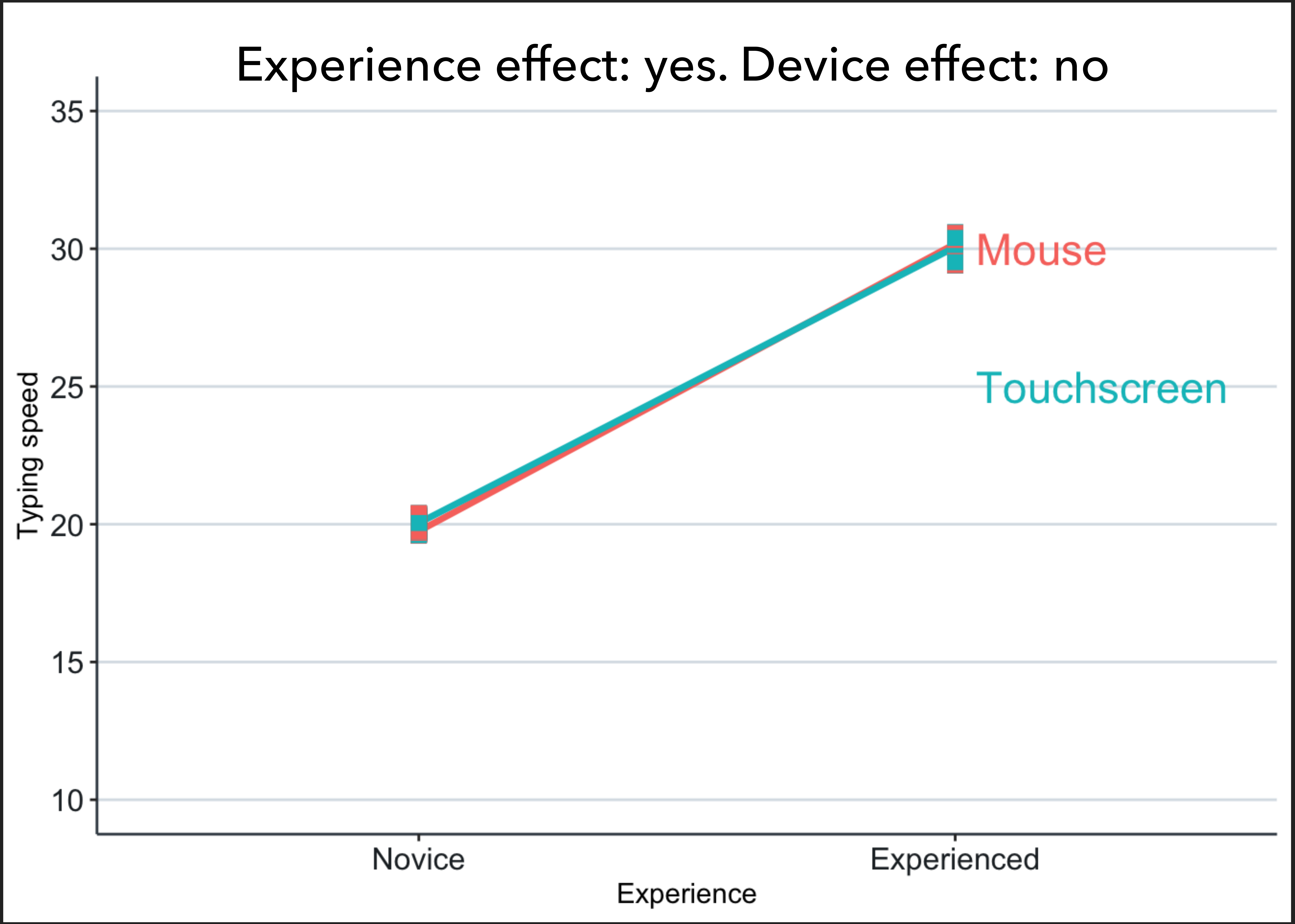
Factorial

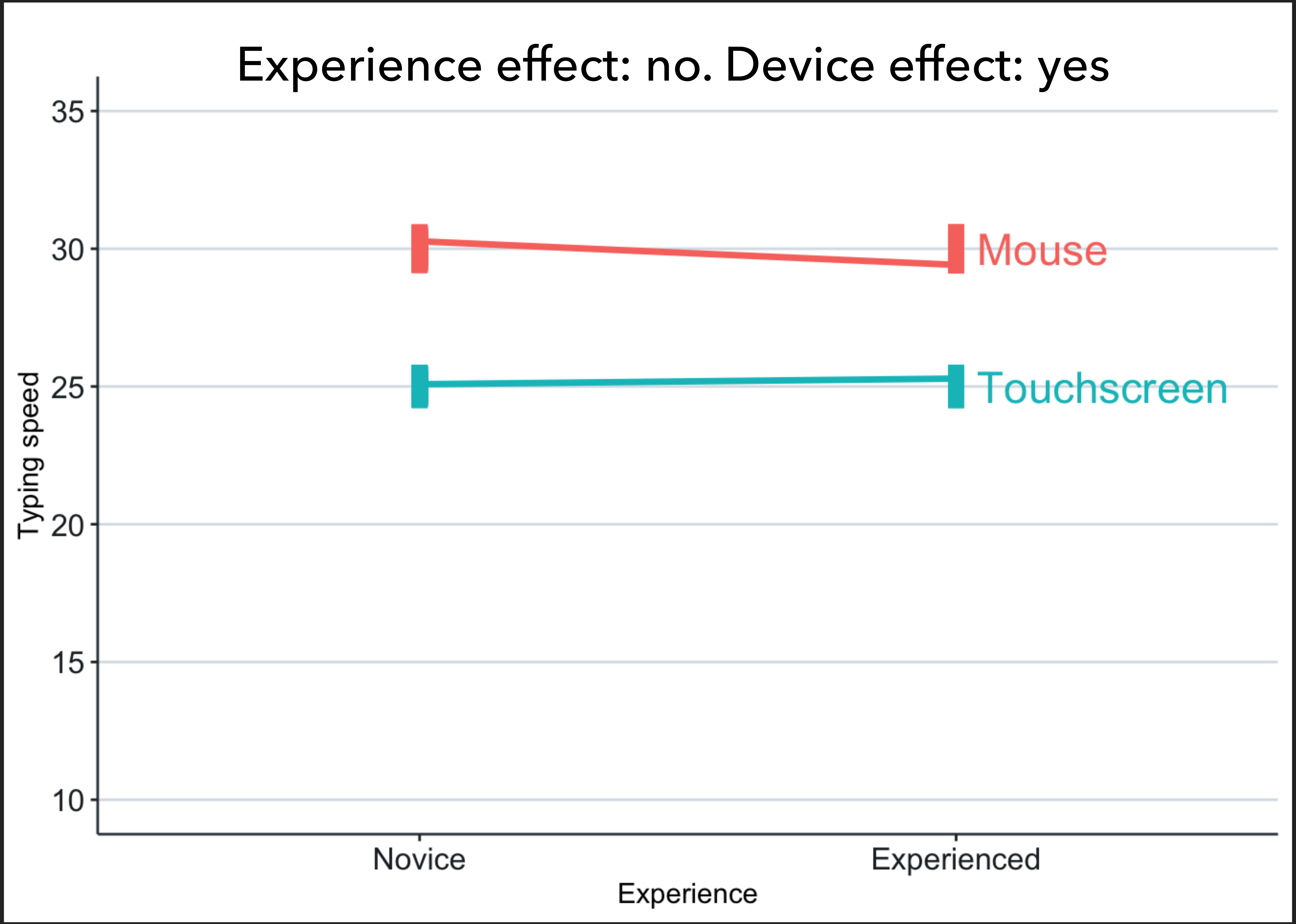
R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

- 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)

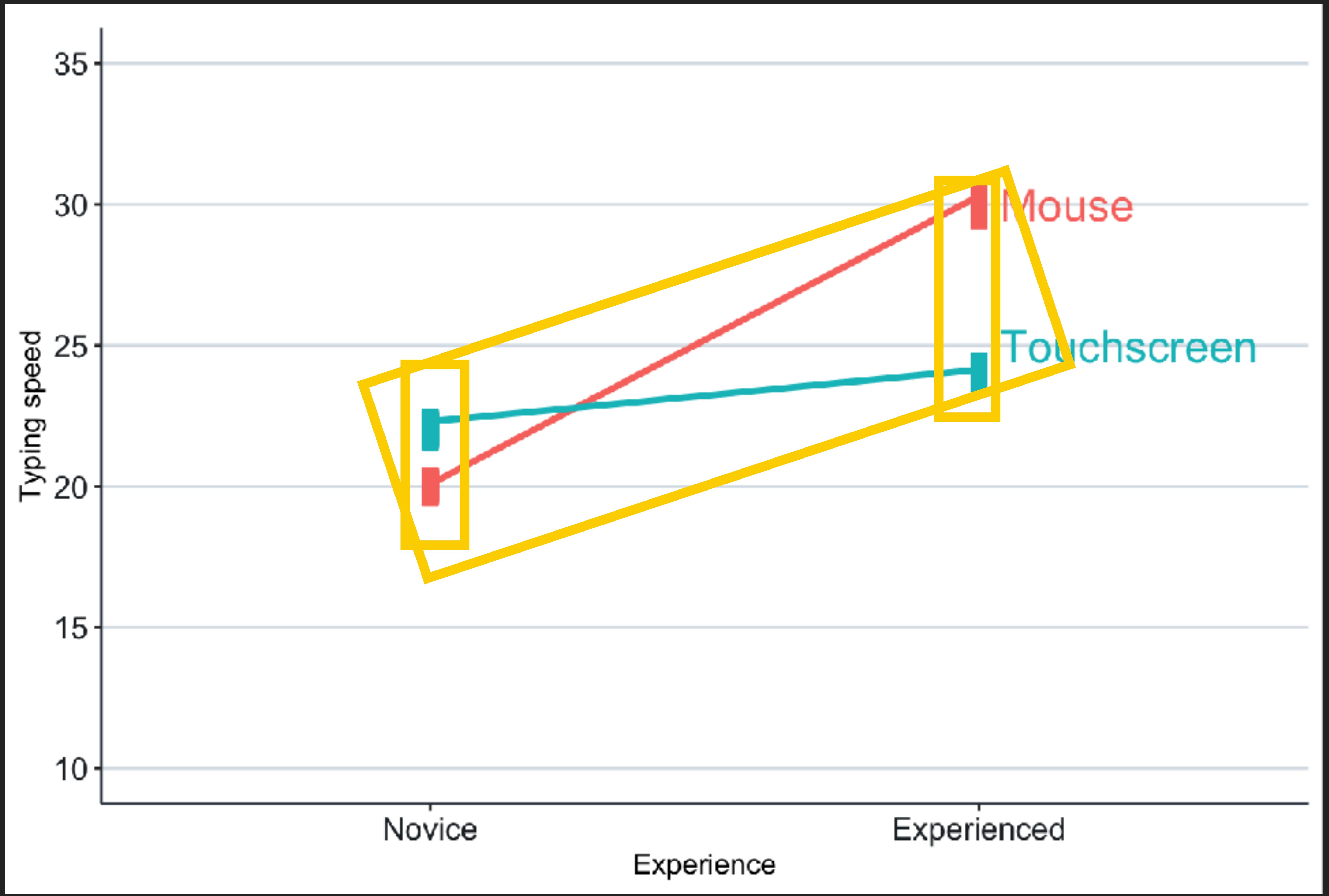

Experience effect: yes. Device effect: yes
35
30
25
20
15
10
Typing speed
Mouse
Touchscreen
Novice
Experienced
Experience

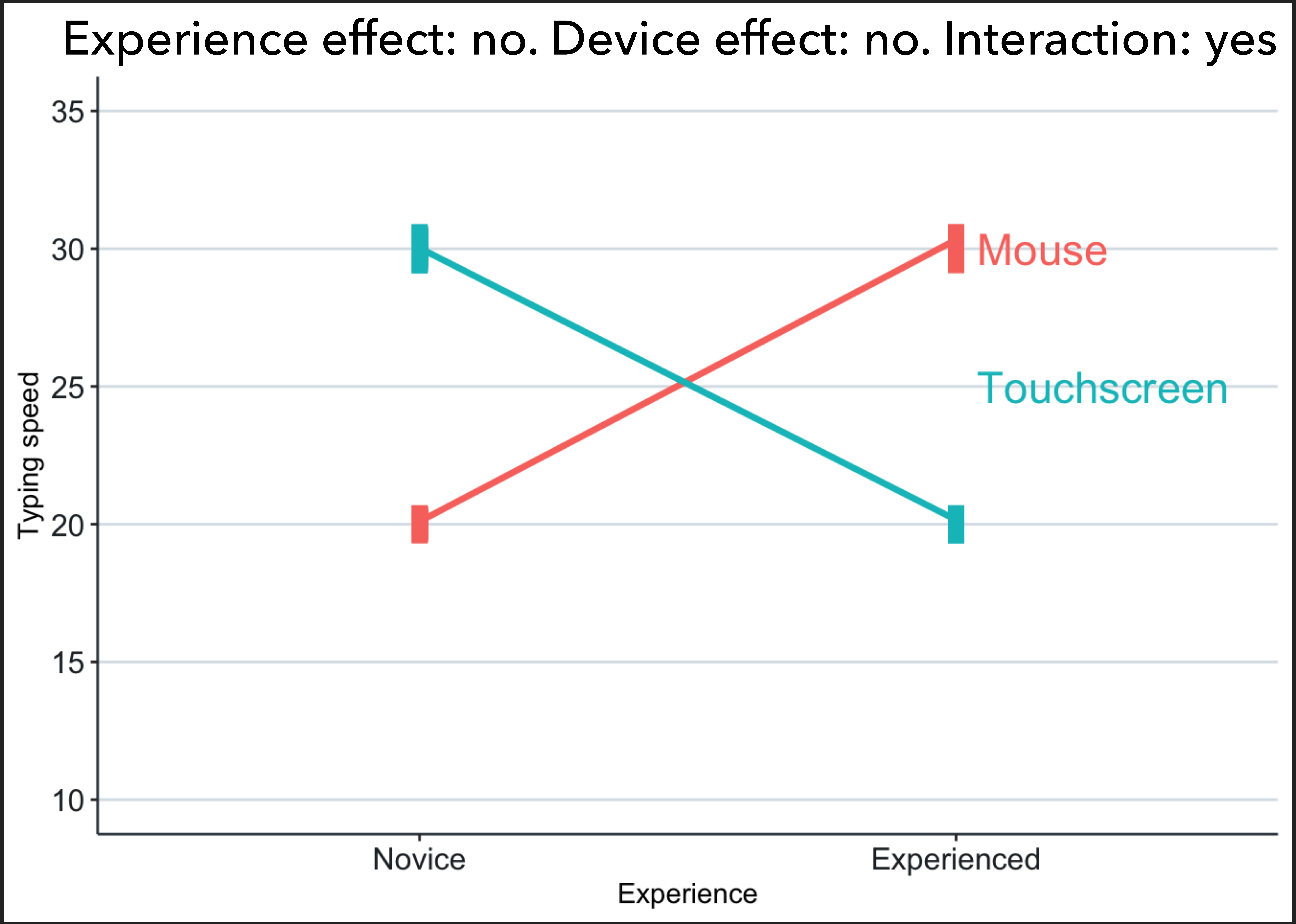

Experience effect: yes. Device effect: no
35
30
25
20
15
10
Typing speed
Mouse
Touchscreen
Novice
Experienced
Experience


Experience effect: no. Device effect: yes
Typing speed
35
30
25
20
15
10
Mouse
Touchscreen
Novice
Experienced
Experience

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
Typing speed
35
30
25
20
15
10
Mouse
Touchscreen
Novice
Experienced
Experience

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

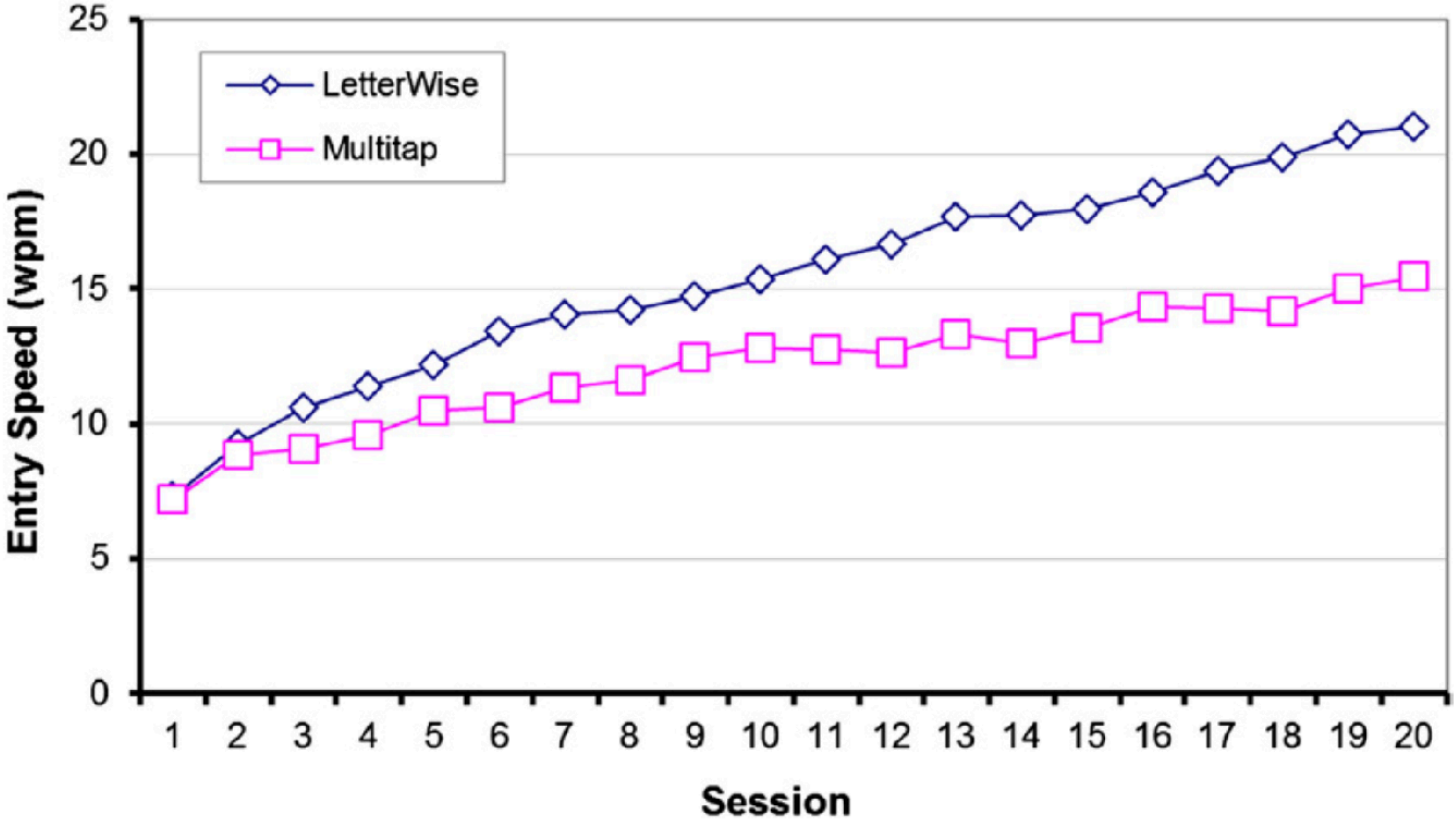


FIGURE 5.16

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.

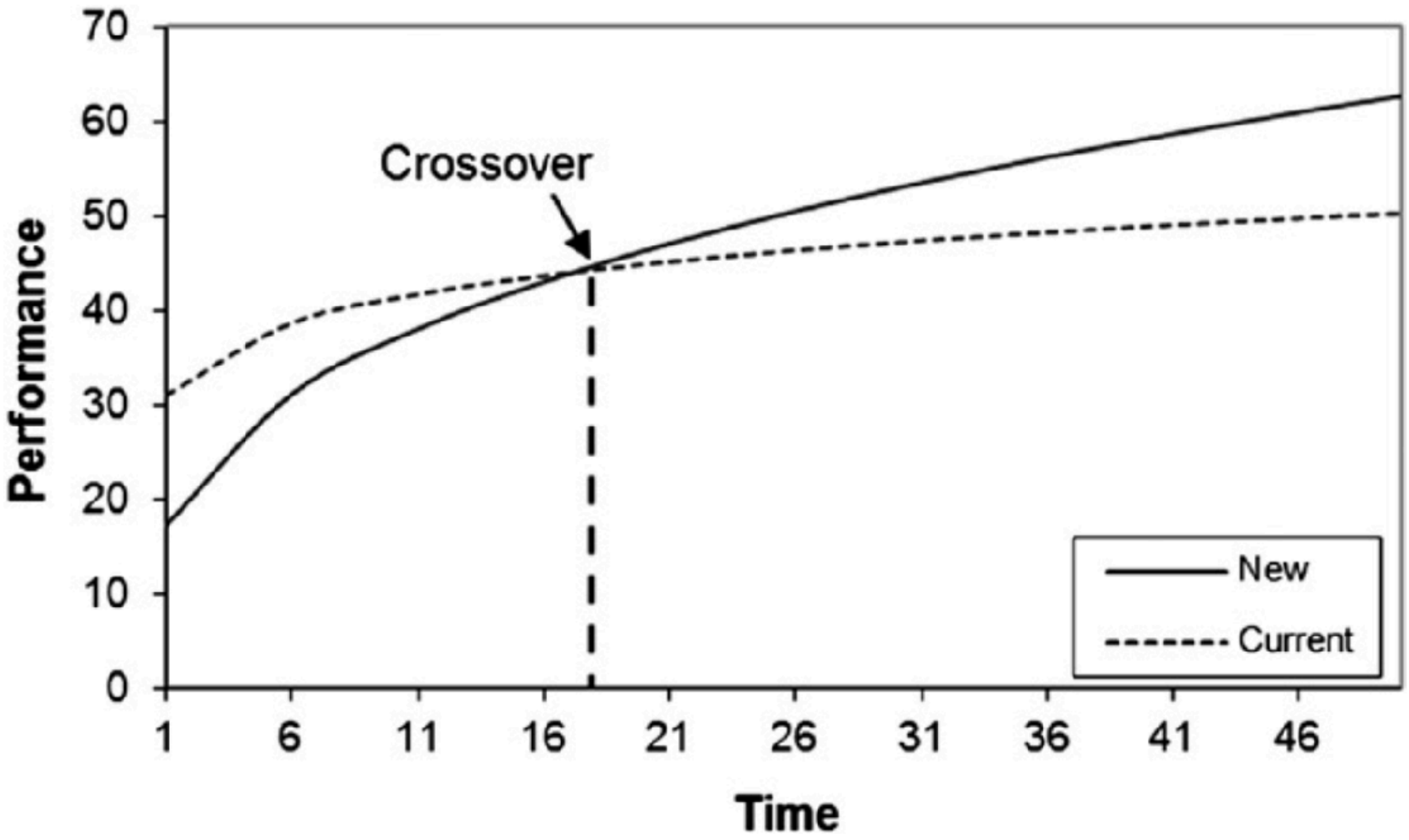


FIGURE 5.17

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

Example paper presentations

WSDM (Conference on Web Search and Data Mining) Experiment

- Setup
  - Four committee members reviewed each paper
  - Two single blind, two double blind
- Results
  - "Reviewers in the single-blind condition [...] preferentially bid for papers from top universities and companies."
  - "Single-blind reviewers are significantly more likely than their double-blind counterparts to recommend for acceptance papers from famous authors [odds multiplier 1.64], top universities [1.58], and top companies [2.10]."

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. Proceedings of the National Academy of Sciences, 114(48), 12708-12713.

NeurIPS (Conference on Neural Information Processing Systems) Experiment

- Setup
  - Organizers split the program committee down the middle
  - Most submitted papers were assigned to a single side
  - 10% of submissions (166) were reviewed by both halves of the committee

- Results
  - "most papers [57%] at NeurIPS would be rejected if one reran the conference review process (with a 95% confidence interval of 40-75%)"

http://blog.mrtz.org/2014/12/15/the-nips-experiment.html

Statistical Conclusion Validity

Hypothesis Tests

- Aka "significance tests"
- Purpose:
  - Could random chance be responsible for an observed effect?
- Null hypothesis (H_0):
  - The hypothesis that chance is to blame.
  - e.g., "There is no difference in the mean time to complete a task using NL2Code vs. writing code from scratch."
- Alternative hypothesis (H_a):
  - Counterpoint to the null (what you hope to prove).
  - e.g., "It takes less time on average to complete a task using NL2Code rather than by writing code from scratch."

Aside: Why Do We Need a Hypothesis? Why Not Just Look at the Outcome of the Experiment and Go With Whichever Treatment Does Better?

- Experiment: invent a series of 50 coin flips.
  - Write down a series of random 1s and 0s: [1, 0, 1, 0, 1, 0, …]

Aside: How Do You Interpret the P-Value?

- H_0: “There is no difference in the mean time to complete a task using NL2Code vs. writing code from scratch.”
- H_a: “It takes less time on average to complete a task using NL2Code rather than writing code from scratch.”

- You run some statistical test (e.g., t-test) and obtain a p-value.

Aside: P-Value Controversy

- What we would like the p-value to convey:
  - (We hope for a low value, so we can conclude that we've proved something.)

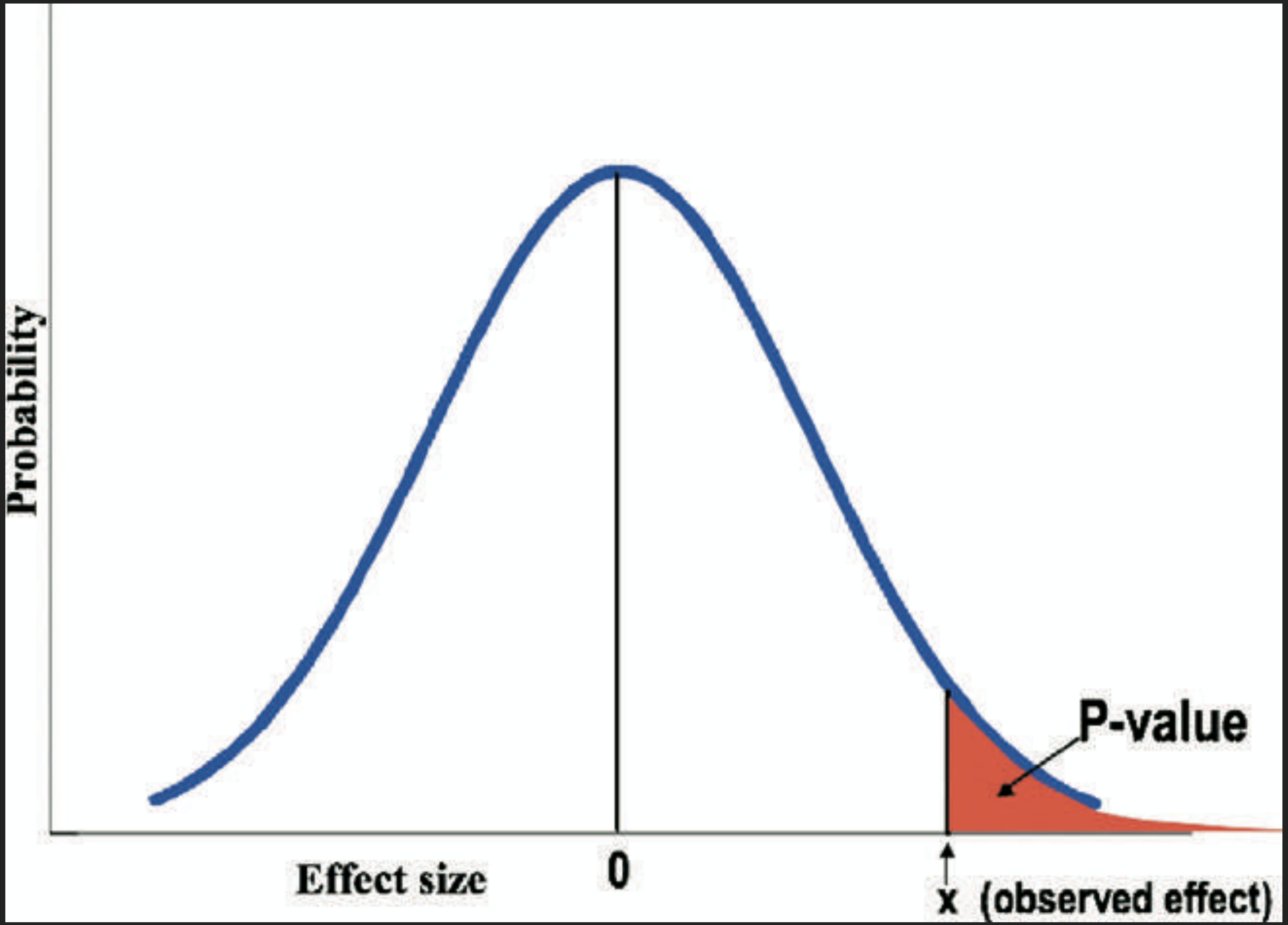
The probability that the result is due to chance: $P(H_0|D)$

- What the p-value actually represents:

The probability that, given a chance model, results as extreme as the observed results could occur: $P(D|H_0)$

Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1105-1114).

The P Value Is the Probability of the Observed Outcome (X) Plus all "More Extreme" Outcomes



Graphical depiction of the definition of a (one-sided) P value. The curve represents the probability of every observed outcome under the null hypothesis.

The P Value Is the Probability of the Observed Outcome (X) Plus all "More Extreme" Outcomes

- Not the probability that the null hypothesis is true!
- Example: Is a coin fair or not?
  - H_0: The coin is fair: P(Heads) = P(Tails) = 1/2
  - H_a: The coin is biased: P(Heads) ≠ 1/2

Consider Four Consecutive Coin Flips:

Probability

- First toss: 

?

Consider Four Consecutive Coin Flips:

Probability

- First toss: 

0.5

- Second toss: 

?

Consider Four Consecutive Coin Flips:

Probability

- First toss: 

0.5

- Second toss: 

0.25

- Third toss: 

0.125

- Fourth toss: 

0.0625

Is Coin Fair?

- Two-sided P = 0.125.

0.0625

0.0625

- This does not mean that the probability of the coin being fair is only 12.5%!

Aside: P-Value Controversy

- What we would like the p-value to convey:
  - (We hope for a low value, so we can conclude that we've proved something.)

The probability that the result is due to chance: $P(H_0|D)$

- What the p-value actually represents:

The probability that, given a chance model, results as extreme as the observed results could occur: $P(D|H_0)$

Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI.
In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1105-1114).

Is Coin Fair?

- Two-sided P = 0.125.

0.0625

0.0625

- This does not mean that the probability of the coin being fair is only 12.5%!

$$P(H_0|D) = \frac{P(D|H_0)\, P(H_0)}{P(D)}$$

Common false belief that the probability of a conclusion being in error can be calculated from the data in a single experiment without reference to external evidence or the plausibility of the underlying mechanism.

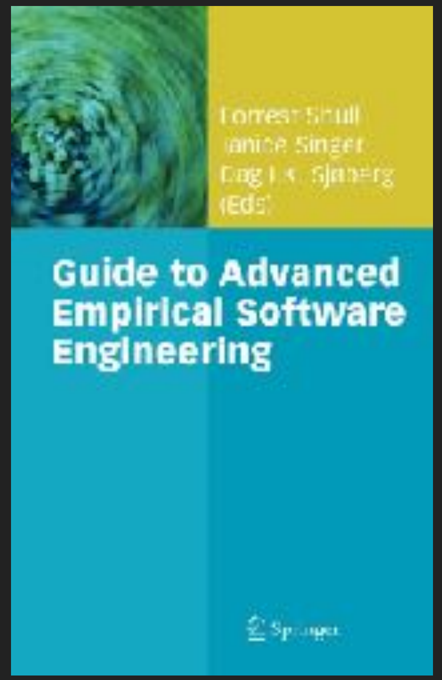
. . . to be continued

Credits

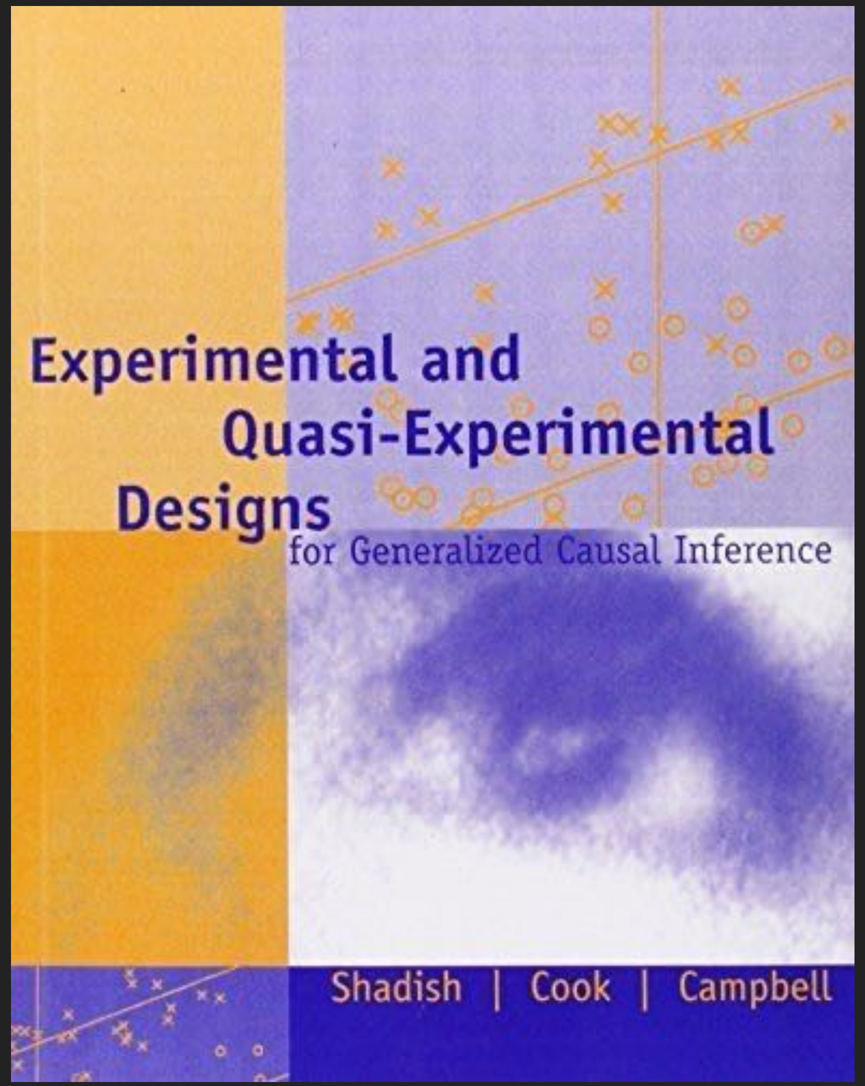
- Graphics: Dave DiCello photography (cover)
- Chapters from Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Publishing
  - Ch1: Experiments and generalized causal inference
  - Ch2: Statistical conclusion 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.
- Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. In Seminars in Hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.
- Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Research methods in human-computer interaction. Morgan Kaufmann.
  - Ch 3: Experimental design
  - Ch 4: Statistical analysis
- MacKenzie, I. S. (2012). Human-computer interaction: An empirical research perspective.
  - Ch 6: Hypothesis testing
- Robertson, J., & Kaptein, M. (Eds.). (2016). Modern statistical methods for HCI. Cham: Springer.
  - Ch 5: Effect sizes and power analysis
  - Ch 13: Fair statistical communication
  - Ch 14: Improving statistical practice
- Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1105-1114).

Read

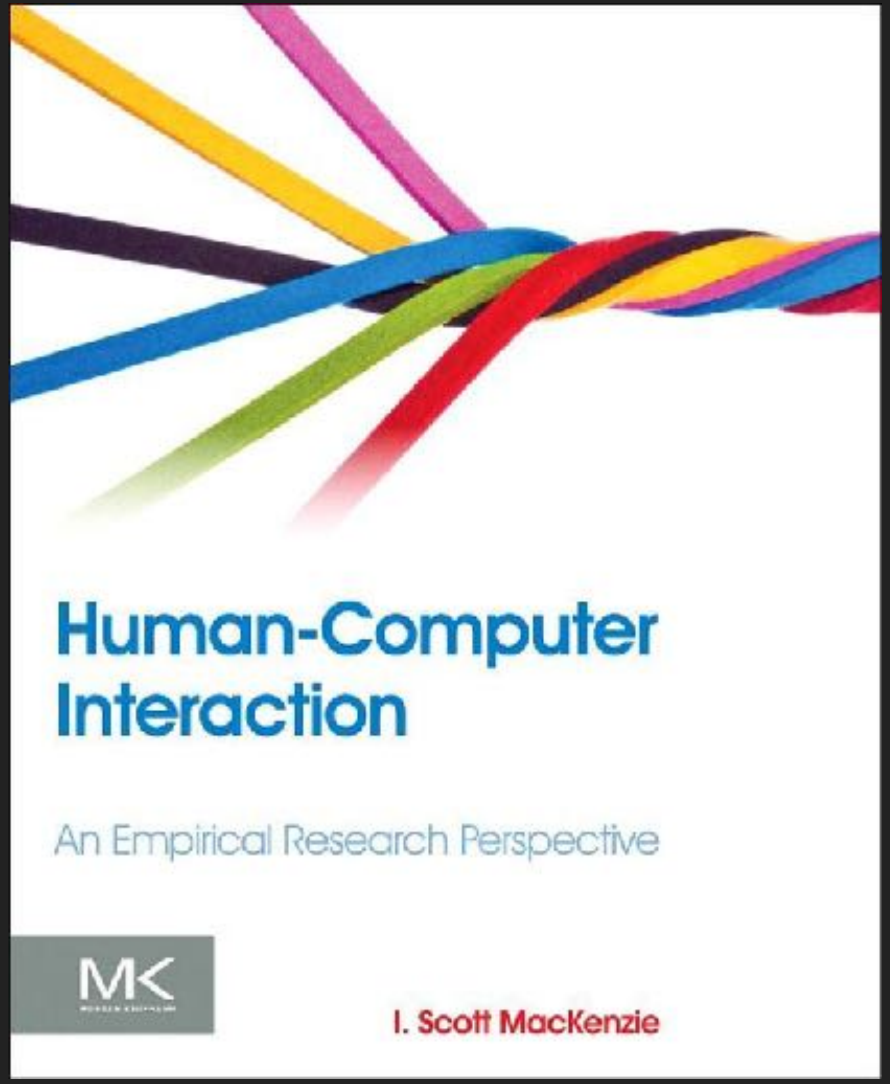
Ch 10 (Analysis and interpretation)



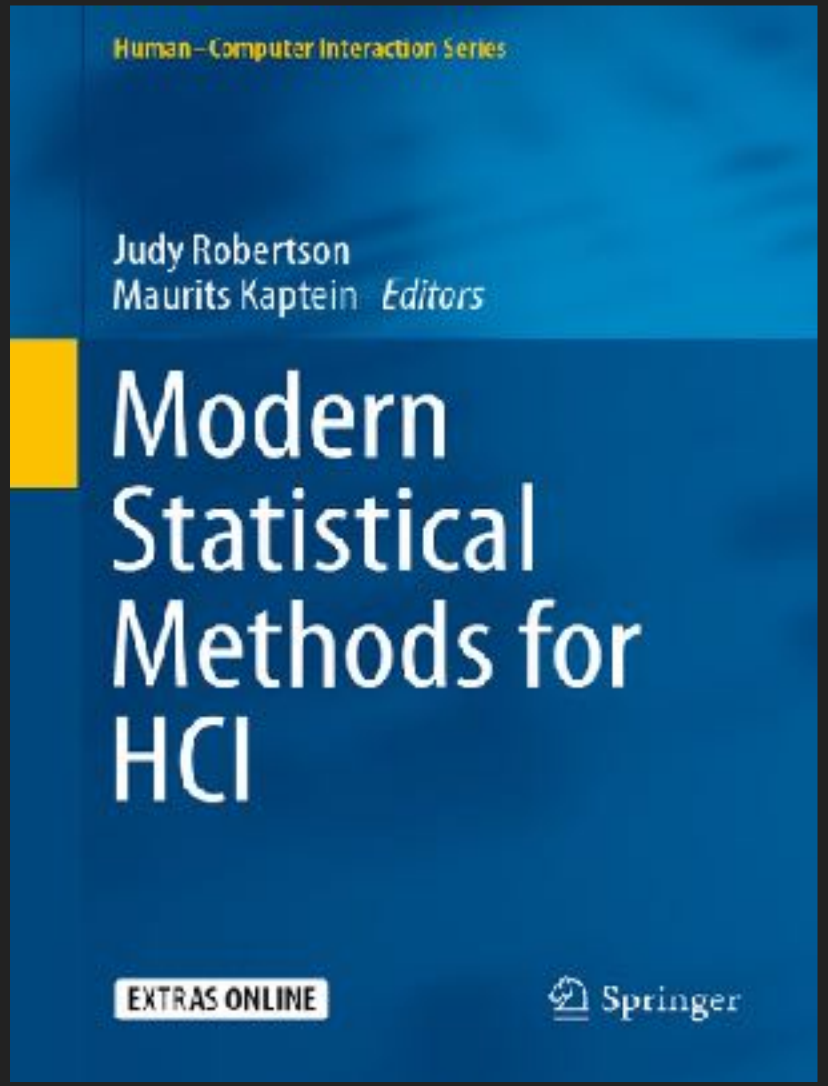
Ch 6 (Statistical methods and measurement)



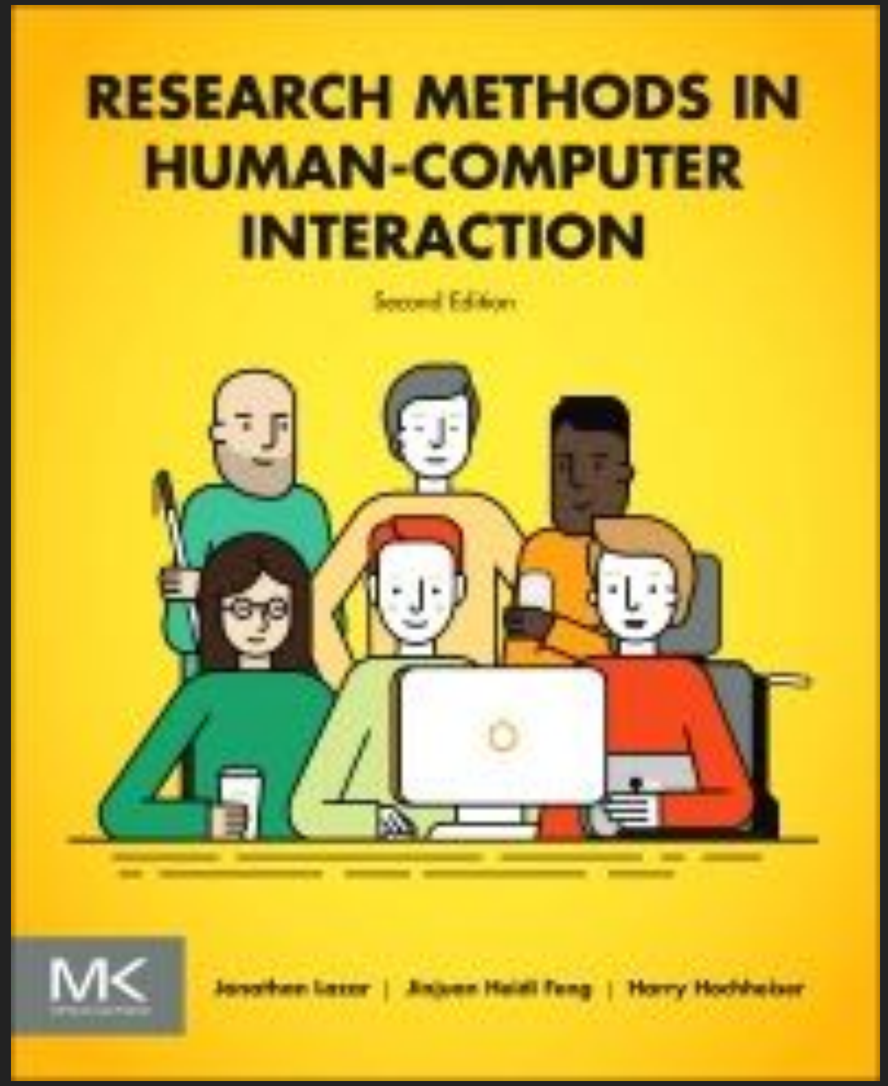
Ch 1 (Experiments and causality)
Ch 2 & 3 (Validity)
Ch 8 (Randomized experiments)



Ch 6 (Hypothesis testing)



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



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