Photo credit: Dave DiCello

17-803 Empirical Methods

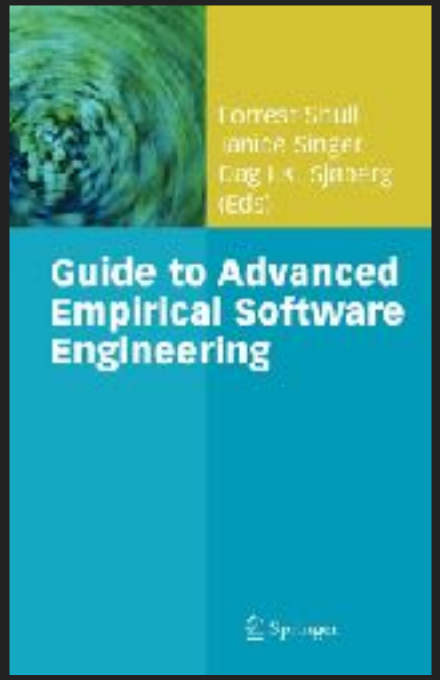
Bogdan Vasilescu, S3D

Designing Experiments (II)

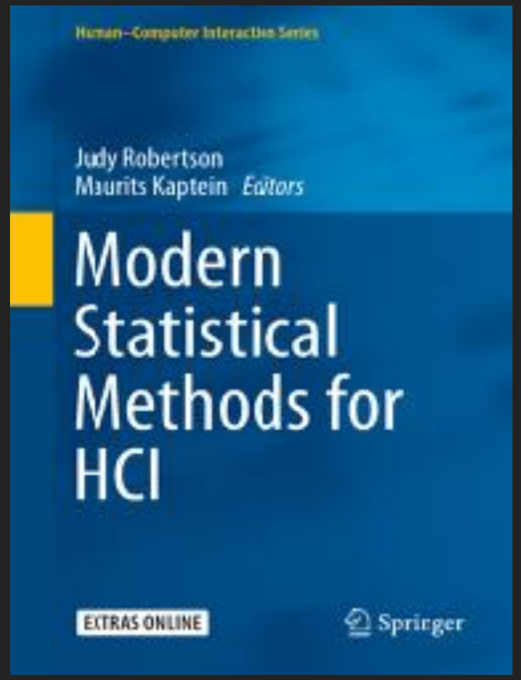
Tuesday, March 12, 2024

Readings

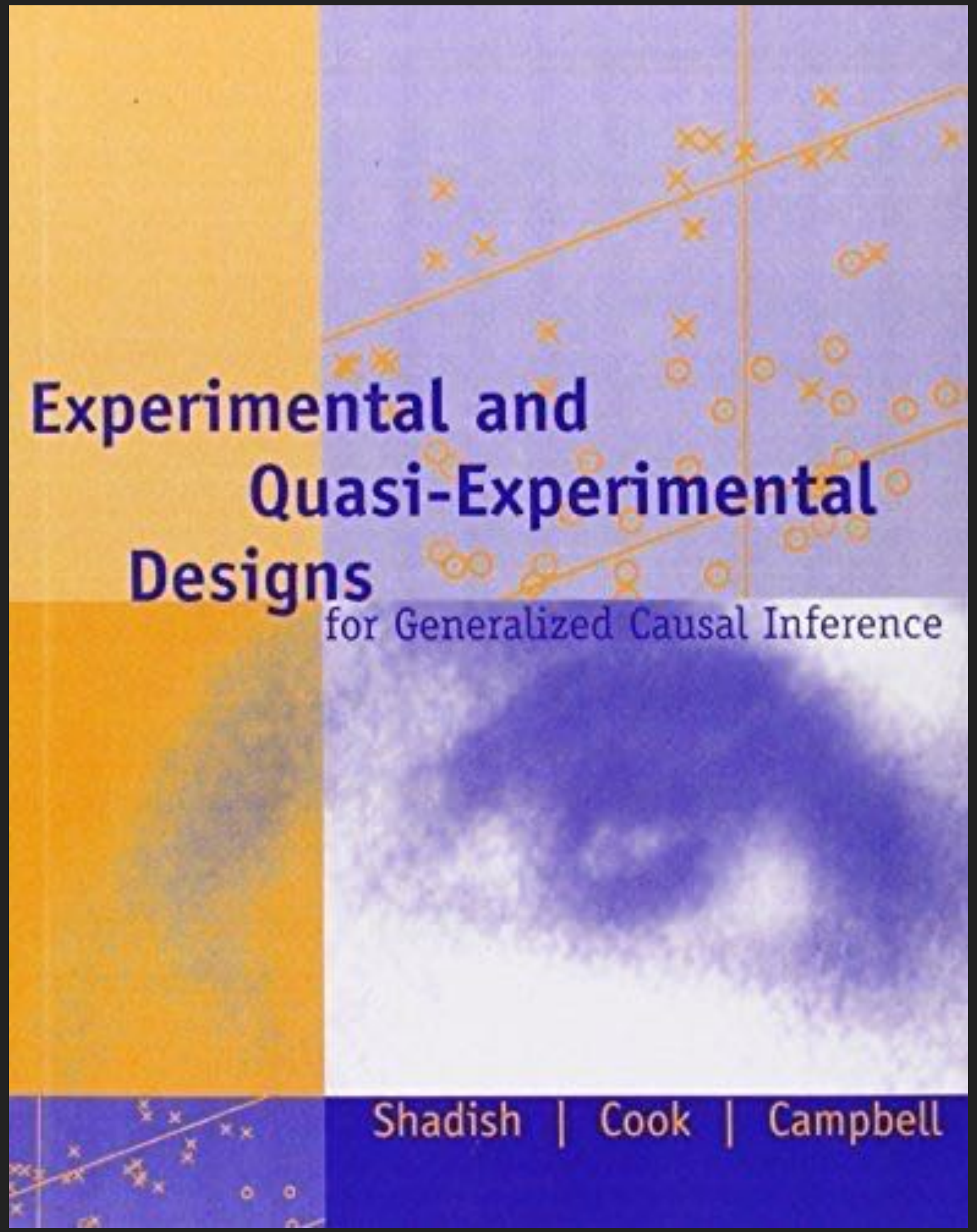
Ch 10 (Analysis and interpretation)



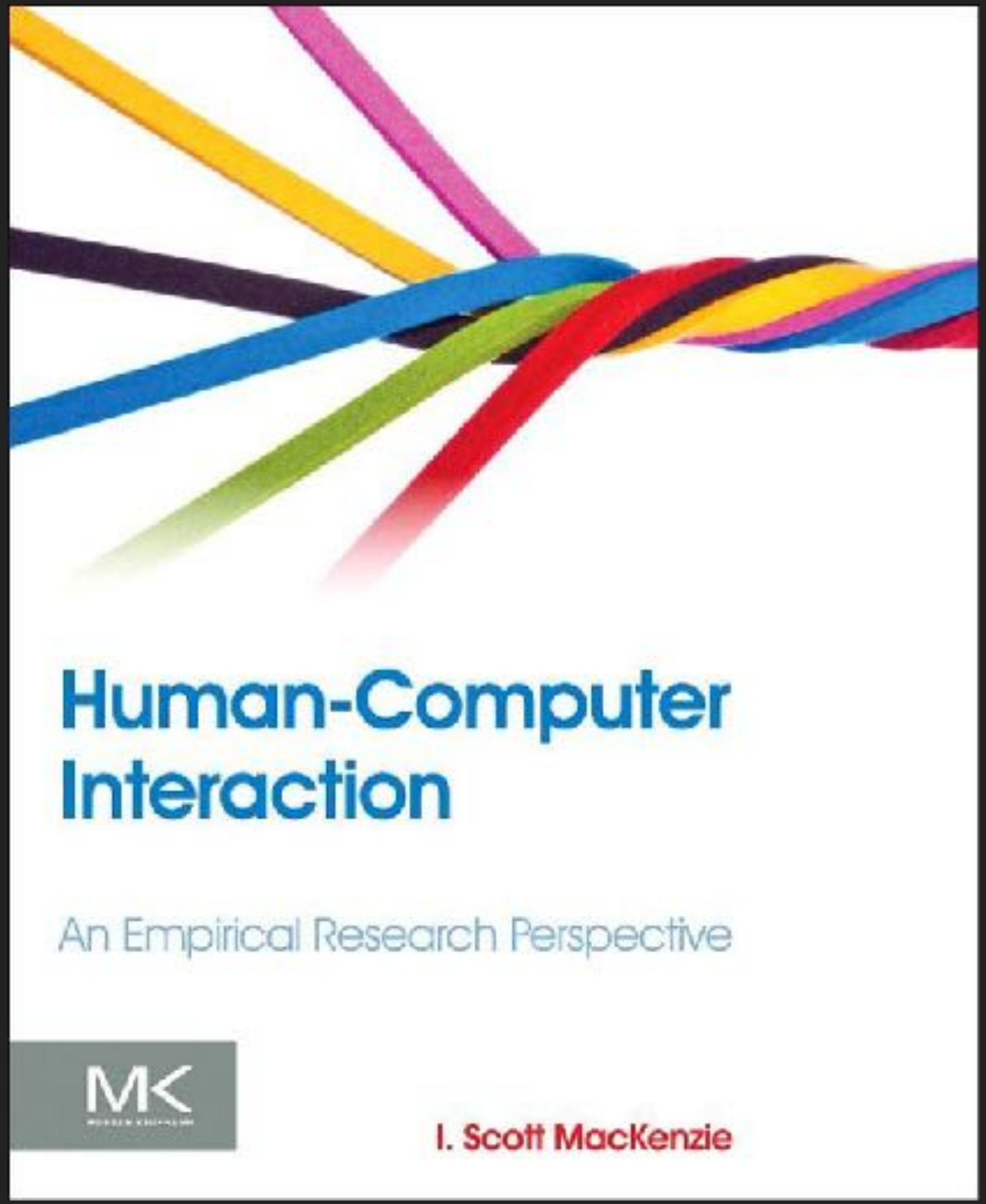
Ch 6 (Statistical methods and measurement)



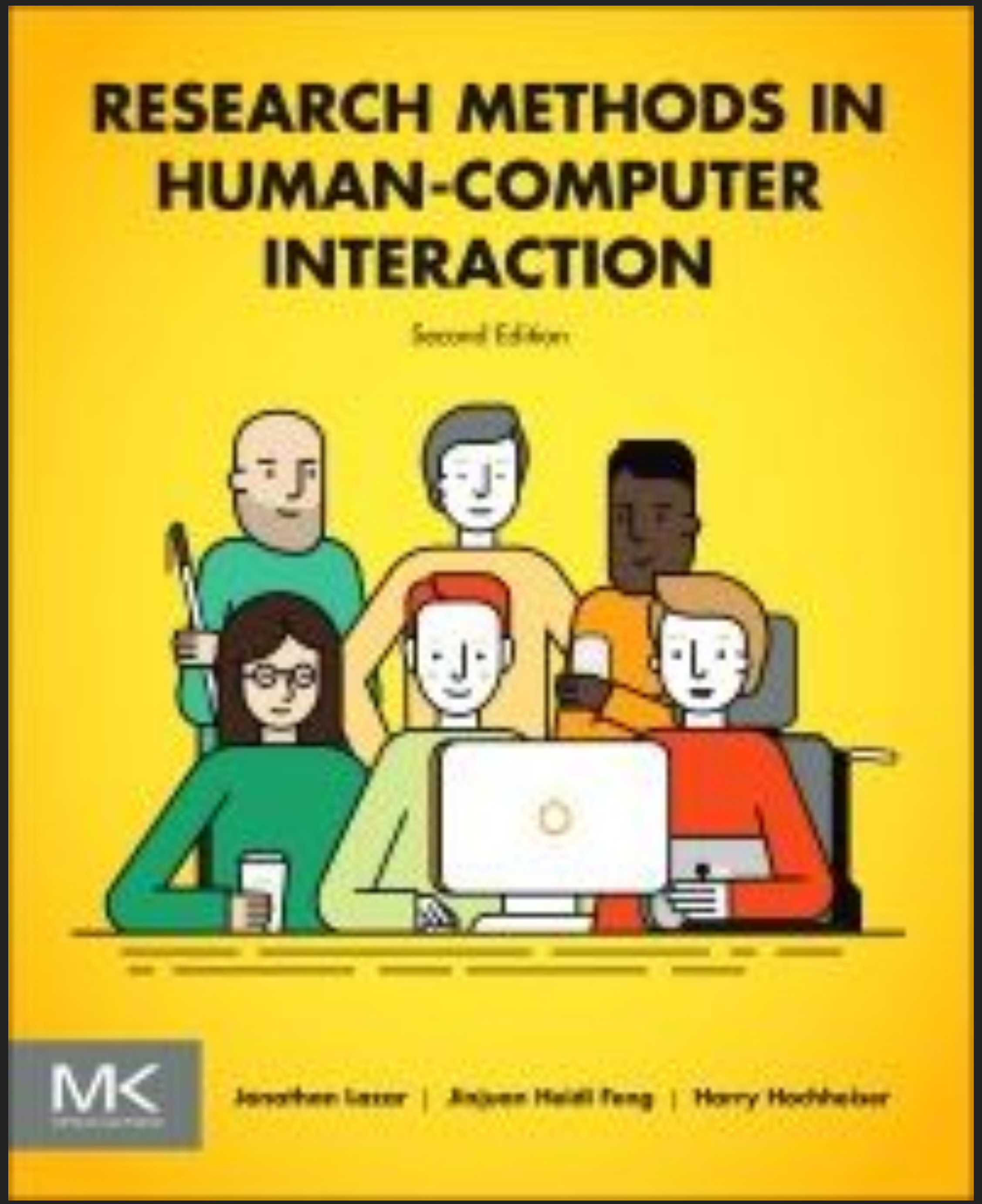
Ch 5 (Effect sizes and power analysis)
Ch 13 (Fair statistical communication)
Ch 14 (Improving statistical practice)



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)

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.

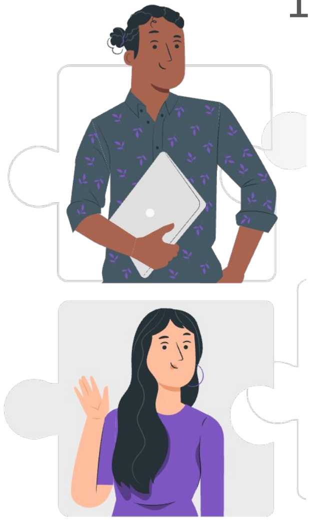
Reviewer bias in single-versus double-blind peer review

By Tomkins, A., Zhang, M., & Heavlin, W. D. (2017)

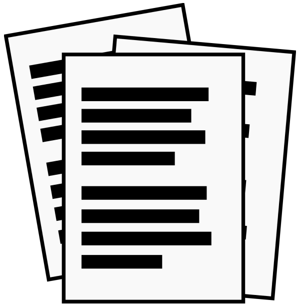
Presentation for course Empirical Methods'24 by Catarina Gamboa

Controlled Experiment

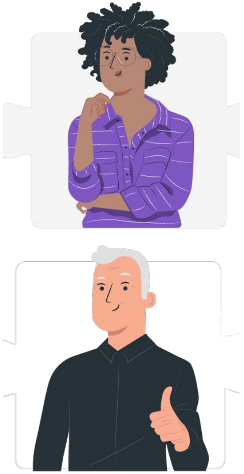
10th ACM International Conference WSDM,
a venue with a 15.6% acceptance rate



Single
Blind

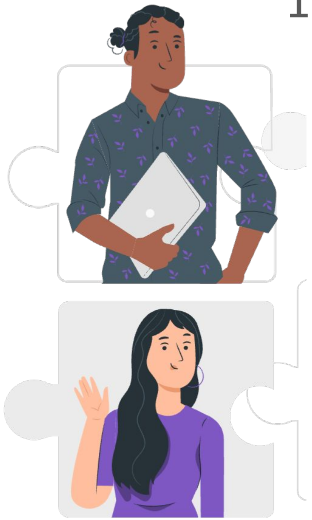


Double
Blind



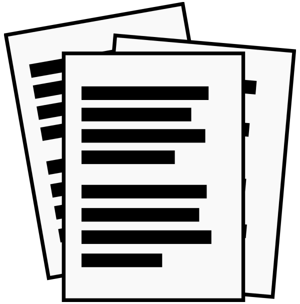
Controlled Experiment

10th ACM International Conference WSDM,
a venue with a 15.6% acceptance rate

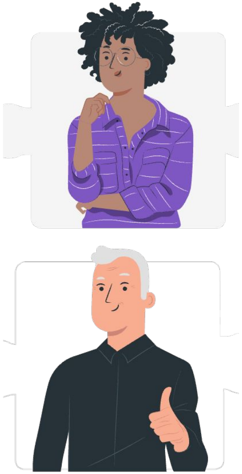


974 pool

Single Blind



Double Blind



983 pool

500 papers

Bidding
Reviewing
Score + Ranking

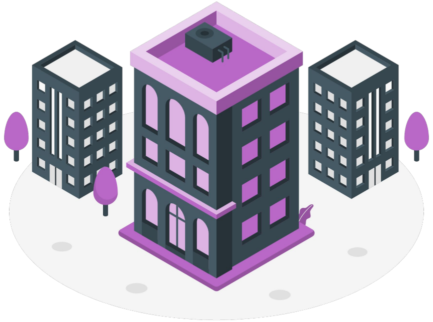
Test hypothesis based on Theories

Matilda Effect (1870)

Female authors receive lower scientific recognition

Matthew Effect (1968)

"the rich get richer, and the poor get poorer."



Institutional Fame/Quality

Controlled Experiment

Original Information

- Author’s name
- Institution
- Country

Covariants

Factor	Feature name	No. of papers	Fraction of Papers, %
Paper from United States	United States	176	35
Same country as reviewer	Same	146	29
Female author	Wom	219	44
Famous author	Fam	81	16
Academic	Aca	370	74
Top university	Uni	135	27
Top company	Com	90	18

Scores

Quality (**b**linded **p**aper **q**uality **s**core): average quality score of the double-blind reviews for that paper

Analysis & Results: Paper Acceptance

Logistic regression analysis to predict the odds that a single-blind reviewer would give a positive (accept) score to a paper.

Top company

Famous author

Top university

Table 2. Learned coefficients and significance for review score prediction

Name	Coefficient	SE	Confidence interval	*P* value	Odds multiplier	bpqs equivalent
Const	−1.83	0.24	[−2.31, −1.36]	0.000	0.16	—
bpqs	0.80	0.08	[0.64, 0.97]	0.000	2.23	1.00
Com	0.74	0.24	[0.27, 1.21]	0.002	2.10	0.92
Fam	0.49	0.22	[0.05, 0.93]	0.027	1.63	0.61
Uni	0.46	0.18	[0.09, 0.83]	0.012	1.58	0.57
Wom	−0.25	0.18	[−0.60, 0.10]	0.160	0.78	−0.31
Same	0.14	0.24	[−0.34, 0.62]	0.564	1.15	0.17
Aca	0.06	0.22	[−0.38, 0.51]	0.775	1.07	0.08
United States	0.01	0.21	[−0.42, 0.44]	0.964	1.01	0.01

Analysis & Results: Bidding

1. Do Single-blind and double-blind reviewers **bid for the same number of papers?**

   Statistical test - Mann-Whitney test

   Single blind bid for fewer papers ($p=0.0002$). On average there is a 22 % decrease in bidding

2. Do they also **bid differently for particular types of papers**?

   Logistic regression

   Company and University features were significant ($p=0.01$ and $p=0.011$)

Test three Bias Theories

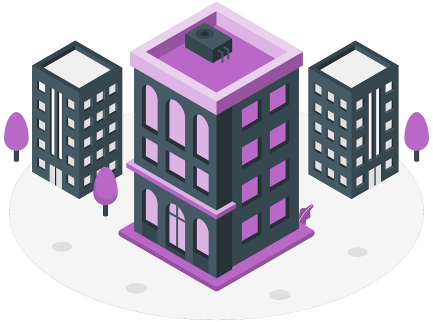
Matilda Effect (1870)

Female authors receive lower scientific recognition

Accept

Matthew Effect (1968)

"the rich get richer, and the poor get poorer."



Institutional Fame/Quality

Bid

Accept

Flaws in Experimental Design

Ian Dardik

Formal Methods Application: An Empirical Tale of Software Development

Ann E. Kelley Sobel, *Member*, *IEEE Computer Society*, and Michael R. Clarkson

Comments on "Formal Methods Application: An Empirical Tale of Software Development"

Daniel M. Berry and Walter F. Tichy

Formal Methods Application:
An Empirical Tale of Software Development

Ann E. Kelley Sobel, *Member*, *IEEE Computer Society*, and Michael R. Clarkson

Goal of the paper:

Show empirically that formal methods yields "better" programs

Using an experiment!

Overview: Experiment to show formal methods are “better”

- Two groups:
  - FM group
  - Control group
- Task: develop an elevator system, class project
  - FM group uses formal methods
  - Control group does not use formal methods
- Main Result (correctness):
  - FM group: 100% programs are correct
  - Control group: 45.5% programs are correct

Claim: the groups are identical except for FM

About the participants:

- College juniors (mostly)
- Computer Science majors
- Took identical classes, except:
  - FM group volunteered for a formal methods curriculum
  - Took two FM classes (control group took no FM classes)
- No statistical difference between the ACT scores of each group
- 6 FM teams, 11 control teams

Task instructions

Control Group

- Hand in source code & executable
- Optional: submit UML diagram (0/11 submitted)

FM Group

- Hand in source code & executable
- Hand in formal specification
- Optional: submit UML diagram (3/6 submitted)

Results (program correctness):

- A program is correct: passes 6 test cases
- 6/6 FM programs correct
- 5/11 control programs correct

Conclusions:

- FM *caused* the FM group’s programs to be more correct
- *Causal* evidence that FM yields “better” programs

Problems?

Comments on "Formal Methods Application: An Empirical Tale of Software Development"

Daniel M. Berry and Walter F. Tichy

Problems: Groups are not identical

- Difference in motivation:
  FM group may be more motivated (self selection)
- Difference in exposure to relevant material:
  FM group took 2 extra classes
  Took a more rigorous Data Structures class
- Differences in learning style:
  Survey identified FM group as “collaborative and competitive”
- Differences in skills:
  FM group self selected, they were ‘up for the challenge’
  Comp Sci GRE scores higher for FM group

Problems: Hawthorne & Novelty Effects

- Hawthorne Effect:
  Subjects act differently when aware of the experiment

- Novelty Effect:
  Subjects act differently when asked to do something new or different

- Subjects likely were aware of the experiment (Hawthorne)

Problems: Other theories may explain results

- Difference in deliverables (FM v. control)
- Lack of design information about the control group (no UML)
- Did the control group perform *any* analysis or design?
- The lack of control leaves room for other theories

Problems: Poor measurements

- 6 tests is not precise enough
- No information provided about these tests
- Ian’s thoughts:
  Binary result (correct / not correct) is not granular enough

Problems: No threats to validity

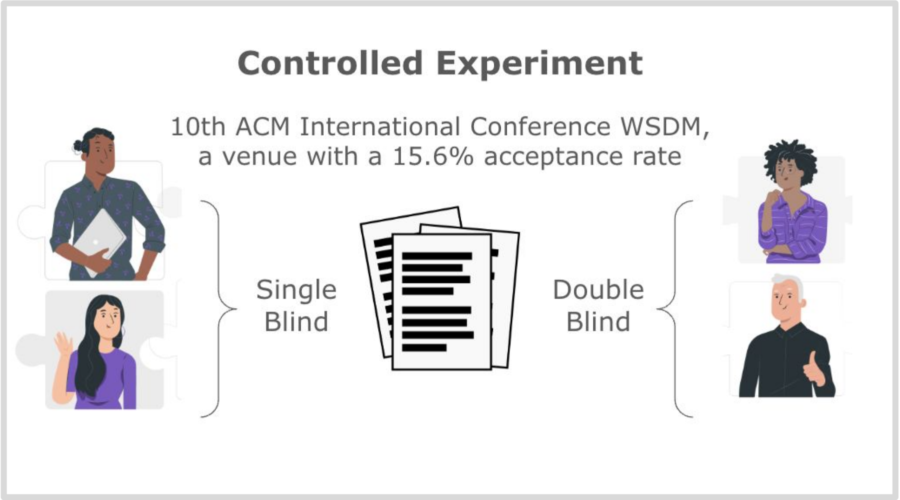
- Construct:
  How well do measurements reflect what we want measured?

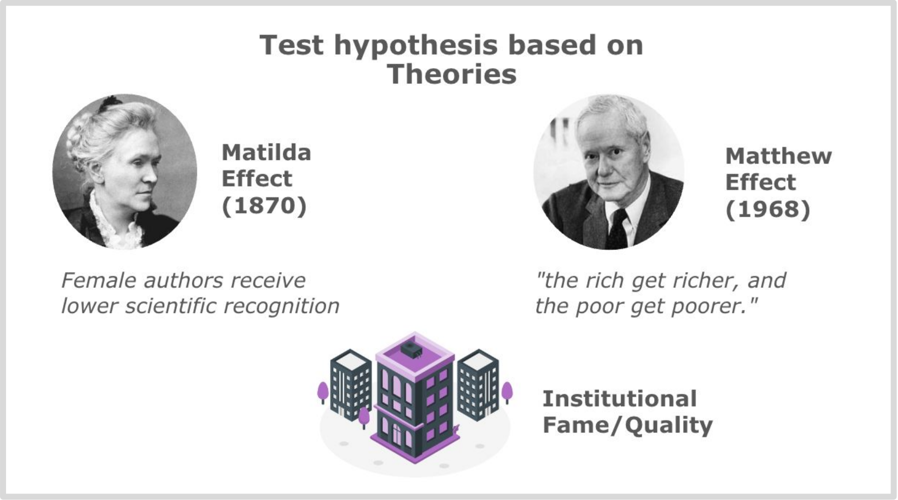
- Internal:
  Is the experiment sound (trustworthy)?

- External:
  Do the results generalize?

Nevertheless, the Sobel paper is a good first step

Takeaways for Empirical Methods





Statistical methods to find evidence in favor of a relationship or effect represented by the coefficients

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

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

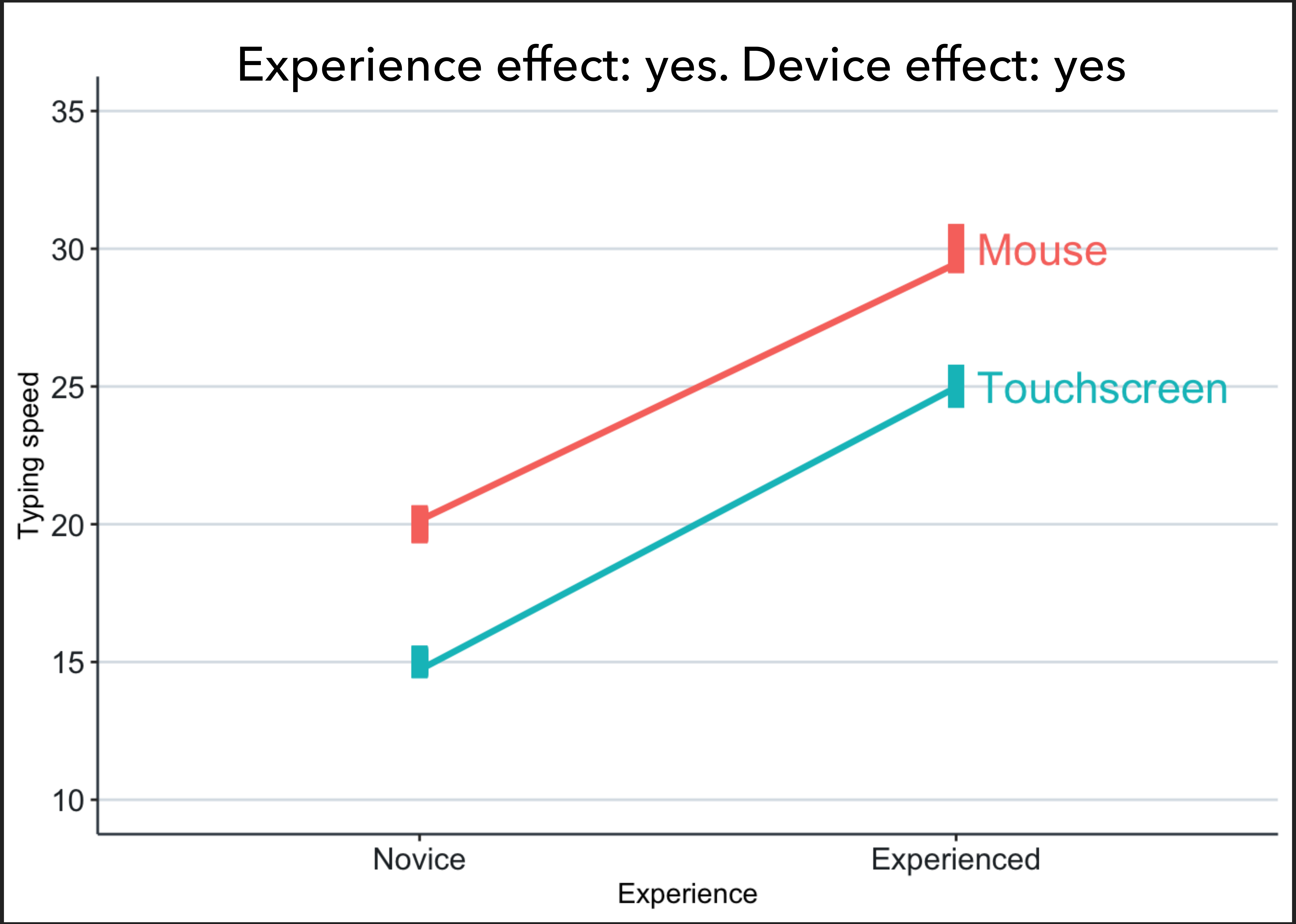
R	O	X_A	O
R	O	X_B	O

Factorial

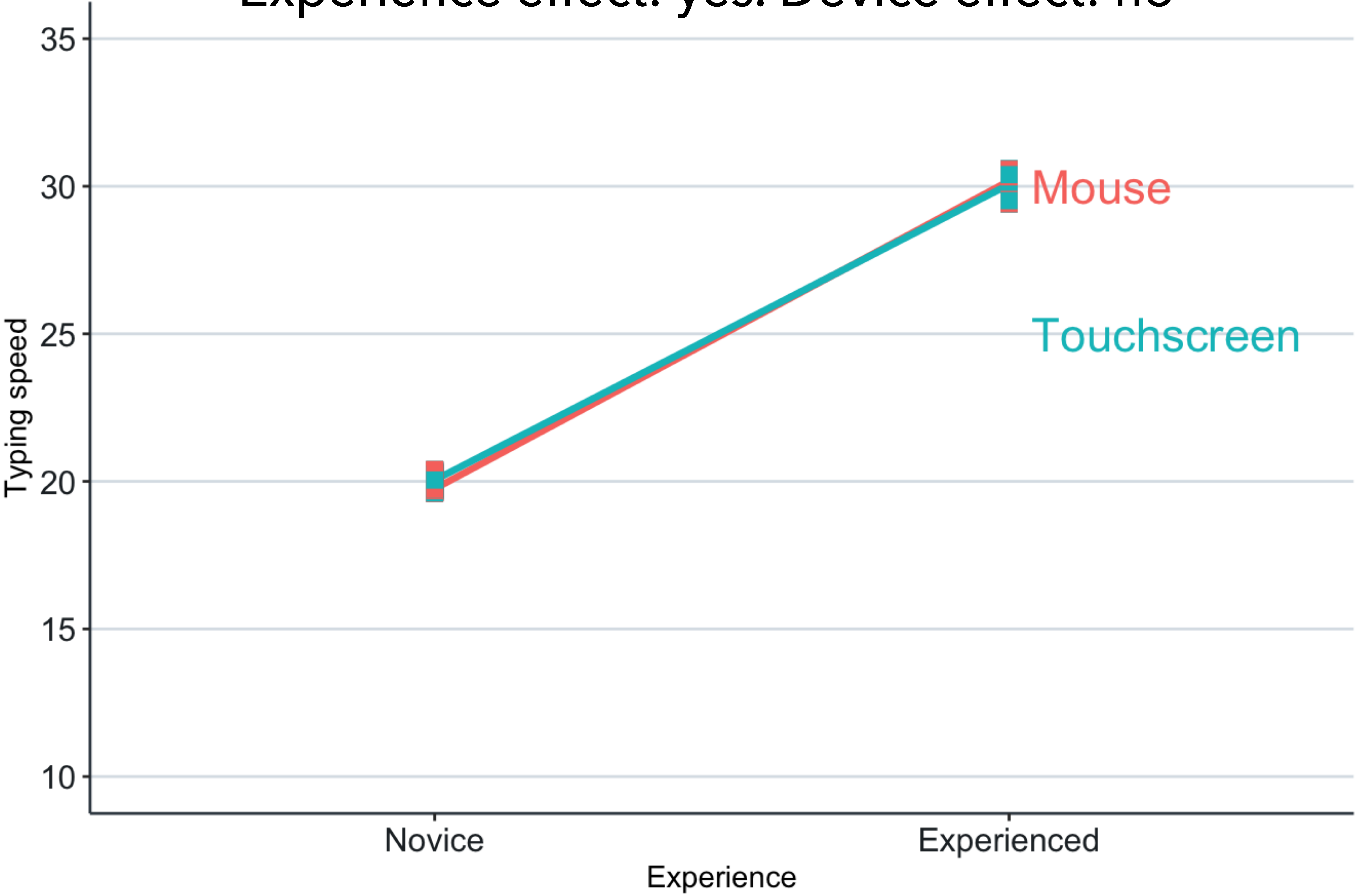
R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

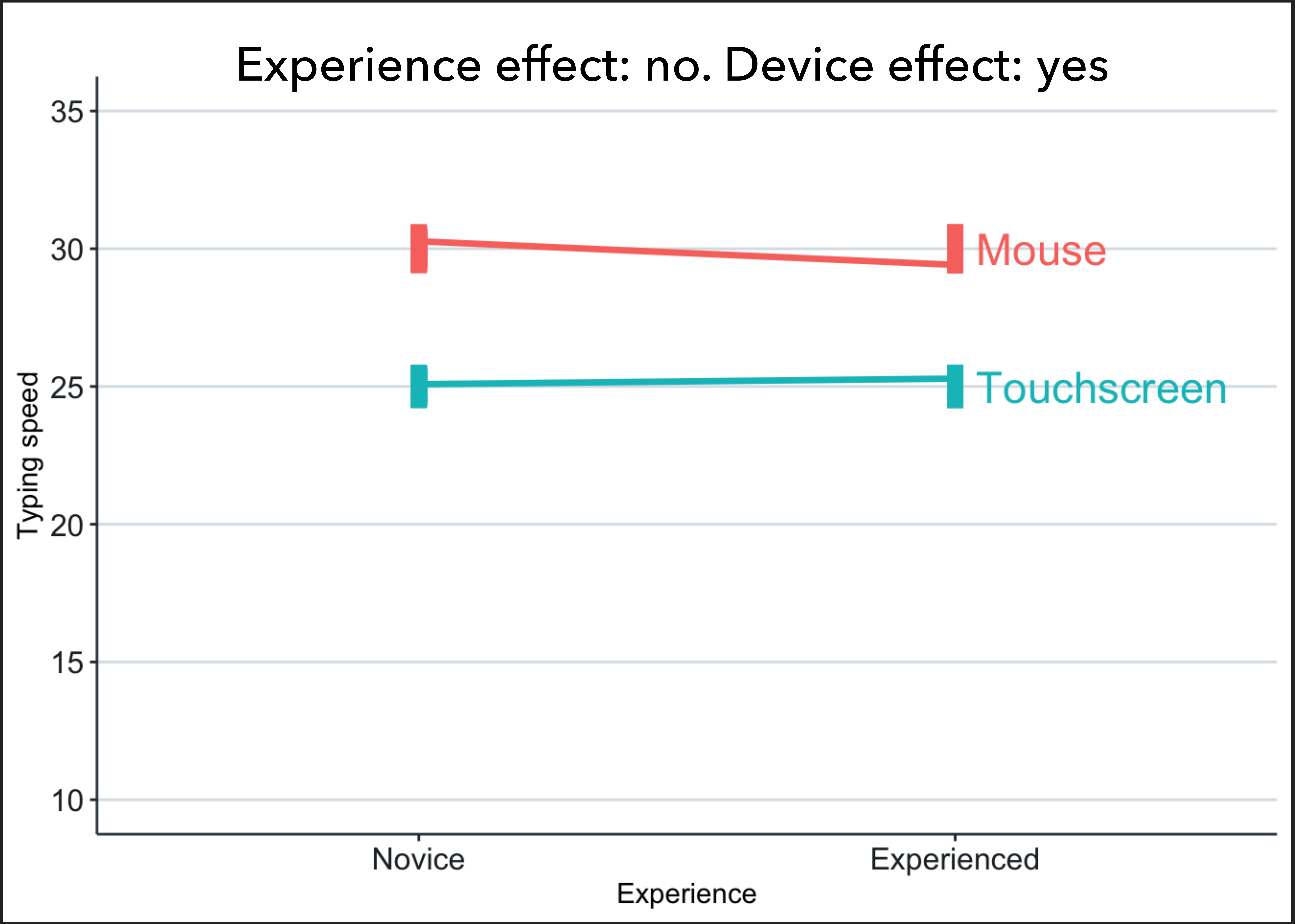
- Three major advantages:
  - They often require fewer units.
  - They allow testing combinations of treatments more easily.
  - They allow testing interactions.

Example: Typing speed = f(Experience, Device)


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

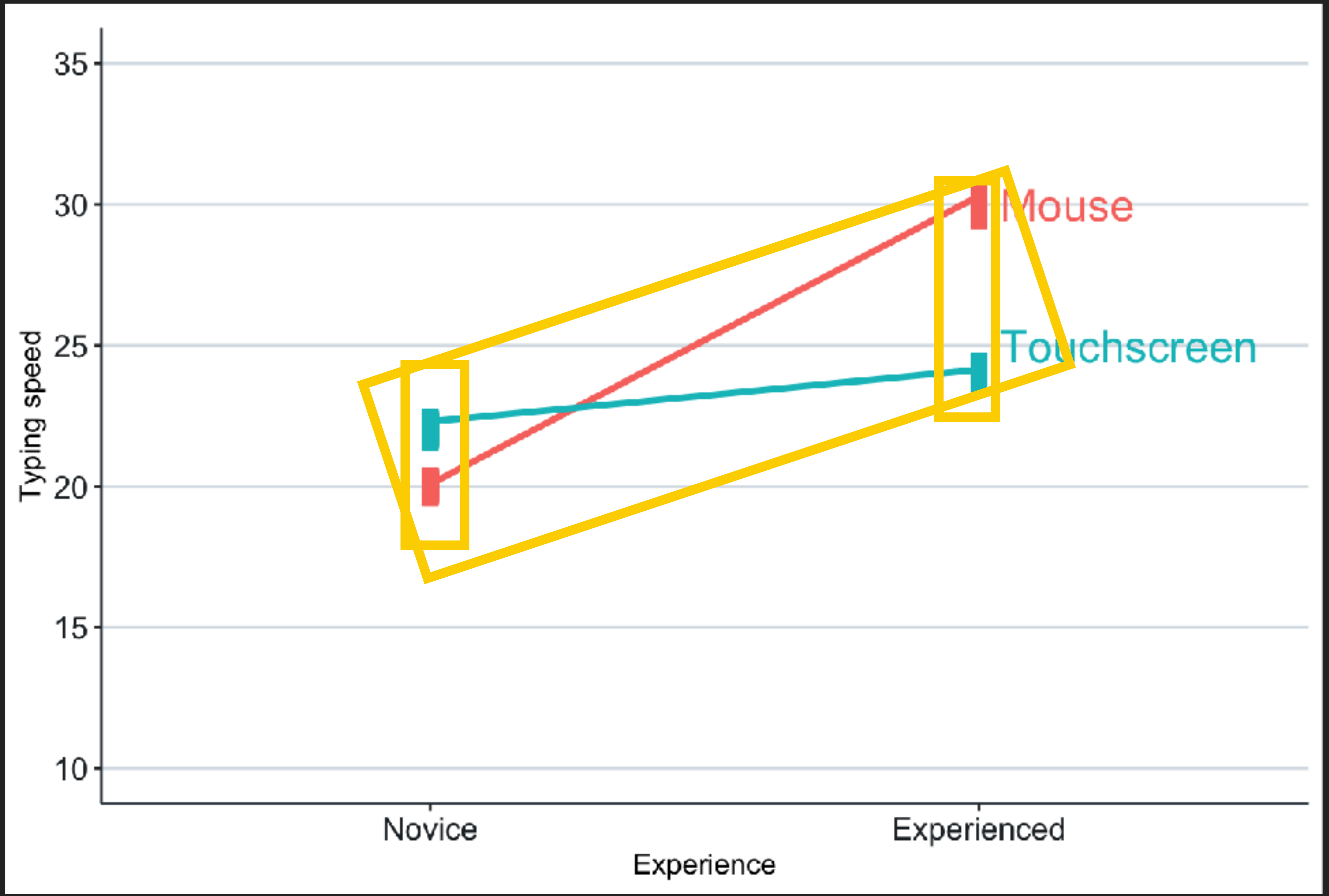
Experience effect: yes. Device effect: no



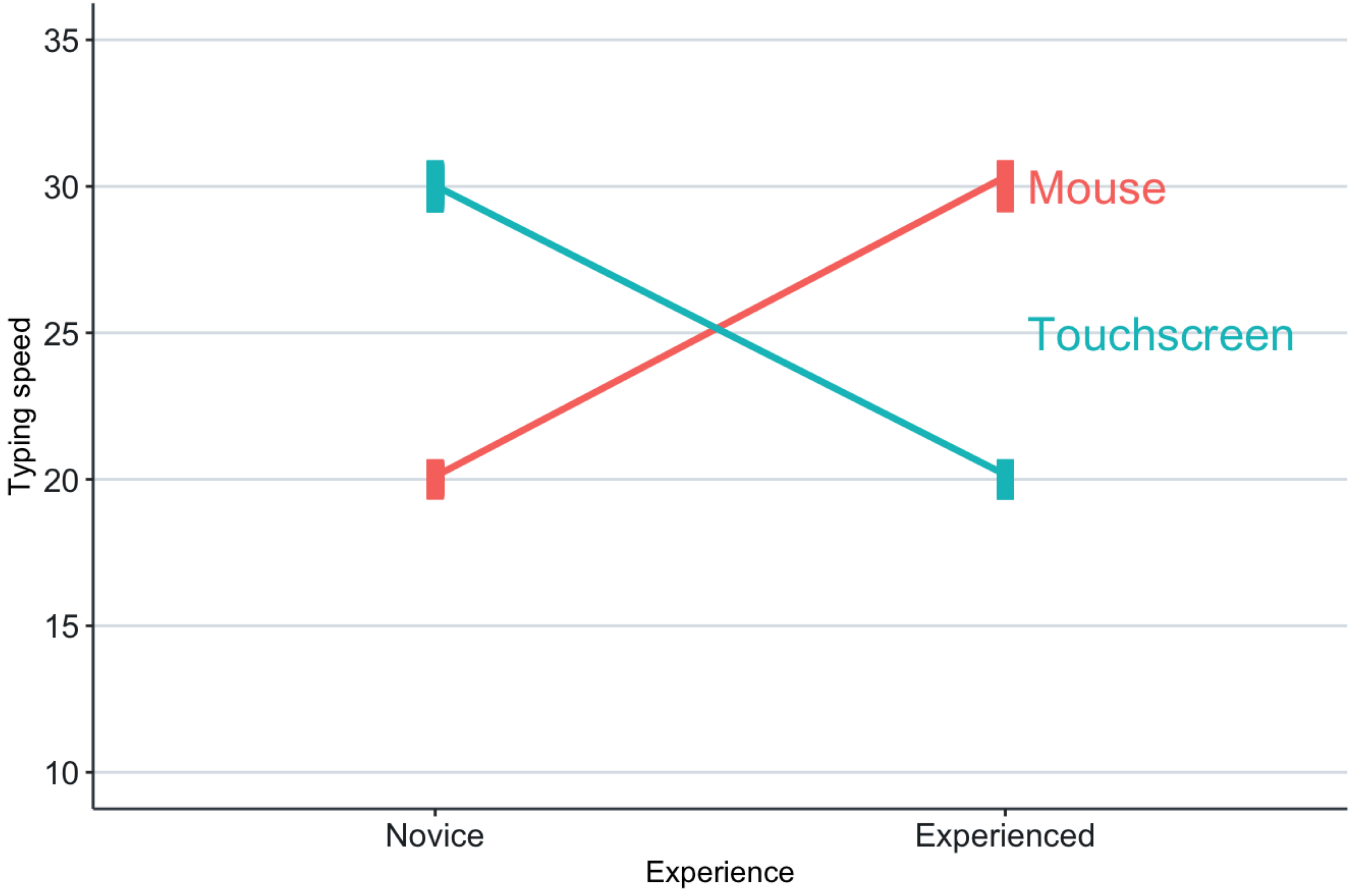

Experience effect: no. Device effect: yes
35
30
25
20
15
10
Typing speed
Mouse
Touchscreen
Novice
Experienced
Experience

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

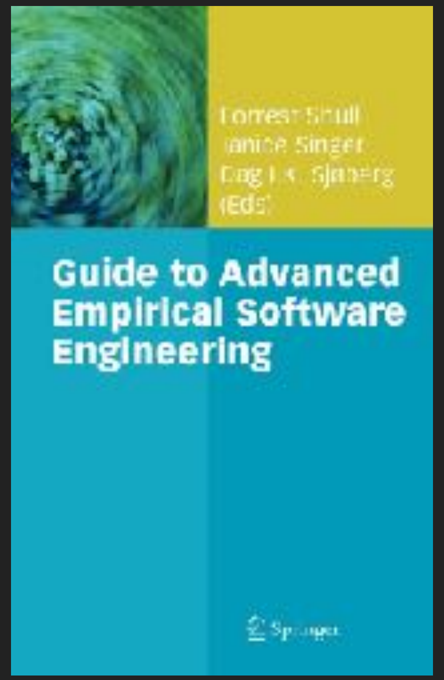


Credits

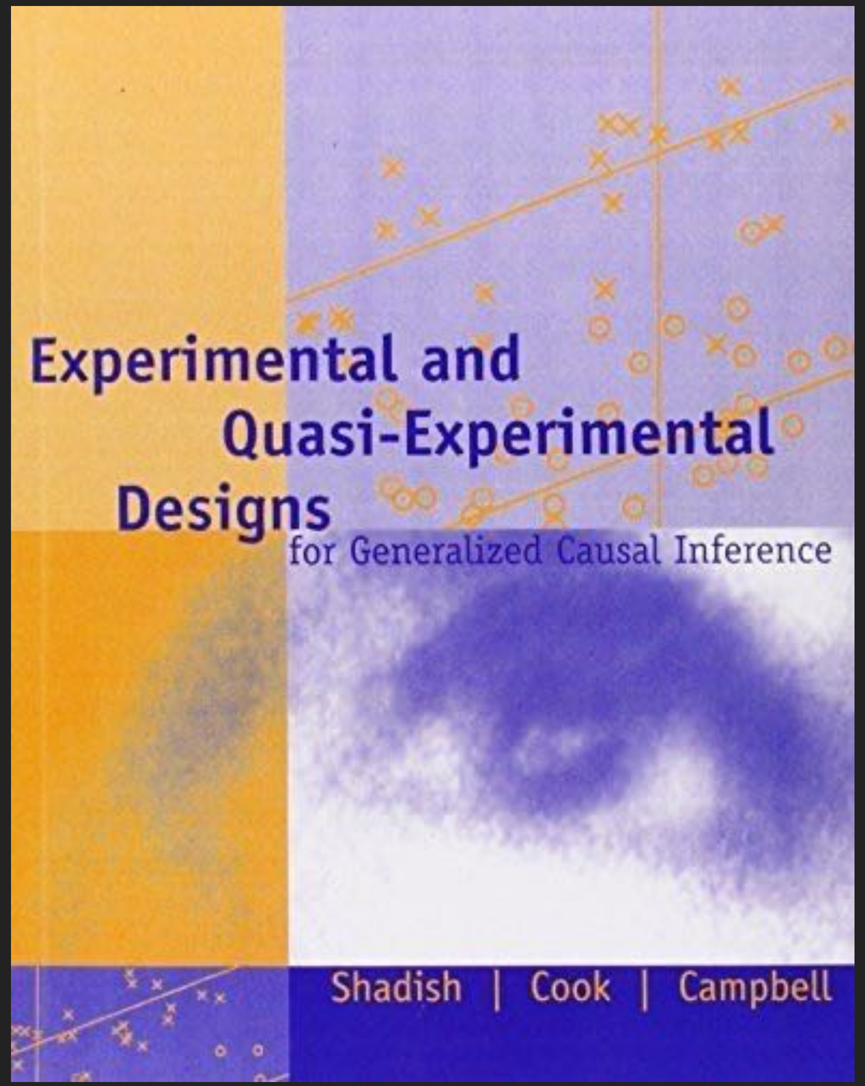
- Graphics: Dave DiCello photography (cover)
- Chapters from Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Publishing
  - Ch1: Experiments and generalized causal inference
  - Ch2: Statistical conclusion 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

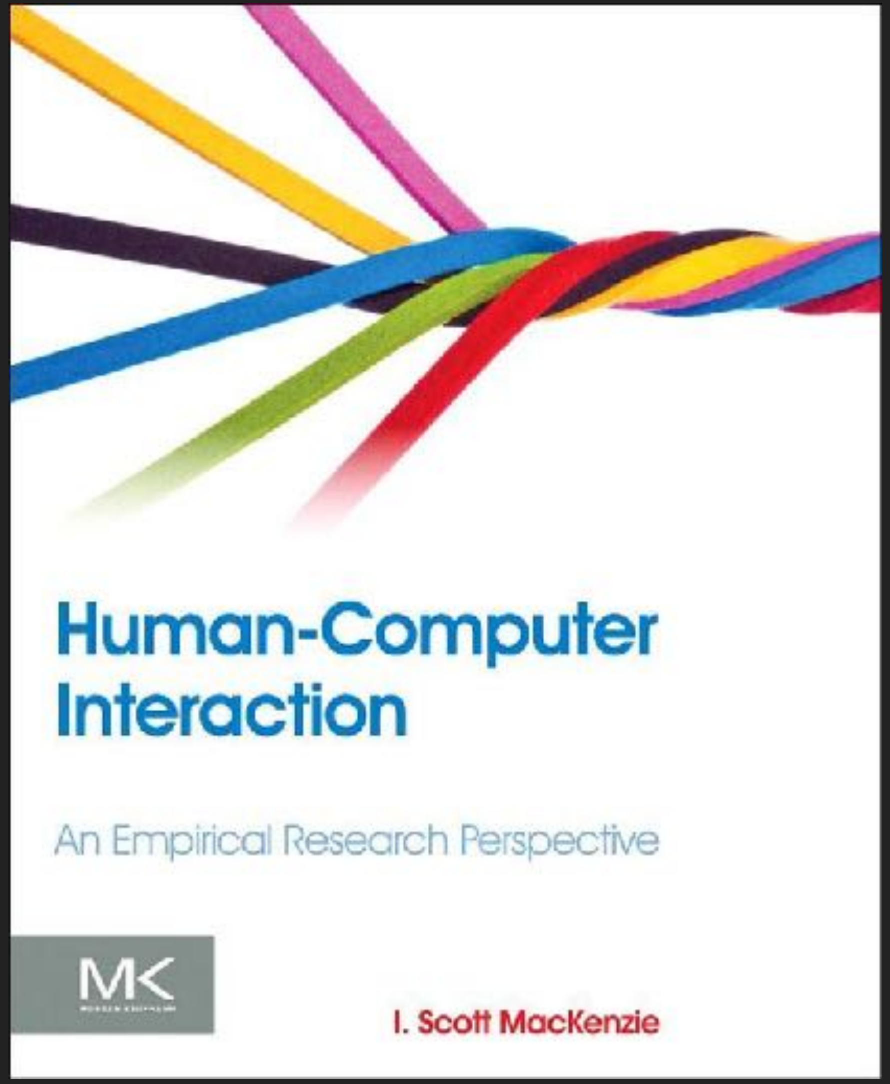
Ch 10 (Analysis and interpretation)



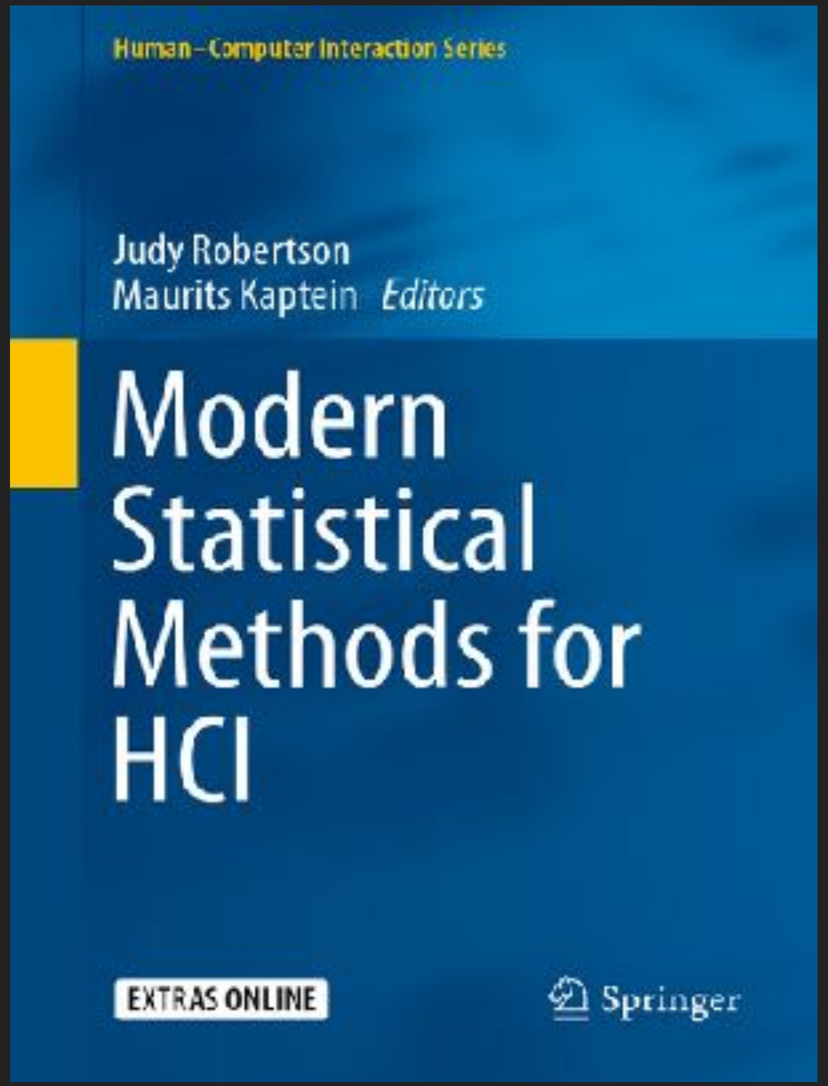
Ch 6 (Statistical methods and measurement)



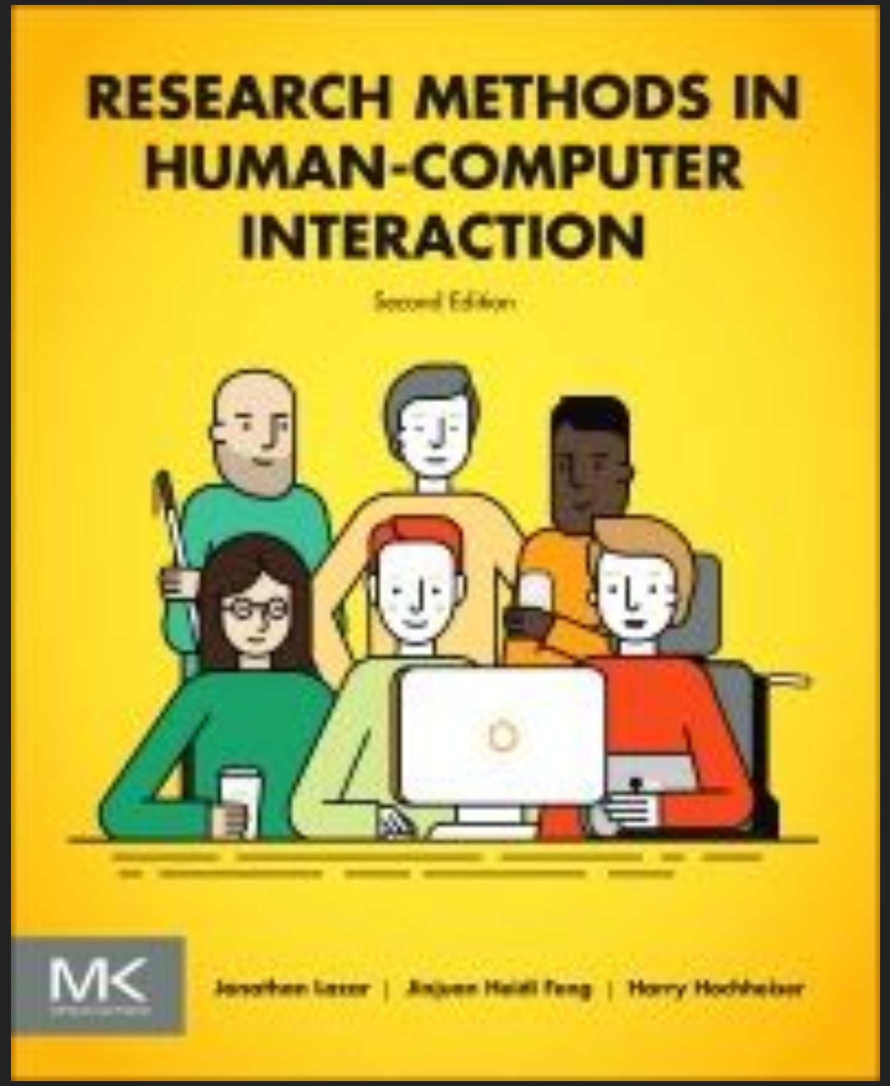
Ch 1 (Experiments and causality)
Ch 2 & 3 (Validity)
Ch 8 (Randomized experiments)



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)