Objecting to Experiments that Compare Two Unobjectionable Policies or Treatments:

Implications for Comparative Effectiveness and Other Pragmatic Trials

Michelle N. Meyer, PhD, JD

Assistant Professor & Associate Director of Research Ethics
Center for Translational Bioethics and Health Care Policy

Faculty Co-Director, Behavioral Insights Team ("nudge unit")
Steele Institute for Health Innovation

@MichelleNMeyer

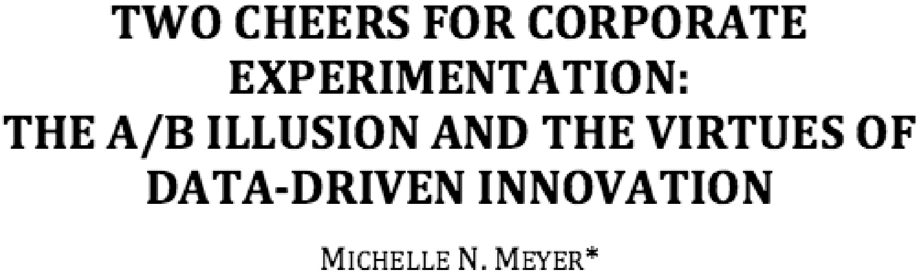
Geisinger

Why A/B tests? (a.k.a. field experiments, pRCTs)

- Increase quality and safety
- Decrease waste/lower costs
- Reduce inequity and injustice (Faden et al., 2011; Faden et al. 2013)

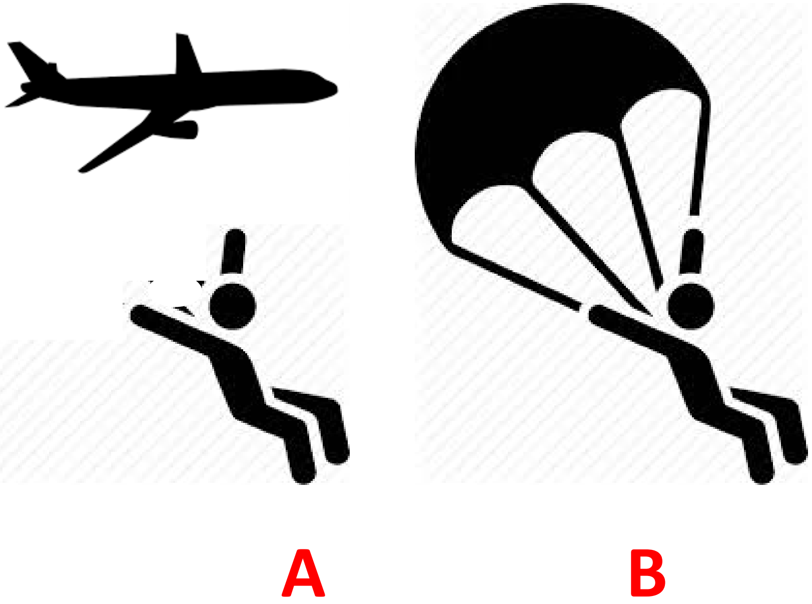
Health systems (& other organizations with captive audiences, e.g., businesses, schools, governments) **control the means of randomization**. They often have an ethical *obligation* to experiment in order to determine the effects of their policies and practices on stakeholders.

CTLJ
COLORADO TECHNOLOGY LAW JOURNAL
Formerly the Journal on Telecommunications and High Technology Law

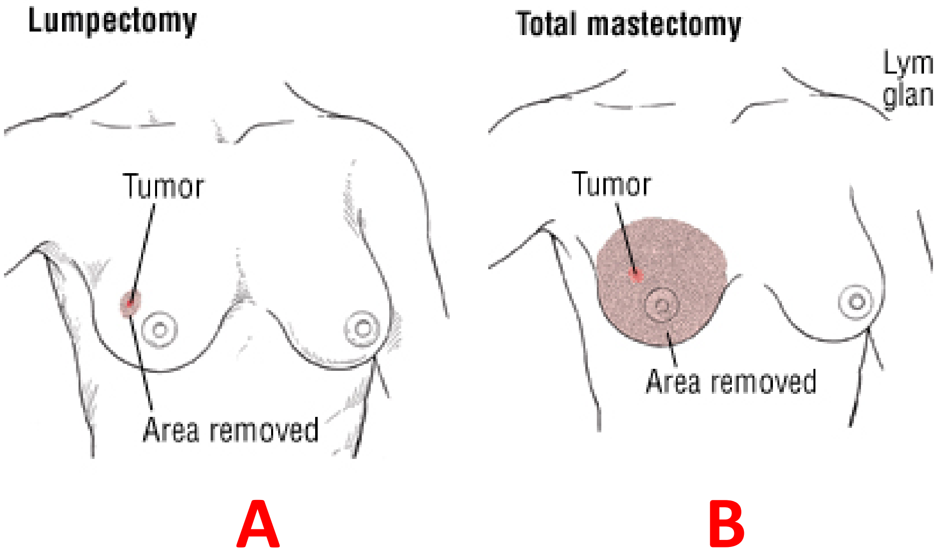


TWO CHEERS FOR CORPORATE EXPERIMENTATION:
THE A/B ILLUSION AND THE VIRTUES OF DATA-DRIVEN INNOVATION

MICHELLE N. MEYER*



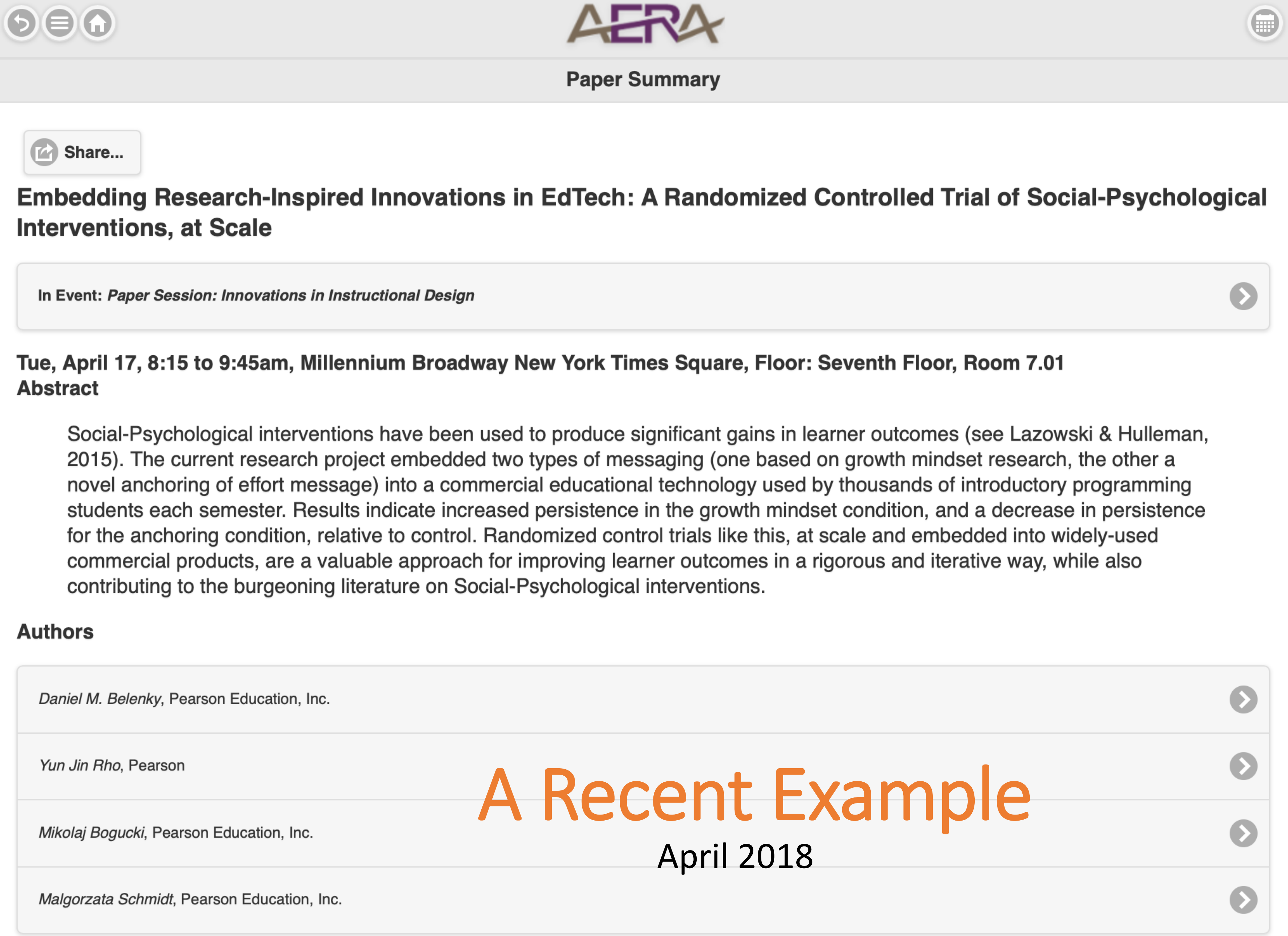
No equipoise



Preference-sensitive decision

Potentially inferior—but uniform—policy preferred to unequal treatment/outcomes

- Equipoise
- Not preference sensitive
- (Temporary) inequality acceptable



Embedding Research-Inspired Innovations in EdTech: A Randomized Controlled Trial of Social-Psychological Interventions, at Scale

In Event: ***Paper Session: Innovations in Instructional Design***

Tue, April 17, 8:15 to 9:45am, Millennium Broadway New York Times Square, Floor: Seventh Floor, Room 7.01

Abstract

Social-Psychological interventions have been used to produce significant gains in learner outcomes (see Lazowski & Hulleman, 2015). The current research project embedded two types of messaging (one based on growth mindset research, the other a novel anchoring of effort message) into a commercial educational technology used by thousands of introductory programming students each semester. Results indicate increased persistence in the growth mindset condition, and a decrease in persistence for the anchoring condition, relative to control. Randomized control trials like this, at scale and embedded into widely-used commercial products, are a valuable approach for improving learner outcomes in a rigorous and iterative way, while also contributing to the burgeoning literature on Social-Psychological interventions.

Authors

Daniel M. Belenky, Pearson Education, Inc.

Yun Jin Rho, Pearson

Mikolaj Bogucki, Pearson Education, Inc.

Malgorzata Schmidt, Pearson Education, Inc.

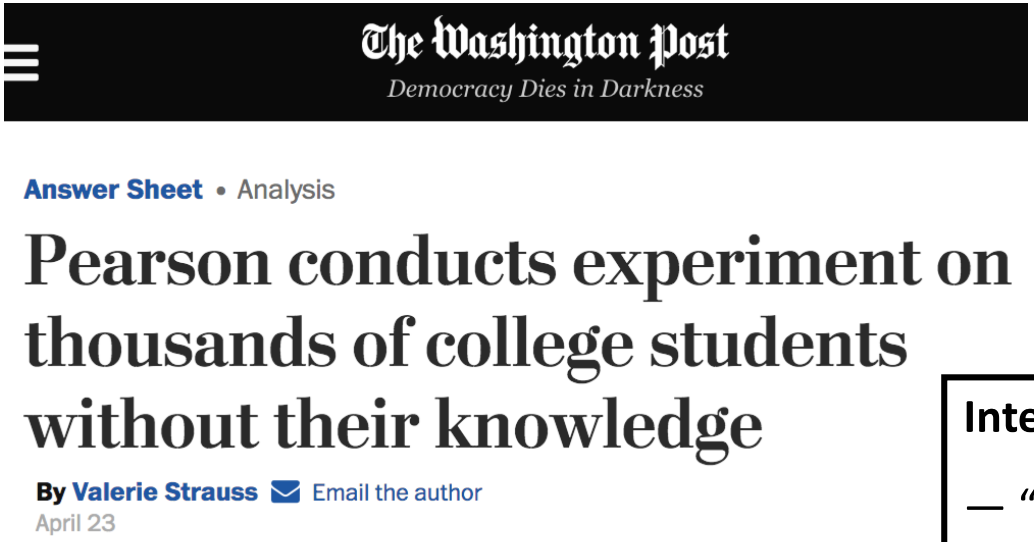
A Recent Example

April 2018

A Recent Example

Nudge	**Problems attempted**
A. Status quo: No encouragement	**212**
B. Anchoring of effort: “Some students tried this question 26 times! Don't worry if it takes you a few tries to get it right.”	**156**
C. Growth mindset: “No one is born a great programmer. Success takes hours and hours of practice.”	**174**

Pearson | MyLab



The Washington Post

Democracy Dies in Darkness

Answer Sheet • Analysis

Pearson conducts experiment on thousands of college students without their knowledge

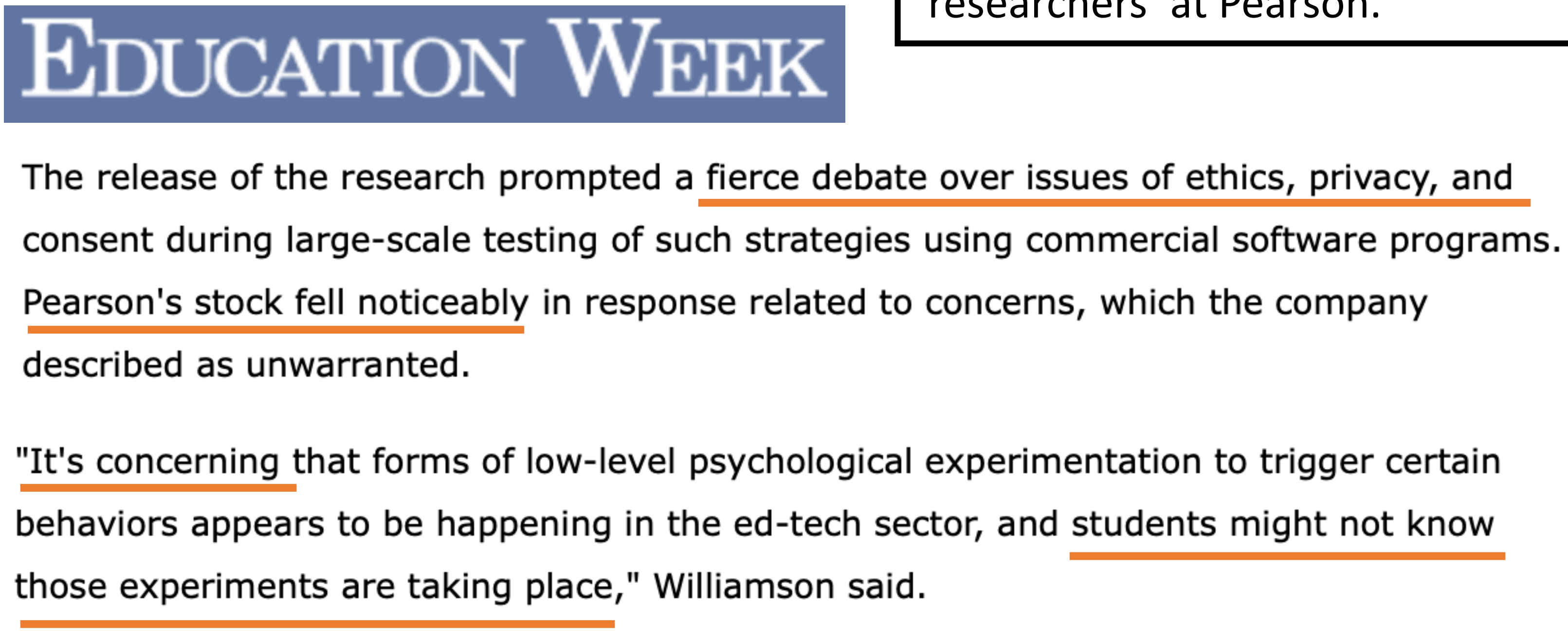
By Valerie Strauss Email the author

April 23

Internet comments:

— “This would be funny if it were not also unethical and outrageous.”

— “[A] completely unethical and possibly illegal breach of scientific protocol by Nazi ‘researchers’ at Pearson.”



EDUCATION WEEK

The release of the research prompted a fierce debate over issues of ethics, privacy, and consent during large-scale testing of such strategies using commercial software programs. Pearson's stock fell noticeably in response related to concerns, which the company described as unwarranted.

"It's concerning that forms of low-level psychological experimentation to trigger certain behaviors appears to be happening in the ed-tech sector, and students might not know those experiments are taking place," Williamson said.

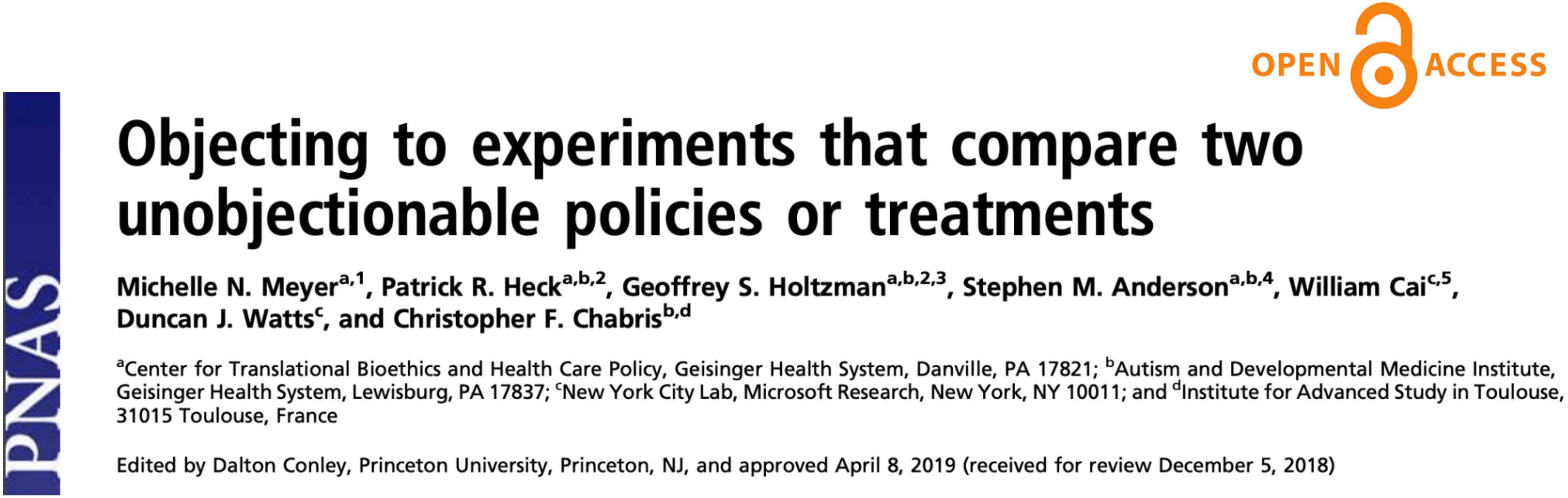
The “A/B Effect”

Viewing an experiment designed to determine the comparative effects of existing or proposed practices (an “A/B test”) as more morally problematic than a universal implementation of either untested practice (A or B).

- IF either treatment A or treatment B would be acceptable if applied to all members of a group on its own,
- AND neither A nor B is objectively superior or subjectively preferred to the other,
- AND temporary inequality is morally acceptable
- THEN randomly assigning those same people to A or B would not impose an unacceptable treatment on anyone, and would have the advantage of generating knowledge about the effects of A and B.

MAIN RESEARCH QUESTION: Can we systematically observe the proposed A/B effect in a variety of domains and populations?

- If so, when and why?
- Are there ways to communicate A/B tests to stakeholders that don’t arouse the A/B effect? E.g., consent documents/processes, LHS notices, published/presented results of learning activities.



Objecting to experiments that compare two unobjectionable policies or treatments

Michelle N. Meyer[a,1], Patrick R. Heck[a,b,2], Geoffrey S. Holtzman[a,b,2,3], Stephen M. Anderson[a,b,4], William Cai[c,5], Duncan J. Watts[c], and Christopher F. Chabris[b,d]

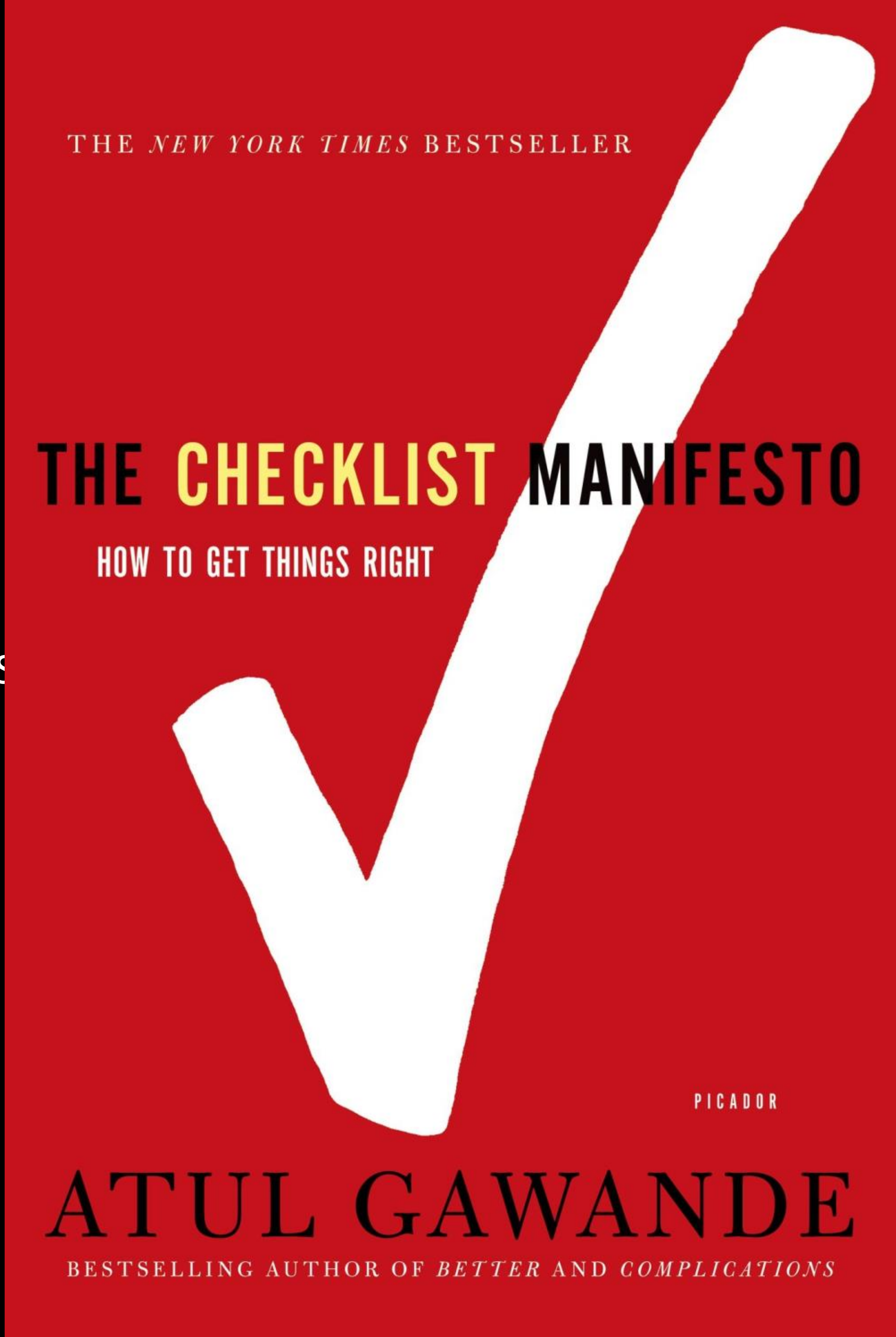
[a]Center for Translational Bioethics and Health Care Policy, Geisinger Health System, Danville, PA 17821; [b]Autism and Developmental Medicine Institute, Geisinger Health System, Lewisburg, PA 17837; [c]New York City Lab, Microsoft Research, New York, NY 10011; and [d]Institute for Advanced Study in Toulouse, 31015 Toulouse, France

Edited by Dalton Conley, Princeton University, Princeton, NJ, and approved April 8, 2019 (received for review December 5, 2018)

General Method

- 16 online, between-subjects vignette experiments & replications (all but the first preregistered)
- Randomization to 1 of 3 (or 4) conditions, in which a well-intentioned agent thinks of 1 (or 2) policies and:
  - implements policy A
  - implements policy B
  - runs a randomized experiment comparing A and B
- DV: “How appropriate is the decision?” (1-5 Likert; neutral midpoint)
- Why? (free response: 28 codes, 2 coders, avg interrater reliability across 4 studies: $\kappa = .83$)
- Total *N* = 5873 unique participants (~100/condition)

Study 1


THE NEW YORK TIMES BESTSELLER
THE CHECKLIST MANIFESTO
HOW TO GET THINGS RIGHT
PICADOR
ATUL GAWANDE
BESTSELLING AUTHOR OF BETTER AND COMPLICATIONS

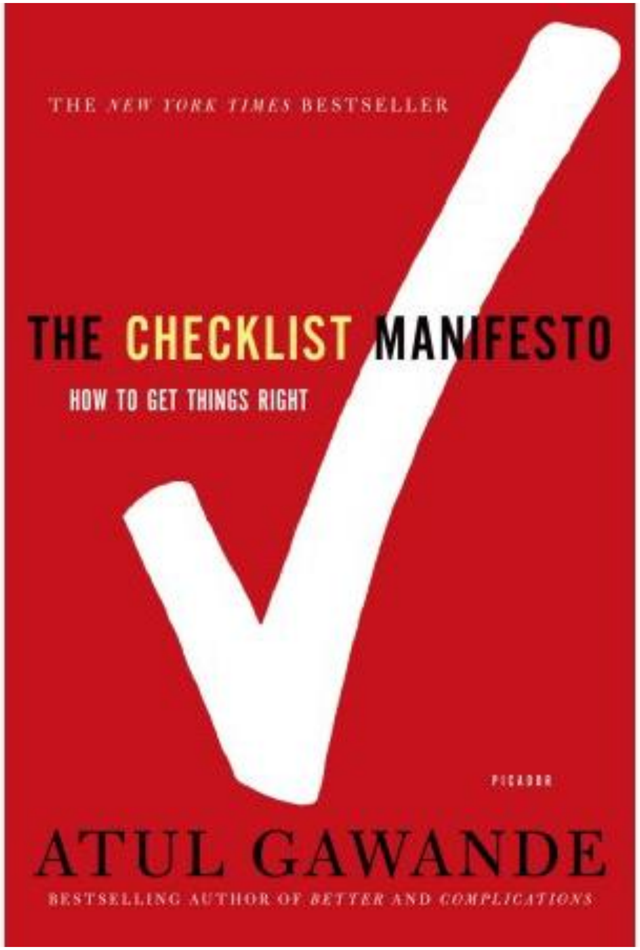
A: Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director wants to reduce these infections, so he decides to **give each doctor who performs this procedure a new ID badge** with a list of standard safety precautions for the procedure **printed on the back**. All patients having this procedure will then be treated by doctors with this list attached to their clothing.

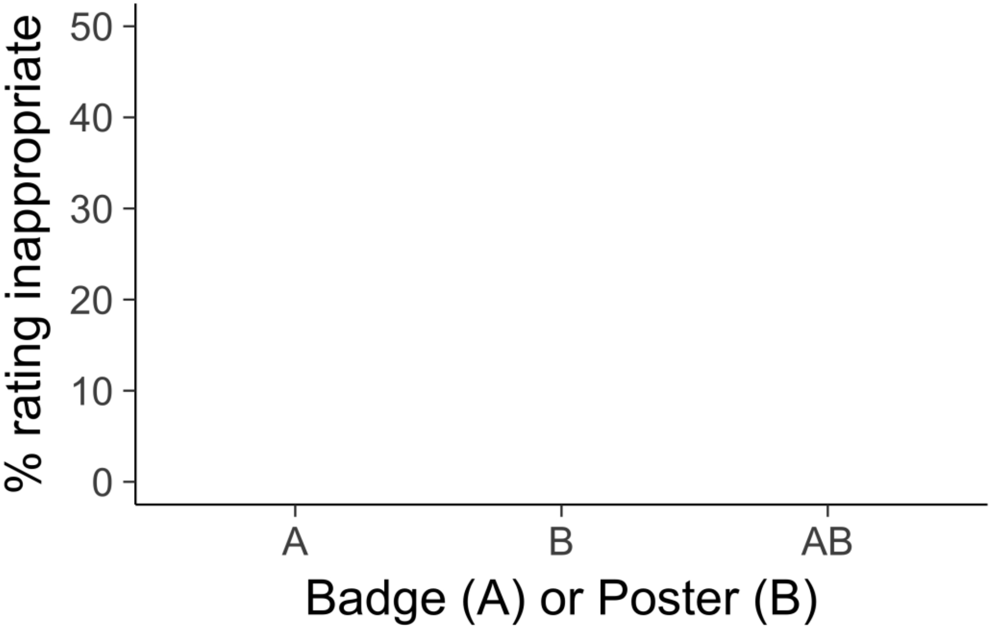
B: . . . A hospital director wants to reduce these infections, so he decides to **hang a poster** with a list of standard safety precautions for this procedure **in all procedure rooms**. All patients having this procedure will then be treated in rooms with this list posted on the wall.

A/B: . . . A hospital director **thinks of two different ways** to reduce these infections, so he decides to **run an experiment by randomly assigning patients to one of two test conditions.** Half of patients will be treated by doctors who have received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in rooms with a poster listing the same precautions hanging on the wall.

A/B Learn: . . . **After a year, the director will have all patients treated in whichever way turns out to have the highest survival rate.**

Study 1: Catheter Checklist (*N* = 338)

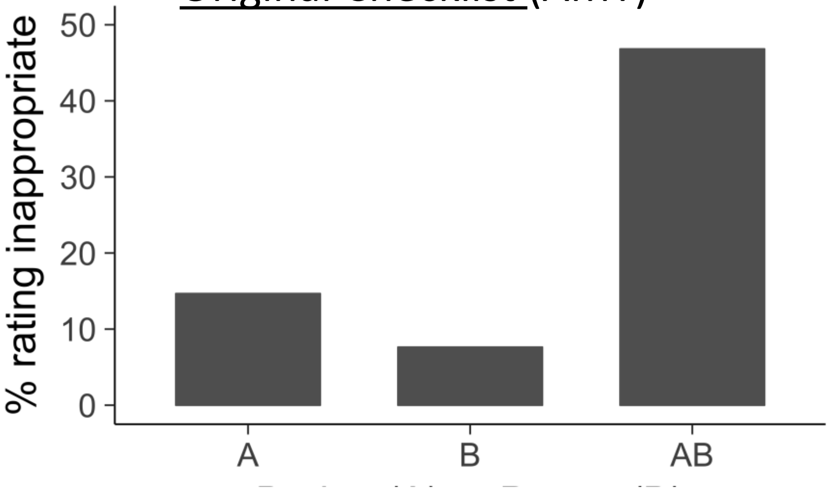

THE NEW YORK TIMES BESTSELLER
THE CHECKLIST MANIFESTO
HOW TO GET THINGS RIGHT
PICADOR
ATUL GAWANDE
BESTSELLING AUTHOR OF BETTER AND COMPLICATIONS



d = 1.08

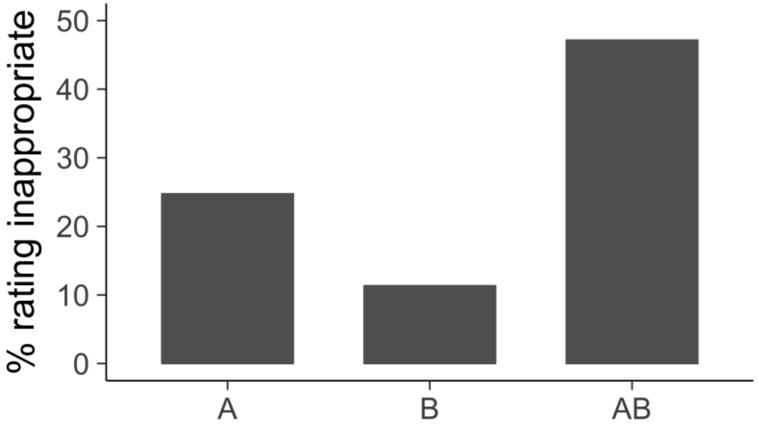
Study 2: Catheter Checklist Replications

Original Checklist (AMT)



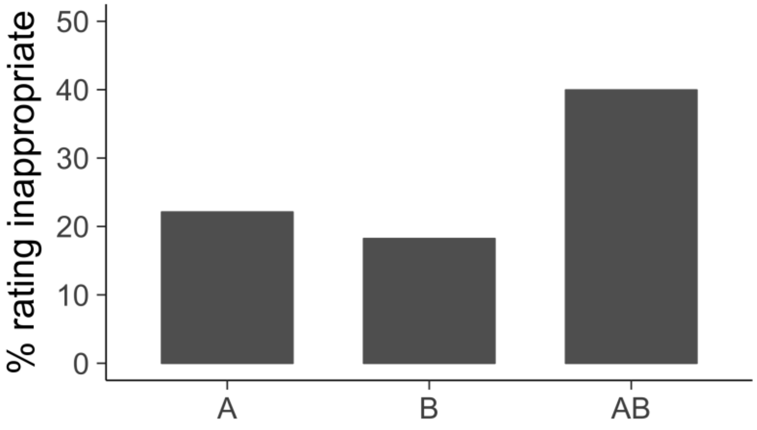
N = 338; *d* = 1.08

Exact Replication (AMT)



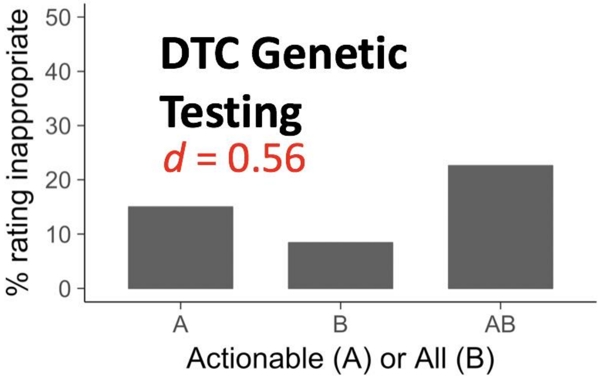
N = 387; *d* = 0.89

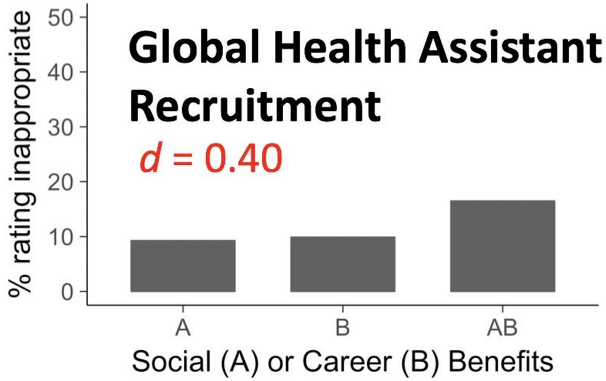
Mobile Replication (Pollfish)

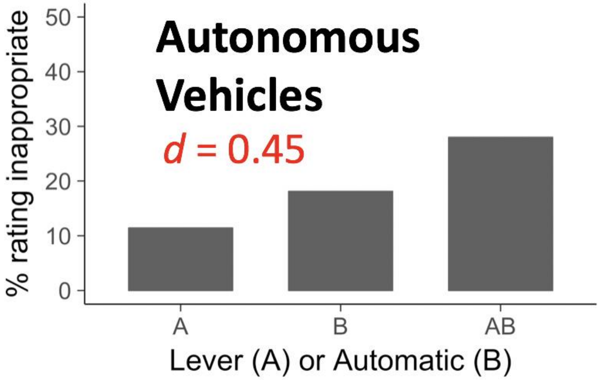


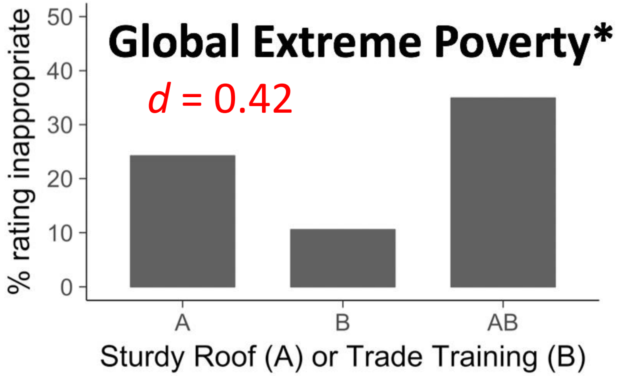
N = 825; *d* = 0.57

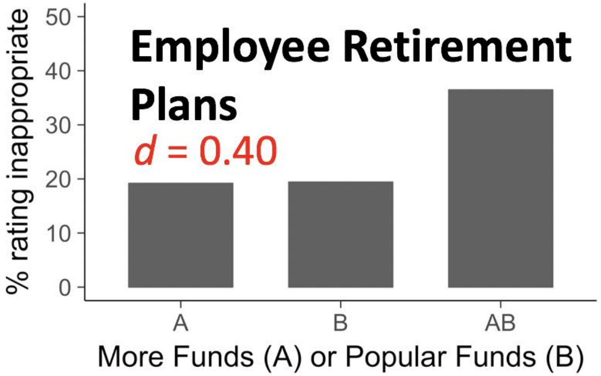
Study 3: Other Domains (*N* = 2312)

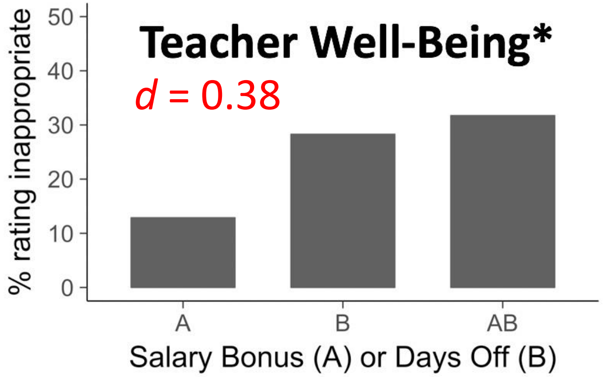

DTC Genetic Testing
d = 0.56


Global Health Assistant Recruitment
d = 0.40


Autonomous Vehicles
d = 0.45


Global Extreme Poverty*
d = 0.42


Employee Retirement Plans
d = 0.40

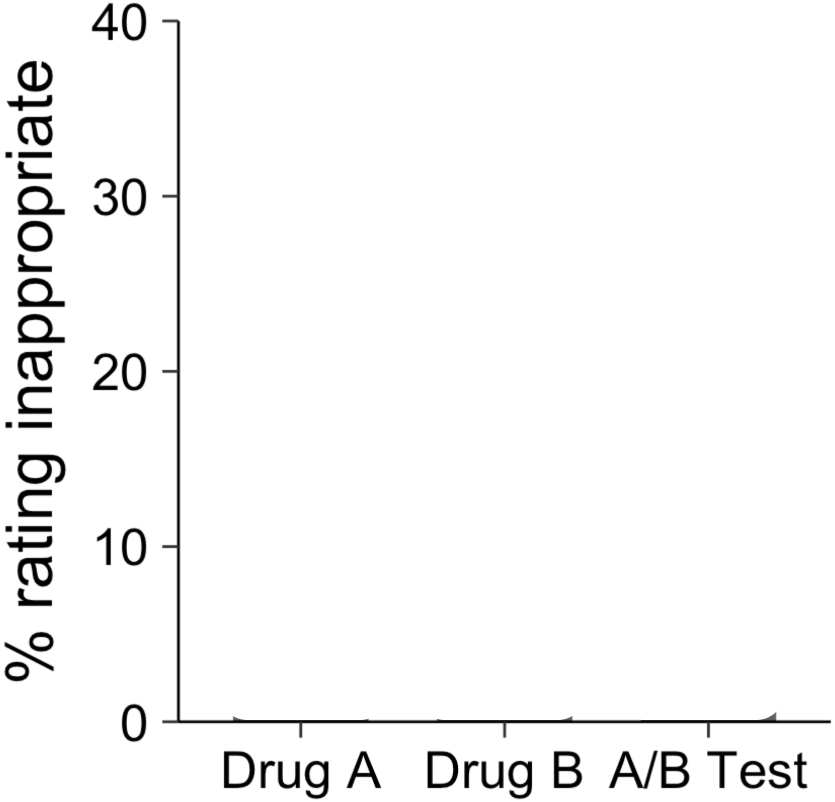

Teacher Well-Being*
d = 0.38

Why Might We Object to A/B Tests of Two Unobjectionable Treatments?

1. Intuitions (possibly dangerously incorrect) about comparative effectiveness of A and B when jointly evaluated
2. Aversion to unequal treatment
3. Aversion to random treatment

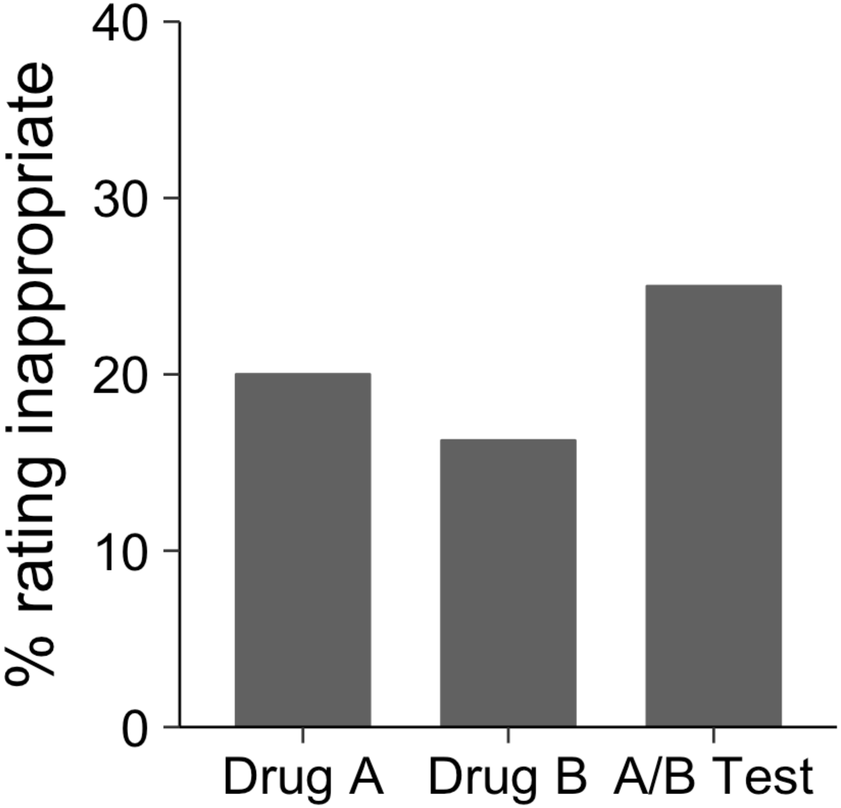
Study 5: Best Drug

The WALK IN CLINIC



N = 307; *d* = 0.64

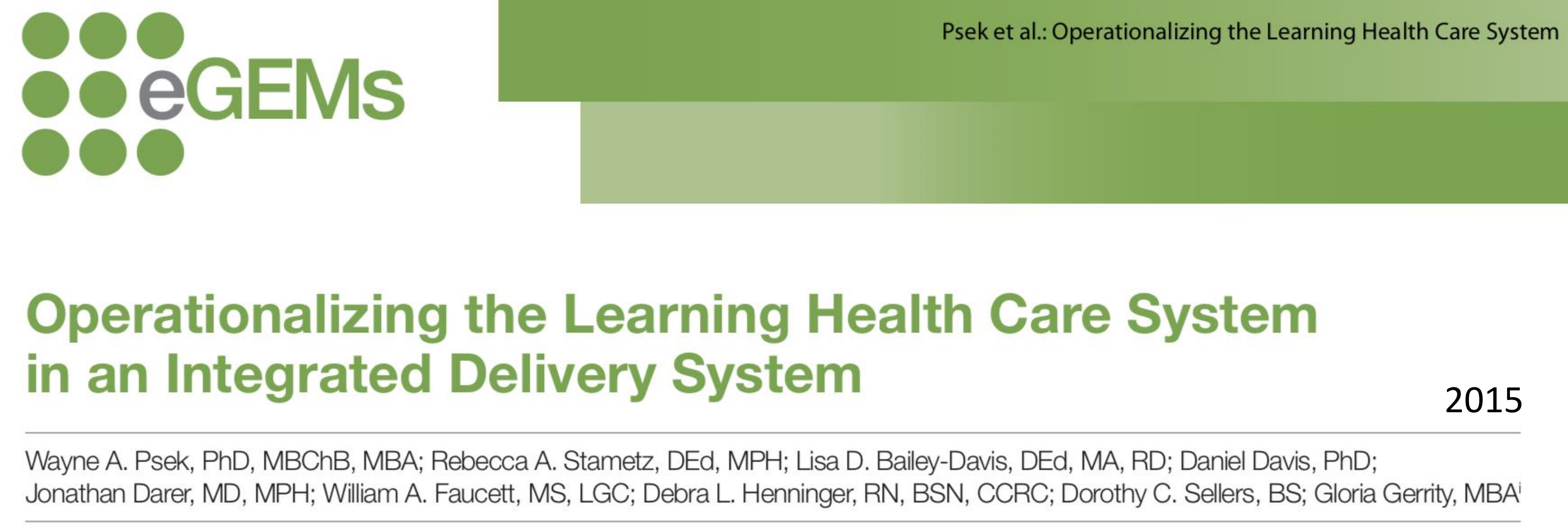
Mobile Replication (Pollfish)



N = 720; *d* = 0.15

Why Might We Object to A/B Tests of Two Unobjectionable Treatments?

1. ~~Intuitions (possibly dangerously incorrect) about comparative effectiveness of A and B when jointly evaluated~~
2. ~~Aversion to unequal treatment~~
3. ~~Aversion to random treatment~~
4. Low science literacy



eGEMs

Psek et al.: Operationalizing the Learning Health Care System

Operationalizing the Learning Health Care System in an Integrated Delivery System

2015

Wayne A. Psek, PhD, MBChB, MBA; Rebecca A. Stametz, DEd, MPH; Lisa D. Bailey-Davis, DEd, MA, RD; Daniel Davis, PhD; Jonathan Darer, MD, MPH; William A. Faucett, MS, LGC; Debra L. Henninger, RN, BSN, CCRC; Dorothy C. Sellers, BS; Gloria Gerrity, MBA[i]

Interviews (n = 41) with Geisinger leadership found unanimous support for “the general concept and goals” of the learning healthcare system and for “enhancing learning across the institution.”

JPCRR Journal of Patient-Centered Research and Reviews

Organizational Learning in an Integrated Health System: Informing Operations for a Learning Health Care System

2017

Deserae Clarke, *Geisinger Institute for Advanced Application, Danville, PA*

Gloria Gerrity, *Geisinger Pediatric Administration, Danville, PA*

Rebecca Stametz, *Geisinger Institute for Advanced Application, Danville, PA*

Amanda Young, *Geisinger Biostatistics Core, Danville, PA*

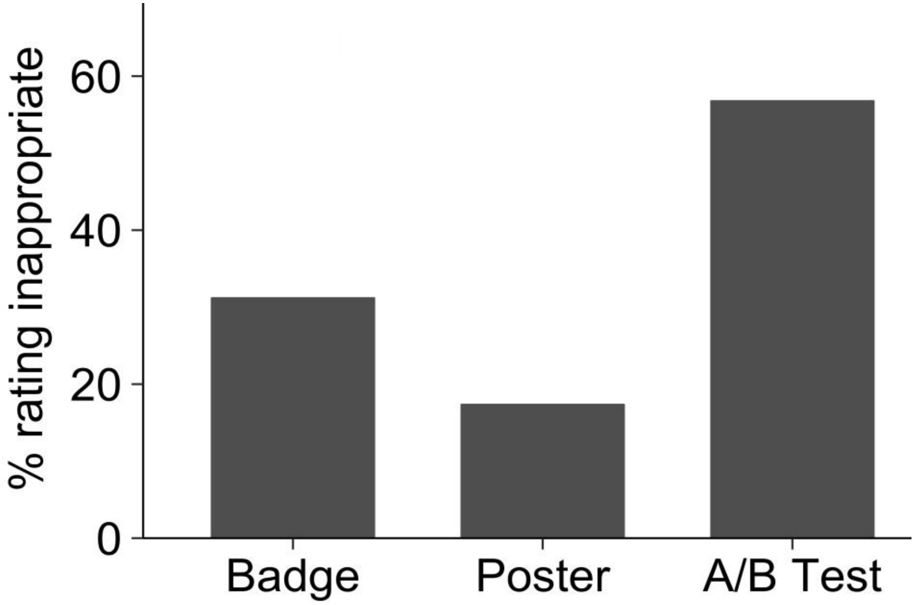
Daniel Davis, *Geisinger Bioethics, Danville, PA*

"Evidence supports the claim that a learning health system is necessary to provide safe, effective, and beneficial patient-centered care at lower cost."

- 98% (n = 126; 64% response rate) of respondents (most of whom were clinicians) agreed
- 53% strongly agreed

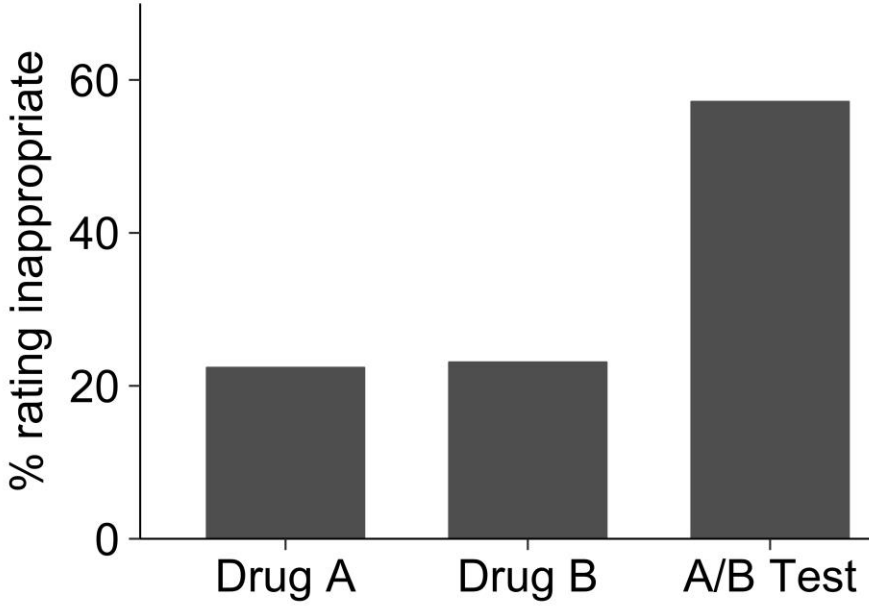
Study 6: Healthcare Providers Sample

<u>Checklist</u> (*N* = 226)



d = 0.86

<u>Best Drug: Walk-In</u> (*N* = 231)



d = 0.87

Why Might We Object to A/B Tests of Two Unobjectionable Treatments?

1. ~~Intuitions (possibly dangerously incorrect) about comparative effectiveness of A and B when jointly evaluated~~
2. ~~Aversion to unequal treatment~~
3. ~~Aversion to random treatment~~
4. ~~Low science literacy~~
5. Low educational attainment
6. Other sociodemographic variables

Why Might We Object to A/B Tests of Two Unobjectionable Treatments?

1. ~~Intuitions (possibly dangerously incorrect) about comparative effectiveness of A and B when jointly evaluated~~
2. ~~Aversion to unequal treatment~~
3. ~~Aversion to random treatment~~
4. ~~Low science literacy~~
5. ~~Low educational attainment~~
6. ~~Other sociodemographic variables~~
7. Lack of consent
    - **18%** of participants in A/B conditions vs. **0.3%** in policy conditions
8. "Experiment" aversion
    - **24%** of participants in A/B conditions vs. **0.1%** in policy conditions
9. Illusion of knowledge
    - Best Drug: 21% of participants who approve policy & 19% of those who object to an A/B test

Conclusions (so far)

- We can observe the “A/B effect” in several domains (e.g., health care, addressing global poverty, autonomous vehicle design, retirement nudges)
- Educational attainment, science literacy, and other demographic variables explain essentially none of the variance among participants
- After controlling for inequality and randomization (Best Drug: Walk-in), several remaining explanations (consent, experiment aversion, illusion of knowledge) appear to contribute to the effect, but none dominates
- “A/B effect” may reflect a heuristic about the ethics of experiments that sometimes leads us astray
- More research needed: causal mechanisms, boundary conditions, debiasing strategies
- Decisionmakers may face less backlash if they implement untested policies/treatments on everyone instead of randomly evaluating them to determine comparative effectiveness

In progress work

(with Chabris, Heck, Pedram Heydari, Anh Huynh)

What if we tell people the agent could have imposed either policy for everyone? (Within-subjects)

<u>Checklist</u>: AB effect 71% as large (*d*=1.19 → *d*=0.84)

- 53% of participants rate A/B test as less appropriate than the average of A & B
- 37% rate the experiment as less appropriate than both policies
- 27% rate both policies not-inappropriate (3, 4, or 5 Likert) & the A/B test inappropriate (1 or 2)
- Ranking: 37% rank A/B test 1st; 46% rank it last

What if we also model clinical equipoise for them?

<u>Best Drug–Walk-In</u>: 61% as large (*d*=0.64 → *d*=0.39)

- 43% of participants rate A/B test as less appropriate than the average of A & B
- 40% rate the experiment as less appropriate than both policies
- 27% rate both policies not-inappropriate (3, 4, or 5 Likert) & the A/B test inappropriate (1 or 2)
- Ranking: 59% rank A/B test 1st; 37% rank it last

Thank you!