CHARTERED COLLEGE OF TEACHING

# The role of research in improving education

Chartered College of Teaching Annual Lecture 2023

Dylan Wiliam (@dylanwiliam)

# What does it mean to be research-based?



- In a ‘research-based’ profession:
  - professionals would, for the majority of decisions they need to take, be able to find and access credible research studies that provided evidence that particular courses of action that would, implemented as directed, be substantially more likely to lead to better outcomes than others.

# Outline



- A framework for thinking about change
  - Understanding why things are as they are
  - Understanding that improvement means making different trade-offs, which
    - Can be planned, argued about and discussed, or
    - Emerge as unintended consequences of other decisions
- What educational research can—and cannot—do
- Five questions to use to ask of research

# A framework for thinking about change

# Why things are as they are: Chesterton's fence



"In the matter of reforming things, as distinct from deforming them, there is one plain and simple principle; a principle which will probably be called a paradox. There exists in such a case a certain institution or law; let us say, for the sake of simplicity, a fence or gate erected across a road. The more modern type of reformer goes gaily up to it and says, "I don't see the use of this; let us clear it away." To which the more intelligent type of reformer will do well to answer: "If you don't see the use of it, I certainly won't let you clear it away. Go away and think. Then, when you can come back and tell me that you do see the use of it, I may allow you to destroy it." (Chesterton, 1929/1990 p. 157)

# There are no perfect solutions; only trade-offs



- Proposals for change are usually
  - clear about what the old approach did badly, and how the proposal will improve those aspects
  - silent about the things that the old approach did well, and that the new proposal will do less well
- Proposals for change should answer two questions:
  - “What will be better if the changes are made?”
  - “What will be worse if the changes are made?”
- If the answer to the second question is “nothing” then the proposer needs to think again.
- There will always be trade-offs
- The question is whether they are explicit or not

# The role of research



- “Evidence-based” is a continuum, not a category
- Research can
  - Indicate areas where improvement efforts are unlikely to improve education
  - Provide information about relevant moderators of effect
  - Provide information about how much improvement is possible or likely
  - Provide information about the costs of the innovation

# Learning from research

# Where should we focus our efforts?



- Inference involves two settings
  - Information is acquired in one setting (learning)
  - Information is applied in others (predictions, choices)
- Two kinds of learning environment
  - *Kind* learning environments
    - close match of informational elements in the two settings
  - *Wicked* learning environments
    - poor match of informational elements in the two settings

Hogarth, Lejarraga, and Soyer (2015)

# Teaching is a wicked learning environment…

- Performance is not learning

# Learning and performance…



- Learning
  - “relatively permanent changes in behavior or knowledge that support long- term retention and transfer.” (p. 276)
- Performance
  - “temporary fluctuations in behavior or knowledge that can be observed and measured during or immediately after the acquisition process.” (loc. cit.)

Soderstrom and Bjork (2015)

# …are different…



"The time-honored distinction between learning and performance dates back decades, spurred by early animal and motor-skills research that revealed that learning can occur even when no discernible changes in performance are observed." (loc. cit.)

Soderstrom and Bjork (2015)

# …and sometimes inversely related…



"More recently, the converse has also been shown—specifically, that improvements in performance can fail to yield significant learning—and, in fact, that certain manipulations can have opposite effects on learning and performance." (loc. cit.)

Soderstrom and Bjork (2015)

# How learning tasks can fail

Students are not engaged

or

Students are engaged but merely complying

or

Students are cognitively active – on the wrong stuff

or

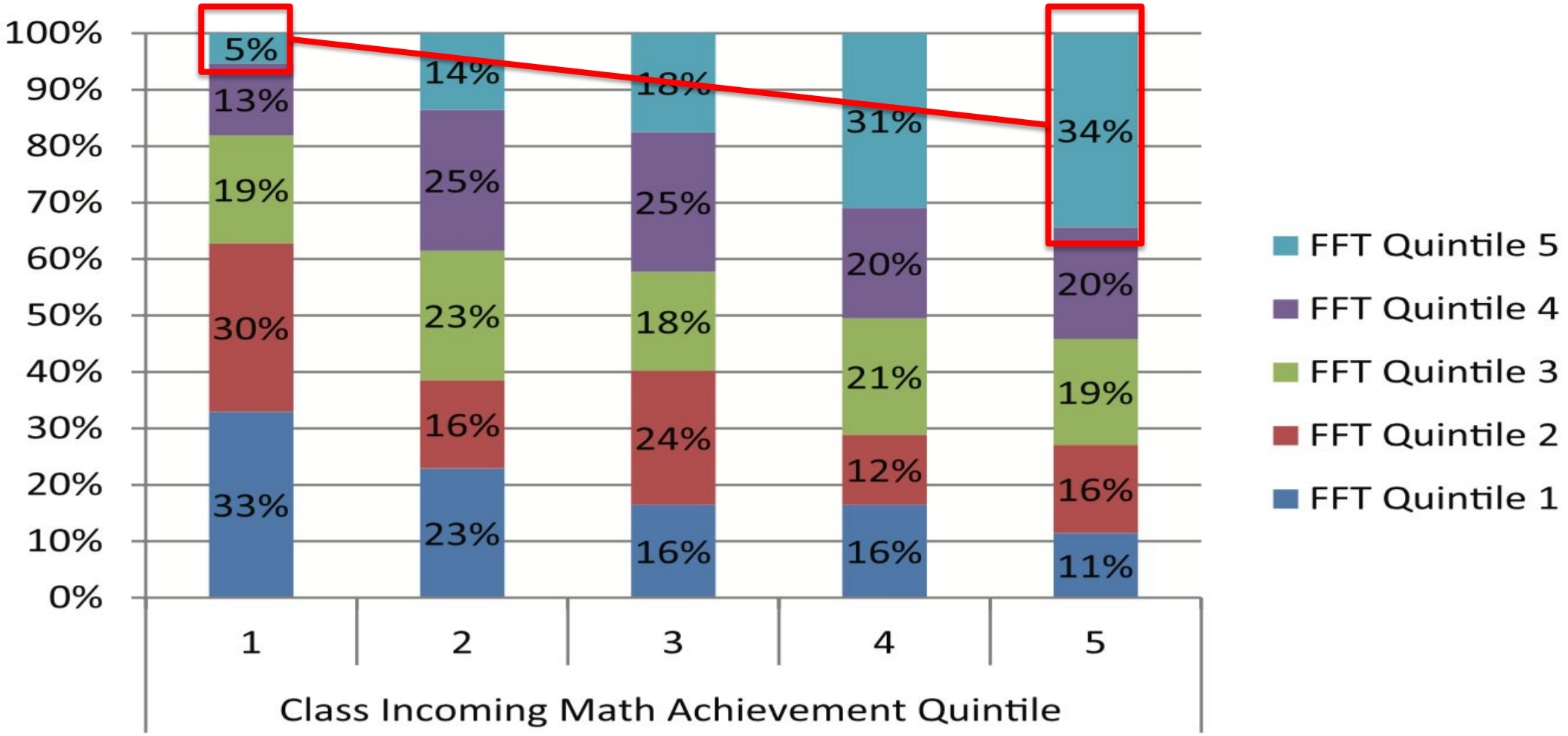
Students are cognitively active – on the right stuff, with cognitive overload

**For learning to occur, students need to be cognitively active, on the right stuff, without cognitive overload**

# Teaching is a wicked learning environment…

- Performance is not learning
- Weak signal; lots of noise

# Bias in lesson observations




Steinberg and Garrett (2016)

# Teaching is a wicked learning environment…

- Performance is not learning
- Weak signal; lots of noise
- Things that work in the short term can be ineffective, or even counter-productive, later

# Short-term and long-term effects



- Data: 10,534 students attending USAFA (2000-2007)
- Students randomly allocated to calculus instructors

| Instructors | |
|---|---|
| Less qualified, less experienced | More qualified, more experienced |
| Higher end of course scores | Lower end of course scores |
| Lower scores on follow-on courses | Higher scores on follow-on courses |
| Higher end of course evaluations | Lower end of course evaluations |

Carrell and West (2010)

# Teaching is a wicked learning environment…

- Performance is not learning
- Weak signal; lots of noise
- Things that work in the short term can be ineffective, or even counter-productive, later
- When things work—or don't work—without good theories, we may well learn the wrong lessons

# Expertise, practice, and teamwork

- Data on 203 surgeons performing coronary artery bypass grafts (CABG) on 38,577 patients in 43 PA hospitals during 1994 and 1995
- Baseline mortality rate: 3.1%
- Average: 37 CABG procedures per surgeon per quarter (standard deviation 17)
- For each extra CABG a surgeon performs each quarter, mortality rate drops by 0.015%
- Surgeon-specific mortality rates
  - High-volume (1 per weekday): 2.5%
  - Occasional (1 per month): 3.3%

Huckman and Pisano (2006)

# Hospital-specific effects



- At a specific hospital
  - For each extra CABG a surgeon performs each quarter *at that hospital*, the mortality rate drops by 0.02%
  - For each extra CABG a surgeon performs each quarter *at a different hospital*, the mortality rate **does not change**

Huckman and Pisano (2006)

# Teaching is a wicked learning environment…

- Performance is not learning
- Weak signal; lots of noise
- Things that work in the short term can be ineffective, or even counter-productive, later
- When things work—or don't work—without good theories, we may well learn the wrong lessons
- Things that work in one setting may not work in a different, but similar, setting

# The Tennessee STAR study



- Kindergarten students randomly assigned to
  - Classes of 22 to 26 with a single teacher
  - Classes of 22 to 26 with a teacher and an aide
  - Classes of 13 to 17 with a single teacher
- Benefits for students assigned to smaller classes
  - 3 months further ahead by end of 2nd grade
  - 2 x benefits for students from poorer homes, minorities
  - high school graduation rates 11 percentage points higher

Molnar et al. (1999); Grissmer (1999); Levin et al. (2007)

# Complications



- Maintaining randomization
  - Migration of high-SES students to smaller classes
- Generalizing to other settings
  - Unrepresentative participating schools

# Pause for reflection

- If you had to ask one question now, what would it be?

# Five questions to ask of research



- Does this solve a problem we have?
- If we do this, how much faster will our students learn?
- What will be the cost:
  - In money?
  - In teacher time?
- Can we implement it here?
- Do we know what to do?

# Does it solve a problem we have?

# “Getting teachers to come to school”



- 113 non-formal education centres run by Seva Mandir
  - In 56 centres, teachers were paid Rs.1,000 pcm
  - In 57 centres, teachers were paid according to attendance (2 time-stamped photos with class at start and end of day)
    - Rs.500 pcm for attendance up to 10 days plus
    - Rs.50 for each each day day over the 10 day threshold
  - Attendance rate:
    - Fixed pay group 58%
    - Incentive group 79%
  - For the incentive group
    - Increase in instructional time: 32%
    - Increase in annual progress: 25%

Duflo, Hanna, and Ryan (2012)

# If we do this, how much faster will our students learn?

# Understanding meta-analysis



- A technique for aggregating results from different studies *that use different outcome measures,* usually by converting results to a common measure, such as effect size
- Standardized effect size is typically defined as:

$$\frac{experimental\ mean - control\ group\ mean}{population\ standard\ deviation}$$

# Problems with meta-analysis

- Inappropriate comparisons
- Aptitude x treatment interactions
- The “file drawer” problem
- Variations in intervention quality
- Variation in population variability
- Selection of studies
- Sensitivity of outcome measures

# Inappropriate comparisons

# Inappropriate comparisons



- Net effects versus gross effects
- Cross-level comparisons
- “Business-as-usual” vs. alternative treatment

# Aptitude x treatment interactions

# Expertise reversal effects in cognitive load theory

- Solving problems in mathematics
  - Improves the performance of experts
  - Less effective with novices
- Worked examples
  - Degrades the performance of experts
  - More effective with novices

Sweller, Kalyuga, and Ayres (2011)

# The file-drawer problem

# Statistical power and effect size



| Average statistical power of experiments | Average effect size of significant studies |
|---|---|
| 100% | 0.40 |
| 80% | 0.47 |
| 60% | 0.53 |
| 40% | 0.59 |
| 20% | 0.68 |

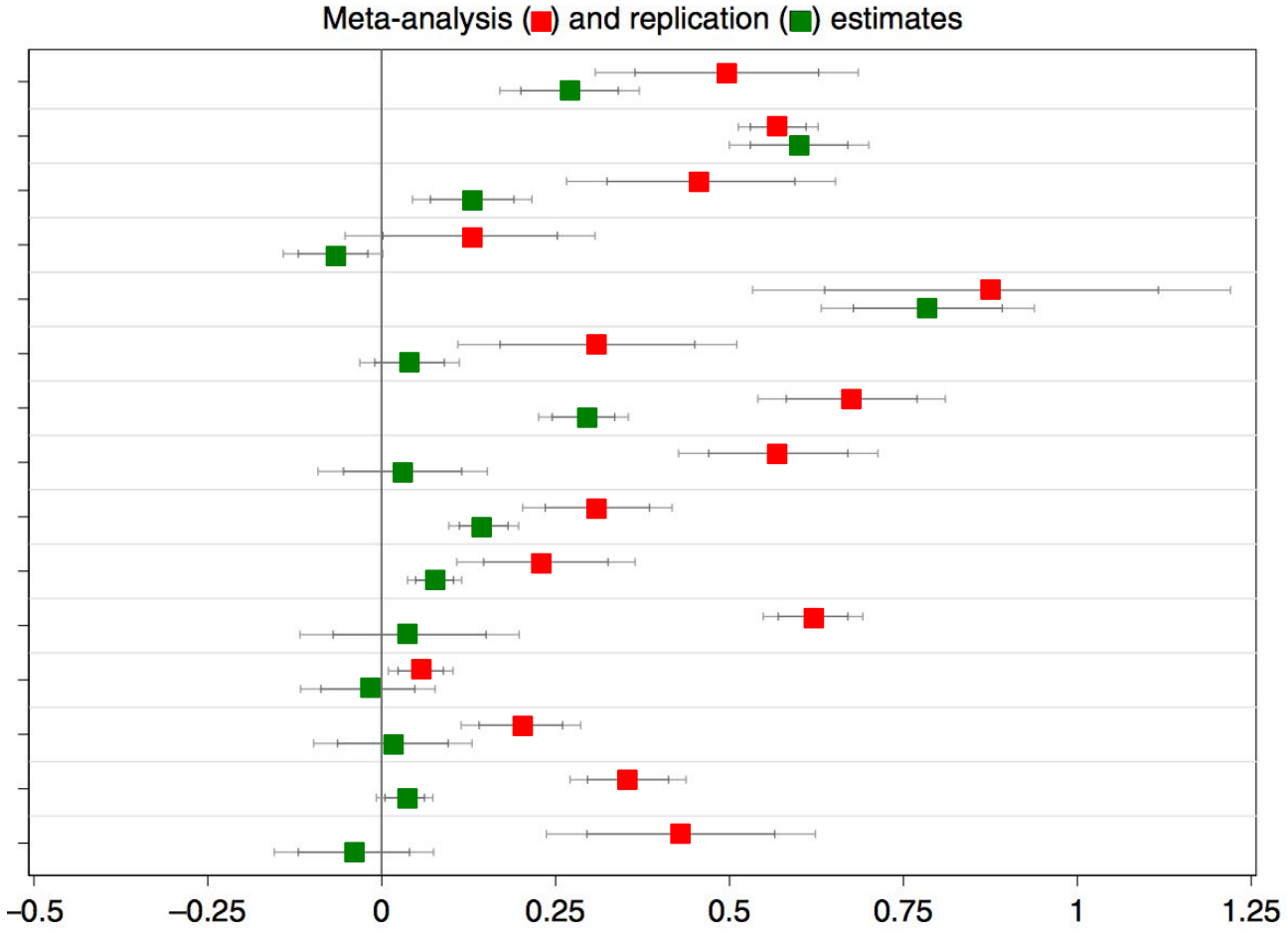
# Meta-analysis and replications

- Review of 15 meta-analyses in different areas of psychology
- Comparison of effect sizes
  - Those reported in the original meta-analysis
  - Those found in pre-registered, multi-site studies

Kvarven, Strømland and Johannesson (2020)

# Meta-analyses and replication studies






Effect sizes in pre-registered multi-site studies were only *one-third* the magnitude of those reported in the published meta-analyses

Kvarven, Strømland and Johannesson (2020)

# Variation in intervention quality

# Quality

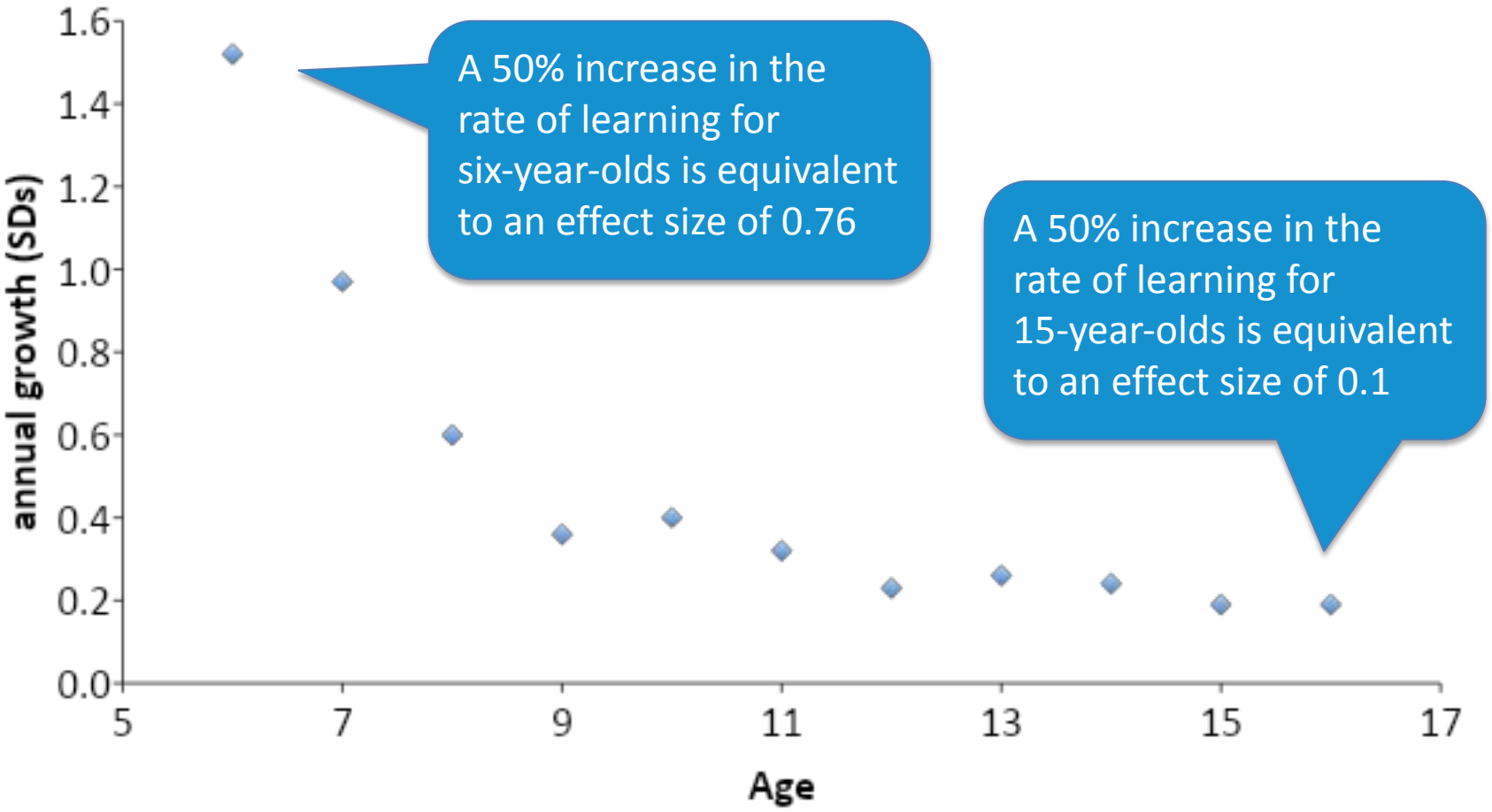


- Interventions vary in their
  - Duration
  - Intensity
    - class size reduction by 20%, 30%, or 50%
    - response to intervention
  - Collateral effects
    - assignment of teachers

# Variation in variability

# Annual growth in achievement, by age






Bloom, Hill, Black, and Lipsey (2008)

# Variation in variability



- Studies with younger children will produce larger effect size estimates
- Studies with restricted populations (e.g., children with special needs, gifted students) will produce larger effect size estimates

# Selection of studies

# “The average effect does not exist”

- A true average effect size needs to include all of the following:
  a) Conducted studies that were significant and reported
  b) Conducted studies that were non-significant and reported
  c) Conducted studies that were significant and not reported
  d) Conducted studies that were non-significant and not-reported
  e) Non-conducted studies that would have been significant if they had been conducted
  f) Non-conducted studies that would have been non-significant if they had been conducted

Leif Nelson (DataColada 2018)

# Sensitivity to instruction

# Sensitivity of outcome measures

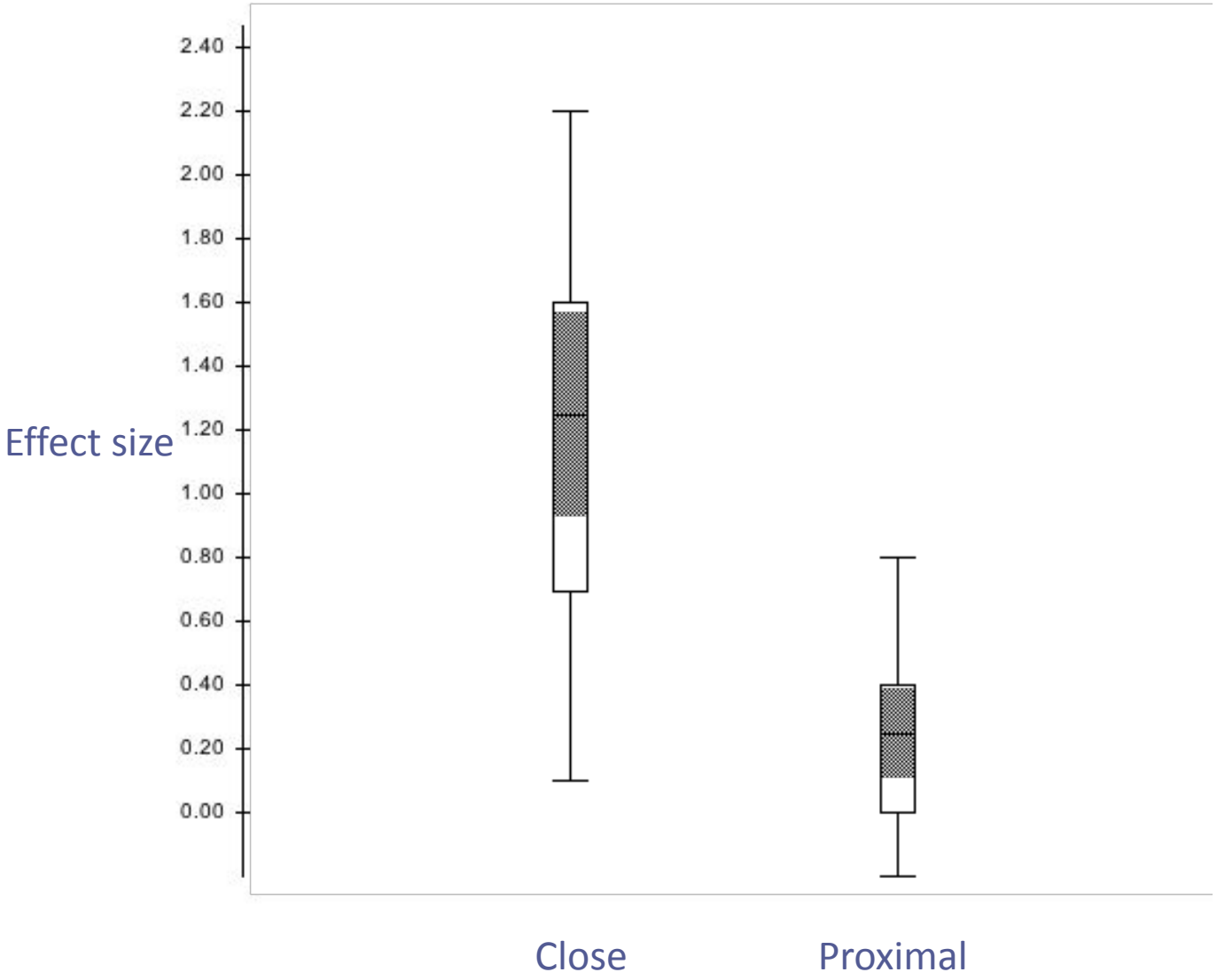


- Distance of assessment from the curriculum
  - Immediate
    - e.g., science journals, notebooks, and classroom tests
  - Close
    - e.g., where an immediate assessment asked about number of pendulum swings in 15 seconds, a close assessment asks about the time taken for 10 swings
  - Proximal
    - e.g., if an immediate assessment asked students to construct boats out of paper cups, the proximal assessment would ask for an explanation of what makes bottles float
  - Distal
    - e.g., where the assessment task is sampled from a different domain and where the problem, procedures, materials and measurement methods differed from those used in the original activities
  - Remote
    - standardized national achievement tests.

Ruiz-Primo, Shavelson, Hamilton, and Klein (2002)

# Impact of sensitivity to instruction





# Outcome measures



- Strictly speaking, effect sizes are justifiable only when they are unnecessary
- Any improvement in educational processes manifests itself by
  - Students learning more in a given time period
  - Students taking less time to learn a given amount
  - In other words, *an increase in the rate of learning*
- The only useful metric is an indication of that rate increase (e.g., extra months of learning per year)

# So, when can you trust meta-analyses?



- When the exact nature of the intervention is clear, and consistent across studies
- When publication bias is investigated, and ideally, corrected
- When moderator analyses are conducted for all plausible moderators of effect, such as
  - Pre-registered vs unregistered studies
  - Age
  - Sensitivity of outcome measures to intervention effects

# And when can you trust meta-meta-analyses?



- When you can trust all or most of the constituent meta-analyses
- Issues that are particularly important
  - Inappropriate comparisons
  - Which studies get done?

# Can we implement it here?

# Back to the Tennessee STAR study



- Kindergarten students randomly assigned to
  - Classes of 22 to 26 with a single teacher
  - Classes of 22 to 26 with a teacher and an aide
  - Classes of 13 to 17 with a single teacher
- Benefits for students assigned to smaller classes
  - 3 months further ahead by end of 2nd grade
  - 2 x benefits for students from poorer homes, minorities
  - high school graduation rates 11 percentage points higher

Molnar et al. (1999); Grissmer (1999); Levin et al. (2007)

# Back to the Tennessee STAR study



- Maintaining randomization
  - Migration of high-SES students to smaller classes
- Generalizing to other settings
  - Unrepresentative participating schools
  - Availability of additional teachers

# No “gold standard” in research

- “What works” versus “what worked”
  - “Hume argued that there is no rational argument to infer knowledge of the unobserved from knowledge of the observed.” (Cartwright, 2019)
- Randomized-control trials
  - *can* show that the difference between the control and treatment groups is probably real
  - *cannot* show that generalizations to other settings are warranted

# Do we know what to do?

# “Not even wrong”

- “…most claims in the literature are so critically underspecified that attempts to empirically evaluate them are doomed to failure — they are *not even wrong.”* (Scheel, 2021)

# Four things to look for

- Adequate theorization
- Standardized interventions
- Pre-registered studies
- Intention-to-treat analyses

# Summary

- Improvement involves trade-offs
  - No perfect solutions
  - Only different trade-offs
- Research can guide these efforts
  - Does it solve a problem we have?
  - How much faster will our students learn?
  - What will be the cost?
  - Can we implement it here?
  - Do we know what to do?
- Educators need to be *critical consumers* of research