Evidence in Education

Matthew Inglis

Loughborough University
Mathematics Education Centre

Plan

- A story about teaching mathematical proofs.
- How can we evaluate teaching innovations, or teaching more generally?
- I will argue that we can't evaluate teaching in real world contexts, so we need to adopt a different approach: learning from research.
- I'll contrast two approaches to learning from research: an approach based on educational findings and an approach based on educational theory.

Proofs in lectures

Understanding proofs from lectures might be difficult because:

- Following live explanation requires rapid recall of basic knowledge, recognition and validation of logical deductions and recognition of larger scale structure.
- Explanation is ephemeral - it is no longer there when a student studies their lecture notes.

Lara Alcock designed a resource...

E-Proof Demo

Students liked them…

Student comments from feedback forms and focus groups:

- “I found hearing the lecturer explaining each line individually helpful in understanding particular parts and how they relate to the entire proof.”
- “e-Proof good for understanding the situation, rather than copying from the board etc.”
- “Having proofs online does make it easier to go at my own pace whilst still having the lecturer explain each part.”

Students liked them…

We used an online feedback form. Eight statements like:

- “e-Proofs helped me understand where different parts of a proof come from and how they fit together”

Each scored 0 (negative) to 4 (positive).

Uniformly positive responses:

Mean score 25.5 (out of 32), 95% CI [24.5, 26.6]

Educational technologists liked them…

Based on feedback from students and colleagues, Lara applied for funding.

There was a robust review process.

And an interview.

Finally Loughborough was awarded an £80k grant by JISC to develop further e-proofs.

Evidence for E-Proofs

Evidence that e-Proofs are effective:

- Qualitative feedback from students
- Quantitative feedback from students
- Peer observations of e-proof use
- Expert peer review of a funding proposal

Unfortunately, e-Proofs don’t work.

Somali Roy's PhD

An example study:

Participants: 49 undergraduates.

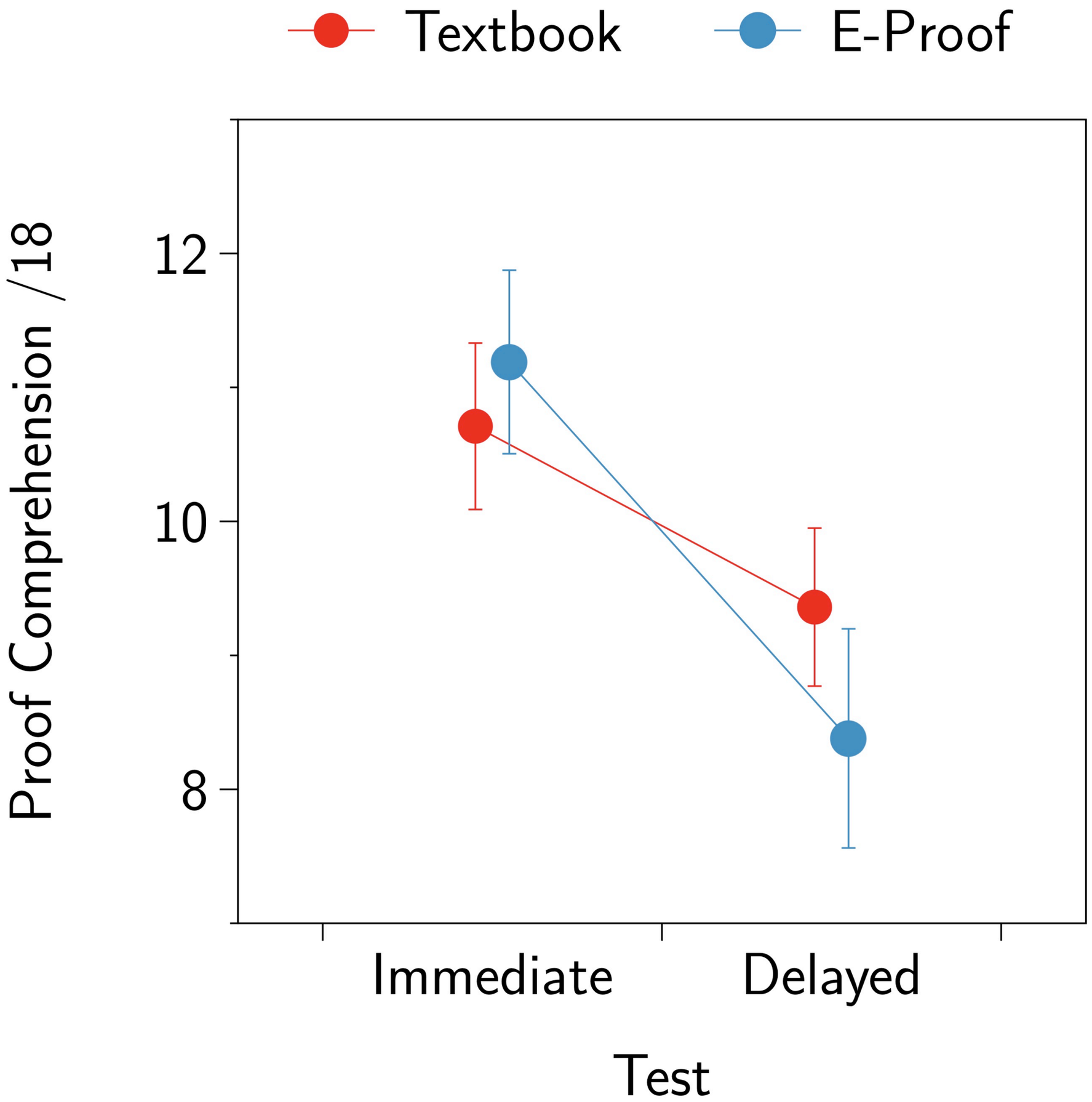
Two proof versions: e-Proof; textbook.

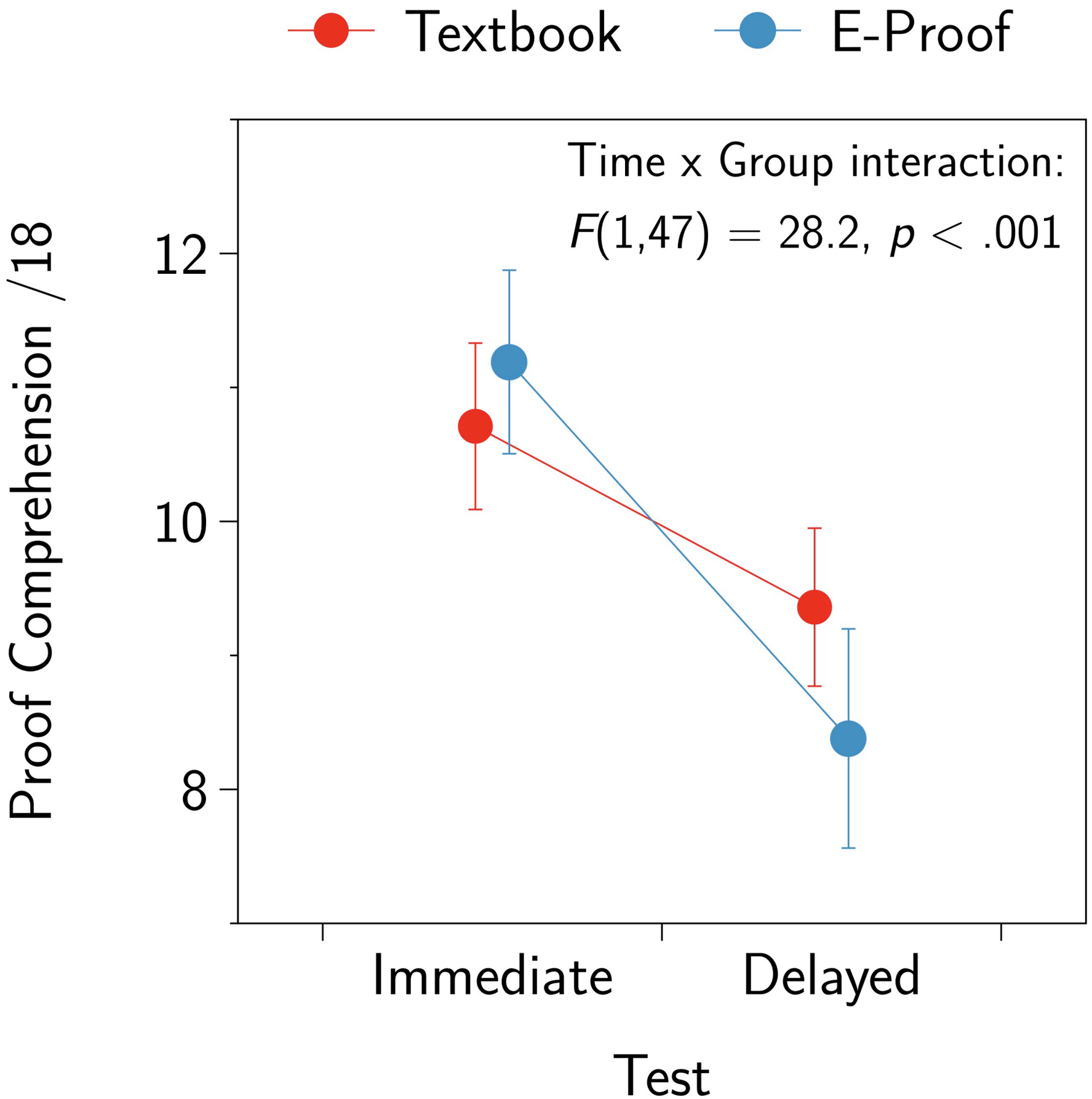
Identical to e-Proof version, except no audio or graphics (just text)

Comprehension test: score out of 18.

Time: 15 minutes to study proof; 30 minutes to complete comprehension test.

Delayed post-test: same comprehension test; 20 minutes to complete. Two weeks later.


Textbook
E-Proof
Proof Comprehension /18
12
10
8
Immediate
Delayed
Test


Textbook
E-Proof
Time x Group interaction:
$F(1,47) = 28.2,\ p < .001$
Proof Comprehension /18
12
10
8
Immediate
Delayed
Test

E-Proofs

- Robust evidence that e-Proofs led to worse retention than reading a textbook.
- Why?
- Both Somali Roy and Mark Hodds worked on answering this question for their PhDs.
- Answer is, roughly, that you benefit from having to construct your own explanations.
- E-Proofs were too easy.
- Students learned less, **but felt like they'd learned more**.
- Full story in *Notices of the AMS* vol 62(7).

Evidence in Education

- When looking at e-proofs, we collected all the standard measures of educational effectiveness used in higher education:
    - qualitative student feedback
    - quantitative student feedback
    - expert peer review
- All pointed in the same direction.
- But actually measuring student learning revealed that all these methods were misleading.

Evidence in Education

This is all very well known to education researchers.

Range of methods used to evaluate educational quality in HE:

- Peer/expert classroom observation
- Student ratings
- Analysis of educational materials
- Lecturer portfolios

I'll review the research on these

Classroom Observations

Do We Know a Successful Teacher When We See One? Experiments in the Identification of Effective Teachers

Journal of Teacher Education
62(4) 367–382

DOI: 10.1177/0022487110390221
http://jte.sagepub.com

Michael Strong[1], John Gargani[2], and Özge Hacifazlioğlu[3]

Abstract

The authors report on three experiments designed to (a) test under increasingly more favorable conditions whether judges can correctly rate teachers of known ability to raise student achievement, (b) inquire about what criteria judges use when making their evaluations, and (c) determine which criteria are most predictive of a teacher's effectiveness. All three experiments resulted in high agreement among judges but low ability to identify effective teachers. Certain items on the established measure that are related to instructional behavior did reliably predict teacher effectiveness. The authors conclude that (a) judges, no matter how experienced, are unable to identify successful teachers; (b) certain cognitive operations may be contributing to this outcome; (c) it is desirable and possible to develop a new measure that does produce accurate predictions of a teacher's ability to raise student achievement test scores.

Keywords

teacher effectiveness, teacher evaluation, classroom observation, value-added

Classroom Observations

Do We Know a Successful Teacher When We See One? Experiments in the Identification of Effective Teachers

Journal of Teacher Education
62(4) 367–382

DOI: 10.1177/0022487110390221
http://jte.sagepub.com

Used value-added measures to identify a group of 'effective' teachers and 'ineffective' teachers (separated by 0.5 SDs in learning gain).

Showed videos of them teaching to **experienced teachers and headteachers** and asked them to identify which was in which group.

Teachers cannot do this.

Across three experiments they performed worse than 50% (score expected if they had guessed).

Mich

Abstr

The a
judges
use wh
three
the es
conclu
operat
produ

Keyw

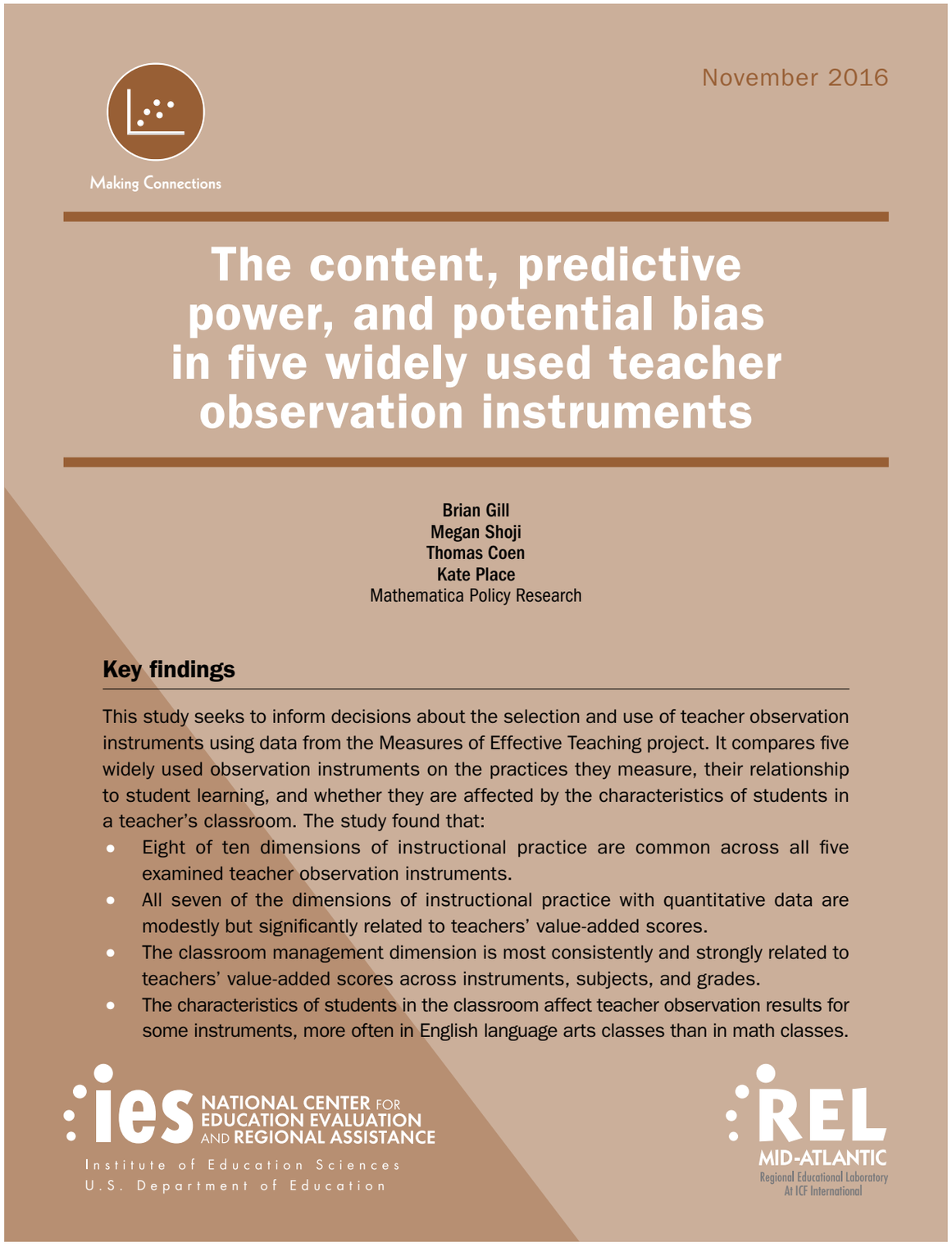
teache

Classroom Observations

Measures of Effective Teaching project, funded by the Gates Foundation.

Large project which aimed to produce and evaluate reliable methods of measuring teacher quality.

Gold standard in lesson observation: much more sophisticated than anything we do in HE.



November 2016

Making Connections

The content, predictive power, and potential bias in five widely used teacher observation instruments

Brian Gill
Megan Shoji
Thomas Coen
Kate Place
Mathematica Policy Research

Key findings

This study seeks to inform decisions about the selection and use of teacher observation instruments using data from the Measures of Effective Teaching project. It compares five widely used observation instruments on the practices they measure, their relationship to student learning, and whether they are affected by the characteristics of students in a teacher's classroom. The study found that:

- Eight of ten dimensions of instructional practice are common across all five examined teacher observation instruments.
- All seven of the dimensions of instructional practice with quantitative data are modestly but significantly related to teachers' value-added scores.
- The classroom management dimension is most consistently and strongly related to teachers' value-added scores across instruments, subjects, and grades.
- The characteristics of students in the classroom affect teacher observation results for some instruments, more often in English language arts classes than in math classes.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

REL MID-ATLANTIC
Regional Educational Laboratory
At ICF International

Classroom Observations

Table 4. Summary results for the strength of relationship between teachers' cross-instrument observation dimension scores and their value added to student learning

Dimension	Adjusted correlation to underlying value added score
Supportive learning environment	.18
Classroom management	.28
Student intellectual engagement with content	.22
Lesson structure and facilitation	.18
Content understanding	.13
Language and discourse	.14
Feedback and assessment	.20

For comparison:

Correlation between height and intelligence ≈ .20

Correlation between swearing frequency and fear of pain ≈ .17

Correlation between SNIP and REF article quality ≈ .33

Evidence in Education

Range of methods used to evaluate educational quality in HE:

- ~~Peer/expert classroom observation~~
- Student ratings
- Analysis of educational materials
- Lecturer portfolios

I'll review the research on these

Student Feedback/Ratings

Studies in Educational Evaluation 54 (2017) 22–42

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related

Bob Uttl*, Carmela A. White[1], Daniela Wong Gonzalez[2]

Department of Psychology, Mount Royal University, Canada

A R T I C L E I N F O

Article history:
Received 23 February 2016
Received in revised form 10 August 2016
Accepted 17 August 2016
Available online 19 September 2016

Keywords:
Meta-analysis of student evaluation of teaching
Multisection studies
Validity
Teaching effectiveness
Evaluation of faculty
SET and learning correlations

A B S T R A C T

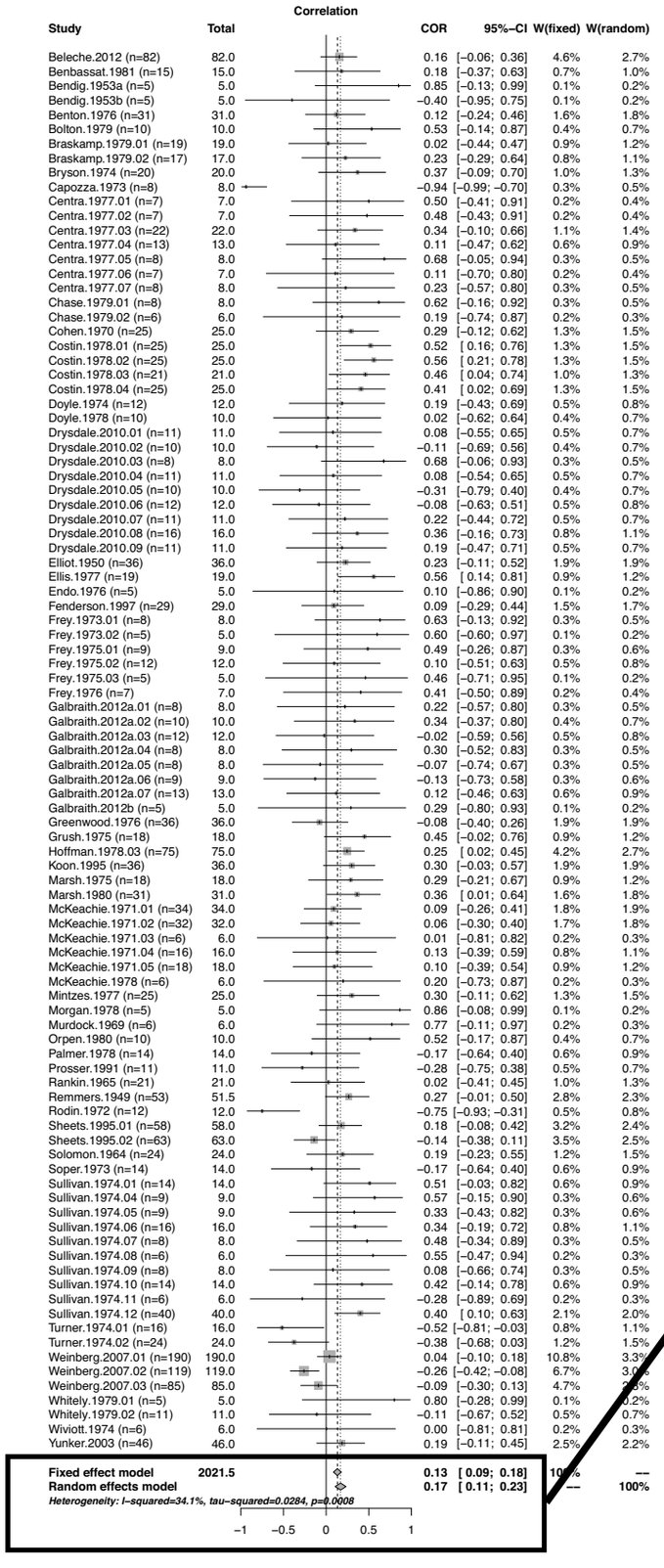
Student evaluation of teaching (SET) ratings are used to evaluate faculty's teaching effectiveness based on a widespread belief that students learn more from highly rated professors. The key evidence cited in support of this belief are meta-analyses of multisection studies showing small-to-moderate correlations between SET ratings and student achievement (e.g., Cohen, 1980, 1981; Feldman, 1989). We re-analyzed previously published meta-analyses of the multisection studies and found that their findings were an artifact of small sample sized studies and publication bias. Whereas the small sample sized studies showed large and moderate correlation, the large sample sized studies showed no or only minimal correlation between SET ratings and learning. Our up-to-date meta-analysis of all multisection studies revealed no significant correlations between the SET ratings and learning. These findings suggest that institutions focused on student learning and career success may want to abandon SET ratings as a measure of faculty's teaching effectiveness.

"*For every complex problem there is an answer that is clear, simple, and wrong.*" H. L. Mencken

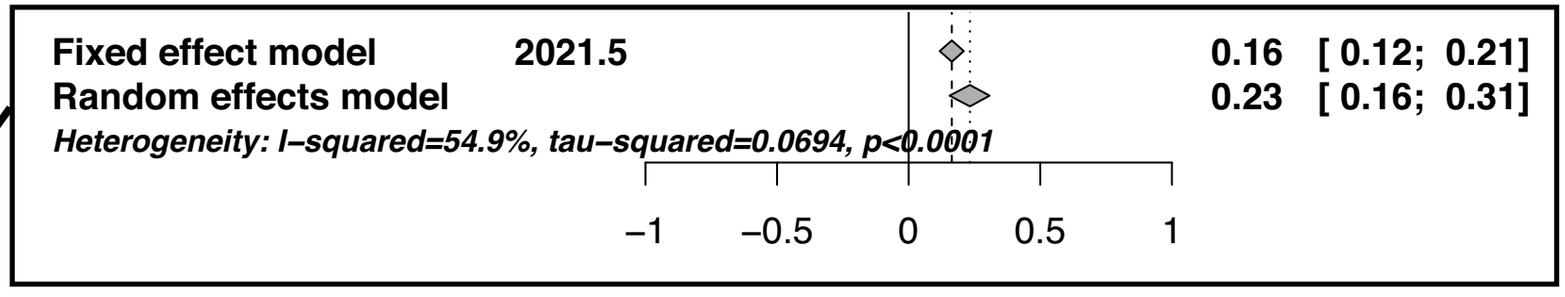
Student Evaluation of Teaching (SET) ratings are used to

and courses (e.g., organization, difficulty) (Feldman, 1989; Spooren, Brockx, & Mortelmans, 2013). The ratings for each course/class are summarized, typically by calculating mean ratings

Student Feedback/Ratings



- Meta-analysis of 97 correlational studies.
- Found a correlation of r ≈ .2 (varies a bit depending on statistical assumptions).
- Re-analysed Cohen's 1980s meta-analysis and found it was misleading due to small samples and publication bias.



Student Feedback/Ratings

- Correlational studies on student feedback are easy to run.
- Much harder to establish that teachers with good feedback scores *cause* better student learning.
- Best design:

    1. Randomly allocate students to lecturers;
    2. Lecturers teach the same material;
    3. Students take the same examination;
    4. Students move to a subsequent follow-on module, study it and take another examination.
    5. Does the lecturer's student evaluations on the prerequisite module predict scores on the subsequent module?

Student Feedback/Ratings

- Correlational studies on student feedback are easy to run.
- Much harder to establish that teachers with good feedback scores *cause* better student learning.

This study has only been done twice

1. Randomly allocate students to lecturers;
2. Lecturers teach the same material;
3. Students take the same examination;
4. Students move to a subsequent follow-on module, study it and take another examination.
5. Does the lecturer's student evaluations on the prerequisite module predict scores on the subsequent module?

Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors

Scott E. Carrell

University of California, Davis and National Bureau of Economic Research

James E. West

U.S. Air Force Academy

In primary and secondary education, measures of teacher quality are often based on contemporaneous student performance on standardized achievement tests. In the postsecondary environment, scores on student evaluations of professors are typically used to measure teaching quality. We possess unique data that allow us to measure relative student performance in mandatory follow-on classes. We compare metrics that capture these three different notions of instructional quality and present evidence that professors who excel at promoting contemporaneous student achievement teach in ways that improve their student evaluations but harm the follow-on achievement of their students in more advanced classes.

Economics of Education Review 41 (2014) 71–88

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Evaluating students' evaluations of professors☆

Michela Braga [a], Marco Paccagnella [b], Michele Pellizzari [c,*]

[a] Bocconi University, Department of Economics, Italy
[b] Bank of Italy, Trento Branch, Italy
[c] University of Geneva, Institute of Economics and Econometrics, Switzerland

ARTICLE INFO

Article history:
Received 2 August 2013
Received in revised form 22 April 2014
Accepted 23 April 2014
Available online 5 May 2014

JEL classification:
I20
M55

Keywords:
Teacher quality
Postsecondary education
Students' evaluations

ABSTRACT

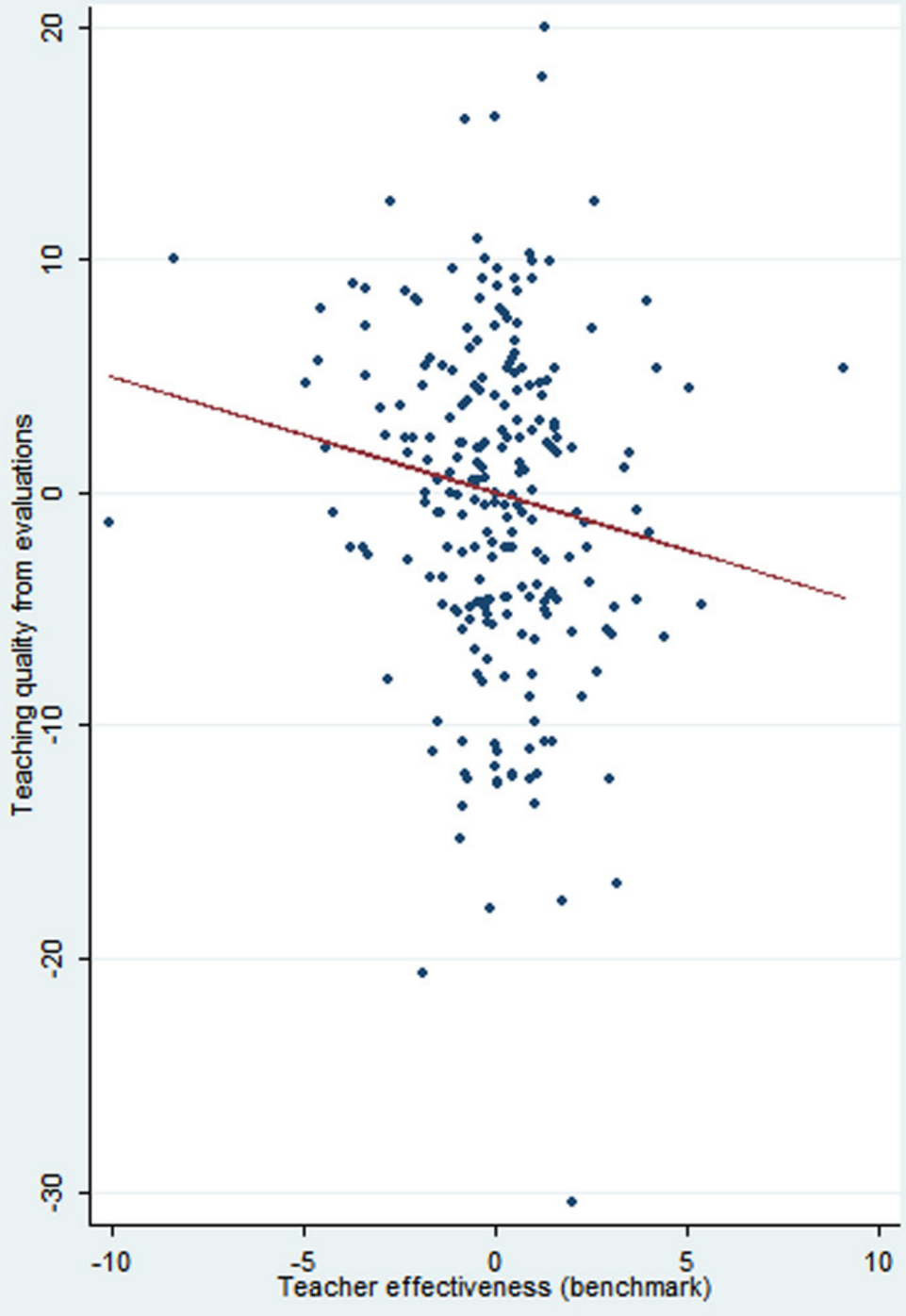
This paper contrasts measures of teacher effectiveness with the students' evaluations for the same teachers using administrative data from Bocconi University. The effectiveness measures are estimated by comparing the performance in follow-on coursework of students who are randomly assigned to teachers. We find that teacher quality matters substantially and that our measure of effectiveness is negatively correlated with the students' evaluations of professors. A simple theory rationalizes this result under the assumption that students evaluate professors based on their realized utility, an assumption that is supported by additional evidence that the evaluations respond to meteorological conditions.

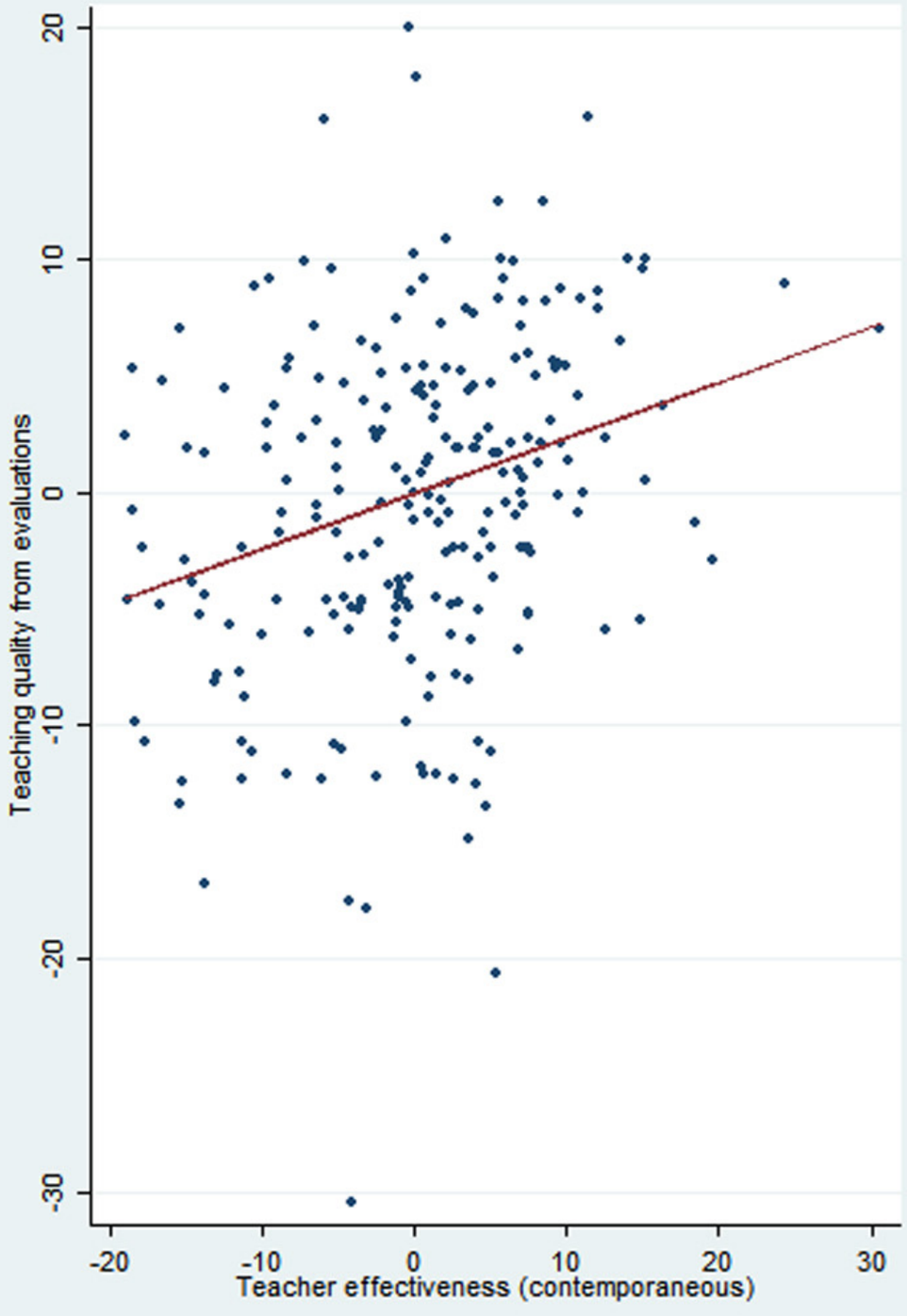
Student Feedback/Ratings

- Braga et al.: “We find that teacher quality matters substantially and that our measure of effectiveness is **negatively** correlated with the students’ evaluations of professors.”
- Carrell & West: “We present evidence that professors who excel at promoting contemporaneous student achievement teach in ways that improve their student evaluations but **harm** the follow-on achievement of their students in more advanced classes.”

Subsequent Module

Current Module





Observations are weighted by squared root of the number of collected questionnaires

Student Feedback/Ratings

Notable that the only two experimental evaluations of student feedback/ratings find the same results:

- student evaluations show weak positive correlations with current achievement;
- student evaluations show weak negative correlations with subsequent achievement.

Evidence in Education

Range of methods used to evaluate educational quality in HE:

- ~~Peer/expert classroom observation~~
- ~~Student ratings~~
- Analysis of educational materials
- Lecturer portfolios

No reason to suppose these are better

Summary

- A fundamental problem with teaching is that we have no valid way of assessing teaching quality, beyond assessing learning gains.
- But assessing learning gains, especially in higher education (no control groups), is really hard. Especially true when some effects are only revealed after a delay (e-Proofs, student feedback).
- The methods we try to use in higher education are known to be invalid.
- What's the implications of this?

A Serious Problem

- What do I do if I want to change my teaching practice?
- If we can't evaluate teaching quality validly, how can I know if my change has improved my students' learning?
- Standard advice is to be a "reflective practitioner" and evaluate whether or not the change has helped my students.
- I think this is very bad advice: it's asking you to do something we know you can't do.

Do not despair!

- So if I can't trust student feedback or advice from observers, what can I do?
- I can compare my proposed intervention with research findings.
- But how to do this?
- Particularly challenging, as there are lots of internal debates within education about the 'right way' of communicating research findings to practitioners.
- There are two broad approaches used to communicate educational research findings to practitioners:

Do not despair!

- So if I can't trust student feedback or advice from observers, what can I do?
- I can compare my proposed intervention with research findings.
- But how to do this?

- Particularly challenging, as there are lots of internal debates within education about the 'right way' of communicating research findings to practitioners.
- There are two broad approaches used to communicate educational research findings to practitioners:

    1. findings-based approach
    2. theory-based approach

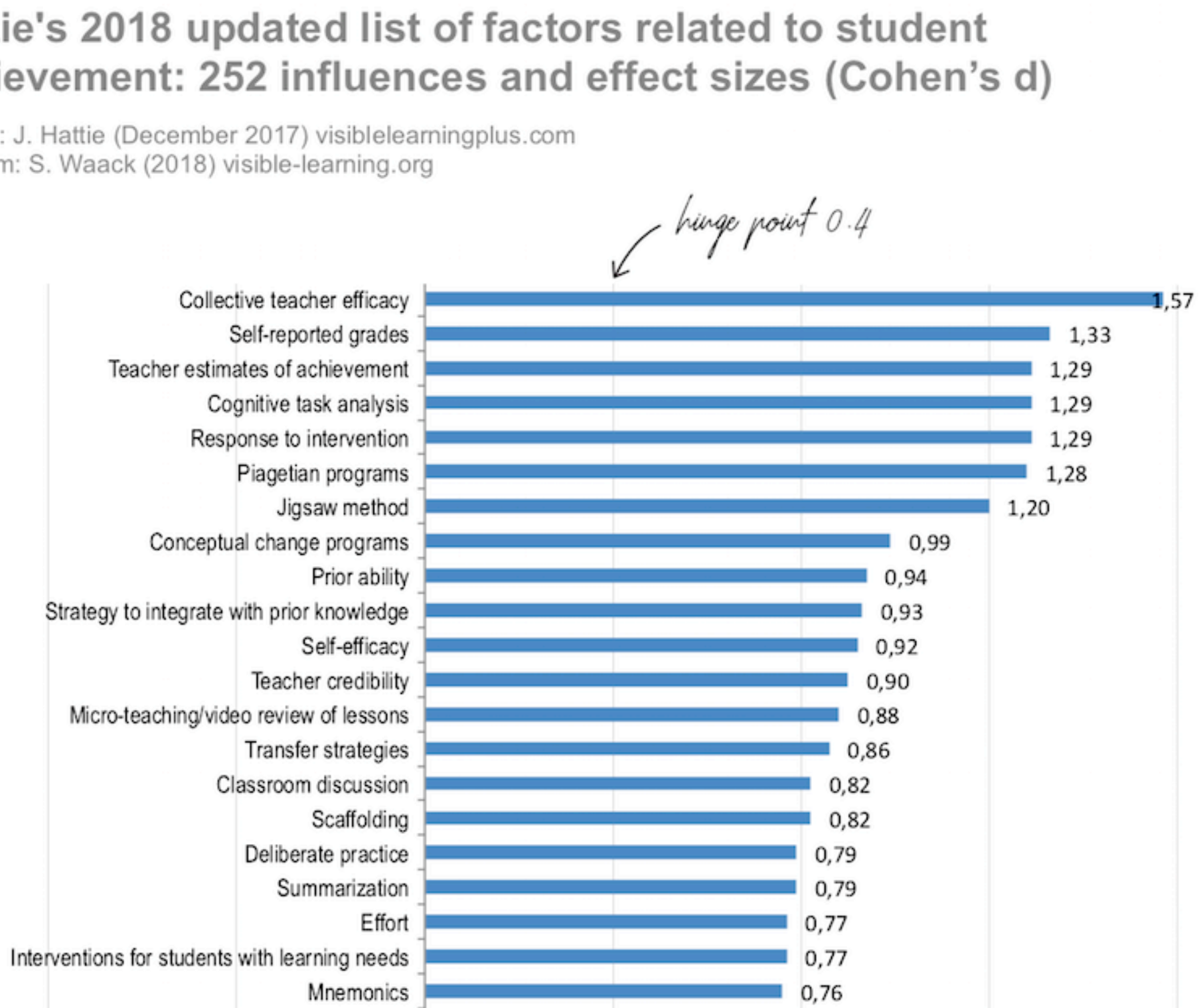
Findings-Based Method

Approach:

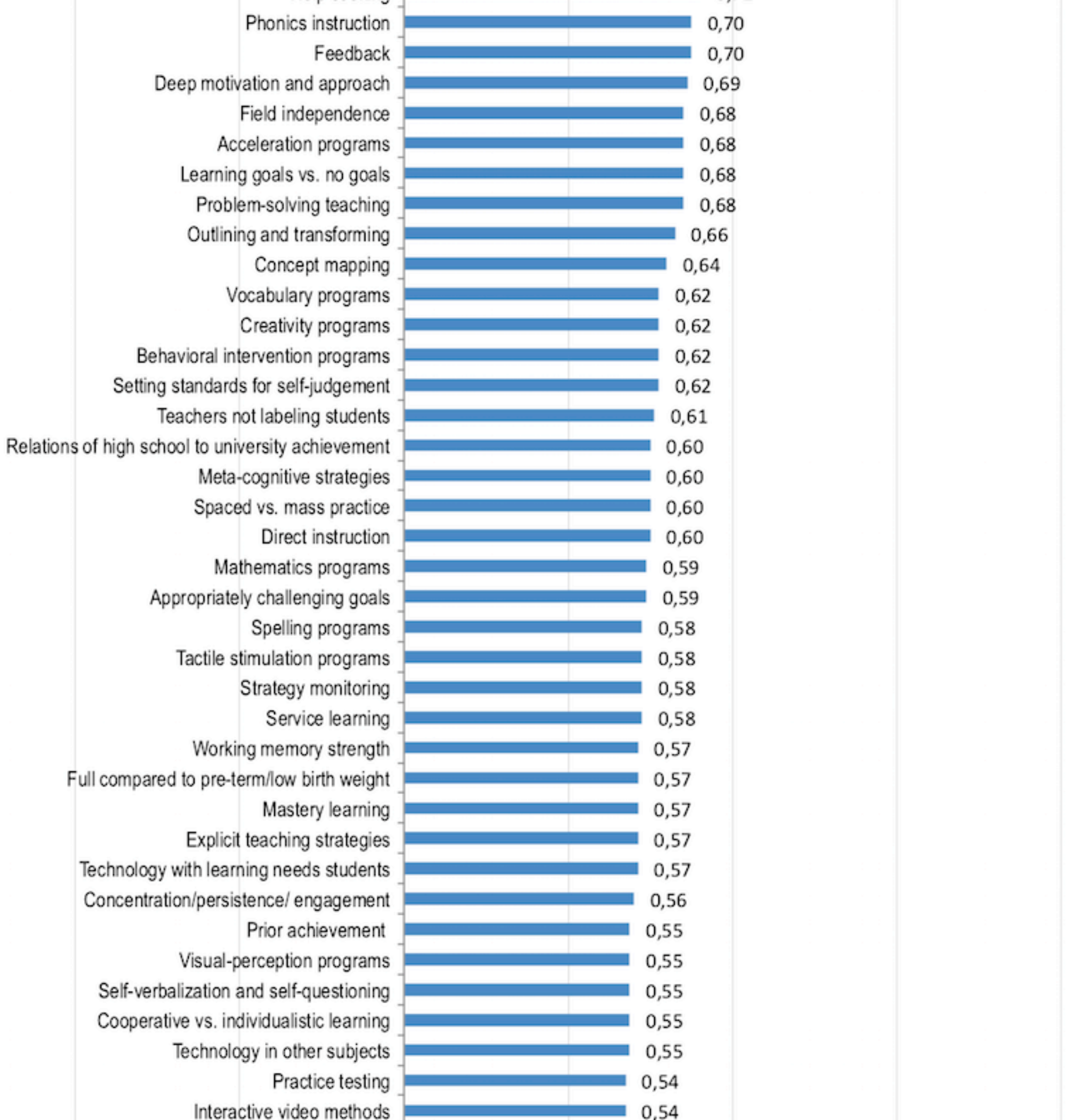
- There have been lots of studies that did evaluate educational interventions properly.
- Let’s group these studies into categories and calculate the average effectiveness of each category.
- Then you as a practitioner can compare your proposed intervention with the effectiveness of the category it falls into.

Findings-Based Approach

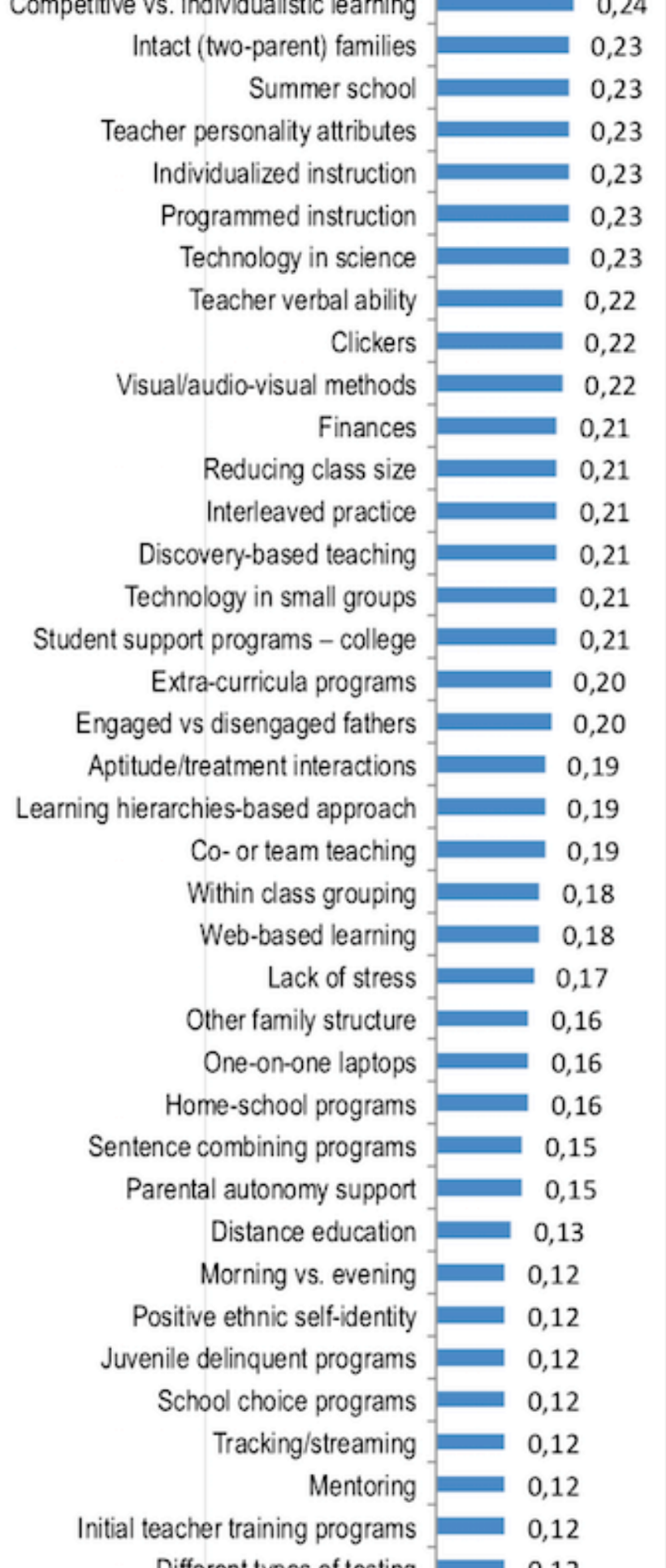
John Hattie's *Visible Learning* is the classic version of this approach.



John Hattie's *Visible Learning* is the classic version of this approach.



John Hattie's *Visible Learning* is the classic version of this approach.



Findings-Based Approach

used by 64% of headteachers

- Other school-level versions of this approach: Sutton Trust and EEF Toolkit.

- Undergraduate-level version: Michael Schneider and Franzis Prekel's *Psychological Bulletin* paper.

used by ??% of Pro-Vice Chancellors for Teaching

Psychological Bulletin

0033-2909/17/$12.00 http://dx.doi.org/10.1037/bul0000098

Variables Associated With Achievement in Higher Education: A Systematic Review of Meta-Analyses

Michael Schneider and Franzis Preckel
University of Trier

The last 2 decades witnessed a surge in empirical studies on the variables associated with achievement in higher education. A number of meta-analyses synthesized these findings. In our systematic literature review, we included 38 meta-analyses investigating 105 correlates of achievement, based on 3,330 effect sizes from almost 2 million students. We provide a list of the 105 variables, ordered by the effect size, and summary statistics for central research topics. The results highlight the close relation between social interaction in courses and achievement. Achievement is also strongly associated with the stimulation of meaningful learning by presenting information in a clear way, relating it to the students, and using conceptually demanding learning tasks. Instruction and communication technology has comparably weak effect sizes, which did not increase over time. Strong moderator effects are found for almost all instructional methods, indicating that how a method is implemented in detail strongly affects achievement. Teachers with high-achieving students invest time and effort in designing the microstructure of their courses, establish clear learning goals, and employ feedback practices. This emphasizes the importance of teacher training in higher education. Students with high achievement are characterized by high self-efficacy, high prior achievement and intelligence, conscientiousness, and the goal-directed use of learning strategies. Barring the paucity of controlled experiments and the lack of meta-analyses on recent educational innovations, the variables associated with achievement in higher education are generally well investigated and well understood. By using these findings, teachers, university administrators, and policymakers can increase the effectivity of higher education.

Keywords: academic achievement, meta-analysis, tertiary education, instruction, individual differences

Supplemental materials: http://dx.doi.org/10.1037/bul0000098.supp

Findings-Based Approach

Relies upon the calculation and comparison of some kind of standardised effect size (typically Cohen's *d*).

$$d = \frac{\bar{x}_{\text{exp}} - \bar{x}_{\text{con}}}{s_{\text{p}}}$$

Findings-Based Approach

- *Basic Problem:* Standardised effect sizes are biased by anything that alters the variance but not the effect.
- Unfortunately, a lot of things alter the variance without altering the effect (e.g., achievement range in the group, length of the test, choice of covariates, etc).

  Example: Inglis & Alcock (2012) found $d = 0.95$ using a test with 14 items; randomly deleting 7 items gives an average $d = 0.81$; randomly deleting 9 gives an average $d = 0.70$ (EEF: difference of “two months progress”).

British Journal of Psychology (2009), 100, 603–617

The British Psychological Society

www.bpsjournals.co.uk

Standardized or simple effect size: What should be reported?

Thom Baguley*

Division of Psychology, Nottingham Trent University, Nottingham, UK

It is regarded as best practice for psychologists to report effect size when disseminating quantitative research findings. Reporting of effect size in the psychological literature is patchy – though this may be changing – and when reported it is far from clear that appropriate effect size statistics are employed. This paper considers the practice of reporting point estimates of standardized effect size and explores factors such as reliability, range restriction and differences in design that distort standardized effect size unless suitable corrections are employed. For most purposes simple (unstandardized) effect size is more robust and versatile than standardized effect size. Guidelines for deciding what effect size metric to use and how to report it are outlined. Foremost among these are: (i) a preference for simple effect size over standardized effect size, and (ii) the use of confidence intervals to indicate a plausible range of values the effect might take. Deciding on the appropriate effect size statistic to report always requires careful thought and should be influenced by the goals of the researcher, the context of the research and the potential needs of readers.

Tukey:

THE SOCIETY FOR THE SUPPRESSION OF THE CORRELATION COEFFICIENT

ABOUT BIBLIOGRAPHY

The society's guiding principle is that most correlation coefficients should never be calculated

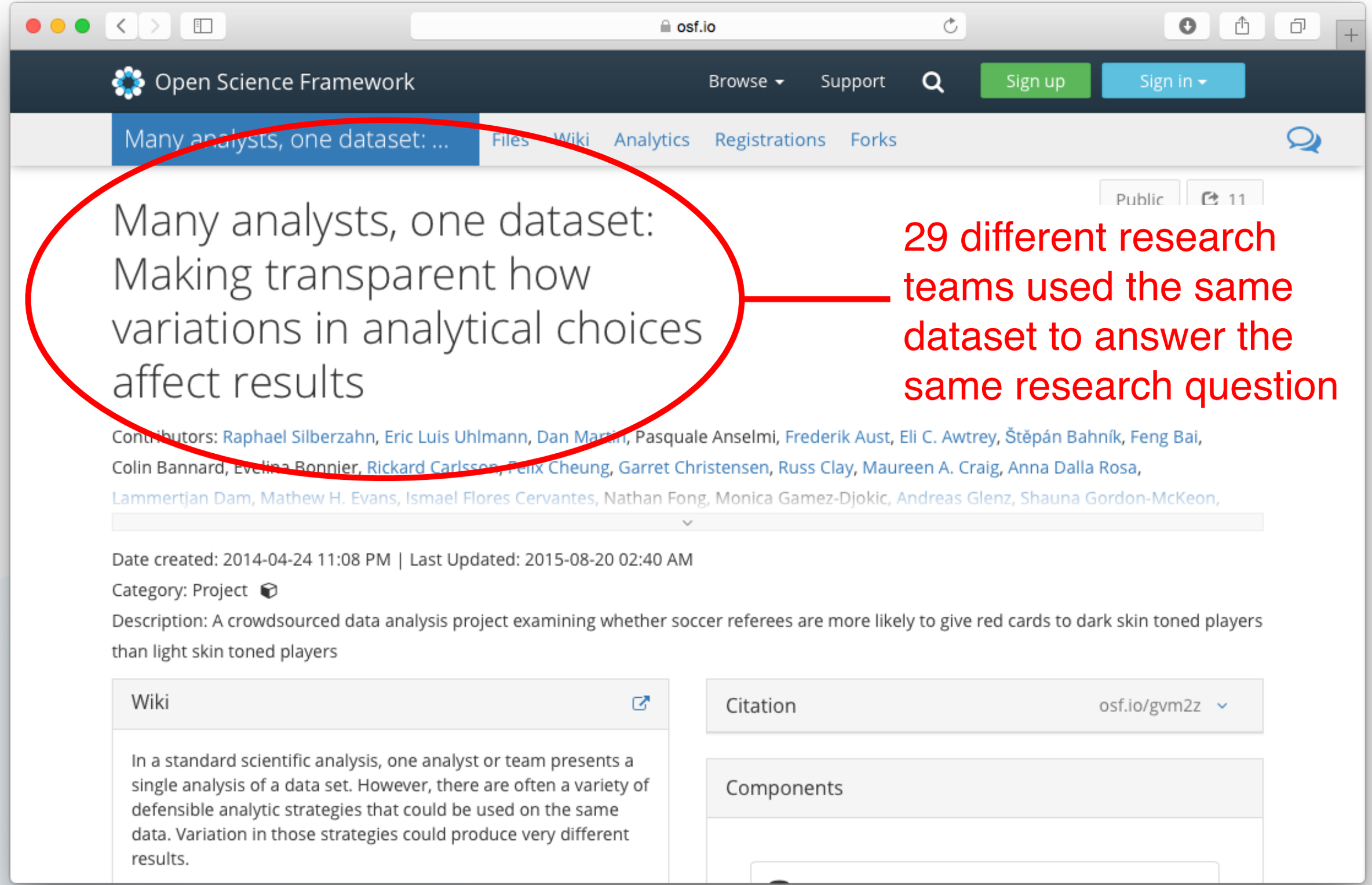
27 critiques of standardised effect sizes

Findings-Based Approach

- Conclusion: even if you don't think they should never be calculated, standardised effect sizes are really hard to validly compare.
- If Category A has a higher effect size in Schneider & Preckel's league table than Category B, this does not imply you should prioritise implementing A over B.
- I think this is a big problem with the findings-based approach to communicating education research.

Another Problem

- When conducting a research study researchers have a large number of analytical choices they can make.
- This is true both at the design stage and at the analysis stage.
- Consider the Open Science Framework's "Crowdsourcing data analysis" project.



29 different research teams used the same dataset to answer the same research question

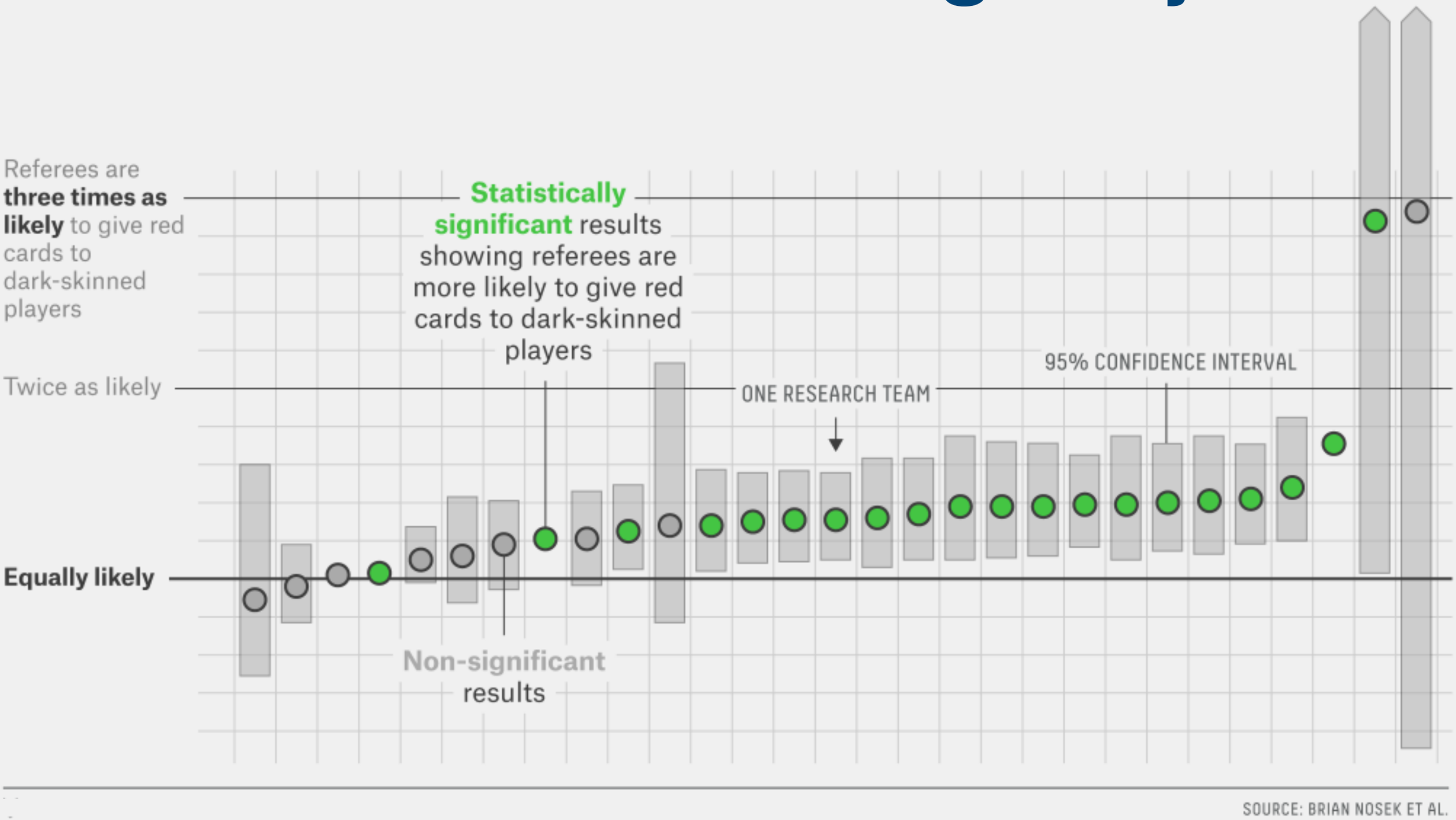
OSF Crowdsourcing Project

- Research Question: Are football referees more likely to give red cards to dark-skinned players than light-skinned players?
- 146,028 player-referee interactions featuring 2053 players and 3147 referees. Skin colour rated from 1 to 5 (‘very light’ to ‘very dark’) by two independent raters.
- Various other variables available in the dataset: player’s position, height, weight, country, age, league, etc. etc.
- Each research team independently came up with analysis strategy and answered the question.

OSF Crowdsourcing Project

- 29 teams produced analyses. Ranged from simple regression models to multilevel models, Poisson models and Bayesian analyses.
- 21 different combinations of covariates from the 29 teams, all were considered “defensible”.
- Each analysis led to a standardised effect size in odds ratio units.
- So, what is ***the*** effect size of skin colour on referee behaviour?

OSF Crowdsourcing Project



SOURCE: BRIAN NOSEK ET AL.

OSF Crowdsourcing Project

Conclusion 1: Standardised effect sizes are highly dependent on subjective analytical choices. ***The*** effect size doesn't exist.

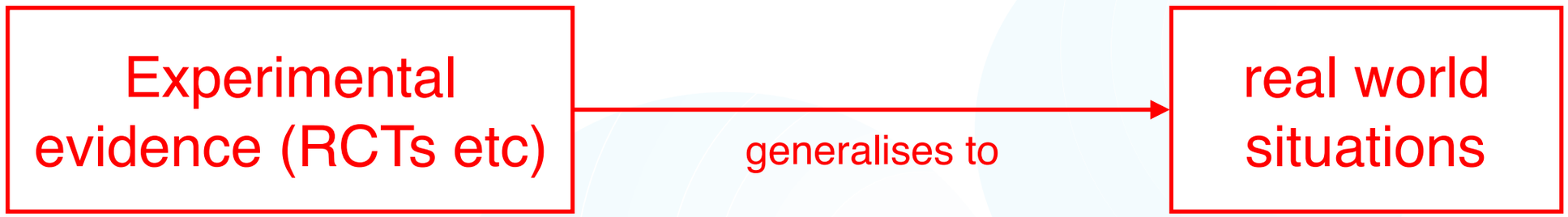
Conclusion 2: Despite this, the 'message' of the data is reasonably clear. Vast majority of teams found that red cards were more likely to be given to dark-skinned players.

We can't say "the effect of skin colour on the frequency of red cards is x" (effect size claim).

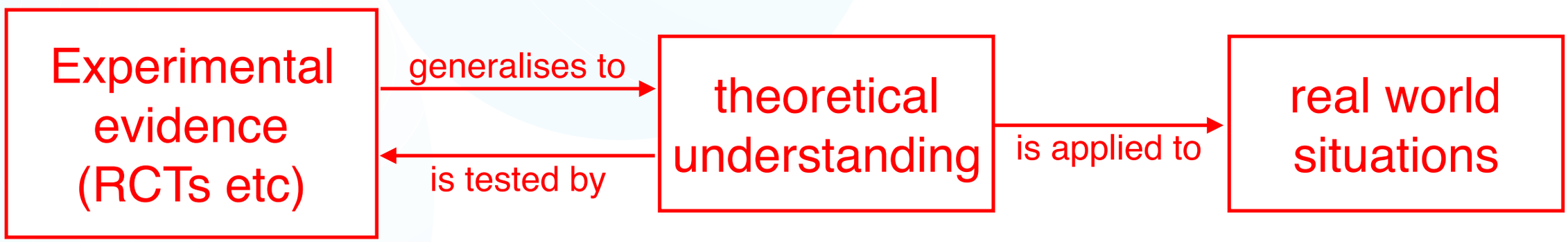
But we can say "skin colour has an effect on the frequency of red cards" (theoretical claim).

Two Types of Research

One type:



Another type:



Mook’s Example: Harlow (1959)

What is motherly love?

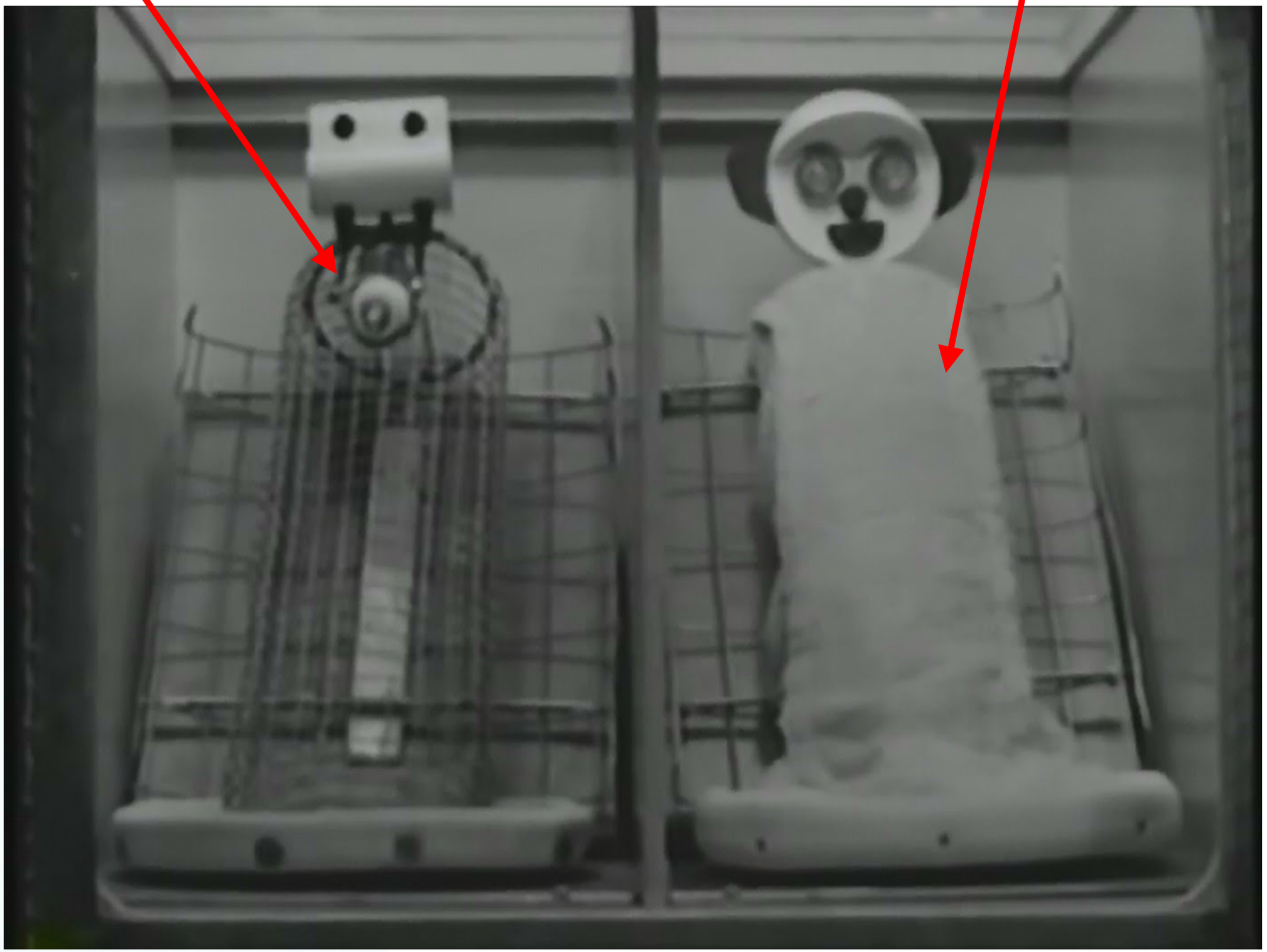
Two theories:

Hunger-reduction theory: Children love their mothers because they give them food.

Attachment theory: Children love their mothers because attachment is important for social and emotional development.

Wire monkey mother with food

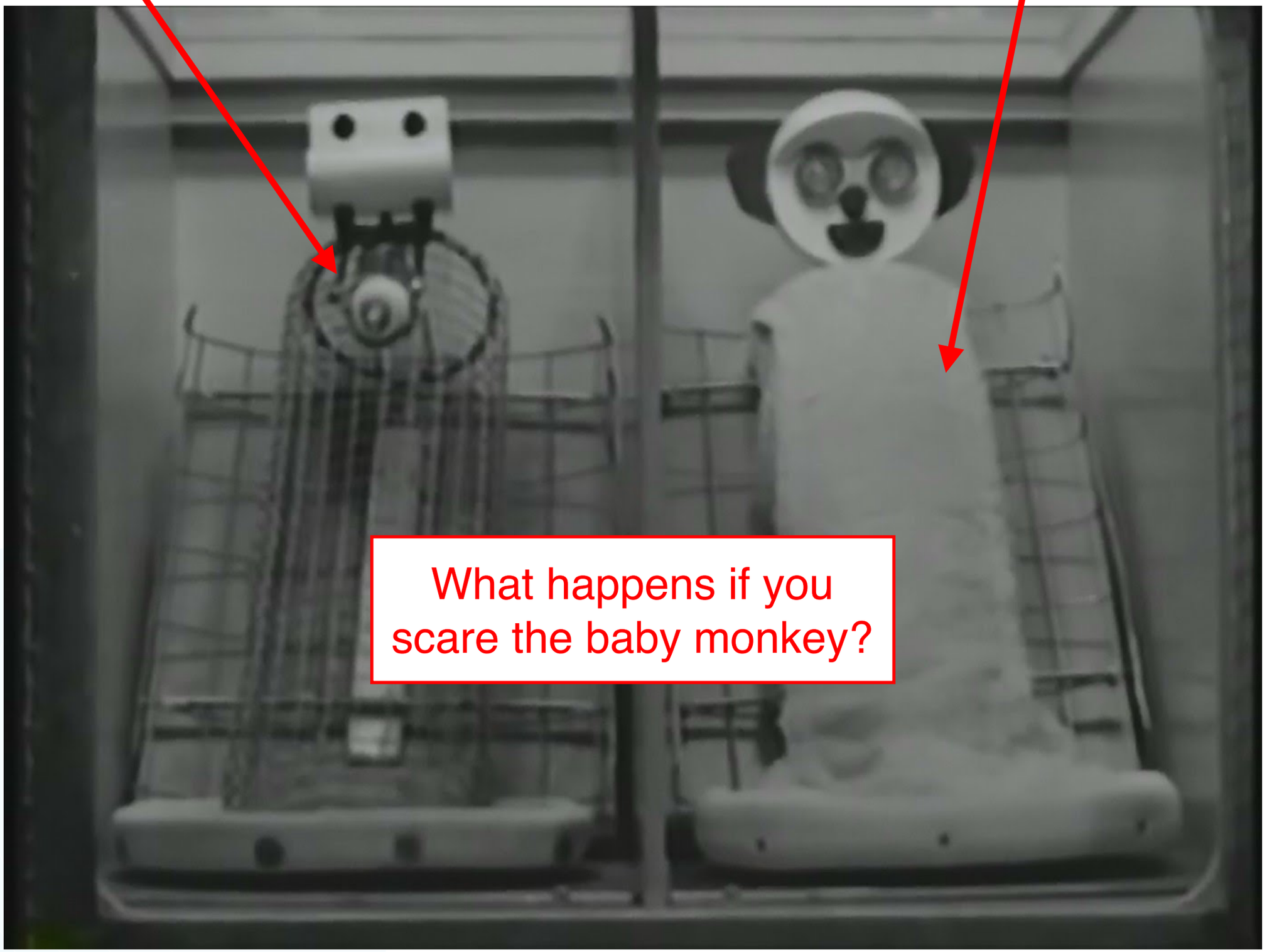
Cloth monkey mother without food



Harlow, H. F. (1959). The Nature of Love. *American Psychologist, 13,* 673-685.

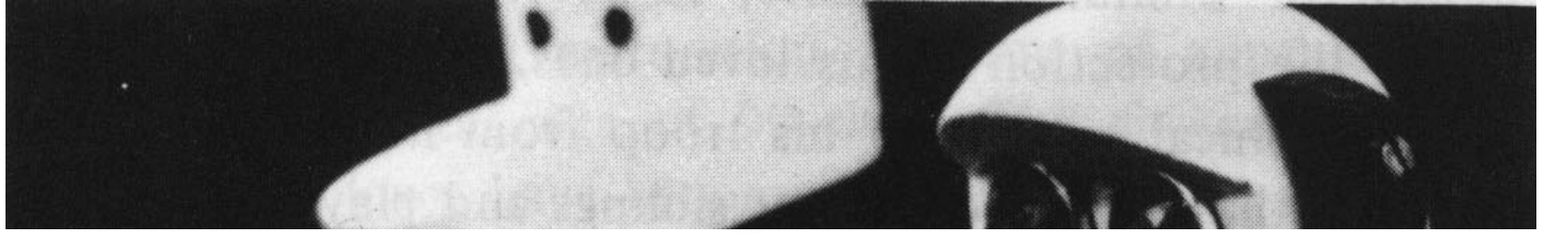
Wire monkey mother with food

Cloth monkey mother without food



Harlow, H. F. (1959). The Nature of Love. *American Psychologist, 13,* 673-685.

Harlow (1959)



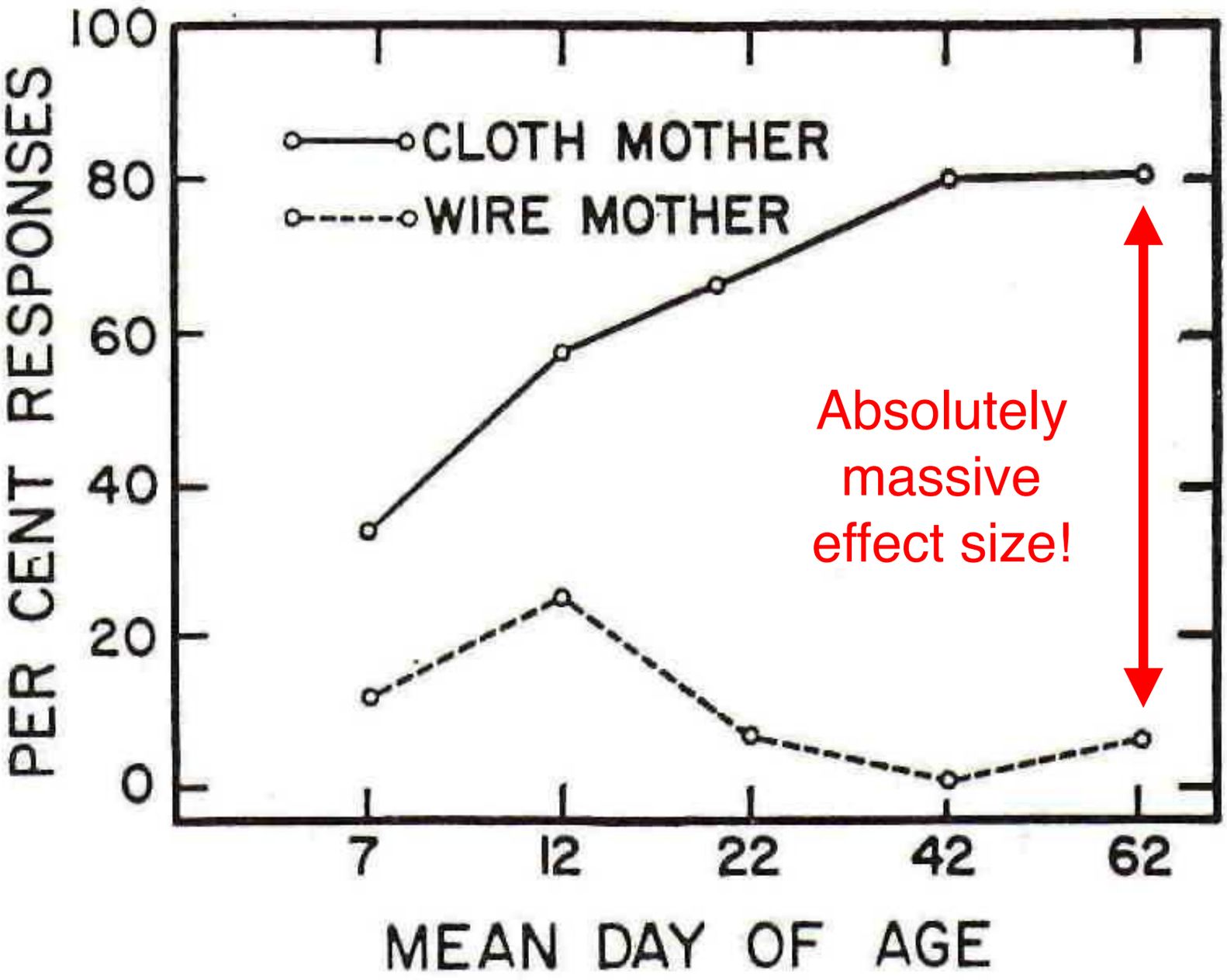


FIG. 15. Differential responsiveness in fear tests.

Harlow (1959)

Mook’s View of Theory

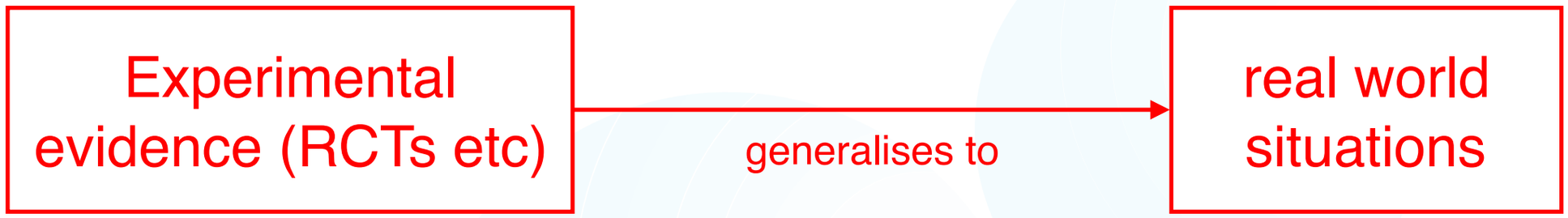
Mook’s point is that Harlow’s studies are **not** considered important because they tell us it would be a bad idea to replace real-world mothers with wire models.

And they are **not** considered important because of their enormous effect sizes.

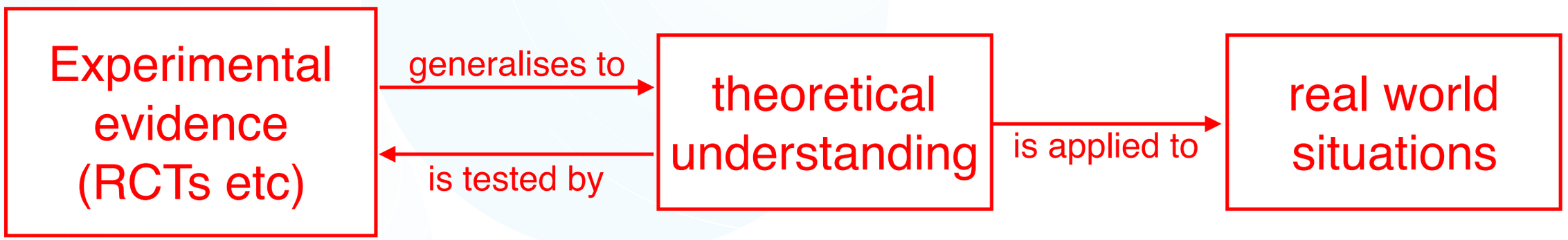
Rather they are important because they allow us to test theories about motherly love. And these theories tell us things about, for example, how child care should be organised.

Two Types of Research

Not very much educational research is like this:



Lots of educational research is like this:



For this second type of research, it's the theoretical understanding that practitioners should care about.

Implication: Communication should be about theory not findings.

An example: the Deans for Impact Science of Learning report

1 HOW DO STUDENTS UNDERSTAND NEW IDEAS?

COGNITIVE PRINCIPLES

To learn, students must transfer information from working memory (where it is consciously processed) to long-term memory (where it can be stored and later retrieved). Students have limited working memory capacities that can be overwhelmed by tasks that are cognitively too demanding. Understanding new ideas can be impeded if students are confronted with too much information at once.[4]

PRACTICAL IMPLICATIONS FOR THE CLASSROOM

- Teachers can use "worked examples" as one method of reducing students' cognitive burdens.[5] A worked example is a step-by-step demonstration of how to perform a task or solve a problem. This guidance – or "scaffolding" – can be gradually removed in subsequent problems so that students are required to complete more problem steps independently.
- Teachers often use multiple modalities to convey an idea; for example, they will speak while showing a graphic. If teachers take care to ensure that the two types of information complement one another – such as showing an animation while describing it aloud – learning is enhanced. But if the two sources of information are split – such as speaking aloud with different text displayed visually – attention is divided and learning is impaired.[6]
- Making content explicit through carefully paced explanation, modeling, and examples can help ensure that students are not overwhelmed.[7] *(Note: "explanation" does not mean teachers must do all the talking.)*

Theoretical understanding of how humans learn

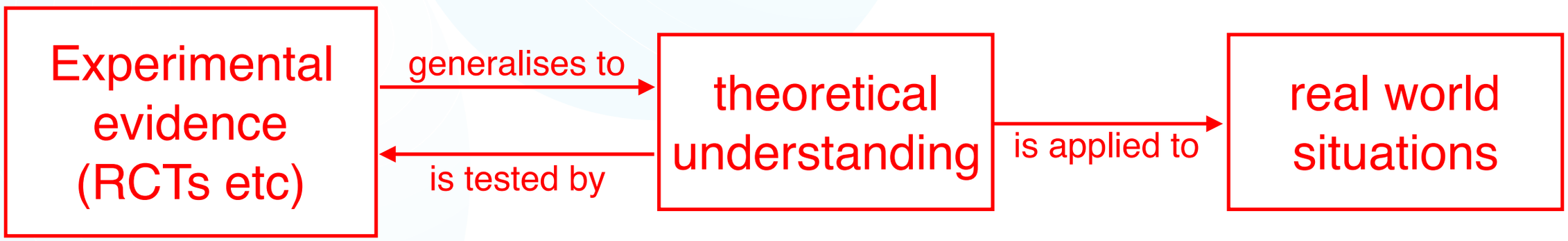
Possible implications for the classroom

References to (externally invalid) lab studies

Two Approaches

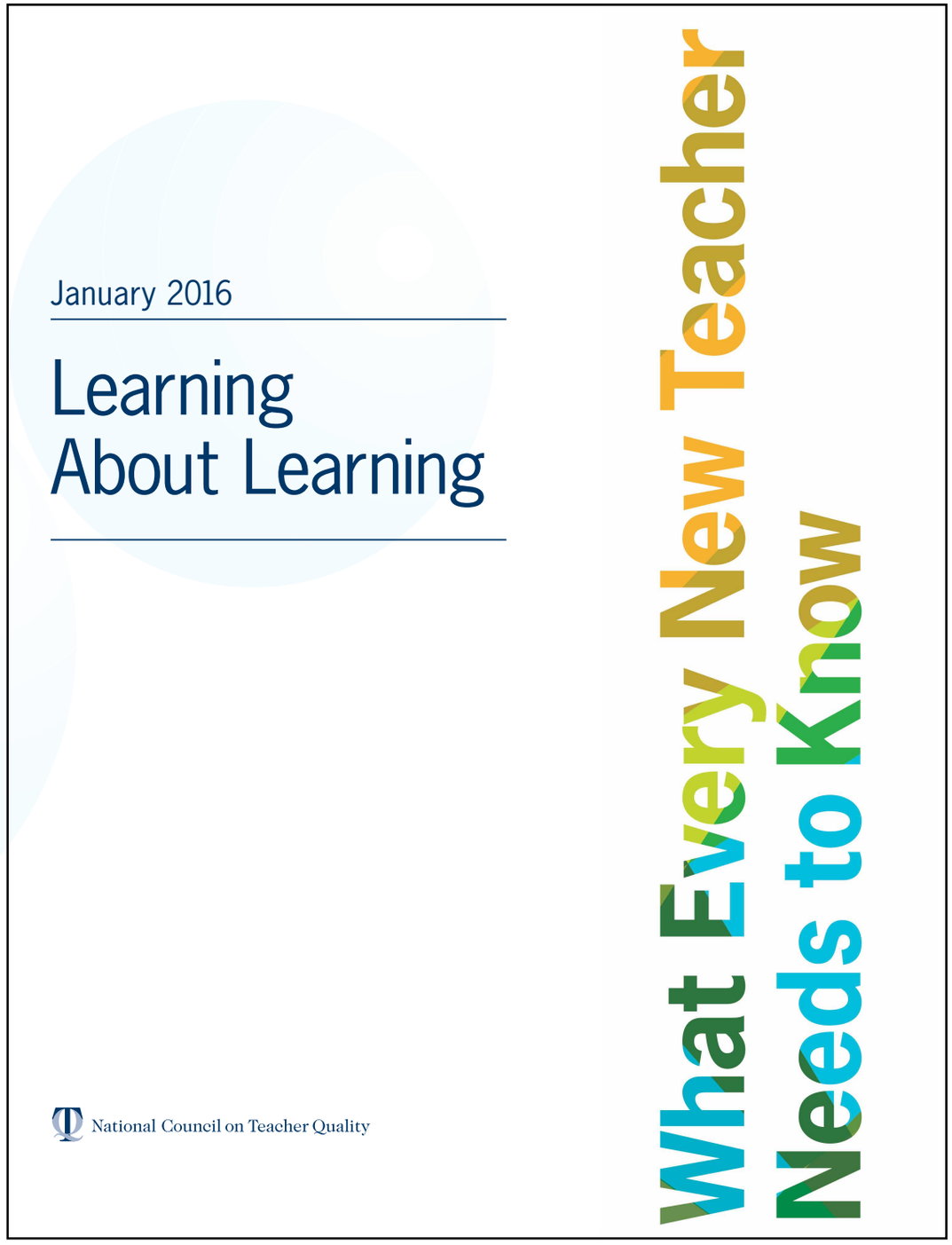
The findings-based approach to communication assumes research is like this:

The theory-based approach to communication assumes research is like this:



Two Approaches

- Apparently we're not very good at the second approach.
- Recent NCTQ report claims that six "research-proven instructional strategies" rarely appear in teacher preparation textbooks.



Summary

- My claim: it's not possible for you to effectively evaluate your teaching practice in day-to-day situations.
- You should resist intuitive judgements about what went well and what didn't in your lectures, because your intuitions, and everyone else's, are known to be unreliable.
- So all you can do is test your practice against what education research tells us about learning.
- There are two quite different approaches to communicating research to practitioners. My claim: the main output of educational research is theoretical understanding, not specific findings or meta-analysed findings.
- The Deans for Impact "Science of Learning" report is a good model.

Thank you

Funding:

THE ROYAL SOCIETY
CELEBRATING 350 YEARS

The Higher Education Academy
Maths, Stats & OR Network

Web: mcg.lboro.ac.uk/mji
Twitter: @mjinglis
Email: m.j.inglis@lboro.ac.uk