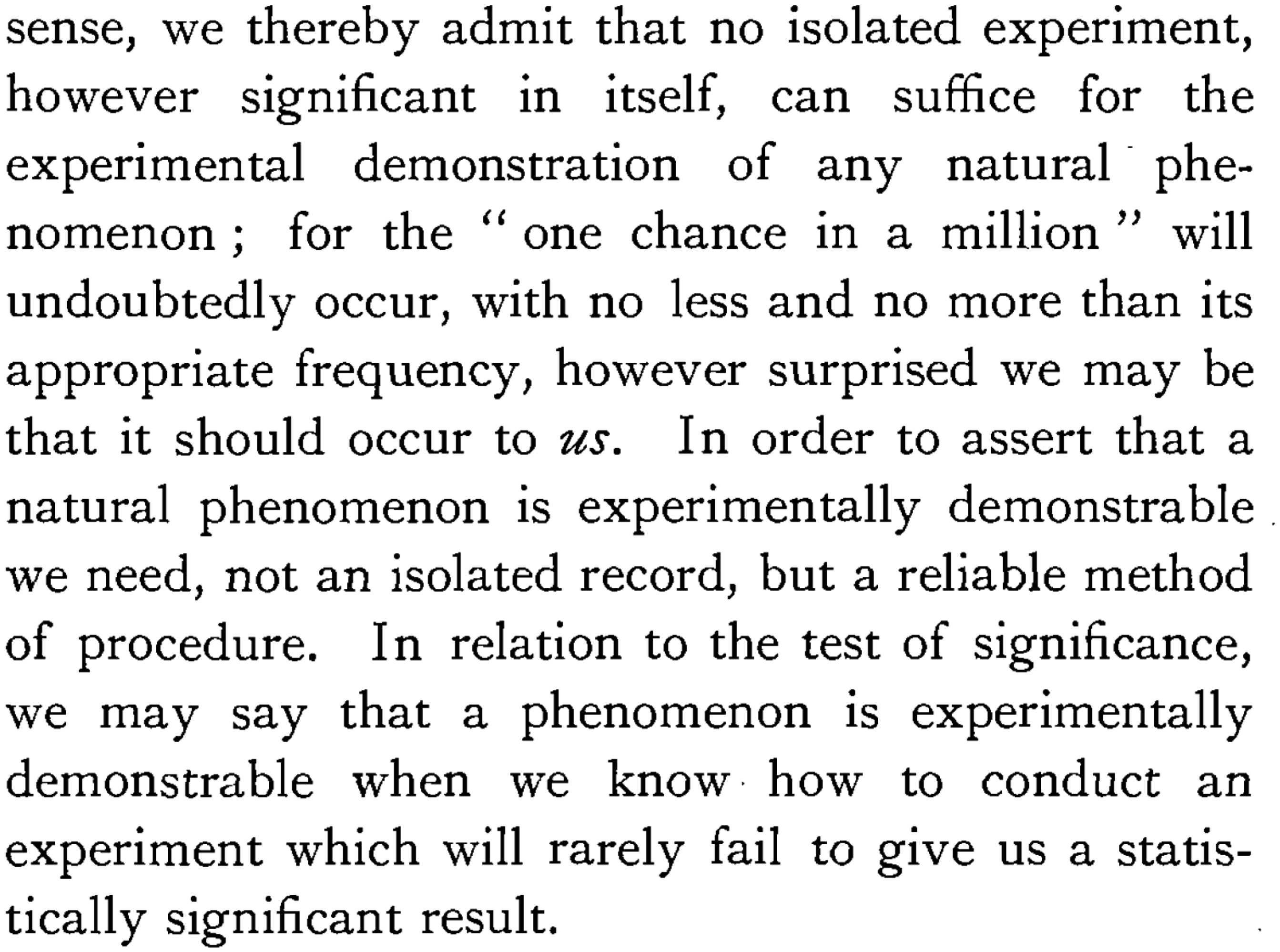
BGSE Development

Replication and Pre-Analysis Plans (Part 1)

Professors: Pamela Jakiela and Owen Ozier

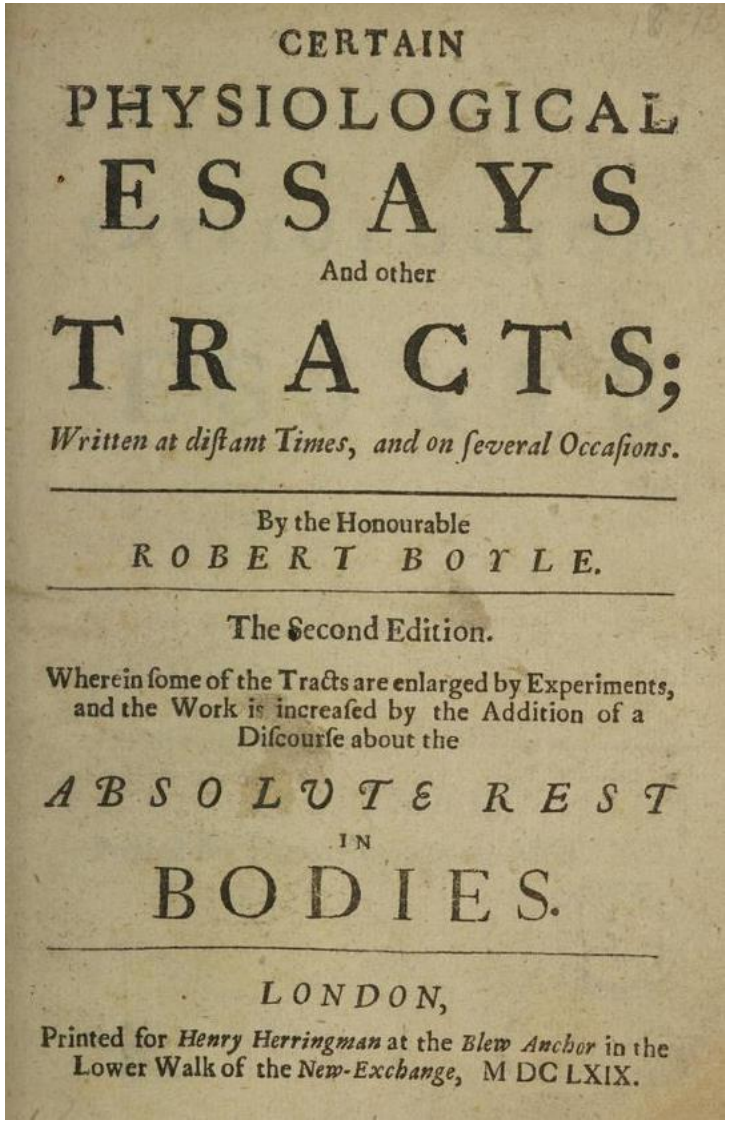
Starting at the beginning

Fisher (20th century)



sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*. In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.

Boyle (1600s)



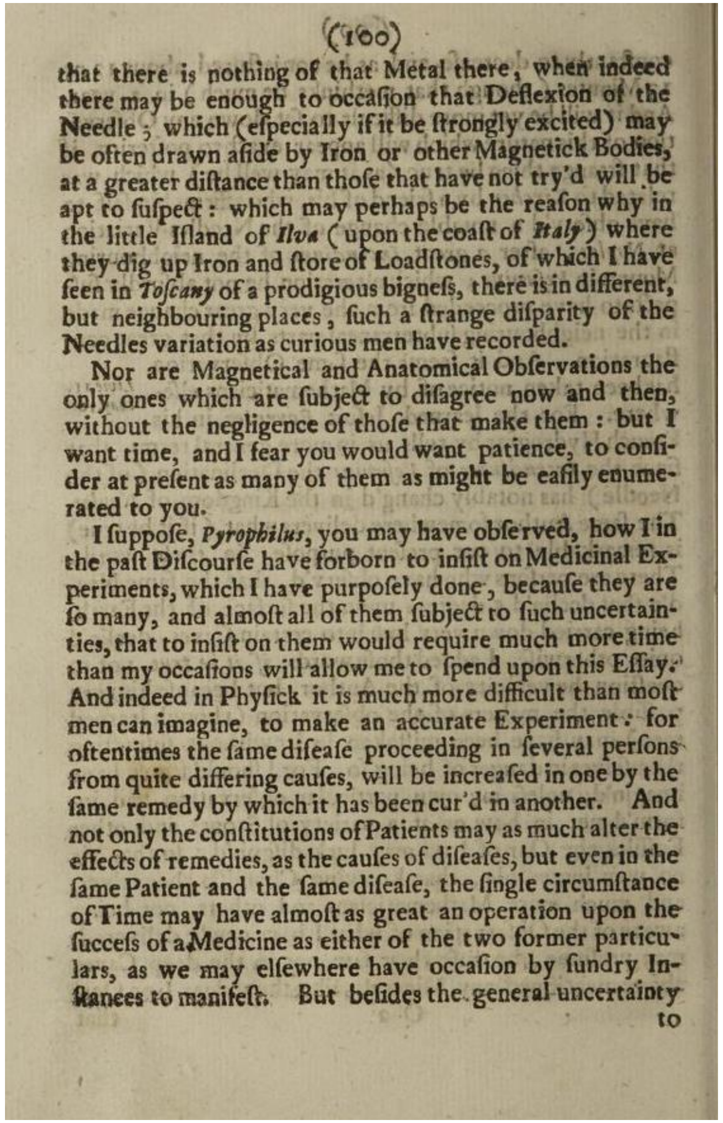
CERTAIN
PHYSIOLOGICAL
ESSAYS
And other
TRACTS;
Written at diſtant Times, and on ſeveral Occaſions.

By the Honourable
ROBERT BOYLE.

The Second Edition.

Wherein ſome of the Tracts are enlarged by Experiments,
and the Work is increaſed by the Addition of a
Diſcourſe about the
ABSOLUTE REST
IN
BODIES.

LONDON,
Printed for *Henry Herringman* at the *Blew Anchor* in the
Lower Walk of the *New-Exchange*, M DC LXIX.



(100)

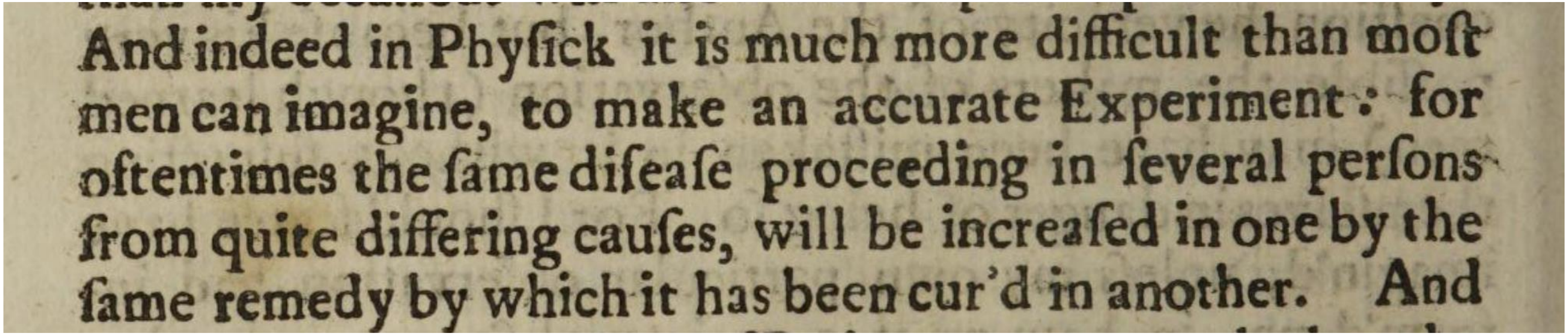
that there is nothing of that Metal there, when indeed there may be enough to occaſion that Deflexion of the Needle; which (eſpecially if it be ſtrongly excited) may be often drawn aſide by Iron or other Magnetick Bodies, at a greater diſtance than thoſe that have not try'd will be apt to ſuſpect: which may perhaps be the reaſon why in the little Iſland of *Ilva* (upon the coaſt of *Italy*) where they dig up Iron and ſtore of Loadſtones, of which I have ſeen in *Toſcany* of a prodigious bigneſs, there is in different, but neighbouring places, ſuch a ſtrange diſparity of the Needles variation as curious men have recorded.

Nor are Magnetical and Anatomical Obſervations the only ones which are ſubject to diſagree now and then, without the negligence of thoſe that make them: but I want time, and I fear you would want patience, to conſider at preſent as many of them as might be eaſily enumerated to you.

I ſuppoſe, *Pyrophilus*, you may have obſerved, how I in the paſt Diſcourſe have forborn to inſiſt on Medicinal Experiments, which I have purpoſely done, becauſe they are ſo many, and almoſt all of them ſubject to ſuch uncertainties, that to inſiſt on them would require much more time than my occaſions will allow me to ſpend upon this Eſſay. And indeed in Phyſick it is much more difficult than moſt men can imagine, to make an accurate Experiment: for oftentimes the ſame diſeaſe proceeding in ſeveral perſons from quite differing cauſes, will be increaſed in one by the ſame remedy by which it has been cur'd in another. And not only the conſtitutions of Patients may as much alter the effects of remedies, as the cauſes of diſeaſes, but even in the ſame Patient and the ſame diſeaſe, the ſingle circumſtance of Time may have almoſt as great an operation upon the ſucceſs of a Medicine as either of the two former particulars, as we may elſewhere have occaſion by ſundry Inſtances to manifeſt. But beſides the general uncertainty

to

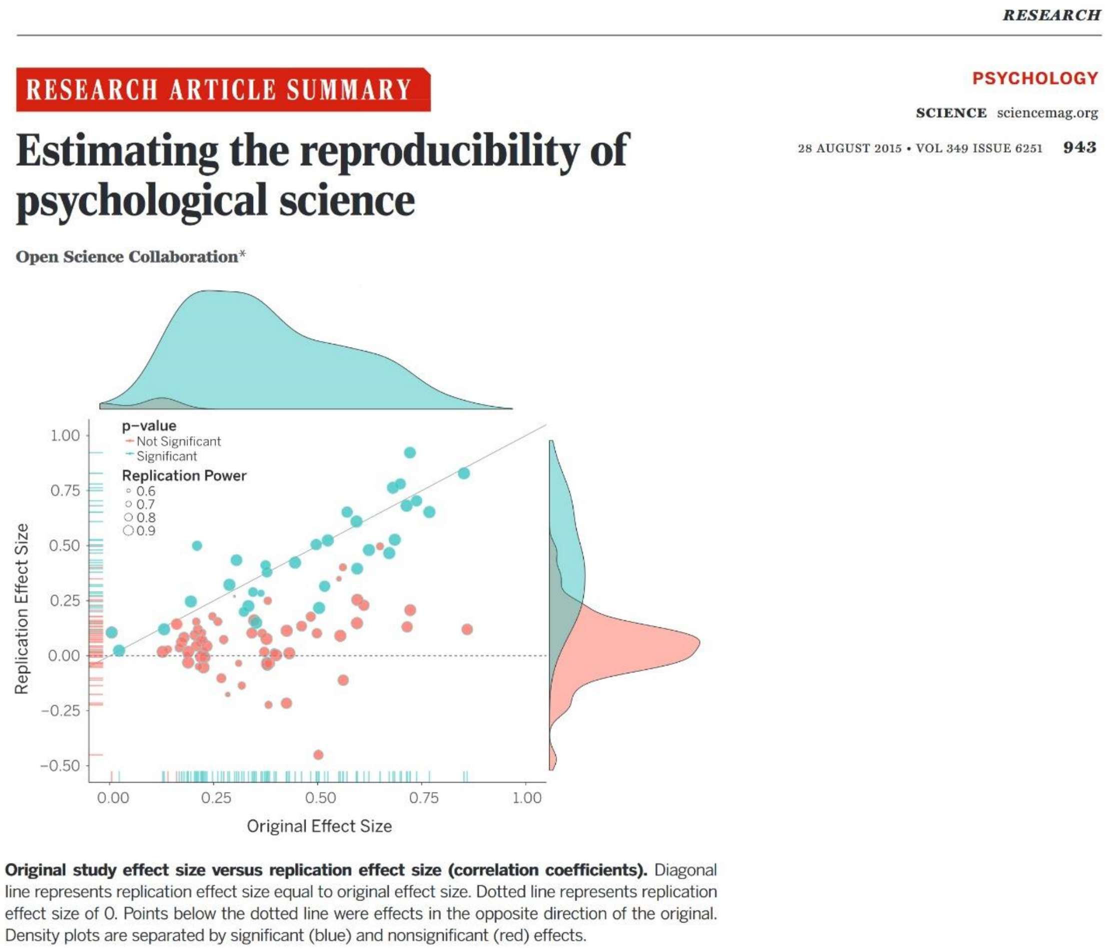
Boyle (1600s)



And indeed in Phyſick it is much more difficult than moſt men can imagine, to make an accurate Experiment: for oftentimes the ſame diſeaſe proceeding in ſeveral perſons from quite differing cauſes, will be increaſed in one by the ſame remedy by which it has been cur'd in another. And

Are scientific results replicable?

What do we know about replicability?



RESEARCH

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

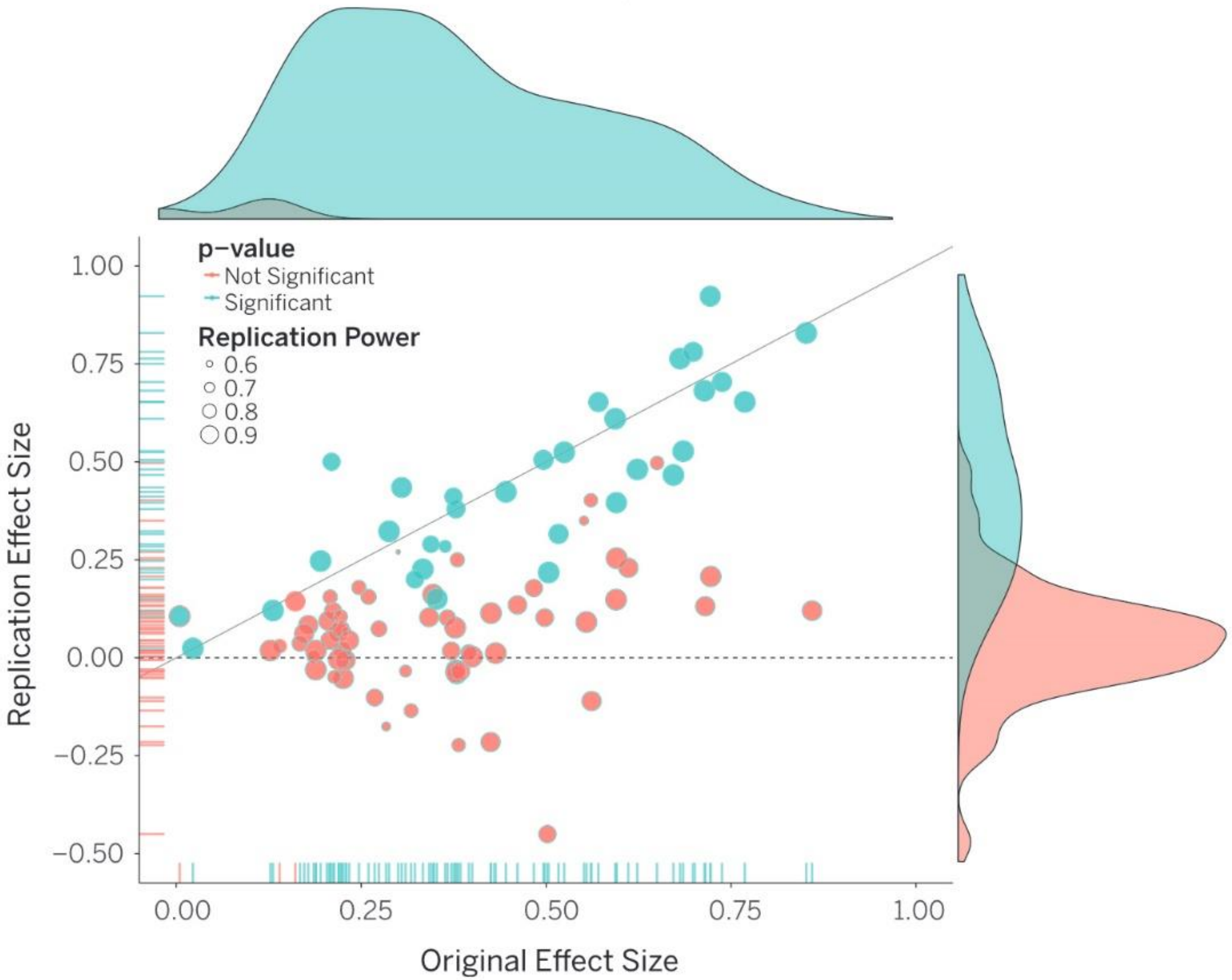
SCIENCE sciencemag.org

28 AUGUST 2015 • VOL 349 ISSUE 6251 943

Estimating the reproducibility of psychological science

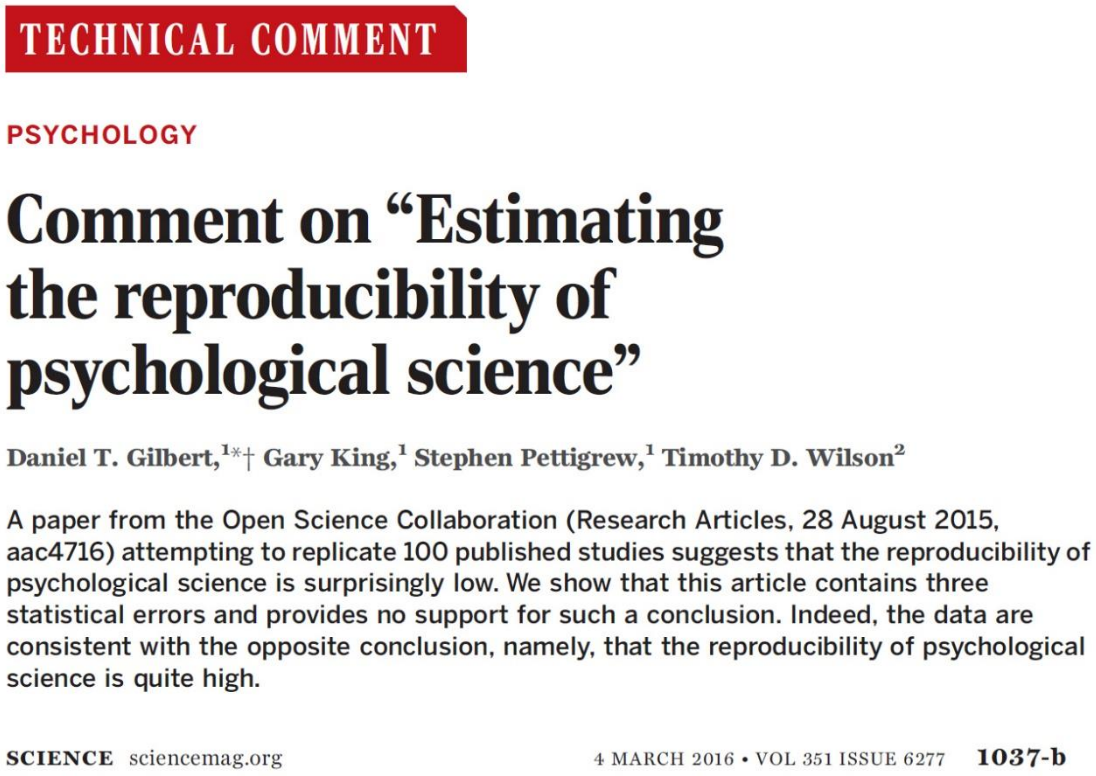
Open Science Collaboration*

Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

What do we know about replicability?



TECHNICAL COMMENT

PSYCHOLOGY

Comment on "Estimating the reproducibility of psychological science"

Daniel T. Gilbert,[1*†] Gary King,[1] Stephen Pettigrew,[1] Timothy D. Wilson[2]

A paper from the Open Science Collaboration (Research Articles, 28 August 2015, aac4716) attempting to replicate 100 published studies suggests that the reproducibility of psychological science is surprisingly low. We show that this article contains three statistical errors and provides no support for such a conclusion. Indeed, the data are consistent with the opposite conclusion, namely, that the reproducibility of psychological science is quite high.

SCIENCE sciencemag.org 4 MARCH 2016 • VOL 351 ISSUE 6277 1037-b

- Fidelity
- Statistical power

"…an original study that asked **college students** to imagine being called on by a professor was replicated with participants who had **never been to college**…

an original study that asked students **who commute** to school to choose between apartments that were short and long drives from campus was replicated with students **who do not commute** to school. …

An original study that asked **Israelis to imagine the consequences of military service** was replicated by asking **Americans to imagine the consequences of a honeymoon**;

an original study that gave **younger children the difficult task** of locating targets on a large screen was replicated by giving **older children the easier task** of locating targets on a small screen;

an original study that showed how a change in the wording of a charitable appeal sent **by mail to Koreans** could boost response rates was replicated by sending 771,408 **email messages to people all over the world** (which produced a response rate of essentially zero in all conditions)."

(Gilbert, et al. 2016)

"…an original study that asked college students to imagine being called on by a professor was replicated with participants who had never been to college…

an original study that asked students who commute to school to choose between apartments that were short and long drives from campus was replicated with students who do not commute to school. …

An original study that asked Israelis to imagine the consequences of military service was replicated by asking Americans to imagine the consequences of a honeymoon;

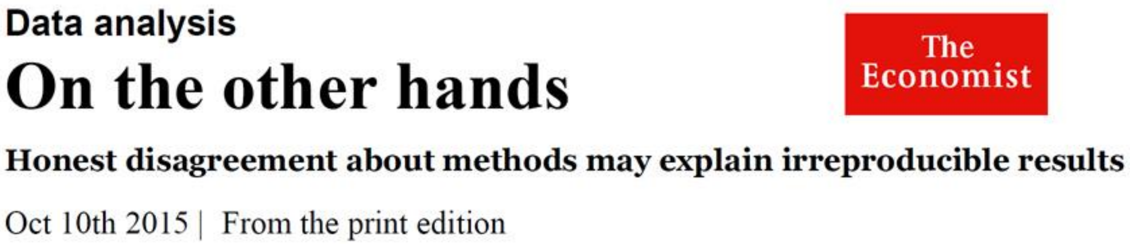
an original study that gave younger children the difficult task of locating targets on a large screen was replicated by giving older children the easier task of locating targets on a small screen;

an original study that showed how a change in the wording of a charitable appeal sent by mail to Koreans could boost response rates was replicated by sending 771,408 e-mail messages to people all over the world (which produced a response rate of essentially zero in all conditions)."

(Gilbert, et al. 2016)

(Caveats: response to response, original study had more details, etc.)

Replication: what do the data really tell us?



Data analysis

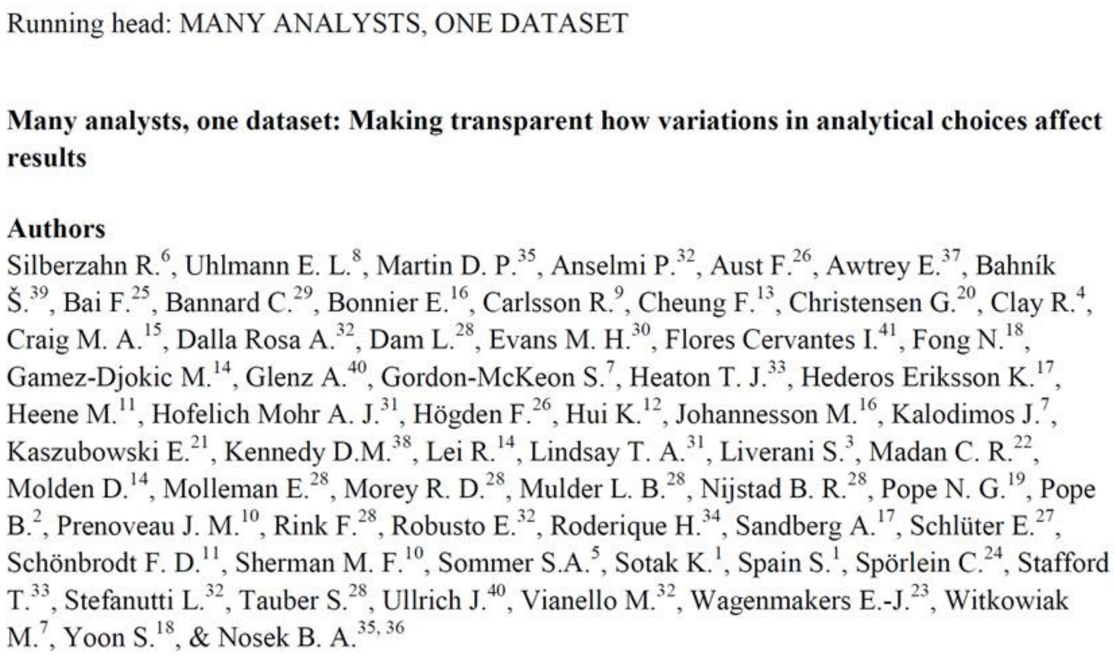
On the other hands

The Economist

Honest disagreement about methods may explain irreproducible results

Oct 10th 2015 | From the print edition

IT SOUNDS like an easy question for any half-competent scientist to answer.



Running head: MANY ANALYSTS, ONE DATASET

Many analysts, one dataset: Making transparent how variations in analytical choices affect results

Authors

Silberzahn R.[6], Uhlmann E. L.[8], Martin D. P.[35], Anselmi P.[32], Aust F.[26], Awtrey E.[37], Bahník Š.[39], Bai F.[25], Bannard C.[29], Bonnier E.[16], Carlsson R.[9], Cheung F.[13], Christensen G.[20], Clay R.[4], Craig M. A.[15], Dalla Rosa A.[32], Dam L.[28], Evans M. H.[30], Flores Cervantes I.[41], Fong N.[18], Gamez-Djokic M.[14], Glenz A.[40], Gordon-McKeon S.[7], Heaton T. J.[33], Hederos Eriksson K.[17], Heene M.[11], Hofelich Mohr A. J.[31], Högden F.[26], Hui K.[12], Johannesson M.[16], Kalodimos J.[7], Kaszubowski E.[21], Kennedy D.M.[38], Lei R.[14], Lindsay T. A.[31], Liverani S.[3], Madan C. R.[22], Molden D.[14], Molleman E.[28], Morey R. D.[28], Mulder L. B.[28], Nijstad B. R.[28], Pope N. G.[19], Pope B.[2], Prenoveau J. M.[10], Rink F.[28], Robusto E.[32], Roderique H.[34], Sandberg A.[17], Schlüter E.[27], Schönbrodt F. D.[11], Sherman M. F.[10], Sommer S.A.[5], Sotak K.[1], Spain S.[1], Spörlein C.[24], Stafford T.[33], Stefanutti L.[32], Tauber S.[28], Ullrich J.[40], Vianello M.[32], Wagenmakers E.-J.[23], Witkowiak M.[7], Yoon S.[18], & Nosek B. A.[35, 36]

190 | NATURE | VOL 526 | 8 OCTOBER 2015

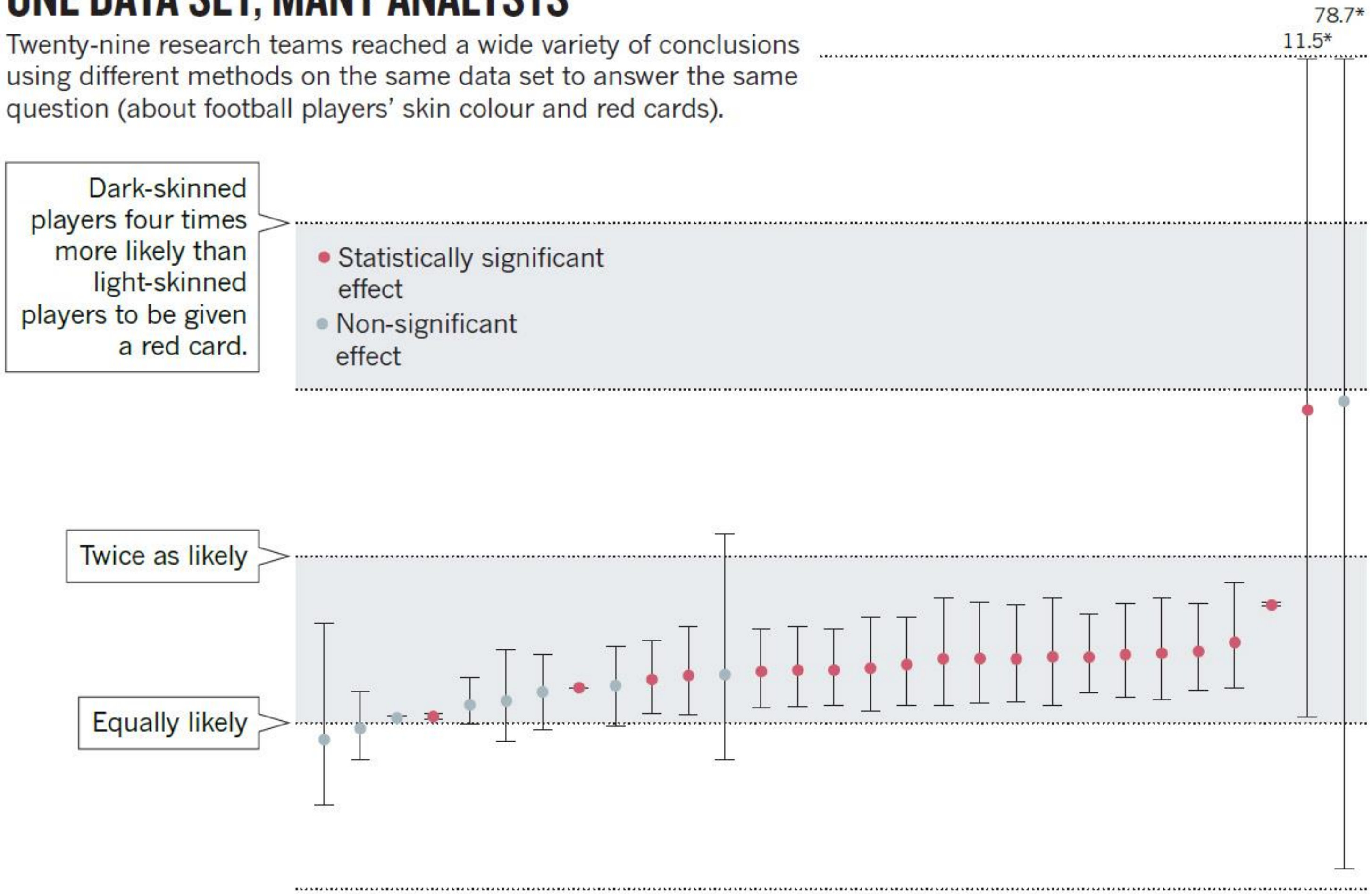
MICHAEL REGAN/GETTY

Mario Balotelli, playing for Manchester City, is shown a red card during a match against Arsenal.

190 | NATURE | VOL 526 | 8 OCTOBER 2015

ONE DATA SET, MANY ANALYSTS

Twenty-nine research teams reached a wide variety of conclusions using different methods on the same data set to answer the same question (about football players' skin colour and red cards).



Point estimates and 95% confidence intervals. *Truncated upper bounds.

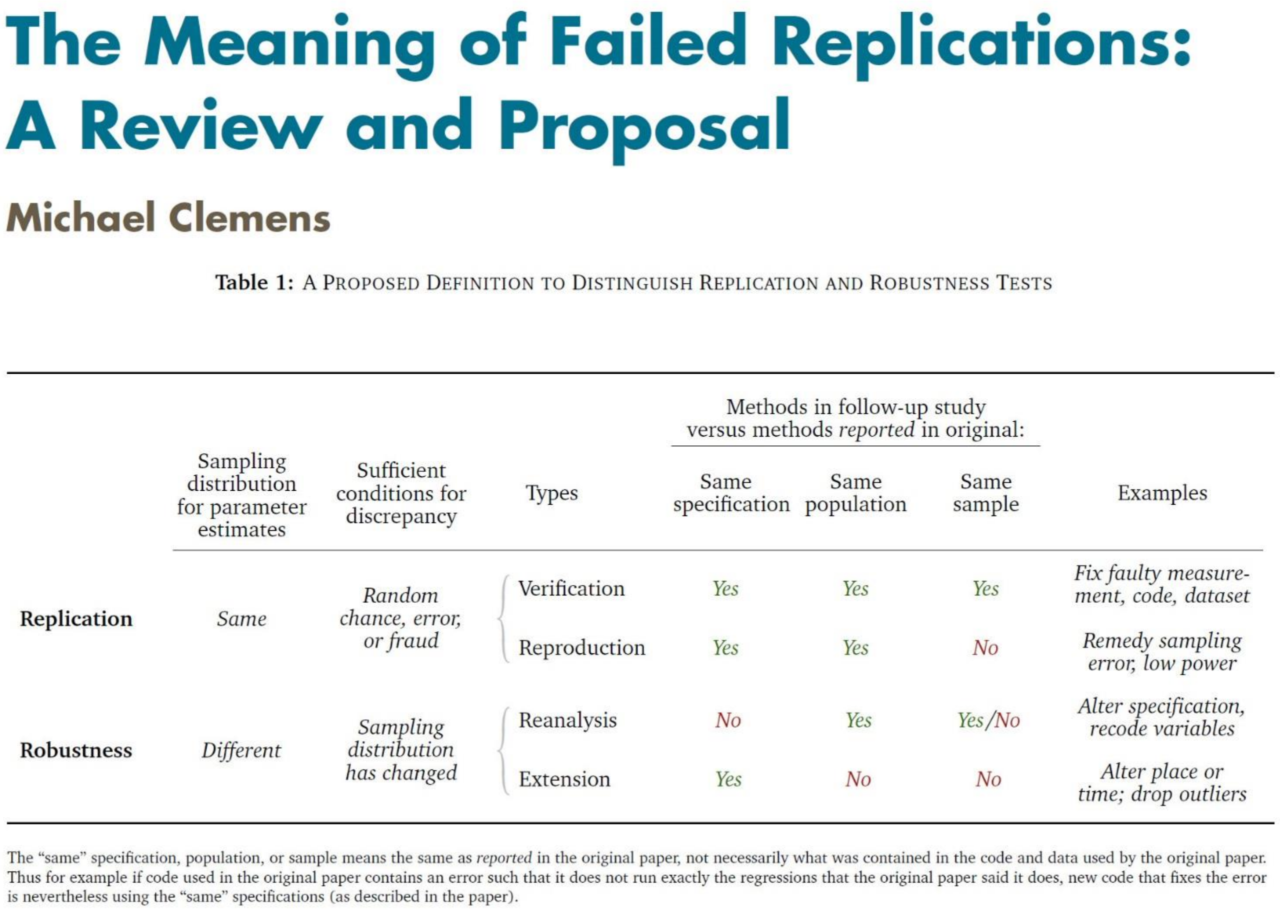
Replication: what do the data really tell us?

- Main question: whether or not soccer referees were more likely to give red cards to dark skin toned players than light skin toned players.
- 29 research teams used 21 unique combinations of covariates
- The word “identification” only appears in an one of 29 team’s description of their work, not in the main study text.
- Twenty teams (69%) found a significant positive relationship and nine teams (31%) observed a non-significant relationship. No team reported a significant negative relationship. Inasmuch as there was a pattern here, perhaps “irreproducible” is an overstatement.
- 32% of respondents were unconfident to somewhat unconfident regarding how appropriate the dataset was for answering the primary research question (whether an association exists between players’ skin tone and referee red card decisions).
- **Not all datasets have an appropriate counterfactual that would permit estimation of effects.**

Replication: terminology

- There is more than one kind of replication/reproducibility.
- Use of terms varies across and within disciplines.
- Implications of “failure” vary by type of replication/reproducibility.

Replication: Michael Clemens' terminology.



The Meaning of Failed Replications: A Review and Proposal

Michael Clemens

Table 1: A Proposed Definition to Distinguish Replication and Robustness Tests

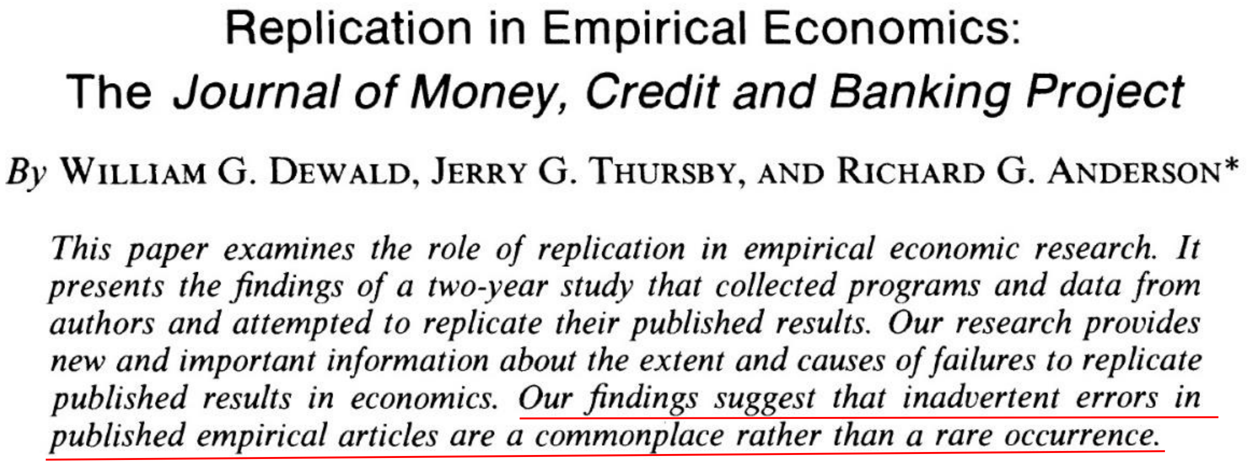
				Methods in follow-up study versus methods *reported* in original:			
	Sampling distribution for parameter estimates	Sufficient conditions for discrepancy	Types	Same specification	Same population	Same sample	Examples
Replication	*Same*	*Random chance, error, or fraud*	Verification	*Yes*	*Yes*	*Yes*	*Fix faulty measurement, code, dataset*
			Reproduction	*Yes*	*Yes*	*No*	*Remedy sampling error, low power*
Robustness	*Different*	*Sampling distribution has changed*	Reanalysis	*No*	*Yes*	*Yes/No*	*Alter specification, recode variables*
			Extension	*Yes*	*No*	*No*	*Alter place or time; drop outliers*

The "same" specification, population, or sample means the same as *reported* in the original paper, not necessarily what was contained in the code and data used by the original paper. Thus for example if code used in the original paper contains an error such that it does not run exactly the regressions that the original paper said it does, new code that fixes the error is nevertheless using the "same" specifications (as described in the paper).

(See also Hamermesh various years, and others!)

What was old is new again / History repeating

1986



Replication in Empirical Economics: The *Journal of Money, Credit and Banking Project*

By WILLIAM G. DEWALD, JERRY G. THURSBY, AND RICHARD G. ANDERSON*

This paper examines the role of replication in empirical economic research. It presents the findings of a two-year study that collected programs and data from authors and attempted to replicate their published results. Our research provides new and important information about the extent and causes of failures to replicate published results in economics. Our findings suggest that inadvertent errors in published empirical articles are a commonplace rather than a rare occurrence.

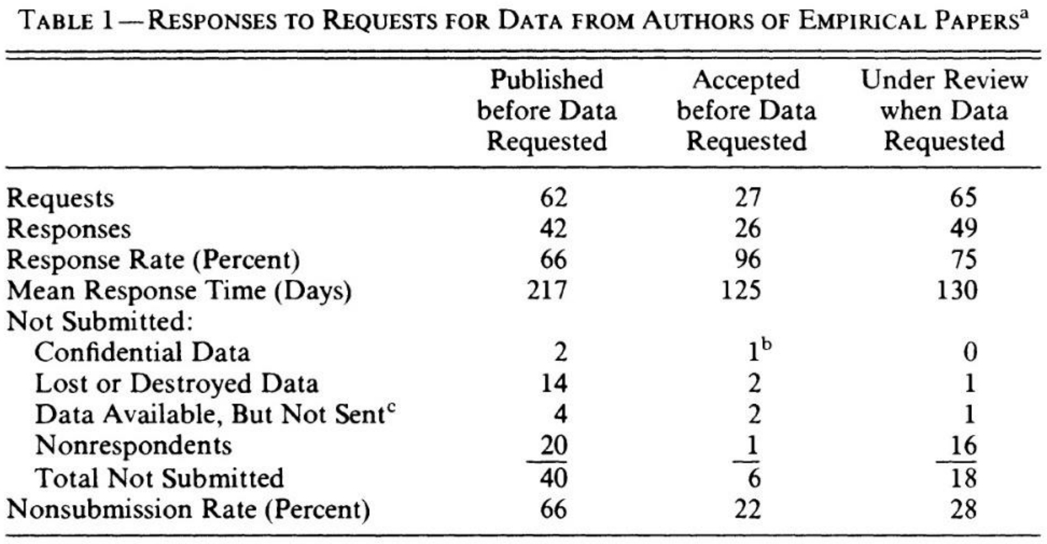


TABLE 1—RESPONSES TO REQUESTS FOR DATA FROM AUTHORS OF EMPIRICAL PAPERS[a]

	Published before Data Requested	Accepted before Data Requested	Under Review when Data Requested
Requests	62	27	65
Responses	42	26	49
Response Rate (Percent)	66	96	75
Mean Response Time (Days)	217	125	130
Not Submitted:			
Confidential Data	2	1[b]	0
Lost or Destroyed Data	14	2	1
Data Available, But Not Sent[c]	4	2	1
Nonrespondents	20	1	16
Total Not Submitted	40	6	18
Nonsubmission Rate (Percent)	66	22	28

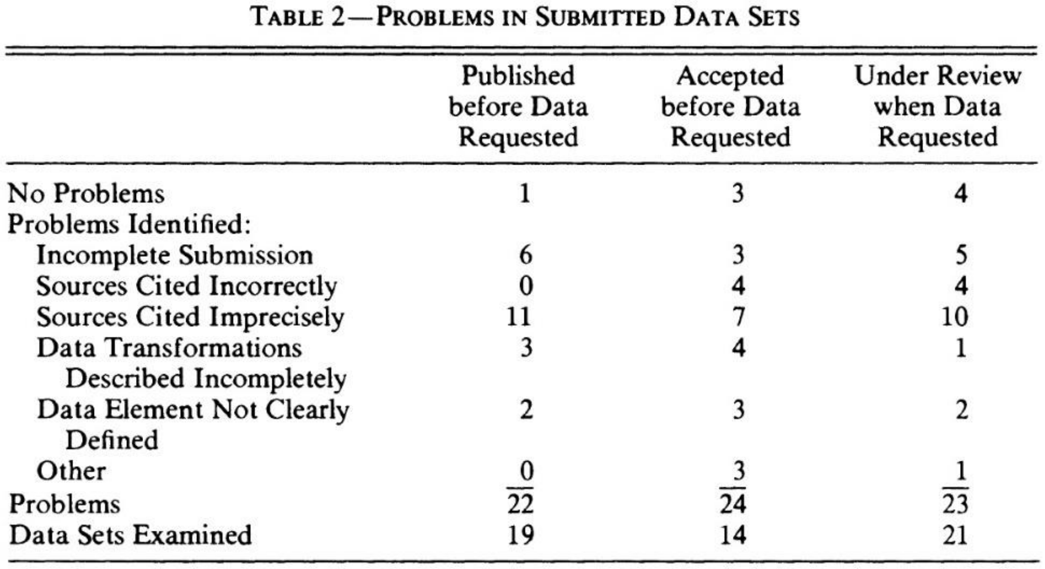
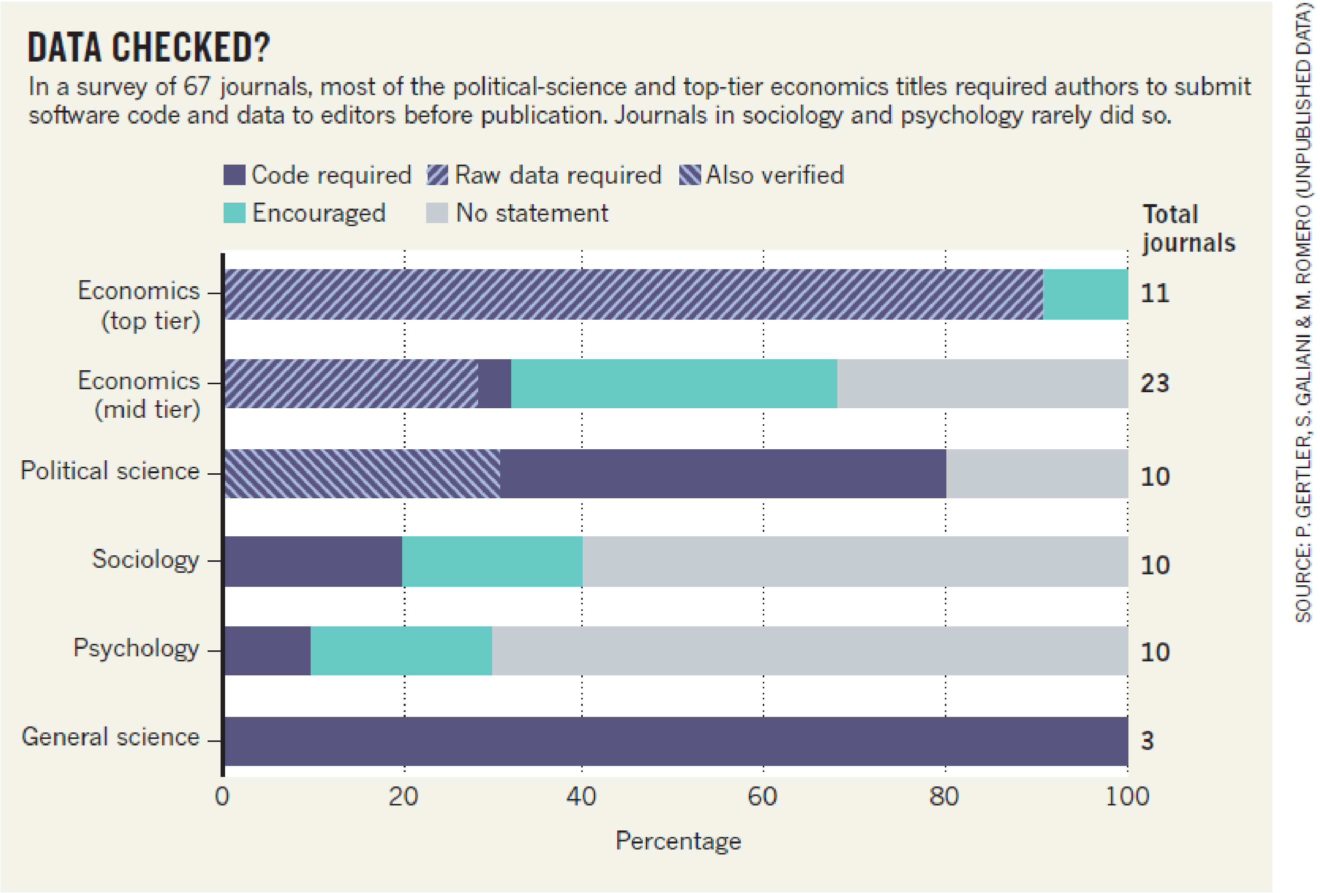


TABLE 2—PROBLEMS IN SUBMITTED DATA SETS

	Published before Data Requested	Accepted before Data Requested	Under Review when Data Requested
No Problems	1	3	4
Problems Identified:			
Incomplete Submission	6	3	5
Sources Cited Incorrectly	0	4	4
Sources Cited Imprecisely	11	7	10
Data Transformations Described Incompletely	3	4	1
Data Element Not Clearly Defined	2	3	2
Other	0	3	1
Problems	22	24	23
Data Sets Examined	19	14	21

Nature (2018): Galiani, Gertler, and Romero



DATA CHECKED?

In a survey of 67 journals, most of the political-science and top-tier economics titles required authors to submit software code and data to editors before publication. Journals in sociology and psychology rarely did so.

SOURCE: P. GERTLER, S. GALIANI & M. ROMERO (UNPUBLISHED DATA)

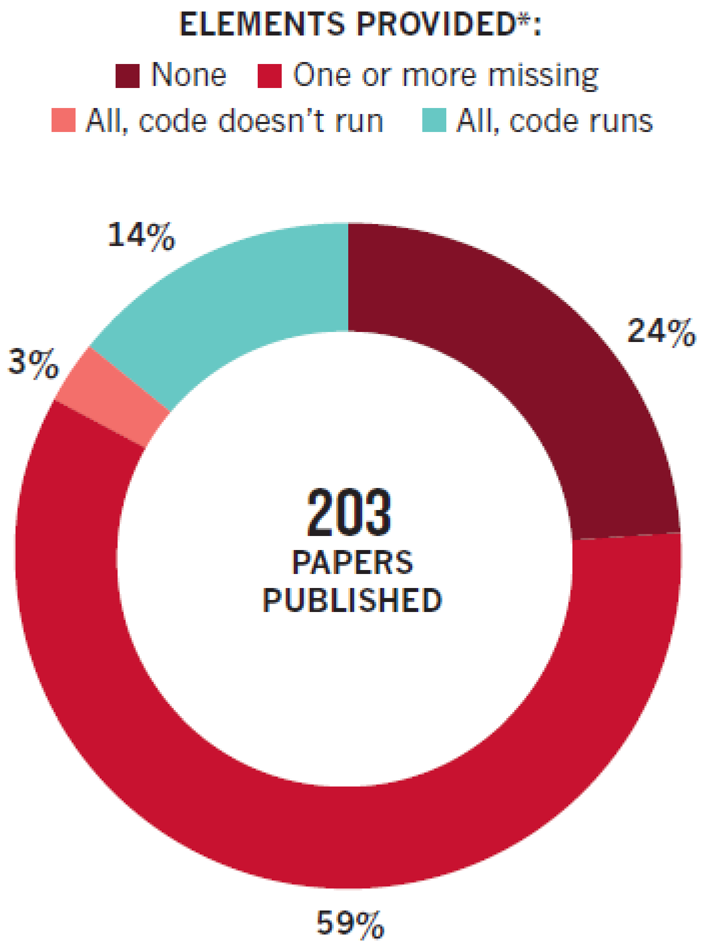
Nature (2018):

SOURCE: P. GERTLER, S. GALIANI & M. ROMERO

22 FEBRUARY 2018 | VOL 554 | NATURE

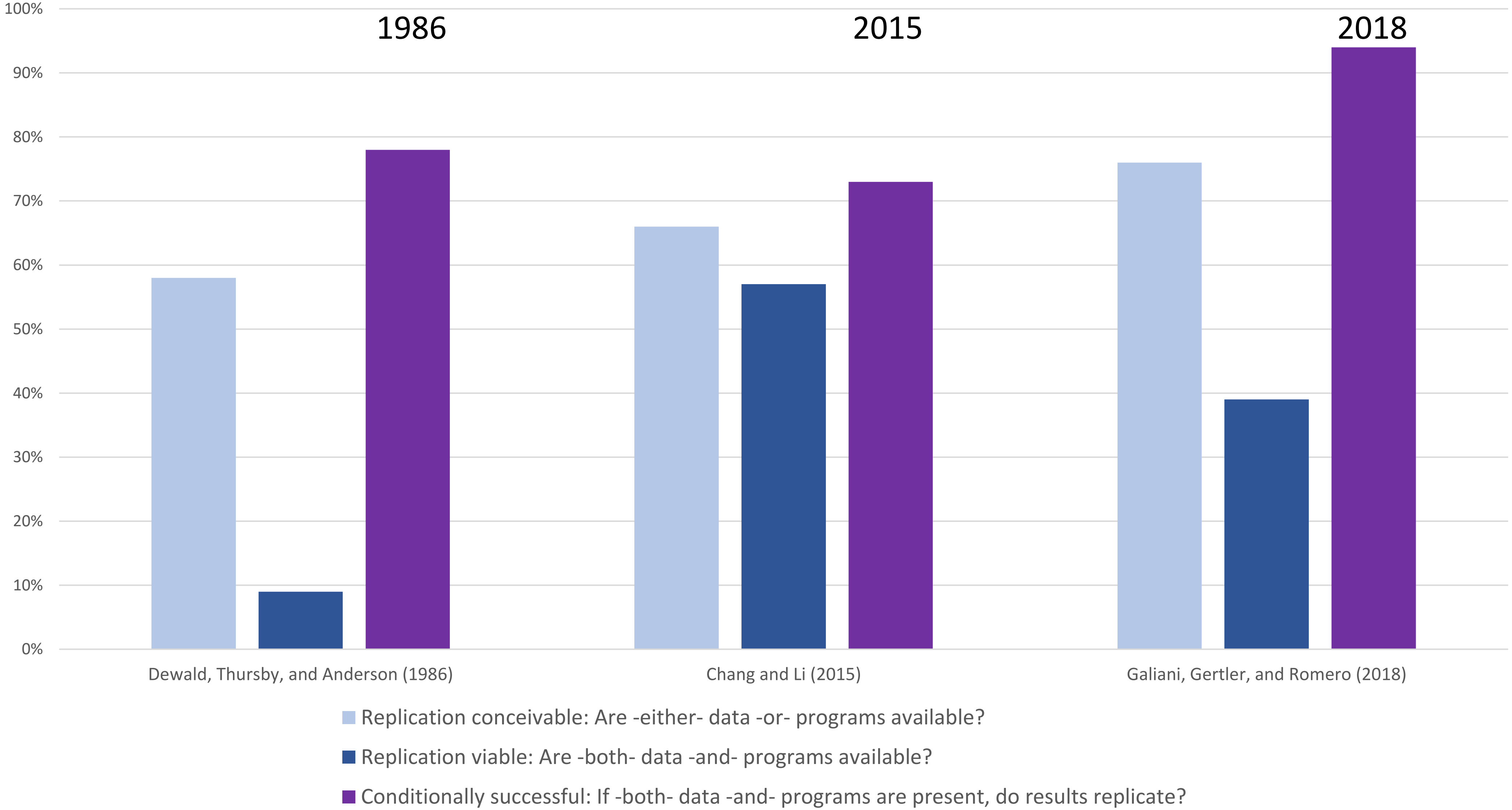
REPLICATION RARELY POSSIBLE

An analysis of 203 economics papers found that fewer than one in seven supplied the materials needed for replication.



*The elements assessed were raw data, raw code, estimation data and estimation code.

How replicable are studies in economics?



REPLICATION

Reproduction

(Fleischmann Pons / Lewis)

Lab experimental economics?

(Camerer, et al, 2016)

REPLICATION

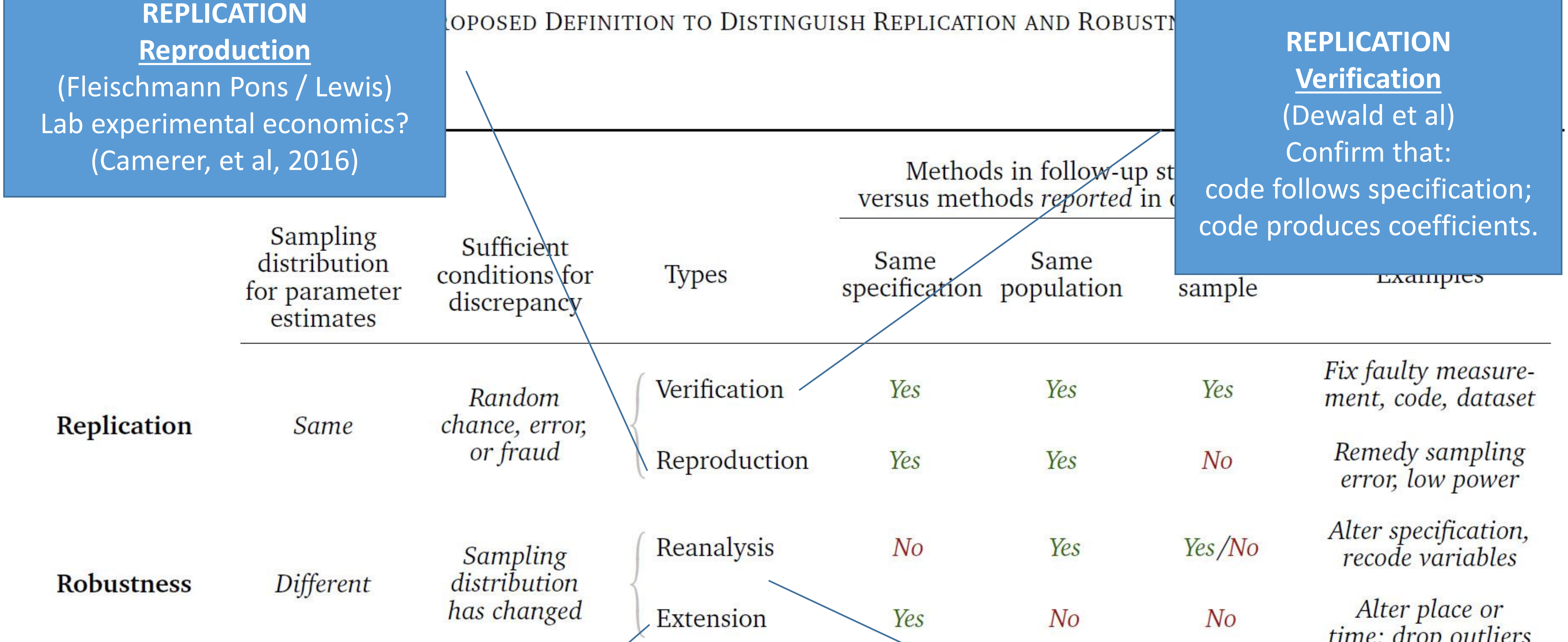
Verification

(Dewald et al)

Confirm that:

code follows specification;

code produces coefficients.



ROPOSED DEFINITION TO DISTINGUISH REPLICATION AND ROBUSTN

	Sampling distribution for parameter estimates	Sufficient conditions for discrepancy	Types	Methods in follow-up st versus methods *reported* in o			
				Same specification	Same population	sample	Examples
Replication	*Same*	*Random chance, error, or fraud*	Verification	*Yes*	*Yes*	*Yes*	*Fix faulty measurement, code, dataset*
			Reproduction	*Yes*	*Yes*	*No*	*Remedy sampling error, low power*
Robustness	*Different*	*Sampling distribution has changed*	Reanalysis	*No*	*Yes*	*Yes/No*	*Alter specification, recode variables*
			Extension	*Yes*	*No*	*No*	*Alter place or time; drop outliers*

ROBUSTNESS

Extension:

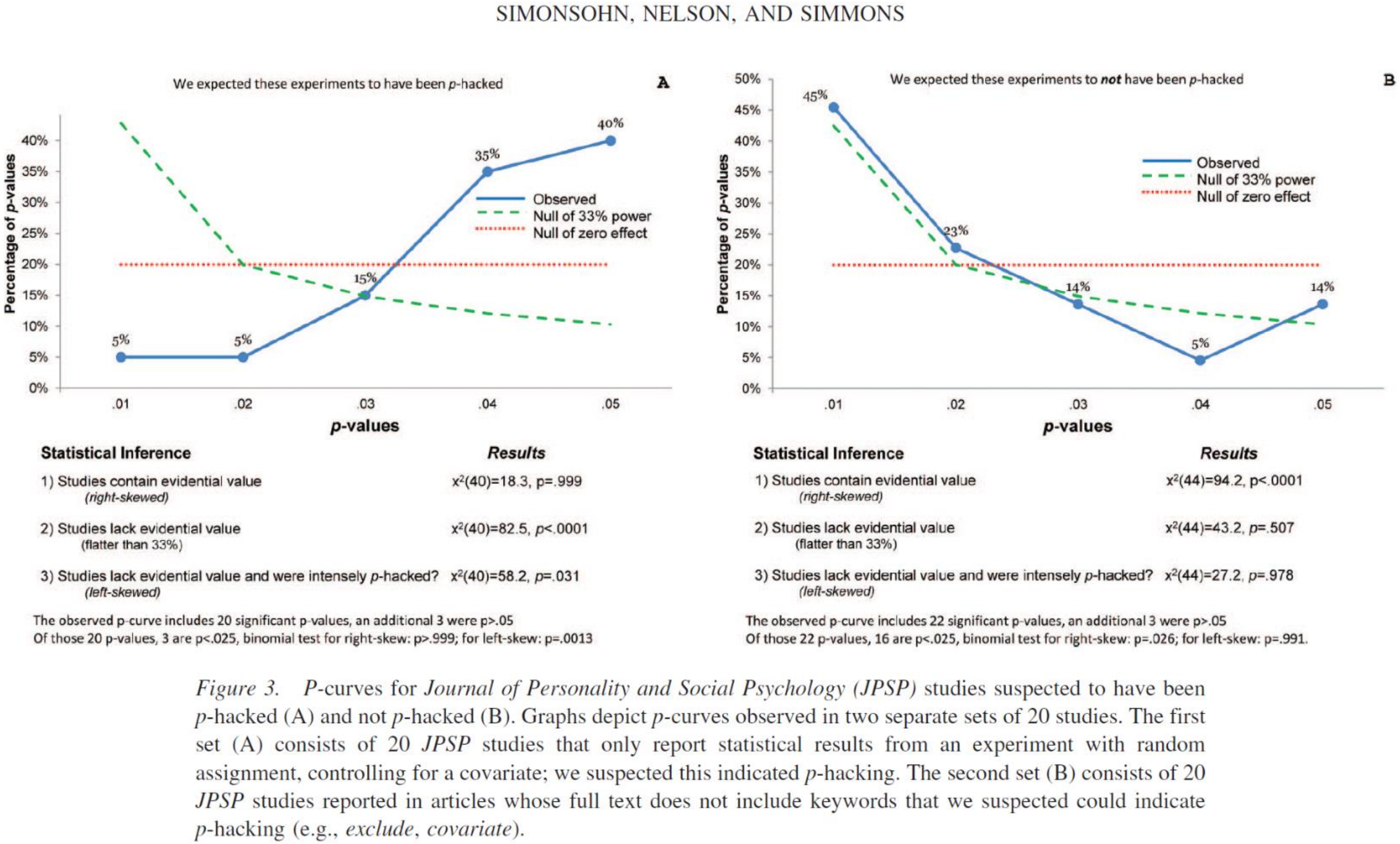
New dataset or different sample restrictions, etc.

The means the same as *reported* in the original paper, not necessarily what was contained in
Thus er contains an error such that it does not run exactly the regressions that the original pa
is ne s described in the paper).

Why would results not be reproducible in a new sample?

p-Hacking: a problem?

p-Hacking: a problem in psychology?



SIMONSOHN, NELSON, AND SIMMONS

Statistical Inference	***Results***
1) Studies contain evidential value *(right-skewed)*	$\chi^2(40)=18.3$, p=.999
2) Studies lack evidential value (flatter than 33%)	$\chi^2(40)=82.5$, *p*<.0001
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	$\chi^2(40)=58.2$, *p*=.031

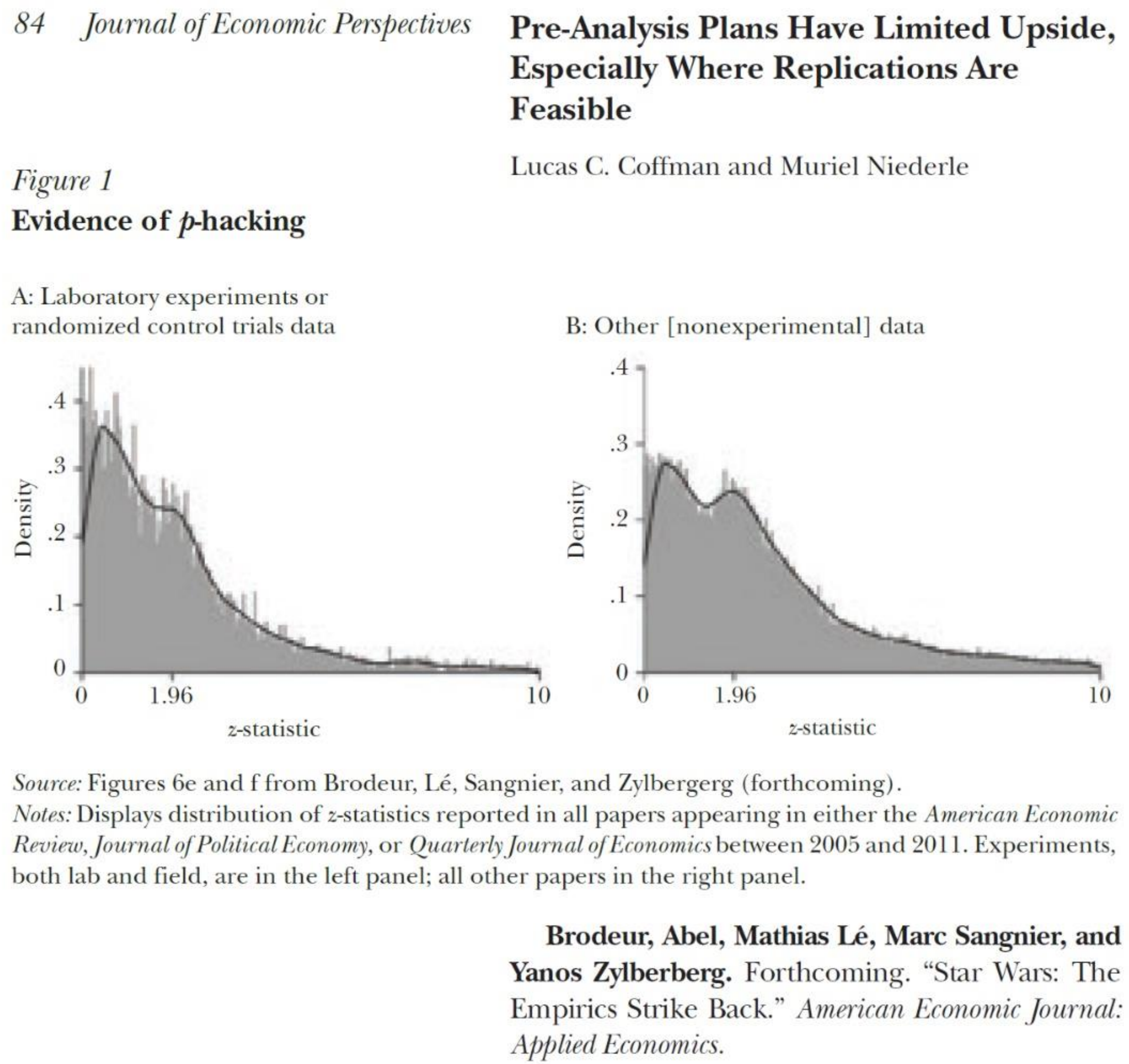
The observed p-curve includes 20 significant p-values, an additional 3 were p>.05
Of those 20 p-values, 3 are p<.025, binomial test for right-skew: p>.999; for left-skew: p=.0013

Statistical Inference	***Results***
1) Studies contain evidential value *(right-skewed)*	$\chi^2(44)=94.2$, p<.0001
2) Studies lack evidential value (flatter than 33%)	$\chi^2(44)=43.2$, *p*=.507
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	$\chi^2(44)=27.2$, *p*=.978

The observed p-curve includes 22 significant p-values, an additional 3 were p>.05
Of those 22 p-values, 16 are p<.025, binomial test for right-skew: p=.026; for left-skew: p=.991.

Figure 3. *P*-curves for *Journal of Personality and Social Psychology (JPSP)* studies suspected to have been *p*-hacked (A) and not *p*-hacked (B). Graphs depict *p*-curves observed in two separate sets of 20 studies. The first set (A) consists of 20 *JPSP* studies that only report statistical results from an experiment with random assignment, controlling for a covariate; we suspected this indicated *p*-hacking. The second set (B) consists of 20 *JPSP* studies reported in articles whose full text does not include keywords that we suspected could indicate *p*-hacking (e.g., *exclude*, *covariate*).

p-Hacking: a problem in economics?



84 *Journal of Economic Perspectives*

Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible

Lucas C. Coffman and Muriel Niederle

Figure 1
Evidence of *p*-hacking

A: Laboratory experiments or randomized control trials data

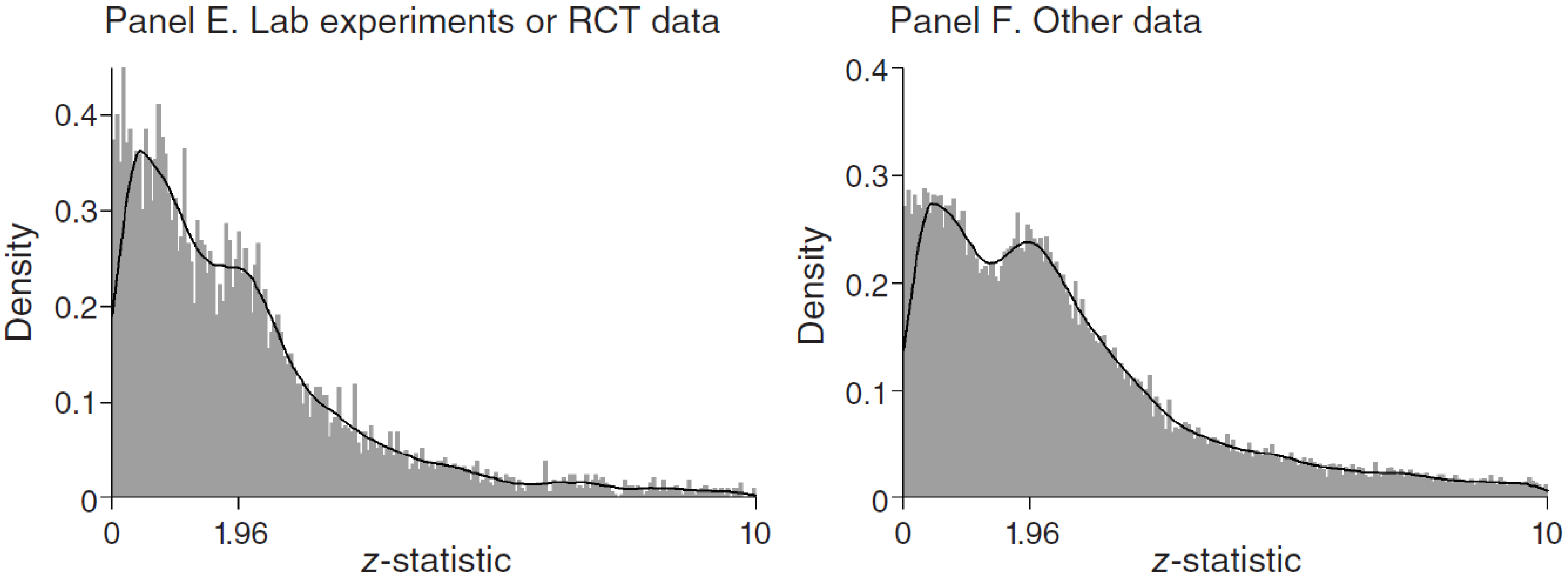
B: Other [nonexperimental] data

.4
.3
.2
.1
0
Density
0
1.96
10
z-statistic
.4
.3
.2
.1
0
Density
0
1.96
10
z-statistic

Source: Figures 6e and f from Brodeur, Lé, Sangnier, and Zylbergerg (forthcoming).
Notes: Displays distribution of *z*-statistics reported in all papers appearing in either the *American Economic Review*, *Journal of Political Economy*, or *Quarterly Journal of Economics* between 2005 and 2011. Experiments, both lab and field, are in the left panel; all other papers in the right panel.

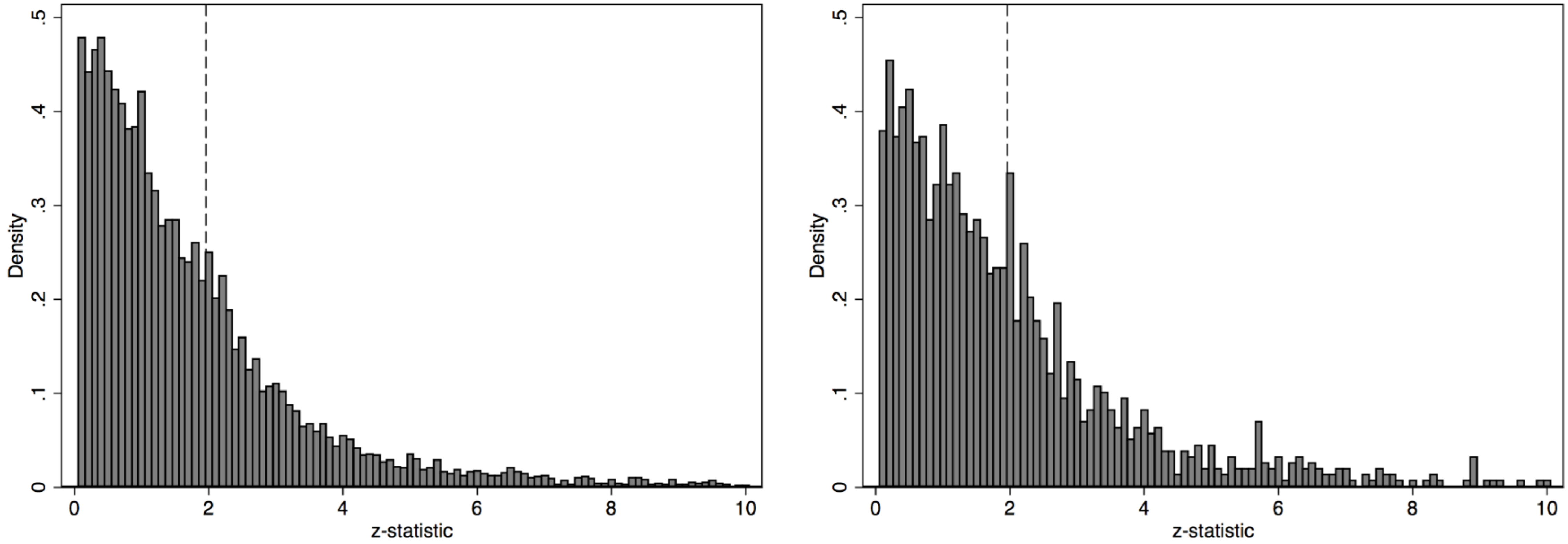
Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Forthcoming. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics.*

p-Hacking: a problem in economics?



From: Brodeur, Lé, Sagnier, and Zylberberg

p-Hacking: a problem in economics?



From: Vivalt

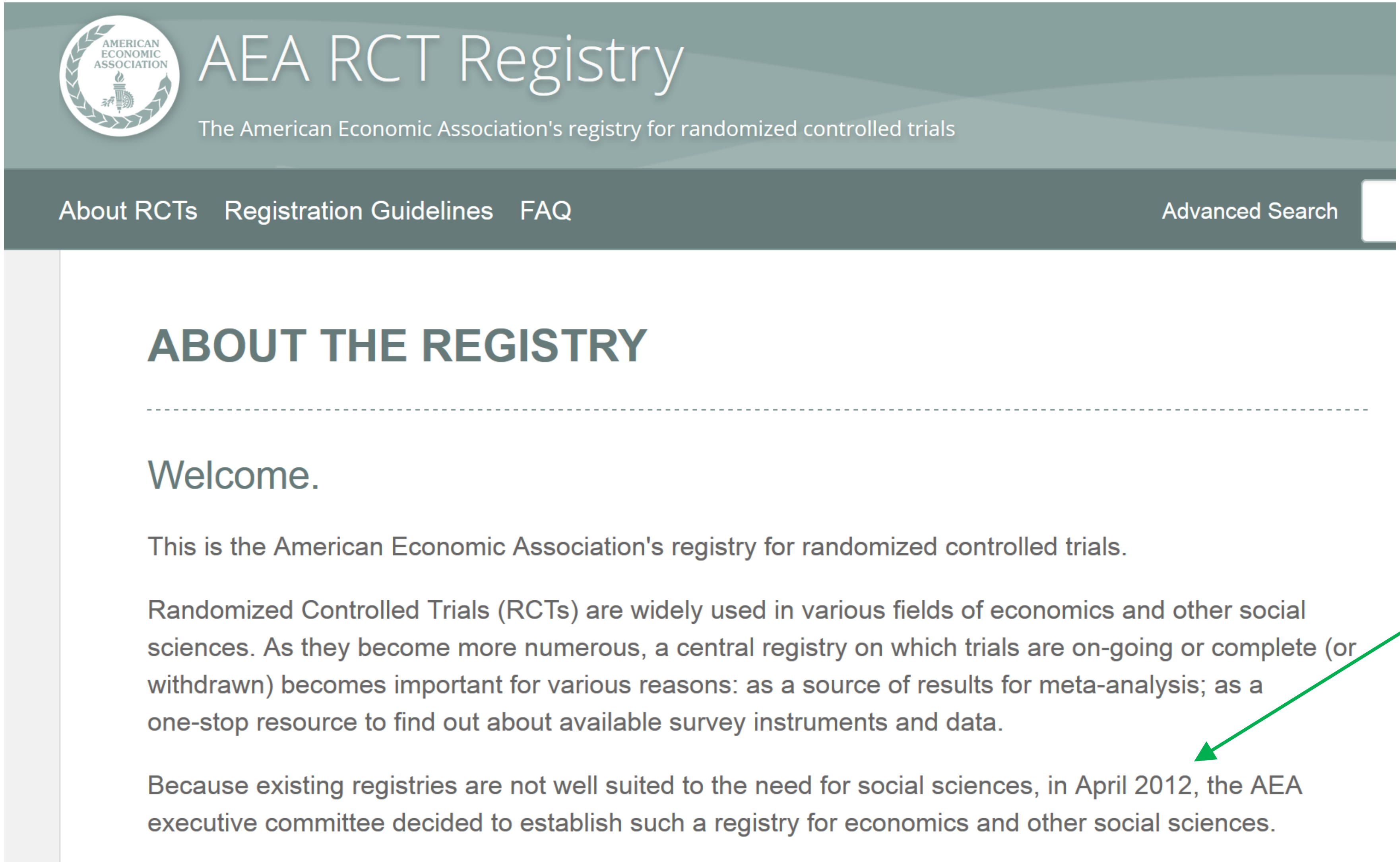
What can we do?

(for any single new study)

Pre-analysis plans: not the simplest thing.

“Pre-specifying the entire chain of logic for every possible realization of the data can quickly become an overwhelming task for even the most committed pre-specifier.” *Olken 2015*

Pre-analysis plans: a short history



AMERICAN ECONOMIC ASSOCIATION
AEA RCT Registry
The American Economic Association's registry for randomized controlled trials
About RCTs
Registration Guidelines
FAQ
Advanced Search

ABOUT THE REGISTRY

Welcome.

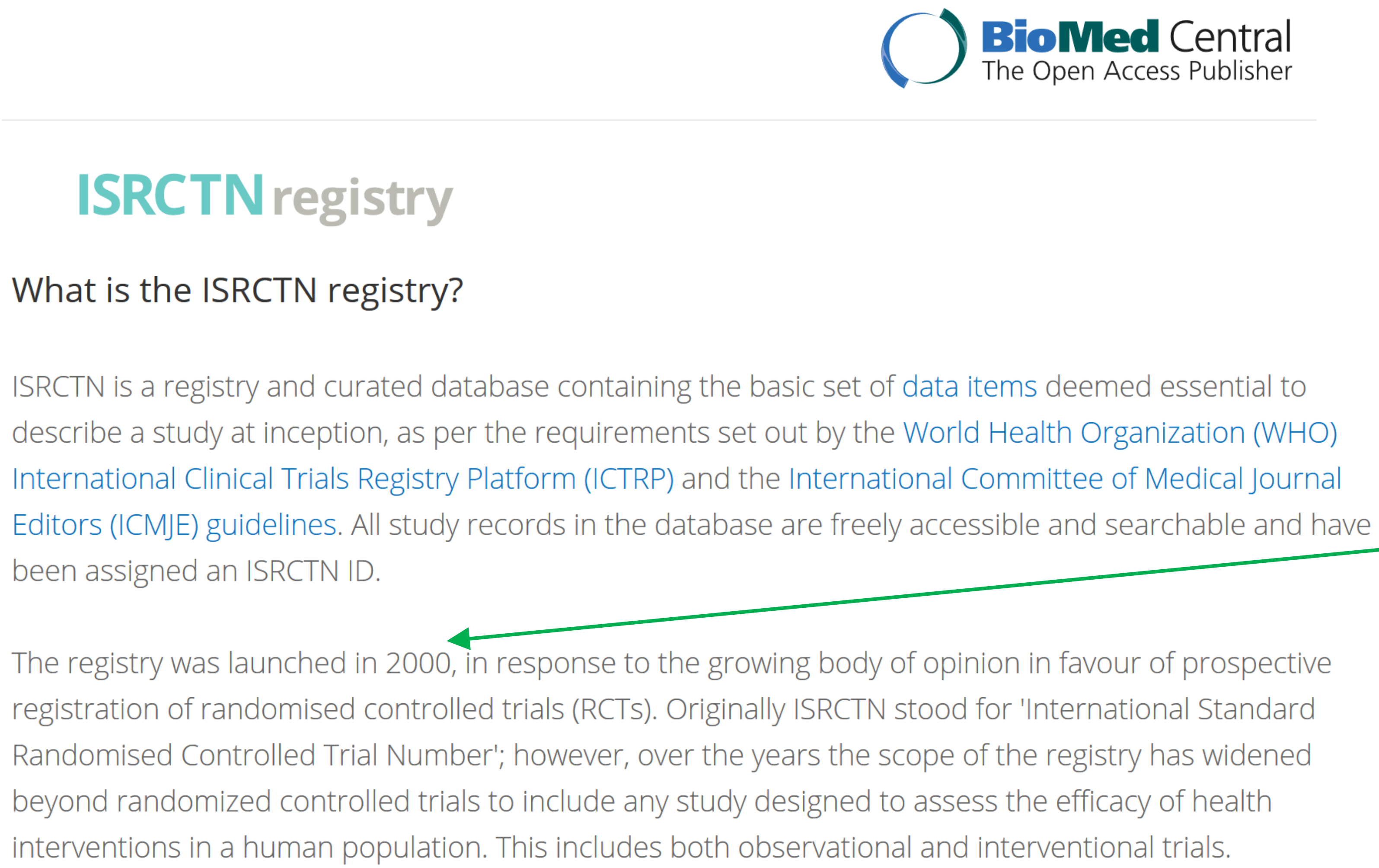
This is the American Economic Association's registry for randomized controlled trials.

Randomized Controlled Trials (RCTs) are widely used in various fields of economics and other social sciences. As they become more numerous, a central registry on which trials are on-going or complete (or withdrawn) becomes important for various reasons: as a source of results for meta-analysis; as a one-stop resource to find out about available survey instruments and data.

Because existing registries are not well suited to the need for social sciences, in April 2012, the AEA executive committee decided to establish such a registry for economics and other social sciences.

2012

Pre-analysis plans: a short history



BioMed Central
The Open Access Publisher

ISRCTN registry

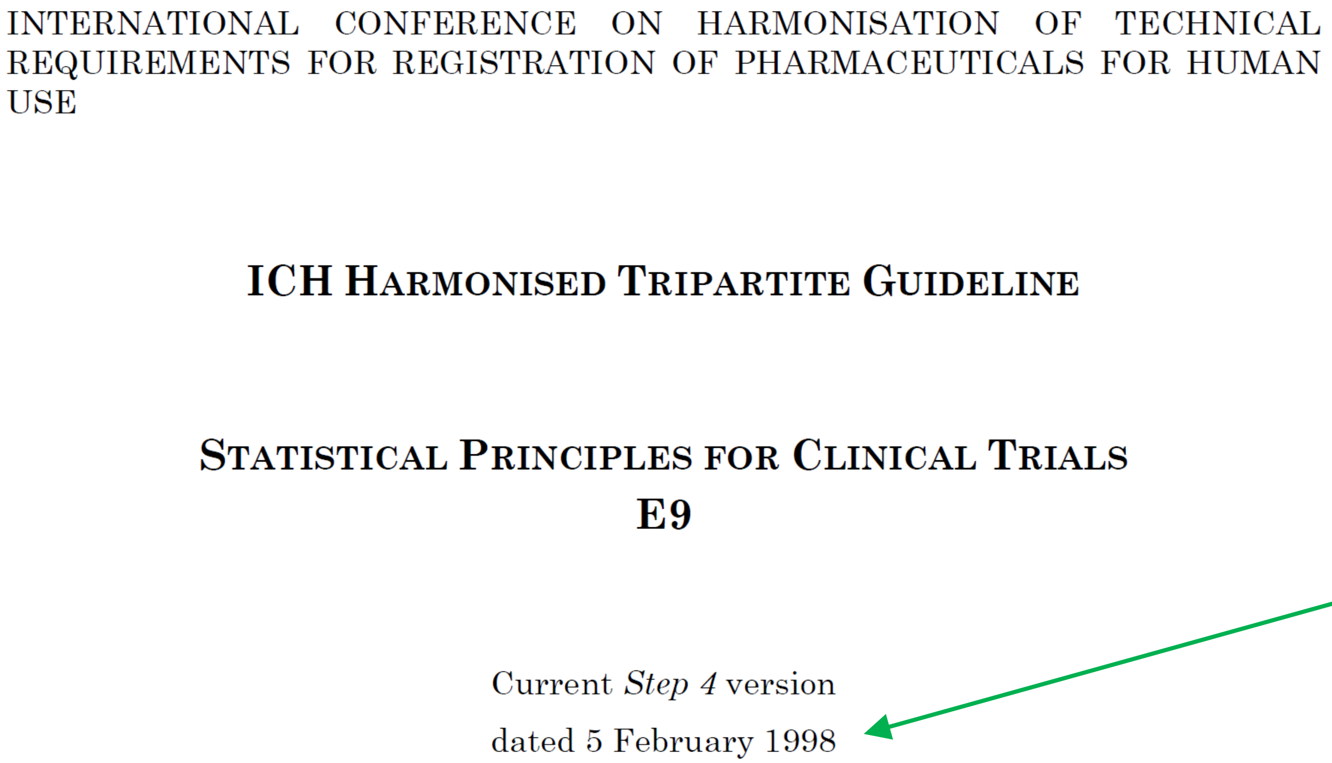
What is the ISRCTN registry?

ISRCTN is a registry and curated database containing the basic set of data items deemed essential to describe a study at inception, as per the requirements set out by the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP) and the International Committee of Medical Journal Editors (ICMJE) guidelines. All study records in the database are freely accessible and searchable and have been assigned an ISRCTN ID.

The registry was launched in 2000, in response to the growing body of opinion in favour of prospective registration of randomised controlled trials (RCTs). Originally ISRCTN stood for 'International Standard Randomised Controlled Trial Number'; however, over the years the scope of the registry has widened beyond randomized controlled trials to include any study designed to assess the efficacy of health interventions in a human population. This includes both observational and interventional trials.

2000

Pre-analysis plans: a short history



INTERNATIONAL CONFERENCE ON HARMONISATION OF TECHNICAL REQUIREMENTS FOR REGISTRATION OF PHARMACEUTICALS FOR HUMAN USE

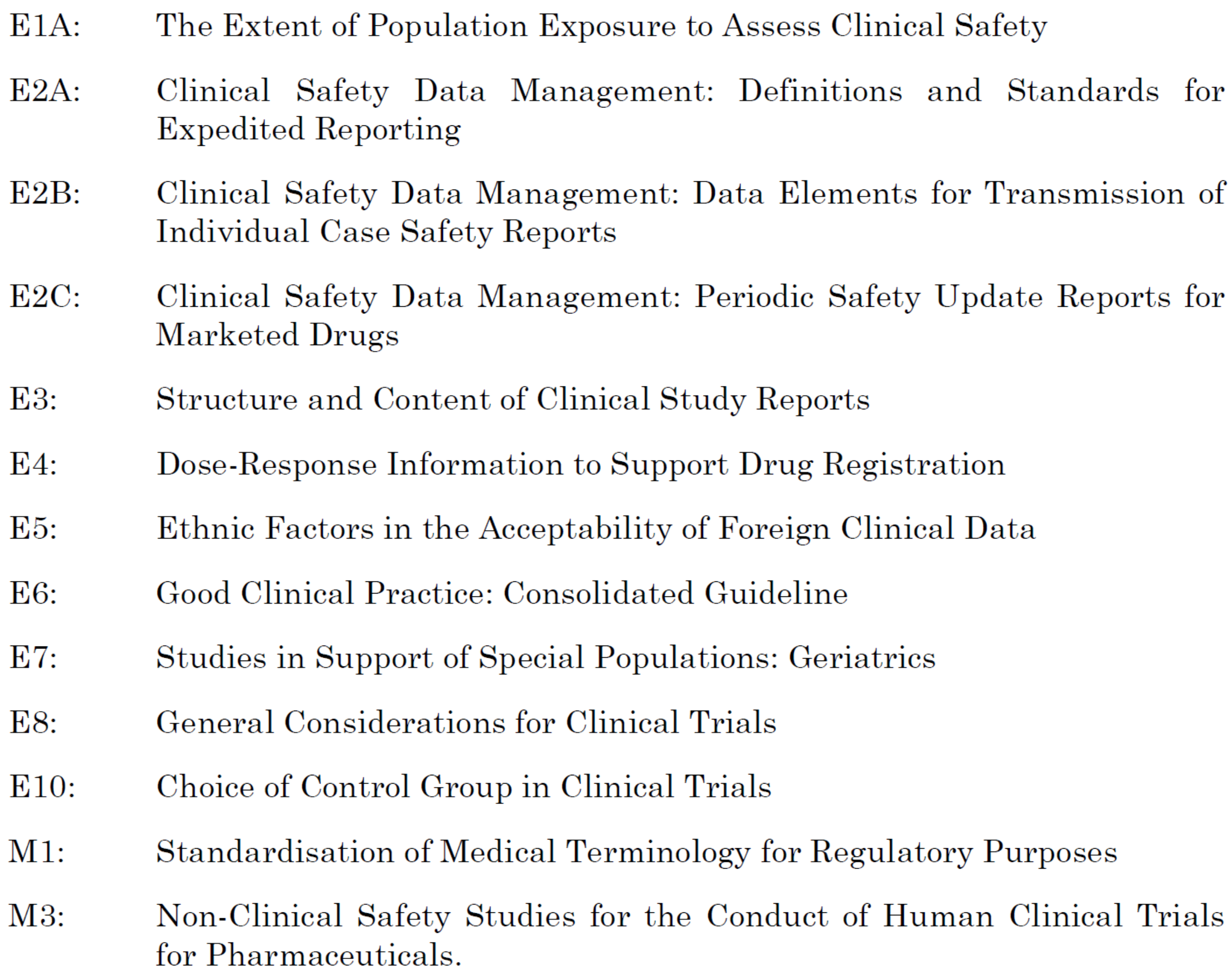
ICH HARMONISED TRIPARTITE GUIDELINE

STATISTICAL PRINCIPLES FOR CLINICAL TRIALS
E9

Current *Step 4* version
dated 5 February 1998

1998

Pre-analysis plans: a short history

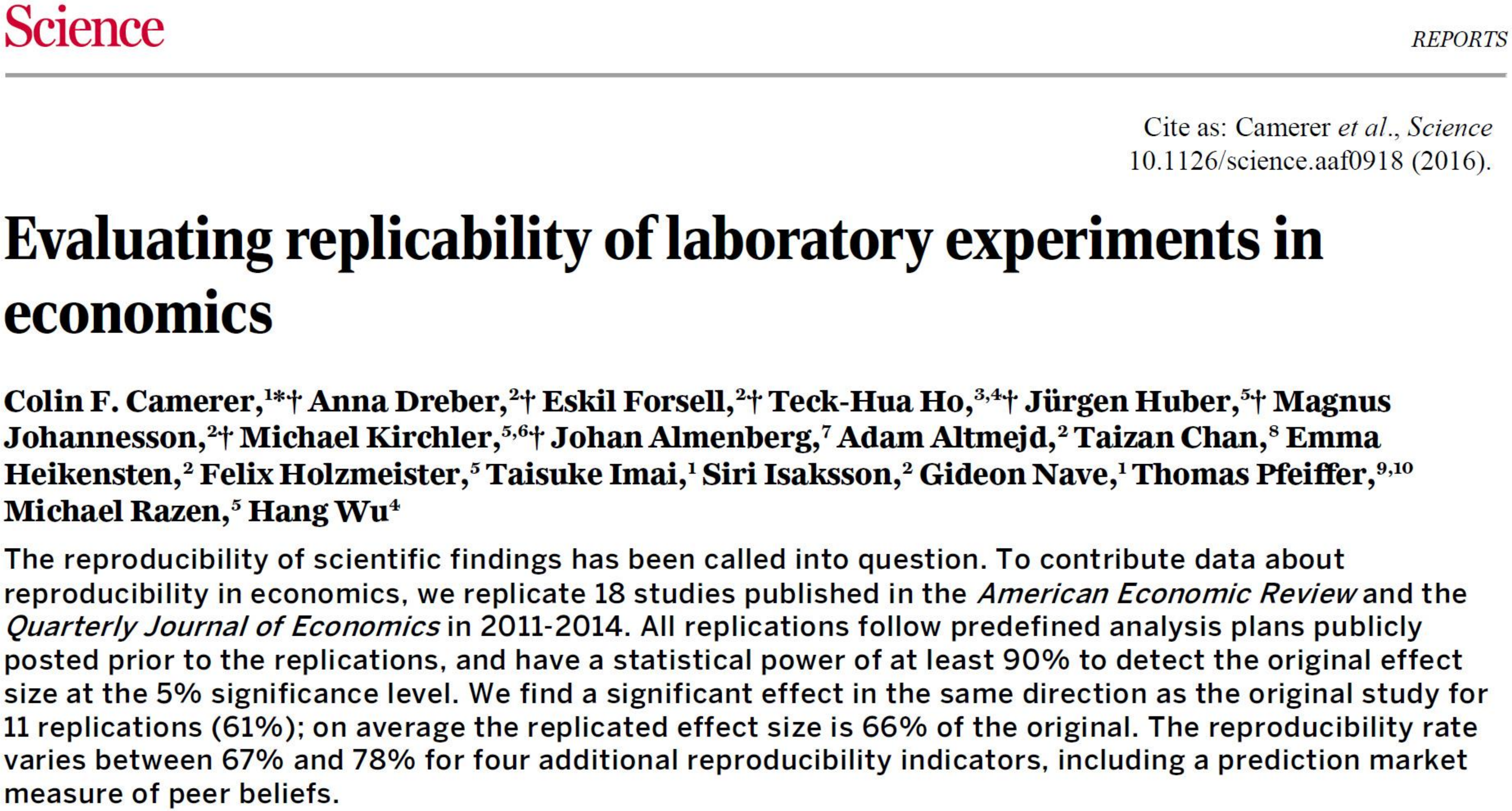


E1A:	The Extent of Population Exposure to Assess Clinical Safety
E2A:	Clinical Safety Data Management: Definitions and Standards for Expedited Reporting
E2B:	Clinical Safety Data Management: Data Elements for Transmission of Individual Case Safety Reports
E2C:	Clinical Safety Data Management: Periodic Safety Update Reports for Marketed Drugs
E3:	Structure and Content of Clinical Study Reports
E4:	Dose-Response Information to Support Drug Registration
E5:	Ethnic Factors in the Acceptability of Foreign Clinical Data
E6:	Good Clinical Practice: Consolidated Guideline
E7:	Studies in Support of Special Populations: Geriatrics
E8:	General Considerations for Clinical Trials
E10:	Choice of Control Group in Clinical Trials
M1:	Standardisation of Medical Terminology for Regulatory Purposes
M3:	Non-Clinical Safety Studies for the Conduct of Human Clinical Trials for Pharmaceuticals.

What can we do?

(across multiple studies)

How to replicate without perverse incentives



Science

REPORTS

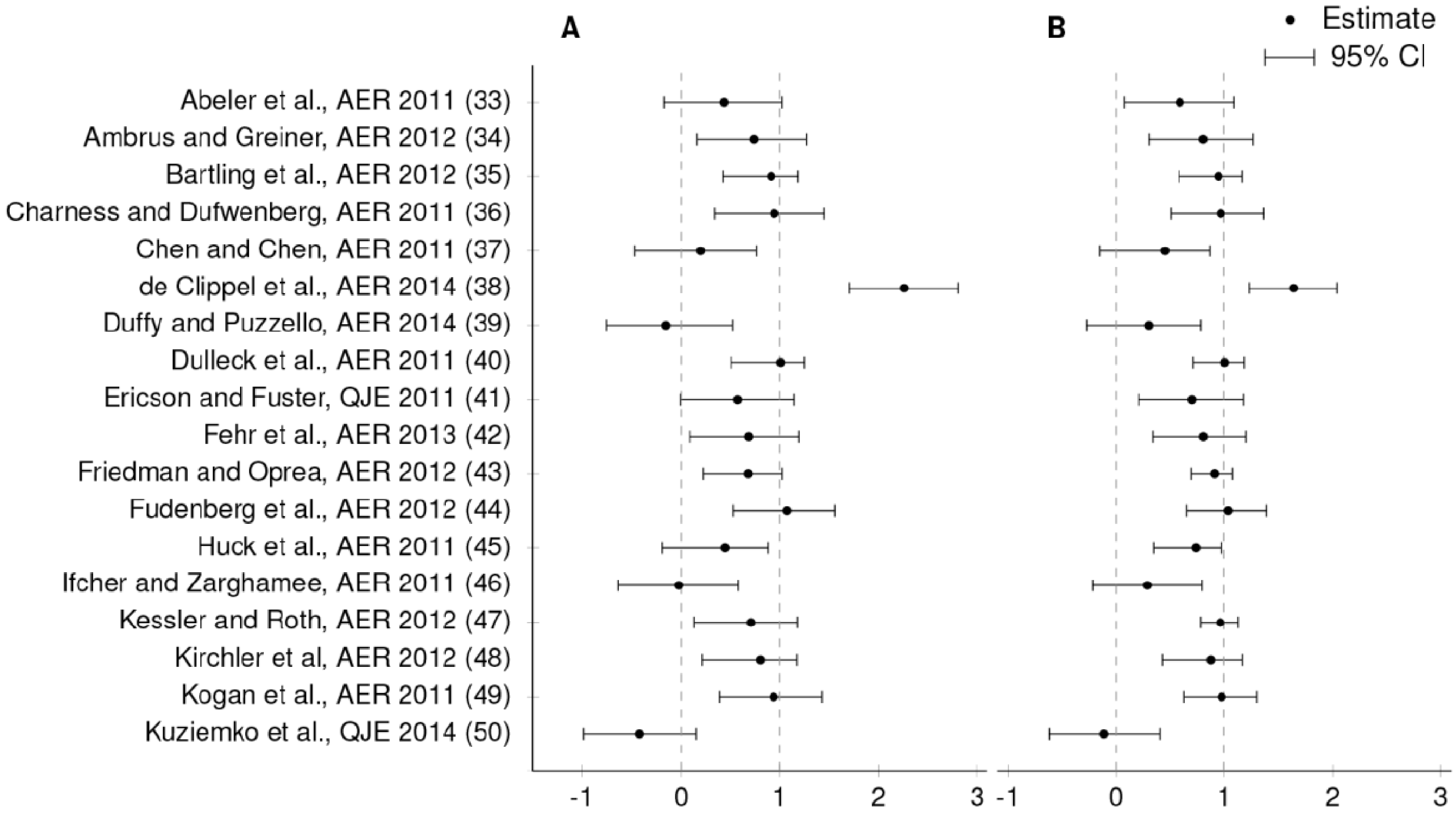
Cite as: Camerer *et al*., *Science*
10.1126/science.aaf0918 (2016).

Evaluating replicability of laboratory experiments in economics

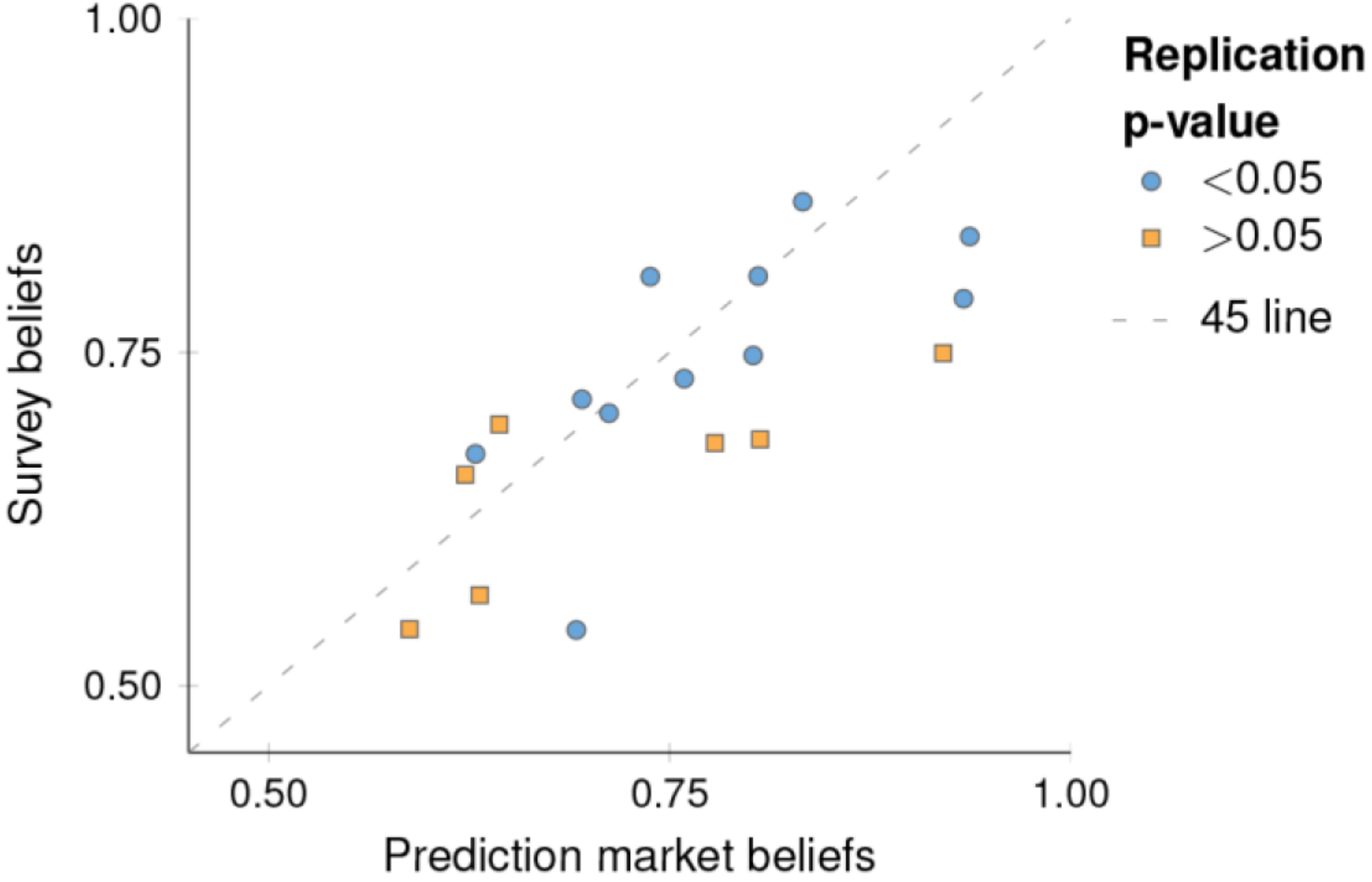
Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10] Michael Razen,[5] Hang Wu[4]

The reproducibility of scientific findings has been called into question. To contribute data about reproducibility in economics, we replicate 18 studies published in the *American Economic Review* and the *Quarterly Journal of Economics* in 2011-2014. All replications follow predefined analysis plans publicly posted prior to the replications, and have a statistical power of at least 90% to detect the original effect size at the 5% significance level. We find a significant effect in the same direction as the original study for 11 replications (61%); on average the replicated effect size is 66% of the original. The reproducibility rate varies between 67% and 78% for four additional reproducibility indicators, including a prediction market measure of peer beliefs.

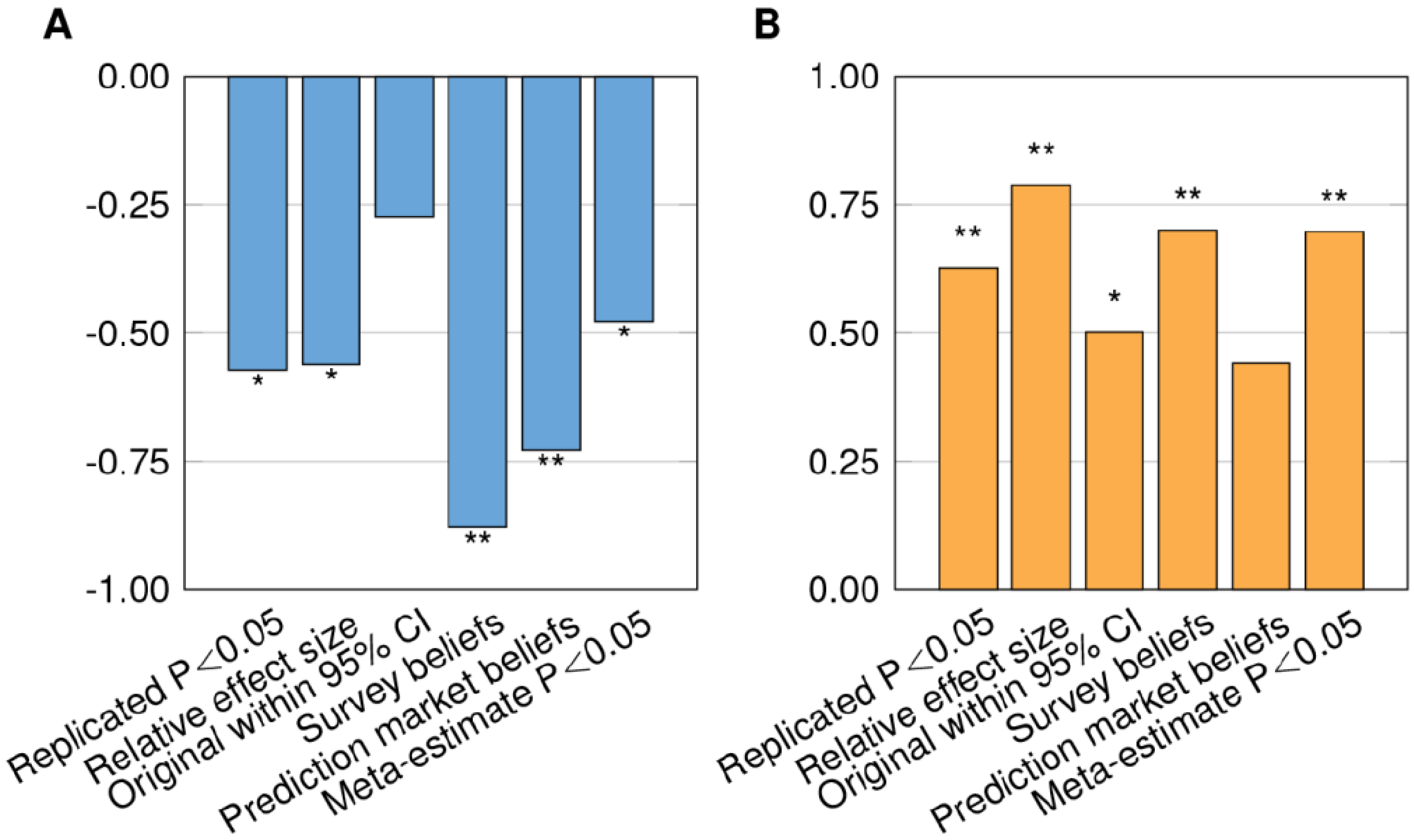
Camerer, et al., 2016



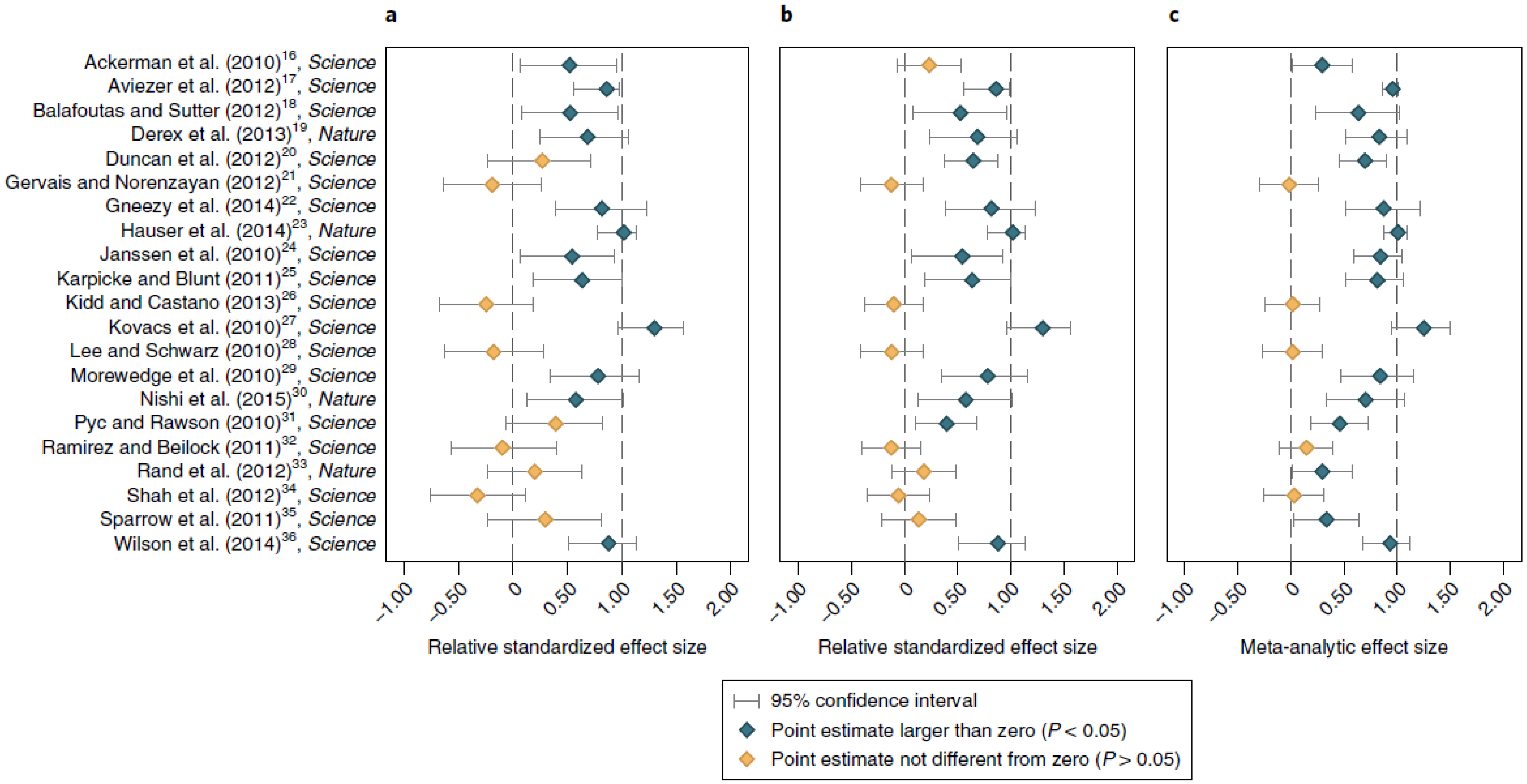
Camerer, et al., 2016



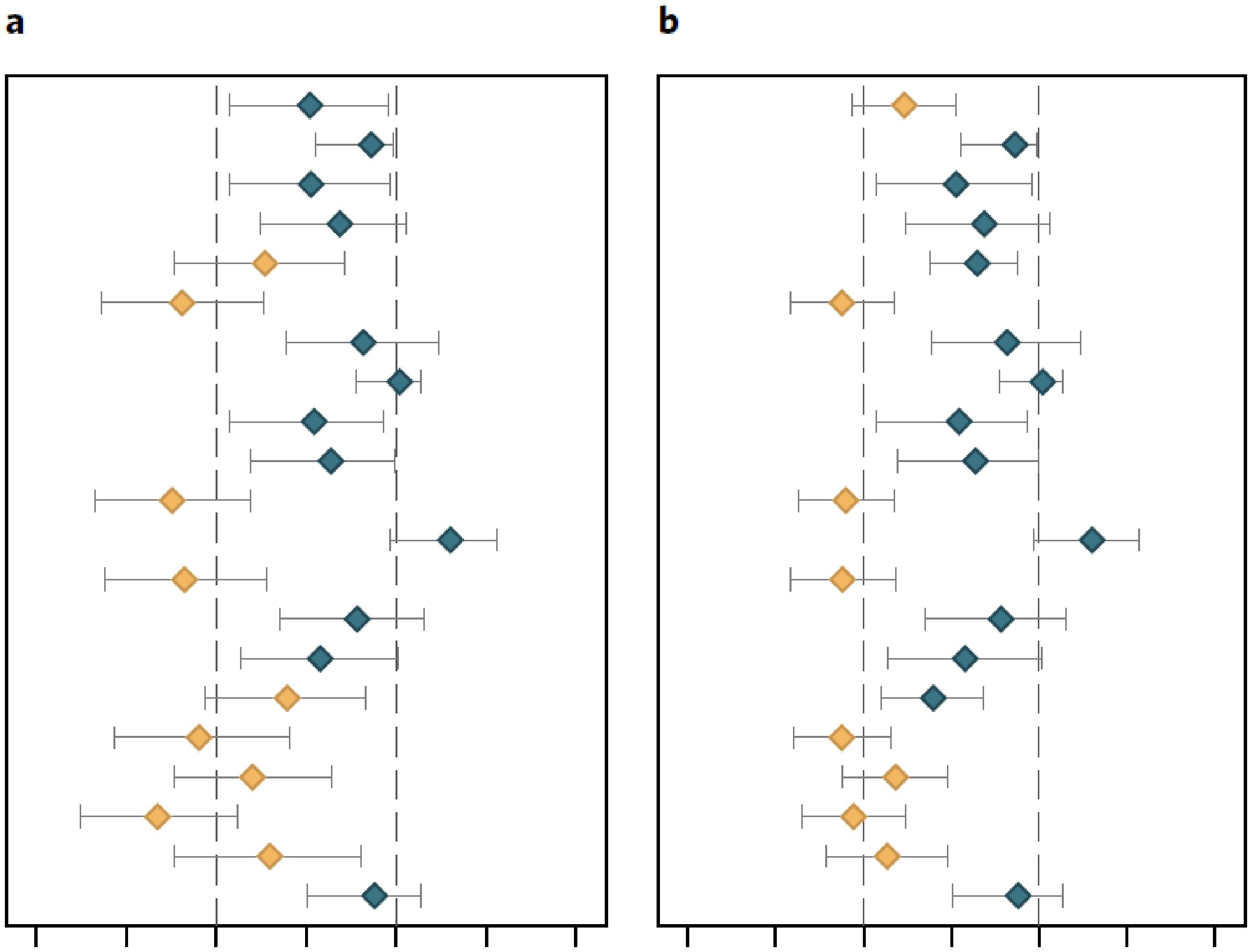
Camerer, et al., 2016



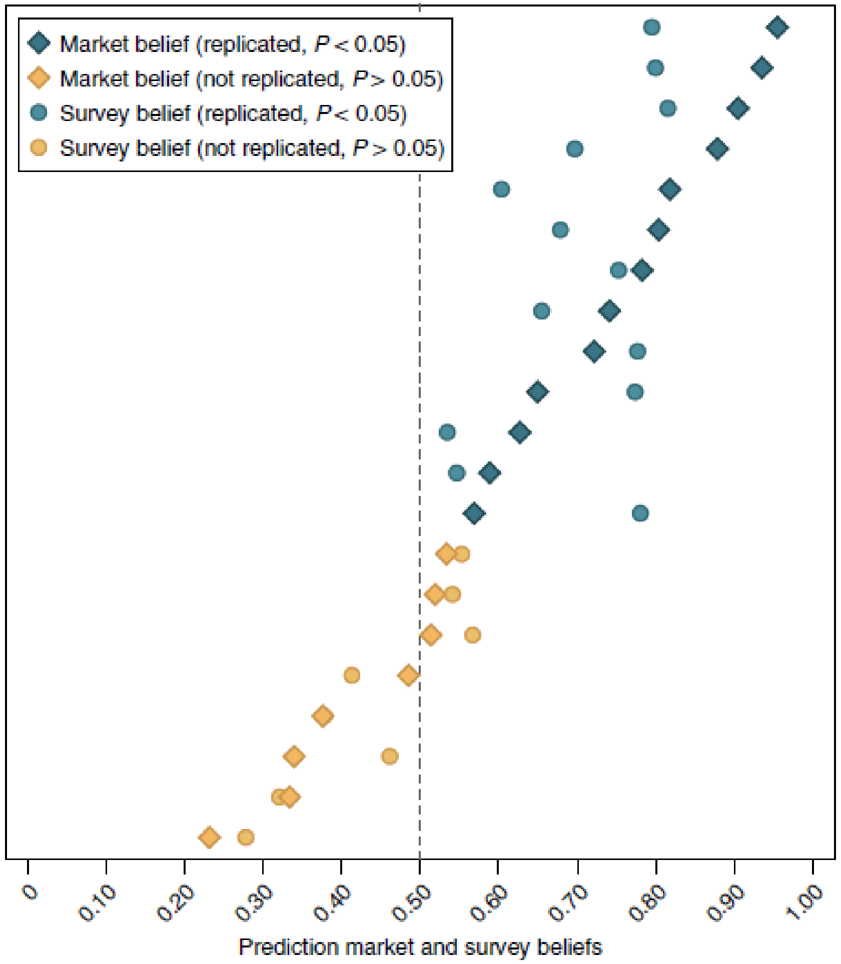
Camerer, et al., **2018**



Camerer, et al., **2018**



Camerer, et al., **2018**



Beyond p-hacking
a “file drawer problem”

What might you expect?

- Suppose 900 hypotheses are tested in which there is no pattern to find – the null holds. In expectation, how many false positives (“statistically significant, nonzero” coefficients, tested at the 5 percent level) will be found?

What might you expect?

- Suppose 900 hypotheses are tested in which there is no pattern to find – the null holds. In expectation, how many false positives (“statistically significant, nonzero” coefficients, tested at the 5 percent level) will be found?
  - 900 x 0.05 = 45
- Suppose 100 hypotheses are tested in which a true effect is present, but the test used has power 0.80 to detect the effect of that magnitude. In expectation, how many of these true effects will be detected (“statistically significant, nonzero” coefficients, tested at the 5 percent level)?

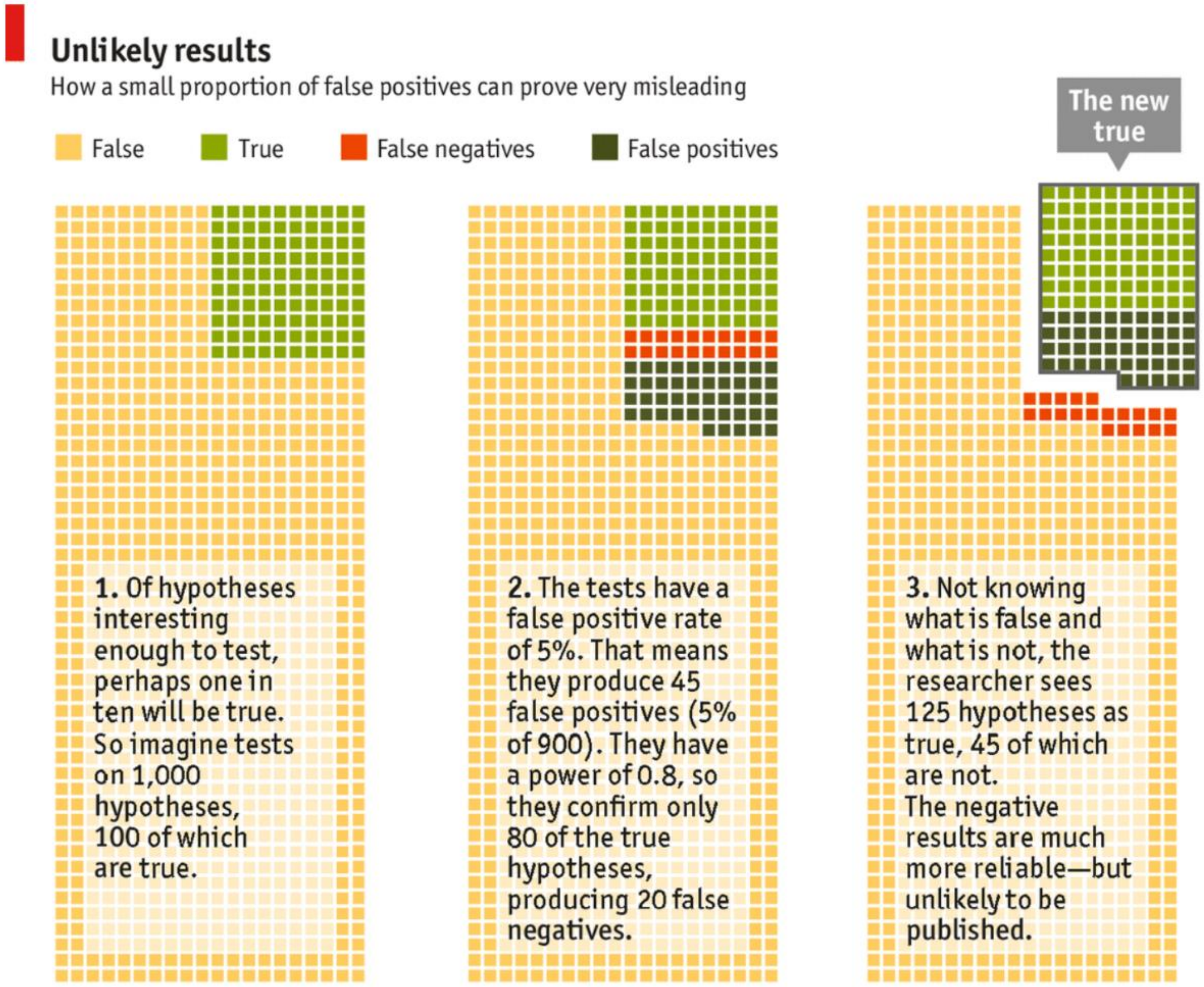
What might you expect?

- Suppose 900 hypotheses are tested in which there is no pattern to find – the null holds. In expectation, how many false positives ("statistically significant, nonzero" coefficients, tested at the 5 percent level) will be found?
  - 900 x 0.05 = 45
- Suppose 100 hypotheses are tested in which a true effect is present, but the test used has power 0.80 to detect the effect of that magnitude. In expectation, how many of these true effects will be detected ("statistically significant, nonzero" coefficients, tested at the 5 percent level)?
  - 100 x 0.80 = 80
- So if there were a file drawer problem in which we only observed significant results, and the hypotheses tested were as described above, what fraction of results would represent "true effects" rather than "false positives" ?

What might you expect?

- Suppose 900 hypotheses are tested in which there is no pattern to find – the null holds. In expectation, how many false positives ("statistically significant, nonzero" coefficients, tested at the 5 percent level) will be found?
    - 900 x 0.05 = 45
- Suppose 100 hypotheses are tested in which a true effect is present, but the test used has power 0.80 to detect the effect of that magnitude. In expectation, how many of these true effects will be detected ("statistically significant, nonzero" coefficients, tested at the 5 percent level)?
    - 100 x 0.80 = 80
- So if there were a file drawer problem in which we only observed significant results, and the hypotheses tested were as described above, what fraction of results would represent "true effects" rather than "false positives" ?
    - 80 / 125 , or about 64 percent.

"Trouble at the lab" -- The Economist, October 19, 2013



Unlikely results

How a small proportion of false positives can prove very misleading

1. Of hypotheses interesting enough to test, perhaps one in ten will be true. So imagine tests on 1,000 hypotheses, 100 of which are true.

2. The tests have a false positive rate of 5%. That means they produce 45 false positives (5% of 900). They have a power of 0.8, so they confirm only 80 of the true hypotheses, producing 20 false negatives.

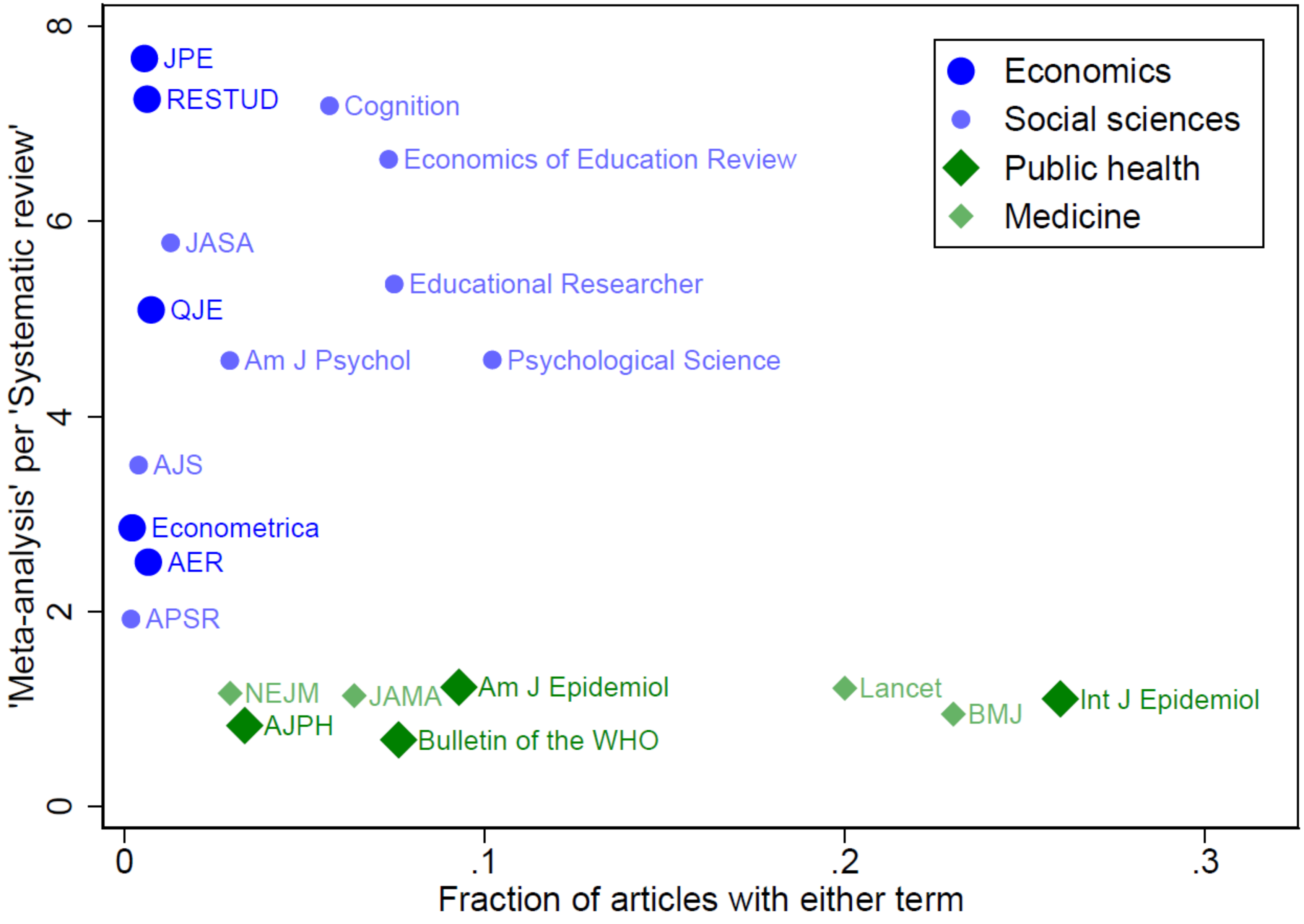
3. Not knowing what is false and what is not, the researcher sees 125 hypotheses as true, 45 of which are not. The negative results are much more reliable—but unlikely to be published.

Source: *The Economist*

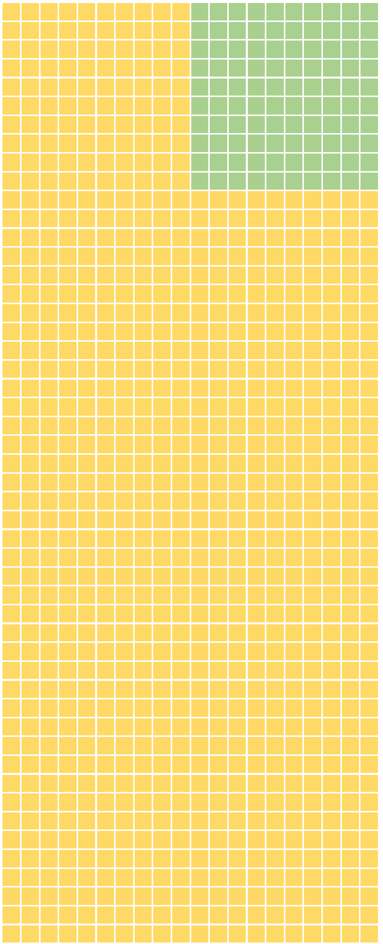
Aggregating evidence

How to do better with more than one study

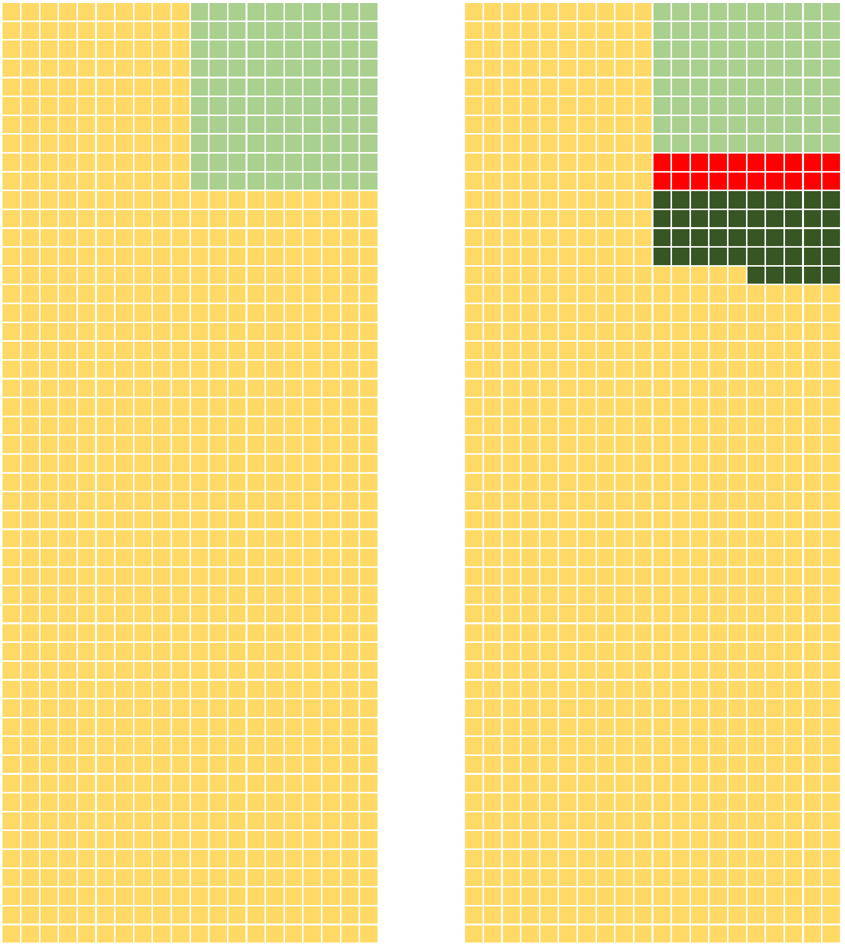
What is a review?



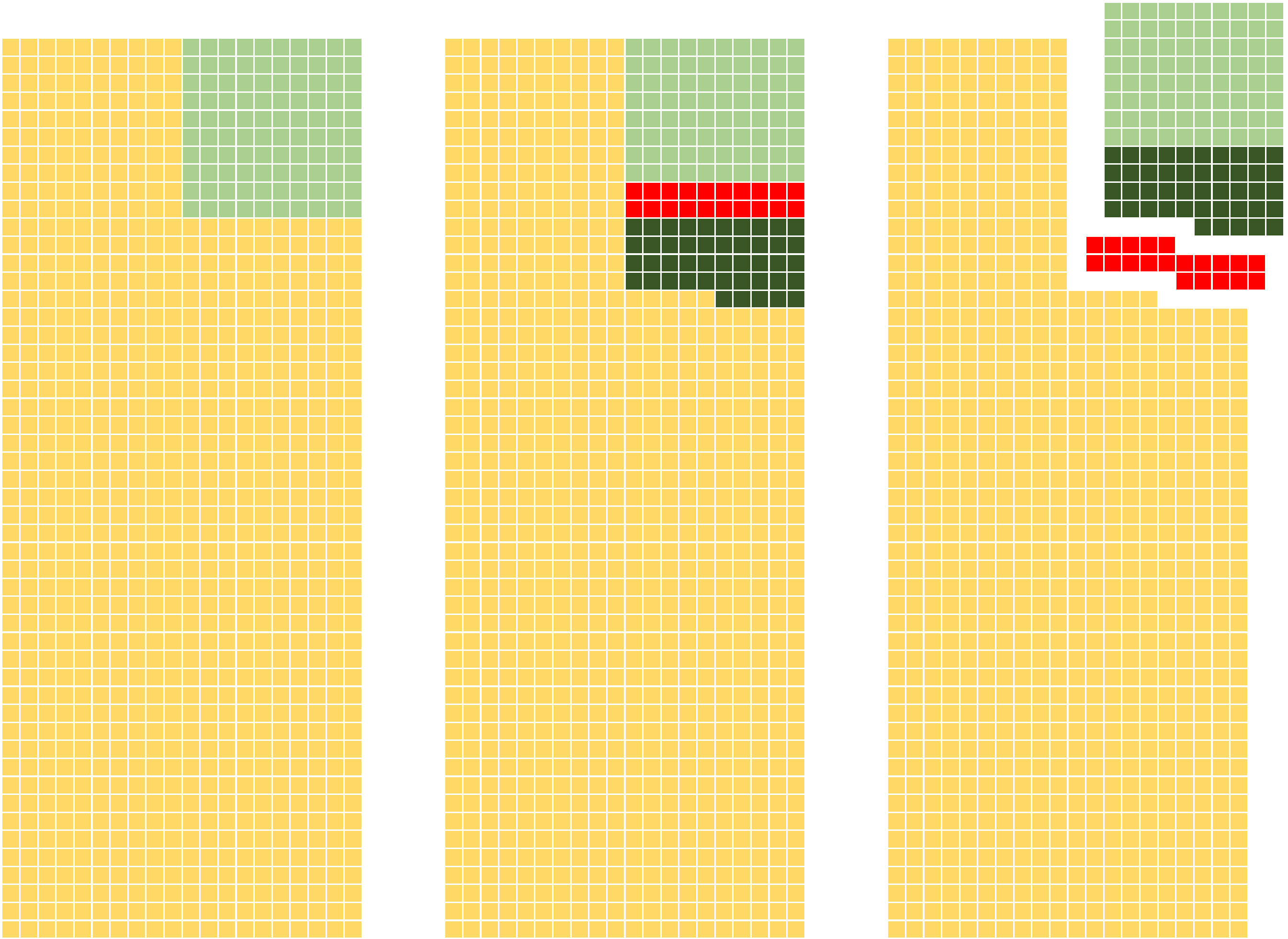
Systematic review and meta-analysis: the answer?



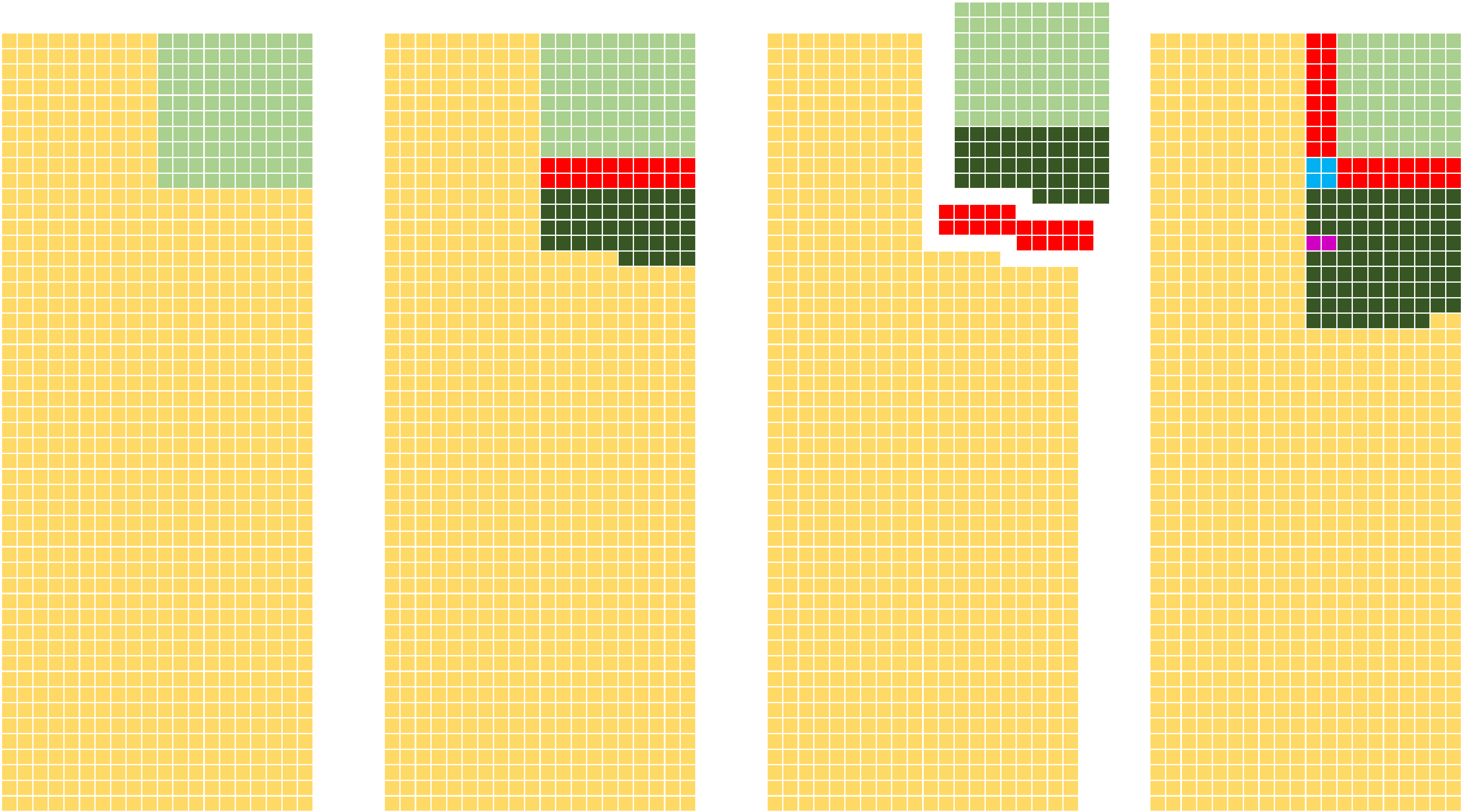
Systematic review and meta-analysis: the answer?



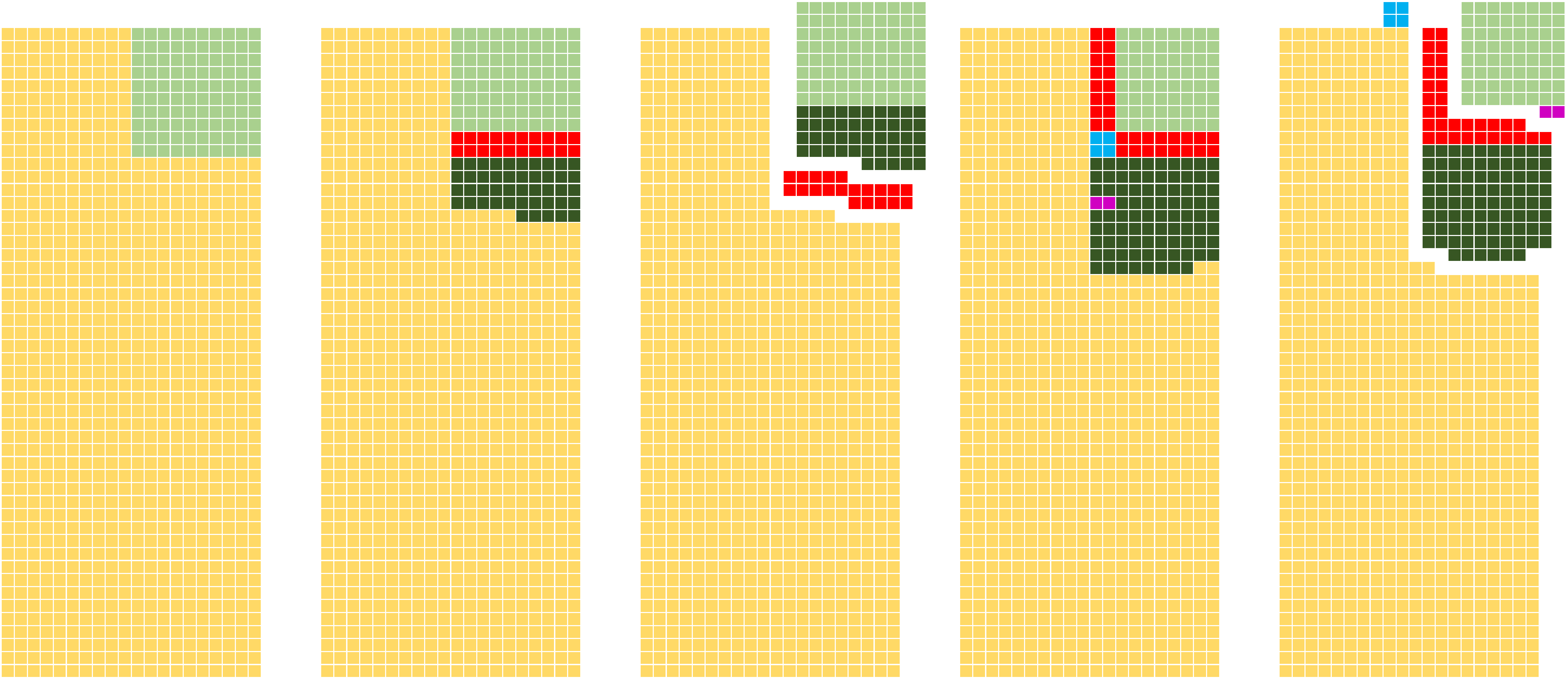
Systematic review and meta-analysis: the answer?



Systematic review and meta-analysis: the answer?

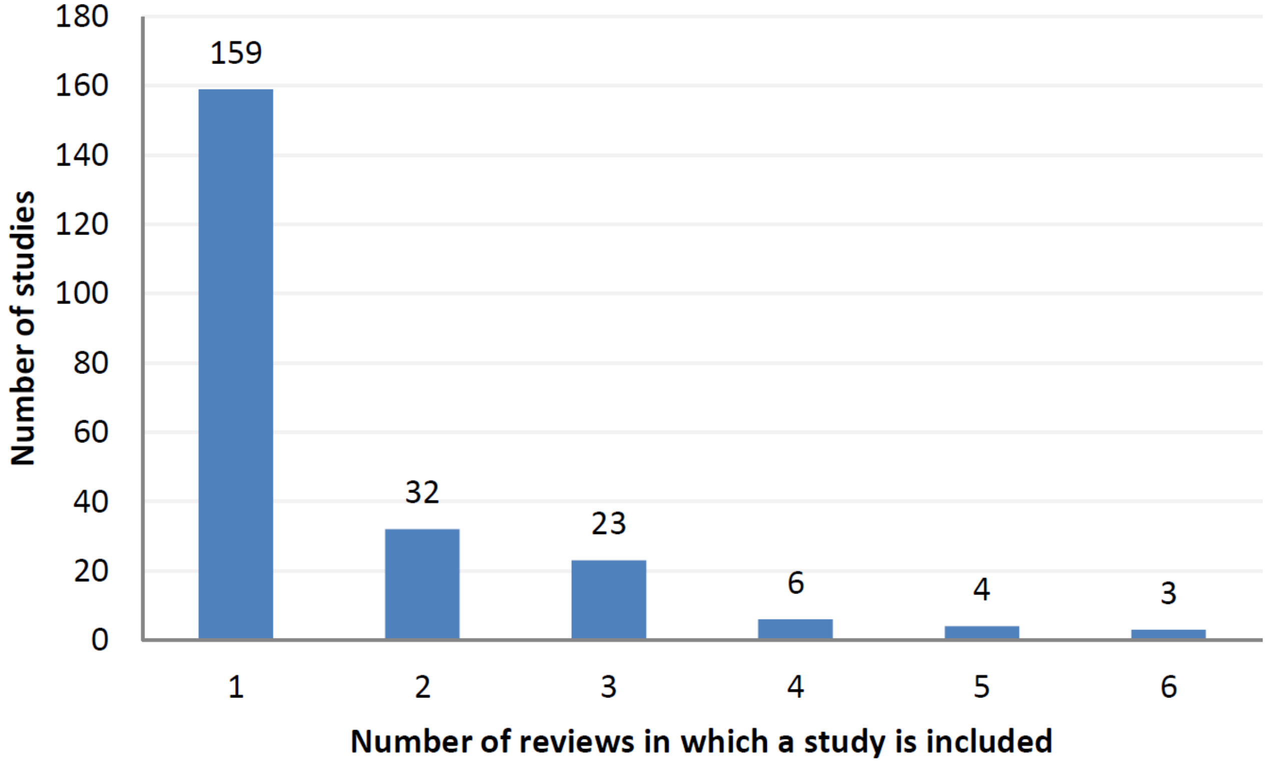


Systematic review and meta-analysis: the answer?



Meta-analyses and systematic reviews

Figure 3: Distribution of Learning Studies across Systematic Reviews



Note: The total number of learning studies is 227.

Evans and Popova 2015

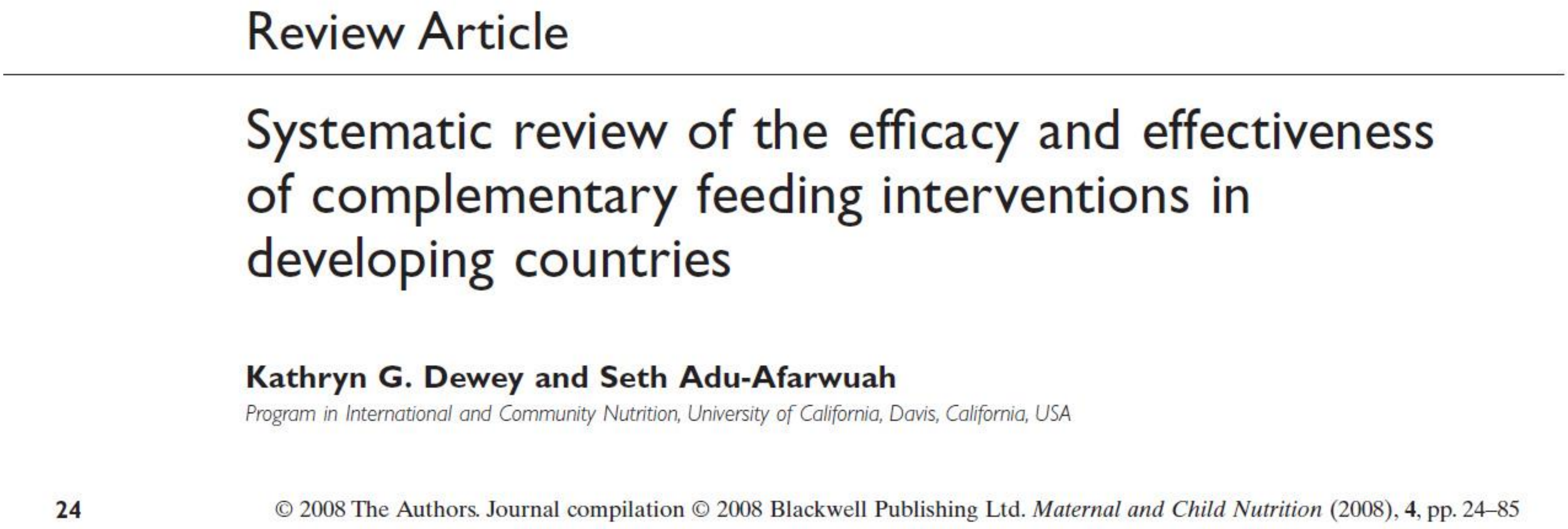
Review Article

Systematic review of the efficacy and effectiveness of complementary feeding interventions in developing countries

Kathryn G. Dewey and Seth Adu-Afarwuah

Program in International and Community Nutrition, University of California, Davis, California, USA

© 2008 The Authors. Journal compilation © 2008 Blackwell Publishing Ltd. *Maternal and Child Nutrition* (2008), **4**, pp. 24–85



Review Article

Systematic review of the efficacy and effectiveness of complementary feeding interventions in developing countries

Kathryn G. Dewey and Seth Adu-Afarwuah

Program in International and Community Nutrition, University of California, Davis, California, USA

24 © 2008 The Authors. Journal compilation © 2008 Blackwell Publishing Ltd. *Maternal and Child Nutrition* (2008), **4**, pp. 24–85

Taken together, these eight efficacy and programme evaluation studies indicate that provision of a complementary food can have a significant impact on growth under well-controlled situations, although the results are somewhat inconsistent: there was a positive impact in Ghana (Lartey et al. 1999; Adu-Afarwuah et al. 2007), Nigeria (Obatolu 2003), Zambia (Owino et al. 2007) and Malawi (Kuusipalo et al. 2006) but no impact in South Africa (Oelofse et al. 2003), Indonesia (Beckett et al. 2000) or Brazil (Santos et al. 2005).

Review Article

Systematic review of the efficacy and effectiveness of complementary feeding interventions in developing countries

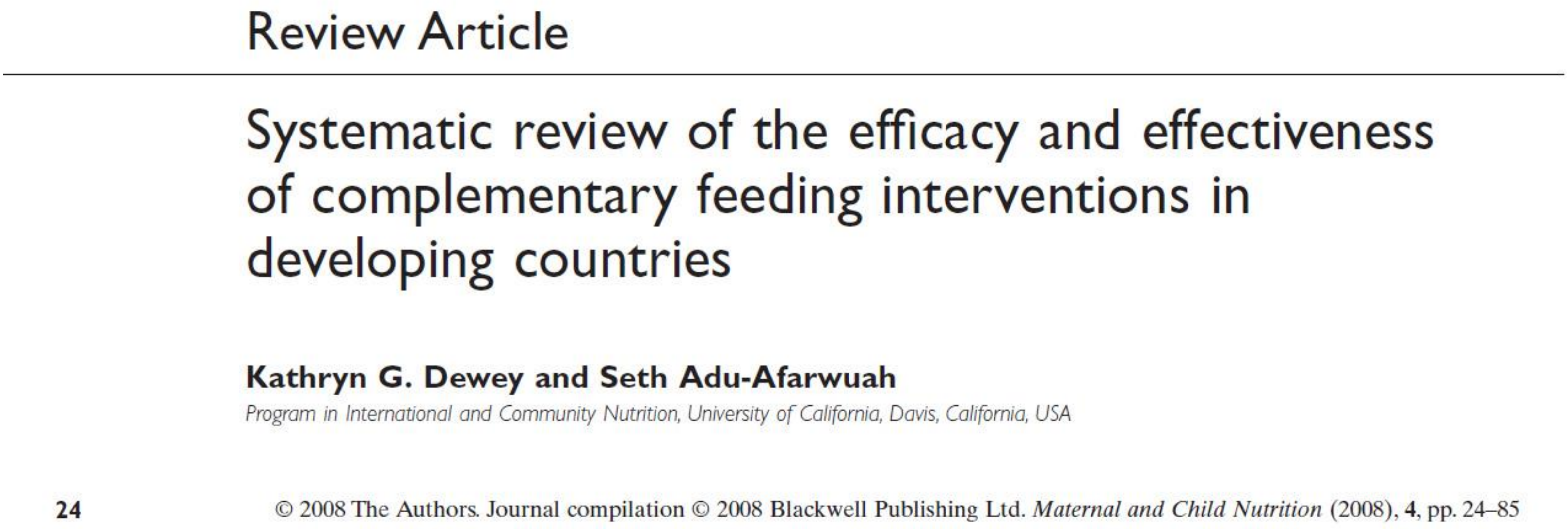
Kathryn G. Dewey and Seth Adu-Afarwuah

Program in International and Community Nutrition, University of California, Davis, California, USA

24

© 2008 The Authors. Journal compilation © 2008 Blackwell Publishing Ltd. *Maternal and Child Nutrition* (2008), **4**, pp. 24–85

Taken together, these eight efficacy and programme evaluation studies indicate that provision of a complementary food can have a significant impact on growth under well-controlled situations, although the results are somewhat inconsistent: there was a positive impact in Ghana (Lartey et al. 1999; Adu-Afarwuah et al. 2007), Nigeria (Obatolu 2003), Zambia (Owino et al. 2007) and Malawi (Kuusipalo et al. 2006) but no impact in South Africa (Oelofse et al. 2003), Indonesia (Beckett et al. 2000) or Brazil (Santos et al. 2005).



Review Article

Systematic review of the efficacy and effectiveness of complementary feeding interventions in developing countries

Kathryn G. Dewey and Seth Adu-Afarwuah

Program in International and Community Nutrition, University of California, Davis, California, USA

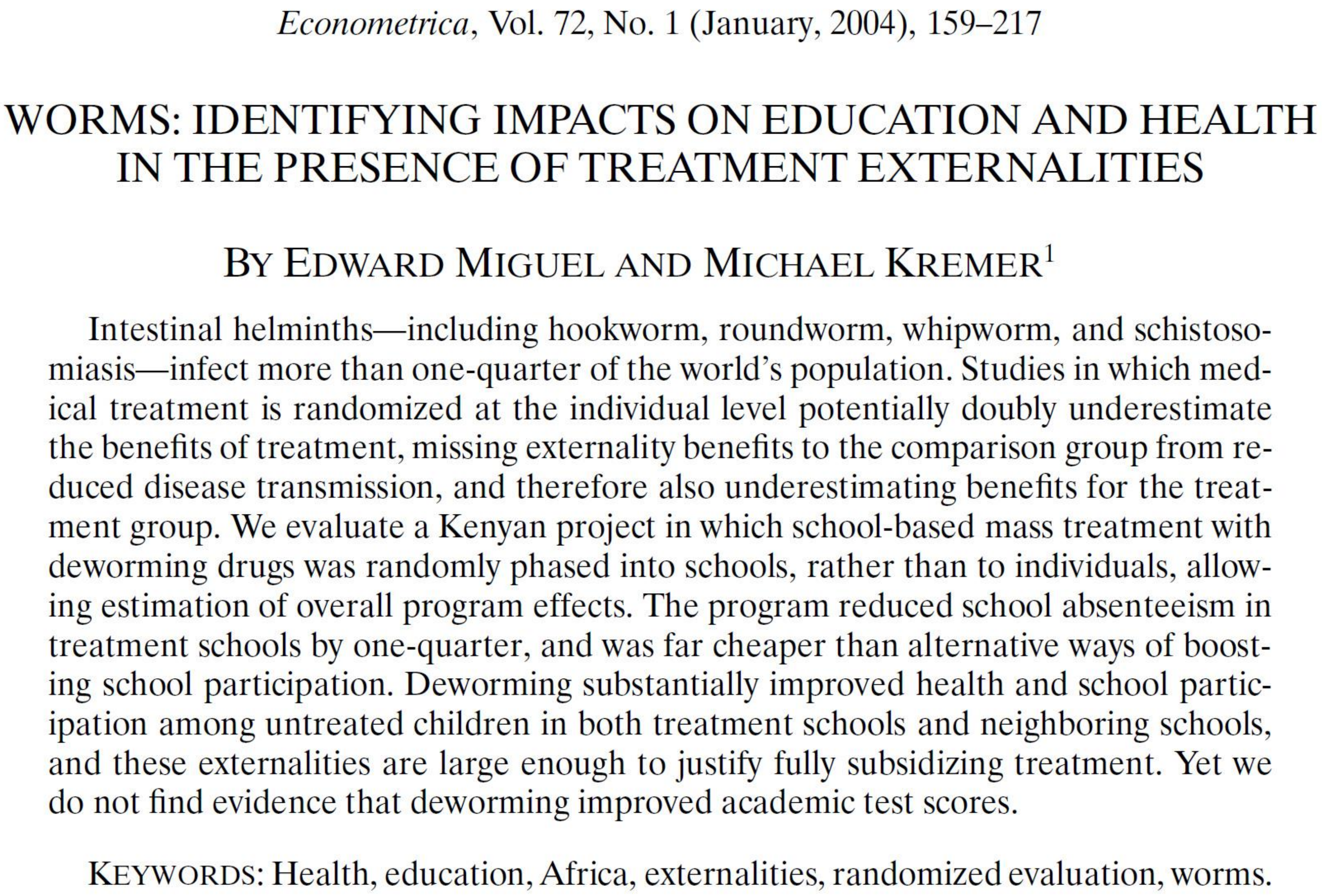
24 © 2008 The Authors. Journal compilation © 2008 Blackwell Publishing Ltd. *Maternal and Child Nutrition* (2008), **4**, pp. 24–85

[An] important aspect… of the Malawi … [study] must be recognized: … the children were malnourished (WAZ < -2 SD; WLZ > -3 SD) at baseline

Case study

Deworming

Worms: the original study



Econometrica, Vol. 72, No. 1 (January, 2004), 159–217

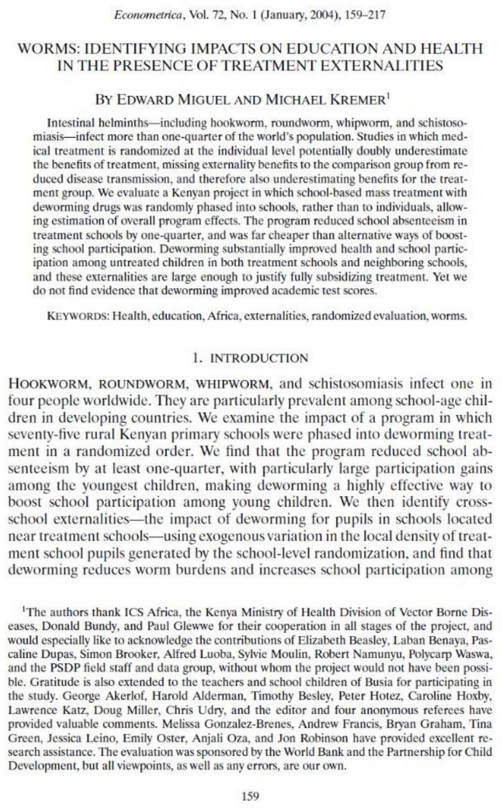
WORMS: IDENTIFYING IMPACTS ON EDUCATION AND HEALTH IN THE PRESENCE OF TREATMENT EXTERNALITIES

BY EDWARD MIGUEL AND MICHAEL KREMER[1]

Intestinal helminths—including hookworm, roundworm, whipworm, and schistosomiasis—infect more than one-quarter of the world's population. Studies in which medical treatment is randomized at the individual level potentially doubly underestimate the benefits of treatment, missing externality benefits to the comparison group from reduced disease transmission, and therefore also underestimating benefits for the treatment group. We evaluate a Kenyan project in which school-based mass treatment with deworming drugs was randomly phased into schools, rather than to individuals, allowing estimation of overall program effects. The program reduced school absenteeism in treatment schools by one-quarter, and was far cheaper than alternative ways of boosting school participation. Deworming substantially improved health and school participation among untreated children in both treatment schools and neighboring schools, and these externalities are large enough to justify fully subsidizing treatment. Yet we do not find evidence that deworming improved academic test scores.

KEYWORDS: Health, education, Africa, externalities, randomized evaluation, worms.

2004 - Worms: the original study



Econometrica, Vol. 72, No. 1 (January, 2004), 159–217

WORMS: IDENTIFYING IMPACTS ON EDUCATION AND HEALTH IN THE PRESENCE OF TREATMENT EXTERNALITIES

BY EDWARD MIGUEL AND MICHAEL KREMER[1]

Intestinal helminths—including hookworm, roundworm, whipworm, and schistosomiasis—infect more than one-quarter of the world's population. Studies in which medical treatment is randomized at the individual level potentially doubly underestimate the benefits of treatment, missing externality benefits to the comparison group from reduced disease transmission, and therefore also underestimating benefits for the treatment group. We evaluate a Kenyan project in which school-based mass treatment with deworming drugs was randomly phased into schools, rather than to individuals, allowing estimation of overall program effects. The program reduced school absenteeism in treatment schools by one-quarter, and was far cheaper than alternative ways of boosting school participation. Deworming substantially improved health and school participation among untreated children in both treatment schools and neighboring schools, and these externalities are large enough to justify fully subsidizing treatment. Yet we do not find evidence that deworming improved academic test scores.

KEYWORDS: Health, education, Africa, externalities, randomized evaluation, worms.

1. INTRODUCTION

HOOKWORM, ROUNDWORM, WHIPWORM, and schistosomiasis infect one in four people worldwide. They are particularly prevalent among school-age children in developing countries. We examine the impact of a program in which seventy-five rural Kenyan primary schools were phased into deworming treatment in a randomized order. We find that the program reduced school absenteeism by at least one-quarter, with particularly large participation gains among the youngest children, making deworming a highly effective way to boost school participation among young children. We then identify cross-school externalities—the impact of deworming for pupils in schools located near treatment schools—using exogenous variation in the local density of treatment school pupils generated by the school-level randomization, and find that deworming reduces worm burdens and increases school participation among

[1] The authors thank ICS Africa, the Kenya Ministry of Health Division of Vector Borne Diseases, Donald Bundy, and Paul Glewwe for their cooperation in all stages of the project, and would especially like to acknowledge the contributions of Elizabeth Beasley, Laban Benaya, Pascaline Dupas, Simon Brooker, Alfred Luoba, Sylvie Moulin, Robert Namunyu, Polycarp Waswa, and the PSDP field staff and data group, without whom the project would not have been possible. Gratitude is also extended to the teachers and school children of Busia for participating in the study. George Akerlof, Harold Alderman, Timothy Besley, Peter Hotez, Caroline Hoxby, Lawrence Katz, Doug Miller, Chris Udry, and the editor and four anonymous referees have provided valuable comments. Melissa Gonzalez-Brenes, Andrew Francis, Bryan Graham, Tina Green, Jessica Leino, Emily Oster, Anjali Oza, and Jon Robinson have provided excellent research assistance. The evaluation was sponsored by the World Bank and the Partnership for Child Development, but all viewpoints, as well as any errors, are our own.

159

Deworming:

Reduces worm infections for treated children
Reduces worm infections for ALL children in treated schools
Reduces worm infections for ALL children NEAR treated schools

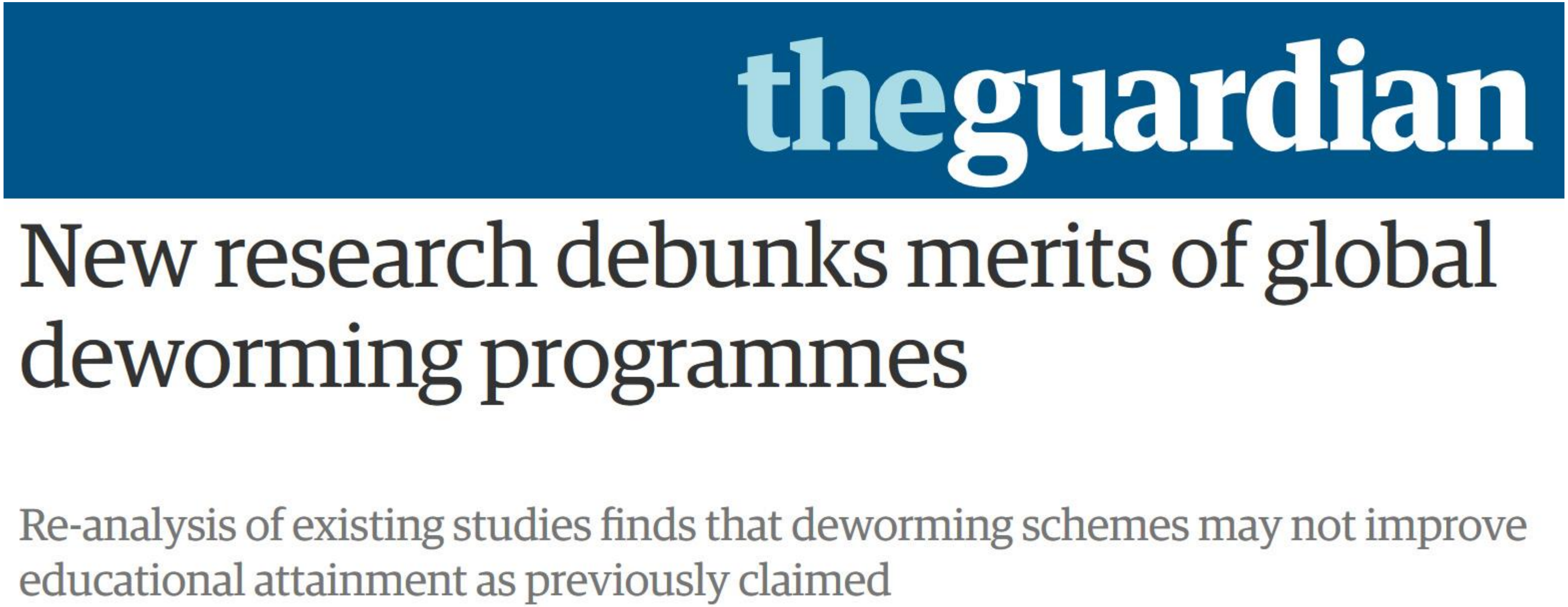
Increases school attendance for treated children
Increases school attendance for ALL children in treated schools
Increases school attendance for ALL children NEAR treated schools

Does not improve academic test scores in the short run

Methodology:

under spillovers, conditionally exogenous regional treatment intensity.

2015 July: replication, re-analysis, and review



New research debunks merits of global deworming programmes

Re-analysis of existing studies finds that deworming schemes may not improve educational attainment as previously claimed

David Evans’ Worm Wars Anthology

THE WORLD BANK

Home About Data Research Learning News Projects & Operations Publications Countries Topics

Development Impact

News, views, methods, and insights from the world of impact evaluation

Bloggers Tags Contact

Choose Blog

Latest Bloggers

David McKenzie

David Evans

Berk Ozler

Markus Goldstein

Development Impact Guest Blogger

Worm Wars: The Anthology

(This post was updated on October 26, 2015.)

Over the last several weeks, tens of thousands of words have been published about a study on the benefits of deworming for Kenyan schoolchildren, about the benefits of deworming more generally, about replication in science and social science, and about the evidence base for development programs. More will surely be written. As you'll see below, several blog posts seek to make sense of the hubbub.

For those engaged or interested in the debate, I provide links to the primary research documents as well as many of the responses.

The key research documents for the current round of the Worm Wars

- The original paper – Miguel & Kremer's Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, *Econometrica* (2004)
- Aiken et al. – Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication, *International Journal of Epidemiology* (2015)
- Davey et al. – Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial, *International Journal of Epidemiology* (2015)

The media articles that prompted a host of current responses

Responses, interpretation, and making sense of the responses and interpretation

Key research documents for the last round of the Worm Wars

The Data

More research papers

Subscribe via RSS

Subscribe by email

E-mail: *

Subscribe

Curated Content

Recent comments

Regular bloggers

Most Popular 3 Months

Most Popular

Guest Bloggers

The timeline

- 2007
  - Miguel and Kremer replication files posted, correcting a number of errors
- 2014 (October)
  - 3ie replication initiative releases
    - "Pure" replication of Miguel and Kremer
      (Clemens: "Verification" type "Replication")
    - "Alternative Scientific/Statistical" replication of Miguel and Kremer
      (Clemens: "Reanalysis" type "Robustness test")
    - Response by Hicks, Kremer, and Miguel
  - Hicks, Kremer, and Miguel update replication files
- 2015 (July-present)
  - IJE, Cochrane, Guardian, Twitter frenzy, Analysis via blogosphere, etc.

A few key documents to examine

A moment to think…

The timeline

- 2007
  - Miguel and Kremer replication files posted, correcting a number of errors
- 2014 (October)
  - 3ie replication initiative releases
    - “Pure” replication of Miguel and Kremer
      (Clemens: “Verification” type “Replication”)
    - “Alternative Scientific/Statistical” replication of Miguel and Kremer
      (Clemens: “Reanalysis” type “Robustness test”)
    - Response by Hicks, Kremer, and Miguel
  - Hicks, Kremer, and Miguel update replication files
- 2015 (July-present)
  - IJE, Cochrane, Guardian, Twitter frenzy, Analysis via blogosphere, etc.

Replication - **Verification**

(Verification type replication – "pure replication")

Aiken, Davey, Hargreaves, and Hayes

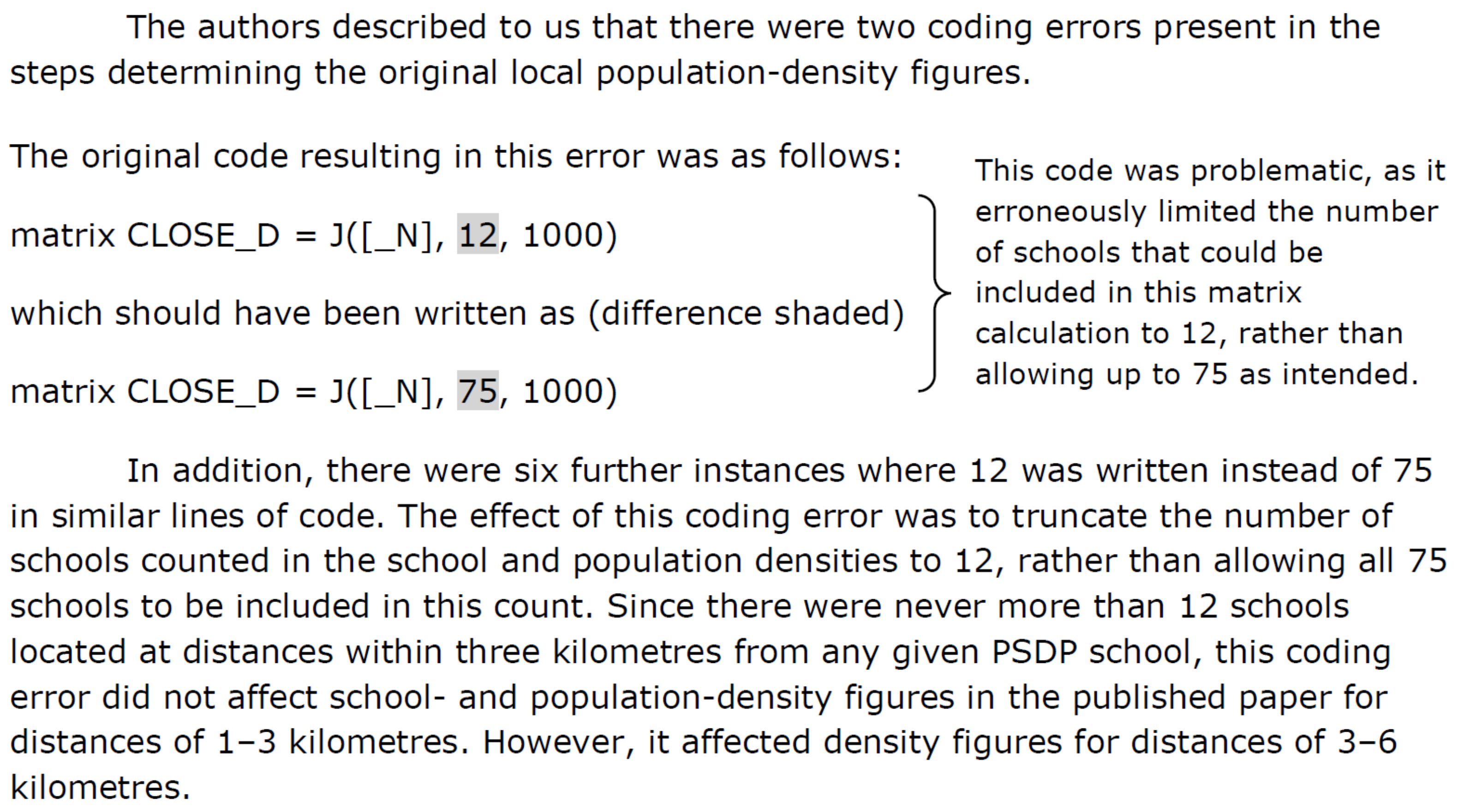
Remember the timeline? Take a look at the 2007-2014 replication files!

A **lot** of typographical glitches and a few data construction mistakes.

Epistemological reflection:
Dewald et al, and Clemens' table, suggest that many verifications basically succeed, though quite often, lots of little mistakes are cleaned up. At what stage of research is this something to do, who should do it, and what should be the reward?

The loop bug

Aiken, et al, 2014
p.17



The authors described to us that there were two coding errors present in the steps determining the original local population-density figures.

The original code resulting in this error was as follows:

```
matrix CLOSE_D = J([_N], 12, 1000)
```

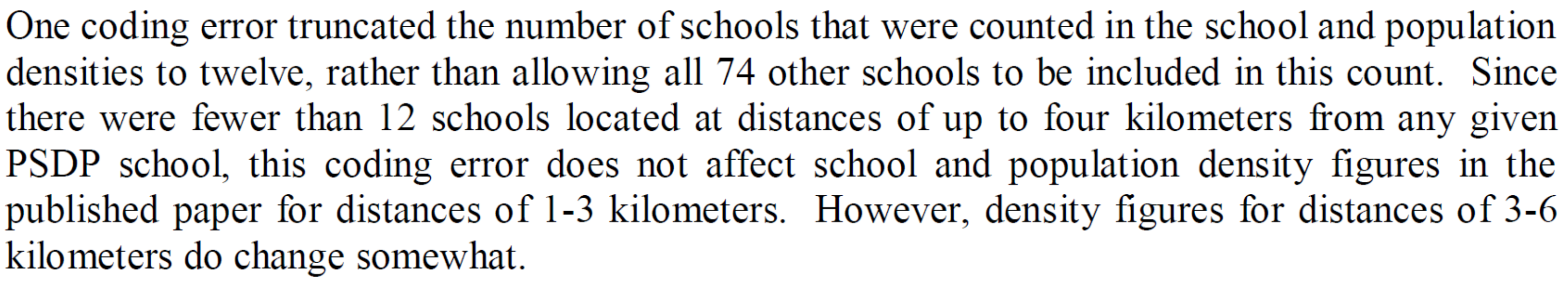
which should have been written as (difference shaded)

```
matrix CLOSE_D = J([_N], 75, 1000)
```

This code was problematic, as it erroneously limited the number of schools that could be included in this matrix calculation to 12, rather than allowing up to 75 as intended.

In addition, there were six further instances where 12 was written instead of 75 in similar lines of code. The effect of this coding error was to truncate the number of schools counted in the school and population densities to 12, rather than allowing all 75 schools to be included in this count. Since there were never more than 12 schools located at distances within three kilometres from any given PSDP school, this coding error did not affect school- and population-density figures in the published paper for distances of 1–3 kilometres. However, it affected density figures for distances of 3–6 kilometres.

Miguel and Kremer, 2008
p.7



One coding error truncated the number of schools that were counted in the school and population densities to twelve, rather than allowing all 74 other schools to be included in this count. Since there were fewer than 12 schools located at distances of up to four kilometers from any given PSDP school, this coding error does not affect school and population density figures in the published paper for distances of 1-3 kilometers. However, density figures for distances of 3-6 kilometers do change somewhat.

Replication

(Verification type replication – "pure replication")

Deworming:

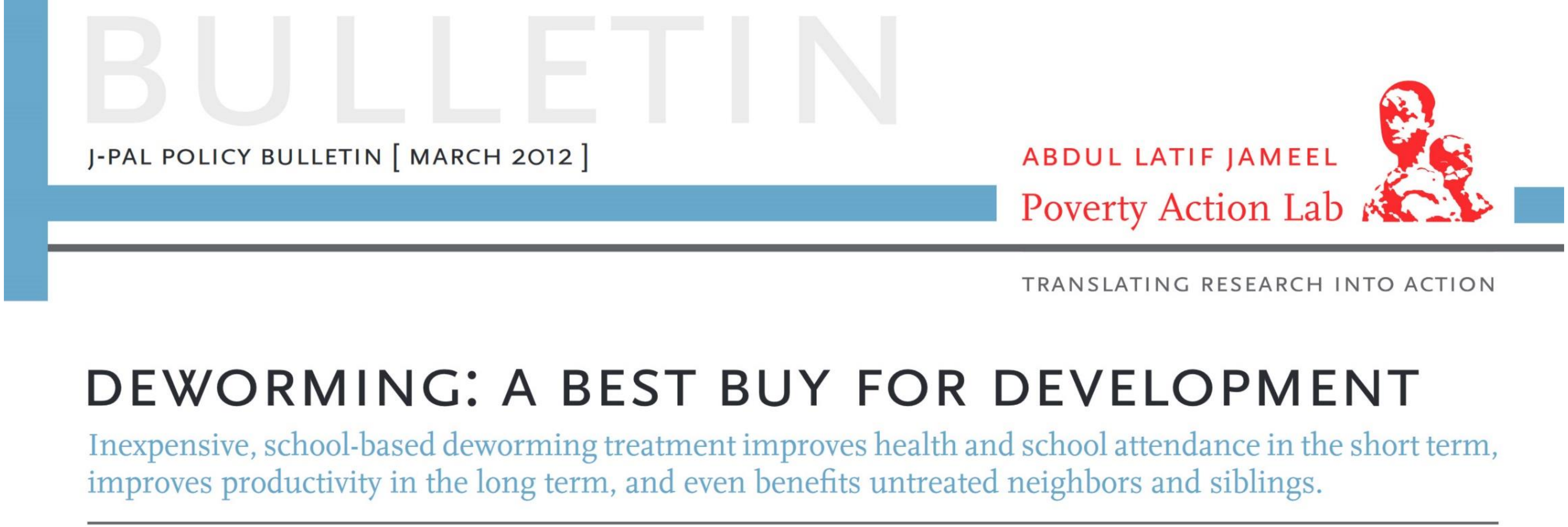
Reduces worm infections for treated children	YES	
Reduces worm infections for ALL children in treated schools	YES	
Reduces worm infections for ALL children NEAR treated schools	YES AND NO	⬅
Increases school attendance for treated children	YES	
Increases school attendance for ALL children in treated schools	YES	
Increases school attendance for ALL children NEAR treated schools	YES AND NO	⬅
Does not improve academic test scores in the short run	YES	

Replicating raw estimated coefficients

	Original	**Revised**
Naïve effect, reduced worm infection	-0.25 (0.05) ***	-0.31 (0.06) ***
Within-school externality on worm infection	-0.12 (0.07) *	-0.18 (0.07) **
Within-school externality on attendance	+.056 (0.02) ***	+.056 (0.02) ***

Table notes: the first row, the "Naïve effect, reduced worm infection," comes from text and tables describing the effect of assignment to treatment on moderate-to-heavy worm infections, in Miguel and Kremer 2004, Table VII, Column 1; and in Aiken et al. 2014 p. 21. The second row concerns what is termed the within-school "indirect" or "externality" on moderate-to-heavy worm infections; Miguel and Kremer 2004, Table VII, Column 2 and Aiken et al. 2014 p. 21. The third row comes from text describing the within-school "indirect" or "externality" effect on what is either termed "school attendance" or "participation;" details in Miguel and Kremer 2004, Table IX, Column 5 and Aiken et al. 2014 p. 30.

Humphreys and the “Headline Number”



BULLETIN

J-PAL POLICY BULLETIN [MARCH 2012]

ABDUL LATIF JAMEEL
Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION

DEWORMING: A BEST BUY FOR DEVELOPMENT

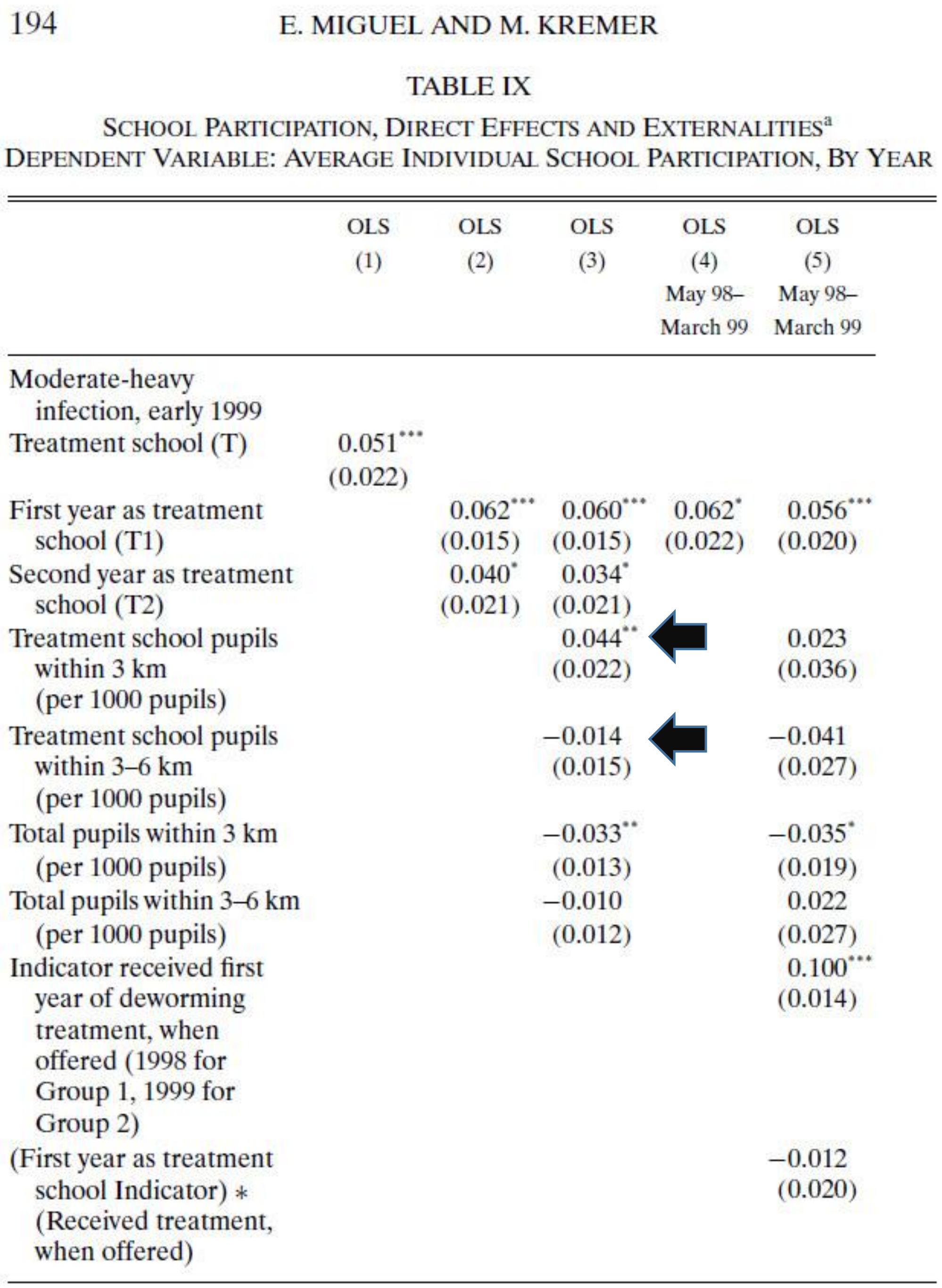
Inexpensive, school-based deworming treatment improves health and school attendance in the short term, improves productivity in the long term, and even benefits untreated neighbors and siblings.

SCHOOL ATTENDANCE INCREASED FOR TREATED AND UNTREATED CHILDREN

Deworming decreased absenteeism at treatment schools by 7.5 percentage points, a one-quarter reduction.

Humphreys and the “Headline Number”

```
. use psdp2014\tmp_o\table9a.dta, clear

. sum pop_3km_original pop_36k_original

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
pop_3km_or~l |      65530    654.6615    628.1794          0   3053.657
pop_36k_or~l |      65530    799.1447    639.1963          0   2515.091


. use psdp2014\tmp_u\table9a.dta, clear

. sum pop_3km_updated pop_36k_updated

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
pop_3km_up~d |      65788    651.4636    621.0725          0   3053.657
pop_36k_up~d |      65788        1724    993.2844          0   4771.587
```



194 E. MIGUEL AND M. KREMER

TABLE IX

SCHOOL PARTICIPATION, DIRECT EFFECTS AND EXTERNALITIES[a]
DEPENDENT VARIABLE: AVERAGE INDIVIDUAL SCHOOL PARTICIPATION, BY YEAR

	OLS (1)	OLS (2)	OLS (3)	OLS (4) May 98– March 99	OLS (5) May 98– March 99
Moderate-heavy infection, early 1999					
Treatment school (T)	0.051*** (0.022)				
First year as treatment school (T1)		0.062*** (0.015)	0.060*** (0.015)	0.062* (0.022)	0.056*** (0.020)
Second year as treatment school (T2)		0.040* (0.021)	0.034* (0.021)		
Treatment school pupils within 3 km (per 1000 pupils)			0.044** (0.022)		0.023 (0.036)
Treatment school pupils within 3–6 km (per 1000 pupils)			−0.014 (0.015)		−0.041 (0.027)
Total pupils within 3 km (per 1000 pupils)			−0.033** (0.013)		−0.035* (0.019)
Total pupils within 3–6 km (per 1000 pupils)			−0.010 (0.012)		0.022 (0.027)
Indicator received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)					0.100*** (0.014)
(First year as treatment school Indicator) ∗ (Received treatment, when offered)					−0.012 (0.020)

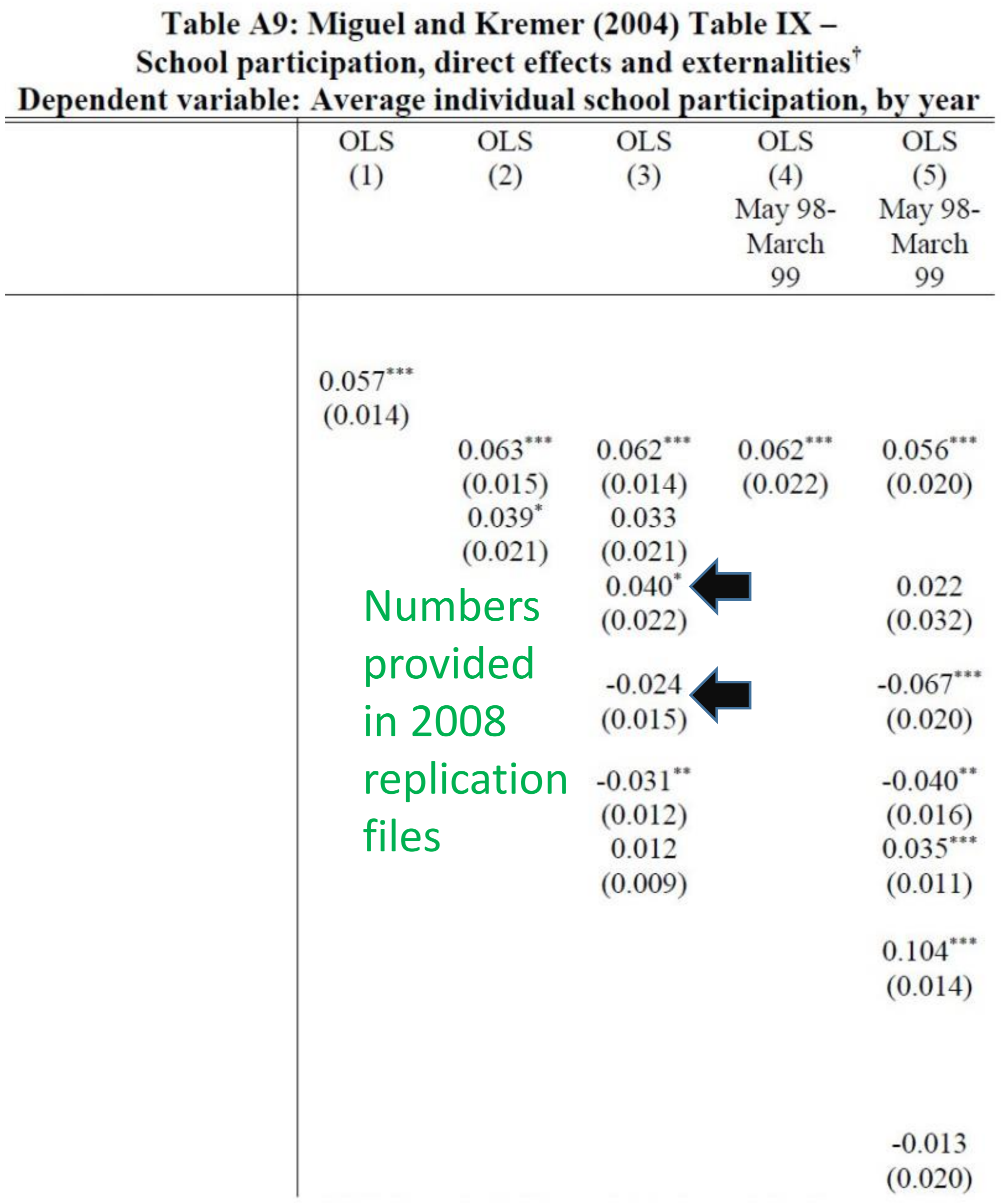


Table A9: Miguel and Kremer (2004) Table IX – School participation, direct effects and externalities[†]
Dependent variable: Average individual school participation, by year

	OLS (1)	OLS (2)	OLS (3)	OLS (4) May 98- March 99	OLS (5) May 98- March 99
	0.057*** (0.014)				
		0.063*** (0.015)	0.062*** (0.014)	0.062*** (0.022)	0.056*** (0.020)
		0.039* (0.021)	0.033 (0.021)		
			0.040* (0.022)		0.022 (0.032)
			-0.024 (0.015)		-0.067*** (0.020)
			-0.031** (0.012)		-0.040** (0.016)
			0.012 (0.009)		0.035*** (0.011)
					0.104*** (0.014)
					-0.013 (0.020)

Numbers provided in 2008 replication files

The Math

		Original		Revised	
		(1)	(2)	(3)	(4)
Coefficient estimates	Treatment (direct effect)	0.0547**	0.0536**	0.0553***	0.0578***
		(0.0232)	(0.0233)	(0.0136)	(0.0139)
	Treatment pupils ('000) 0-3km	0.04797**	0.04567**	0.03801*	0.04461**
		(0.0192)	(0.0182)	(0.0209)	(0.0207)
	Treatment pupils ('000) 3-6km	-0.01268		-0.02429	
		(0.0153)		(0.0149)	
Means	Treatment pupils 0-3km	608.3046	608.3046	605.6553	605.6553
	Treatment pupils 3-6km	726.8933		1631.4675	
Externality averages	Average externalities 0-3km	0.0292**	0.0278**	0.0230*	0.0270**
		(0.0117)	(0.0111)	(0.0127)	(0.0125)
	Average externalities 3-6km	-0.0092		-0.0396	
		(0.0111)		(0.0243)	
Externality totals	Total externalities above	0.0200	0.0278**	-0.0166	0.0270**
		(0.0135)	(0.0111)	(0.0300)	(0.0125)
	Overall deworming effect	0.0747***	0.0814***	0.0387	0.0848***
		(0.0273)	(0.0258)	(0.0321)	(0.0172)

2014 Replication guide, Table B2

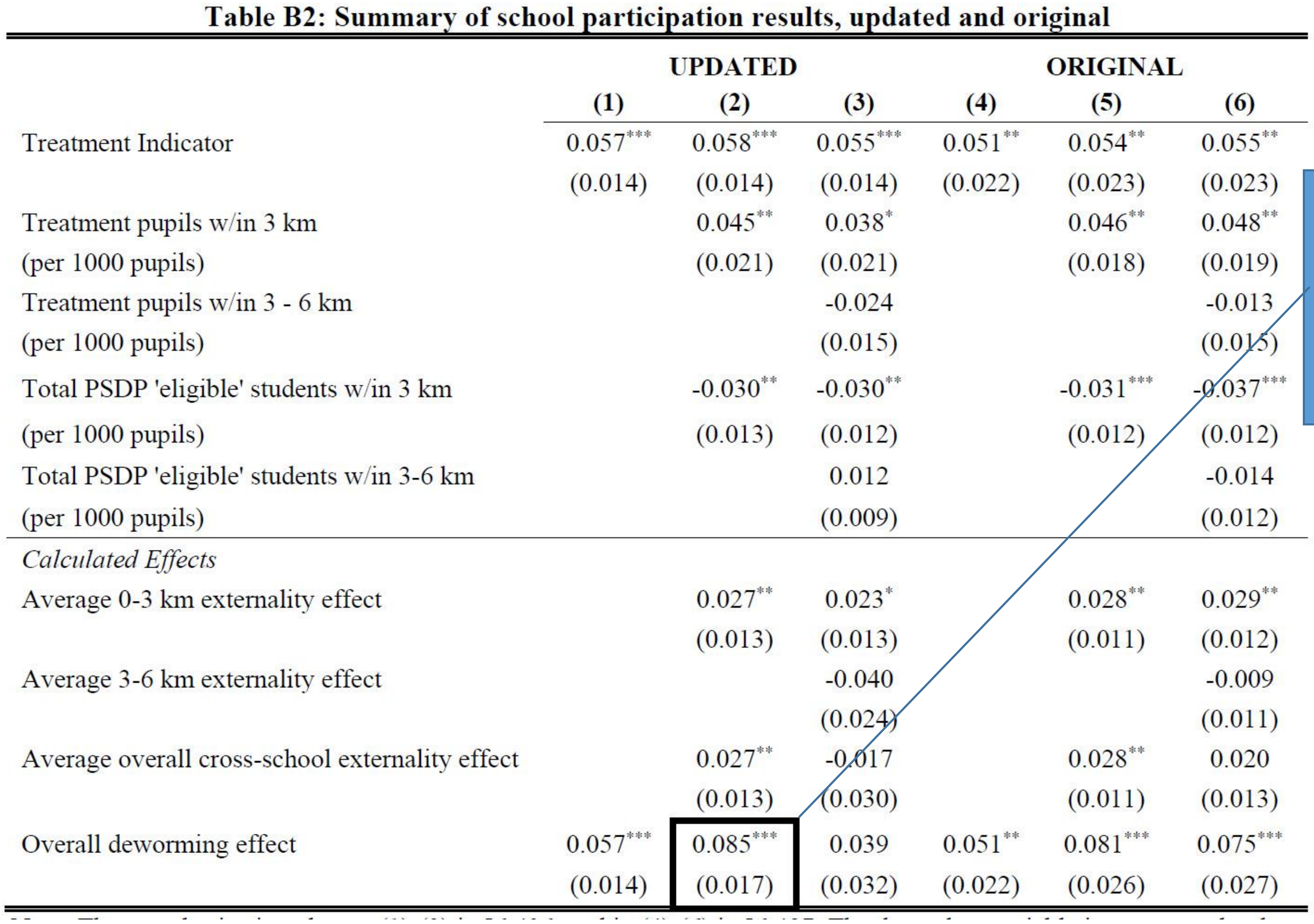


Table B2: Summary of school participation results, updated and original

	UPDATED			ORIGINAL		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Indicator	0.057***	0.058***	0.055***	0.051**	0.054**	0.055**
	(0.014)	(0.014)	(0.014)	(0.022)	(0.023)	(0.023)
Treatment pupils w/in 3 km		0.045**	0.038*		0.046**	0.048**
(per 1000 pupils)		(0.021)	(0.021)		(0.018)	(0.019)
Treatment pupils w/in 3 - 6 km			-0.024			-0.013
(per 1000 pupils)			(0.015)			(0.015)
Total PSDP 'eligible' students w/in 3 km		-0.030**	-0.030**		-0.031***	-0.037***
(per 1000 pupils)		(0.013)	(0.012)		(0.012)	(0.012)
Total PSDP 'eligible' students w/in 3-6 km			0.012			-0.014
(per 1000 pupils)			(0.009)			(0.012)
Calculated Effects						
Average 0-3 km externality effect		0.027**	0.023*		0.028**	0.029**
		(0.013)	(0.013)		(0.011)	(0.012)
Average 3-6 km externality effect			-0.040			-0.009
			(0.024)			(0.011)
Average overall cross-school externality effect		0.027**	-0.017		0.028**	0.020
		(0.013)	(0.030)		(0.011)	(0.013)
Overall deworming effect	0.057***	0.085***	0.039	0.051**	0.081***	0.075***
	(0.014)	(0.017)	(0.032)	(0.022)	(0.026)	(0.027)

Multiple test correction?

Multiple test corrections

Carlo Emilio Bonferroni

Olive Jean Dunn

Photo credits: *Wikipedia/Dr. Ruth Mickey (creative commons licenses)*

Multiple test corrections

Carlo Emilio Bonferroni

Olive Jean Dunn

Photo credits: *Wikipedia/Dr. Ruth Mickey (creative commons licenses)*

2014 Replication guide, Table B2

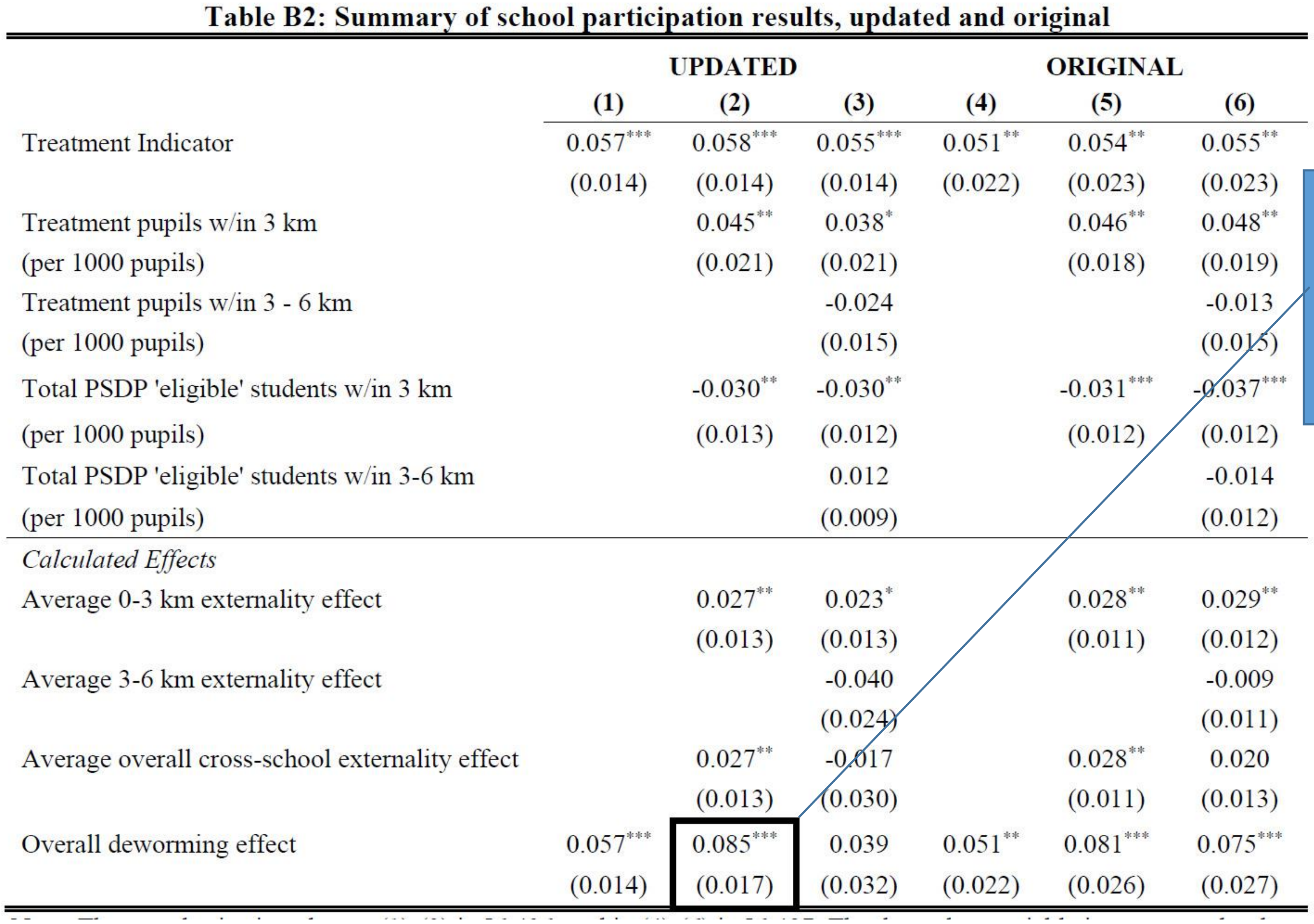
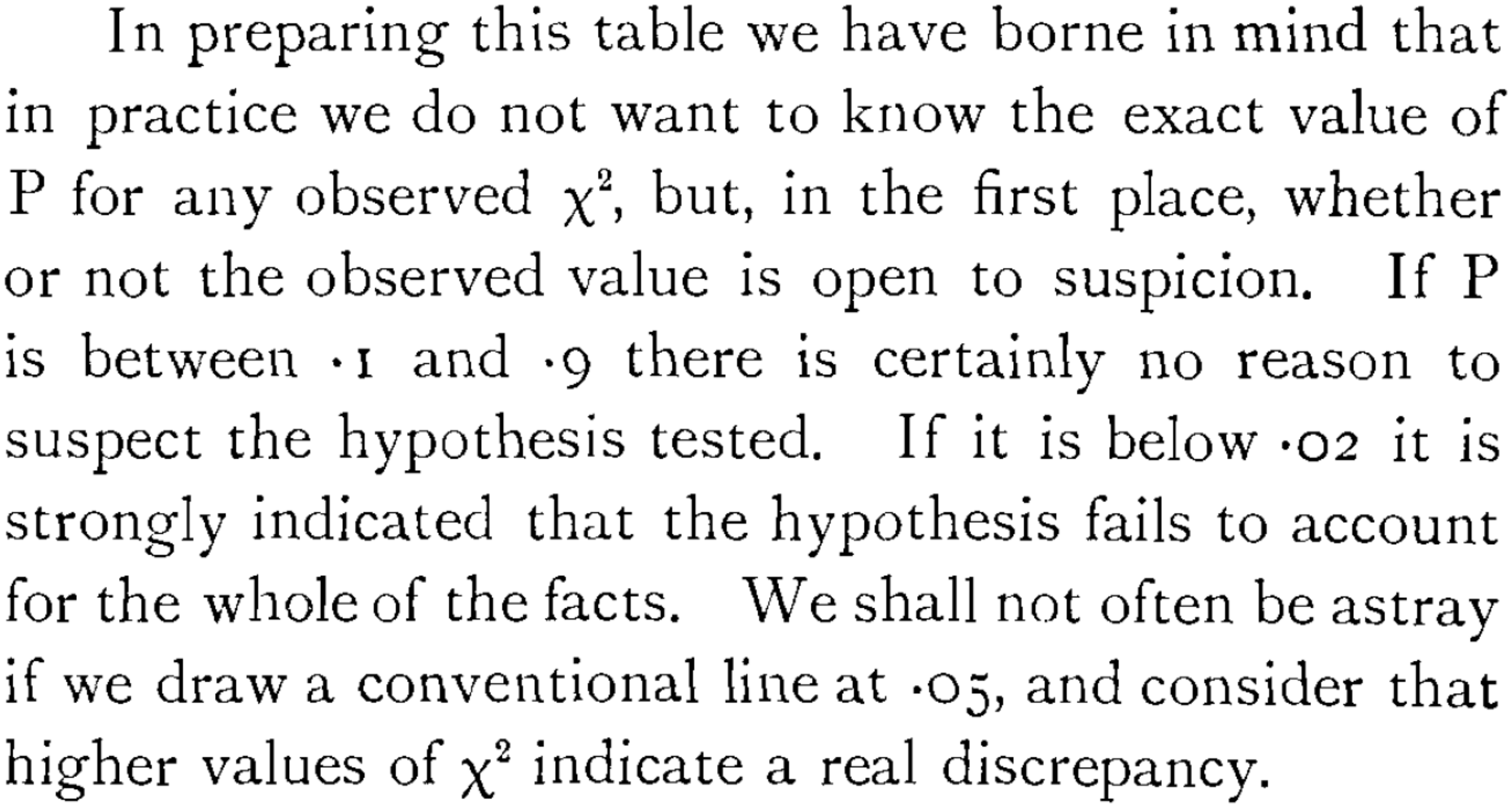


Table B2: Summary of school participation results, updated and original

	UPDATED			ORIGINAL		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Indicator	0.057***	0.058***	0.055***	0.051**	0.054**	0.055**
	(0.014)	(0.014)	(0.014)	(0.022)	(0.023)	(0.023)
Treatment pupils w/in 3 km		0.045**	0.038*		0.046**	0.048**
(per 1000 pupils)		(0.021)	(0.021)		(0.018)	(0.019)
Treatment pupils w/in 3 - 6 km			-0.024			-0.013
(per 1000 pupils)			(0.015)			(0.015)
Total PSDP 'eligible' students w/in 3 km		-0.030**	-0.030**		-0.031***	-0.037***
(per 1000 pupils)		(0.013)	(0.012)		(0.012)	(0.012)
Total PSDP 'eligible' students w/in 3-6 km			0.012			-0.014
(per 1000 pupils)			(0.009)			(0.012)
Calculated Effects						
Average 0-3 km externality effect		0.027**	0.023*		0.028**	0.029**
		(0.013)	(0.013)		(0.011)	(0.012)
Average 3-6 km externality effect			-0.040			-0.009
			(0.024)			(0.011)
Average overall cross-school externality effect		0.027**	-0.017		0.028**	0.020
		(0.013)	(0.030)		(0.011)	(0.013)
Overall deworming effect	0.057***	0.085***	0.039	0.051**	0.081***	0.075***
	(0.014)	(0.017)	(0.032)	(0.022)	(0.026)	(0.027)

Whatever multiple test correction you are inclined to use (if any), a T-statistic of 5 will withstand it.

Aside: why 0.05? Fisher (20th century)



In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between ·1 and ·9 there is certainly no reason to suspect the hypothesis tested. If it is below ·02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at ·05, and consider that higher values of χ^2 indicate a real discrepancy.

The timeline

- 2007
  - Miguel and Kremer replication files posted, correcting a number of errors
- 2014 (October)
  - 3ie replication initiative releases
    - "Pure" replication of Miguel and Kremer
      (Clemens: "Verification" type "Replication")
    - "Alternative Scientific/Statistical" replication of Miguel and Kremer
      (Clemens: "Reanalysis" type "Robustness test")
    - Response by Hicks, Kremer, and Miguel
  - Hicks, Kremer, and Miguel update replication files
- 2015 (July-present)
  - IJE, Cochrane, Guardian, Twitter frenzy, Analysis via blogosphere, etc.

Ozler (not me) reading of **Reanalysis**

(Reanalysis-type robustness test, "alternative statistical and scientific replication")

Davey, Aiken, Hayes, and Hargreaves

"In their reanalysis of the data from the original study, [Davey (et al)] make some choices that are significantly different than the ones made by the original study authors. There are many departures but four of them are key:

(i) definition of treatment;

(ii) ignoring the longitudinal data in favor of cross-sectional analysis of treatment effects by year;

(iii) weighting observations differently; and

(iv) ignoring spillovers from treatment to control"

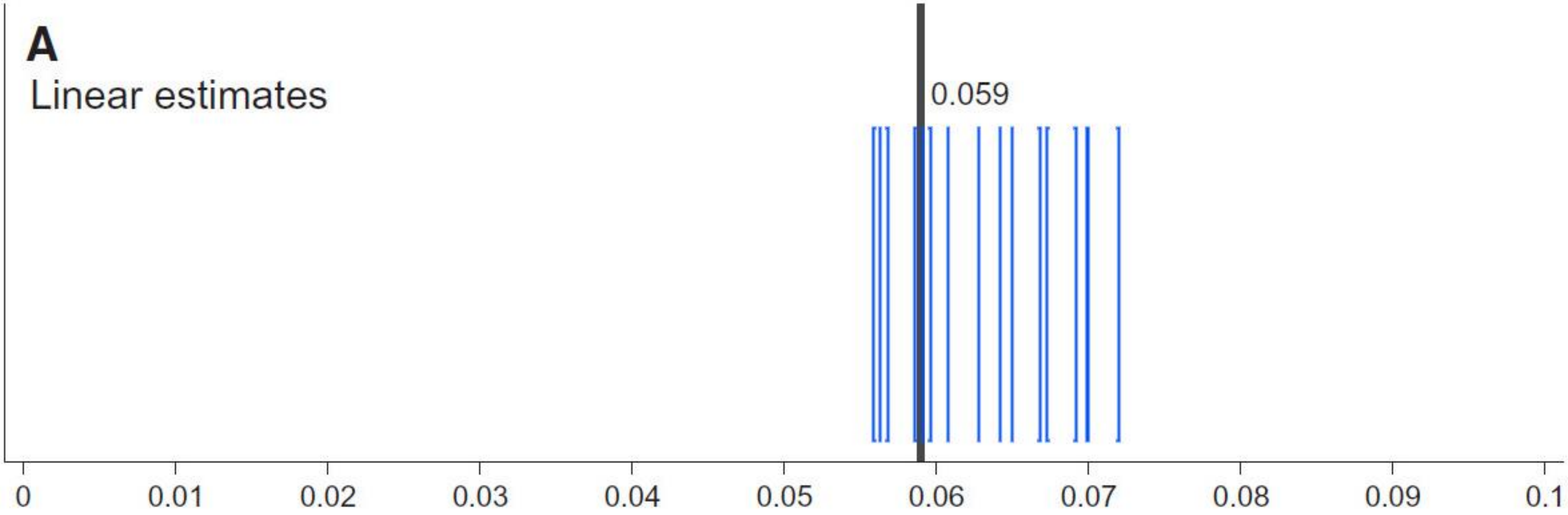
The danger of (and incentives to carry out) a reverse p-hack? (Galiani, Gertler, and Romero 2017)

Reanalysis

16 permutations

Not splitting the dataset

Sample – full or eligible
Covariates – include or not
Weighting – attendance vs pupil
Timing – intended vs actual



Reanalysis

16 permutations
In each of two frameworks
Not splitting the dataset

Sample – full or eligible
Covariates – include or not
Weighting – attendance vs pupil
Timing – intended vs actual

Davey et al abstract:
"When both years were combined, there was strong evidence of an effect on attendance."

(So the Guardian headline didn't follow directly from the study)

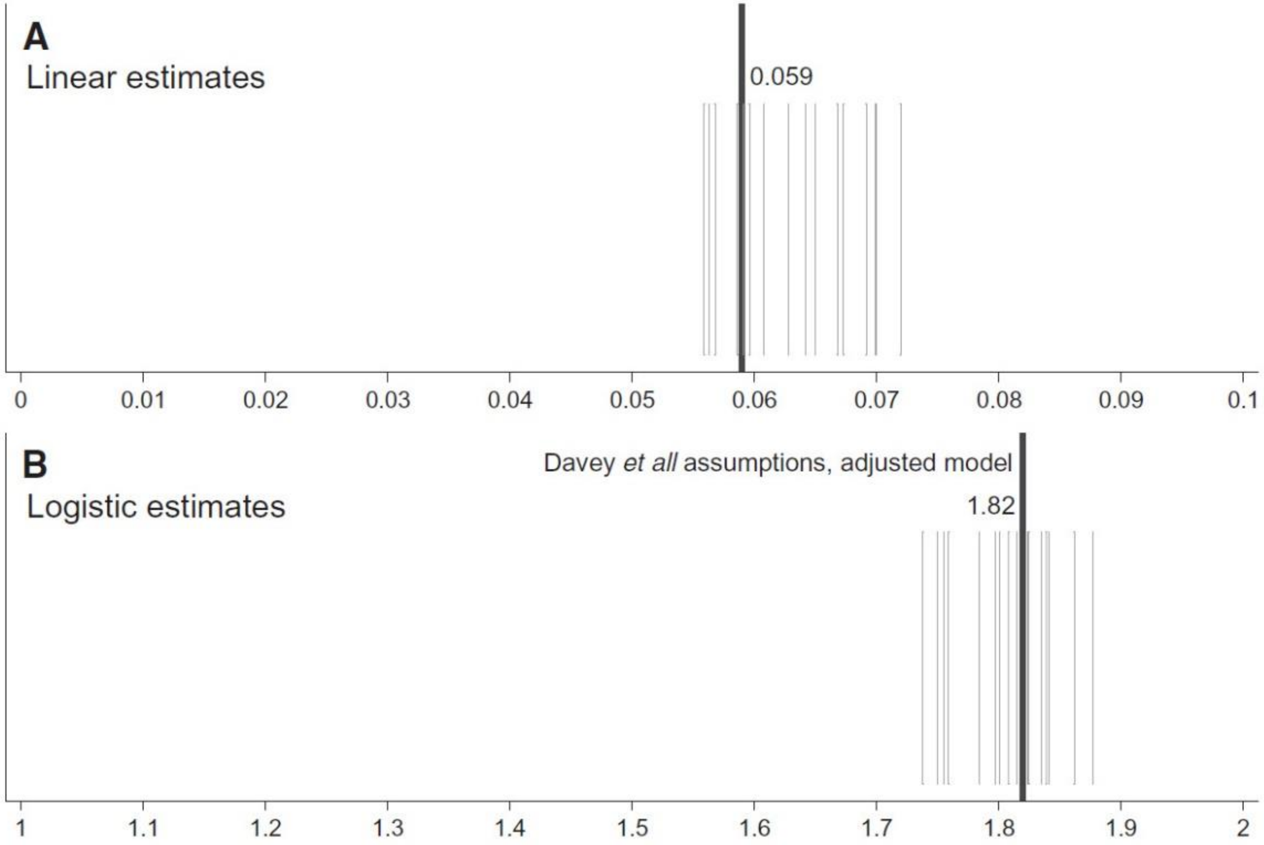
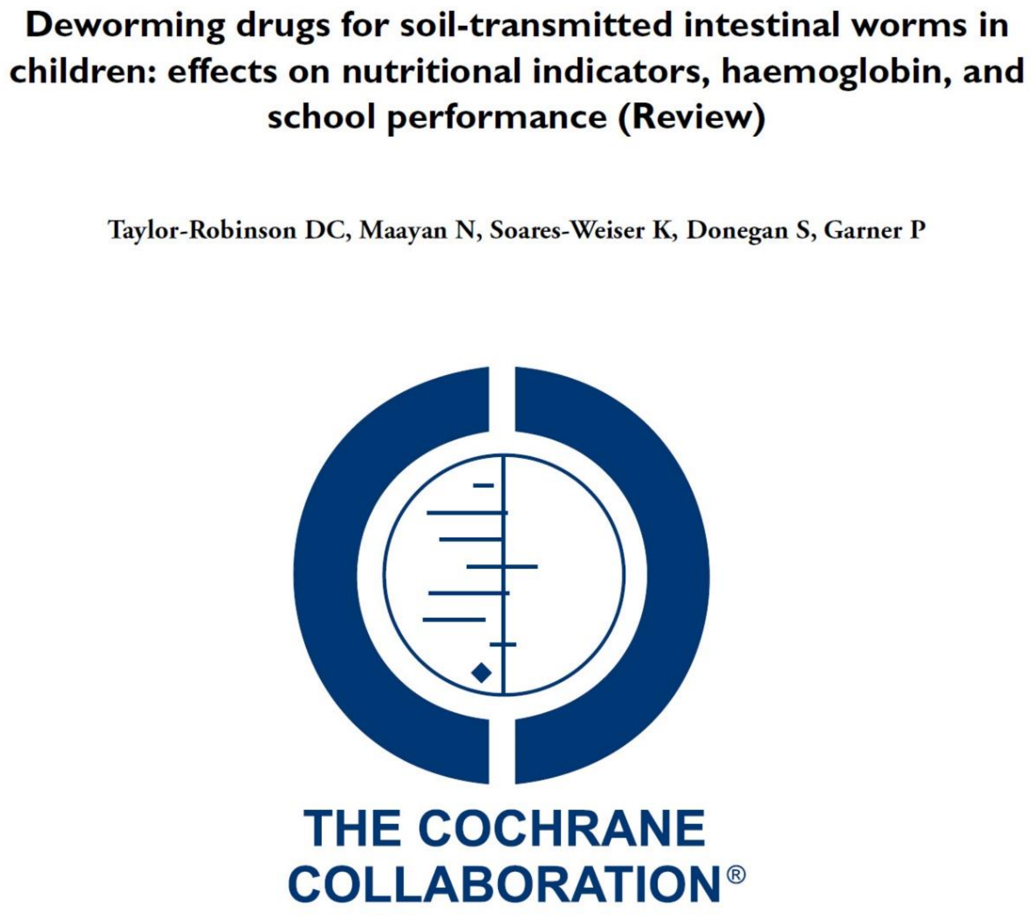


Figure 2. Deworming treatment effect estimates on school participation. Each vertical grey line denotes a coefficient estimate of the effect of deworming on school participation. The estimates use both years of data, and differ in: (i) statistical model (the original linear regression model in Panel A, and random effects logistics regression from Davey *et al.*[2] in Panel B); (ii) sample (the original full sample, and the sample eligible for treatment in Davey *et al.*); (iii) regression models adjusted for covariates and unadjusted; (iv) approaches to weighting observations (each attendance observation equally, and each pupil equally); and (v) the dataset that in Davey *et al.* employ in their analysis, which incorrectly defines treatment and makes additional missing data assumptions (Appendix B), vs data that correctly define treatment. All 16 coefficient estimates in Panel A are significant at $P < 0.01$; all 16 estimates in Panel B are significant at $P < 0.001$. The bold vertical lines denote the adjusted model estimate using Davey *et al.*'s[2] data; the Panel B estimate is from their Table 2, top right panel.

The timeline

- 2007
    - Miguel and Kremer replication files posted, correcting a number of errors
- 2014 (October)
    - 3ie replication initiative releases
        - “Pure” replication of Miguel and Kremer
          (Clemens: “Verification” type “Replication”)
        - “Alternative Scientific/Statistical” replication of Miguel and Kremer
          (Clemens: “Reanalysis” type “Robustness test”)
        - Response by Hicks, Kremer, and Miguel
    - Hicks, Kremer, and Miguel update replication files
- 2015 (July-present)
    - IJE, Cochrane, Guardian, Twitter frenzy, Analysis via blogosphere, etc.

Worms: the “review”



Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance (Review)

Taylor-Robinson DC, Maayan N, Soares-Weiser K, Donegan S, Garner P

THE COCHRANE COLLABORATION®

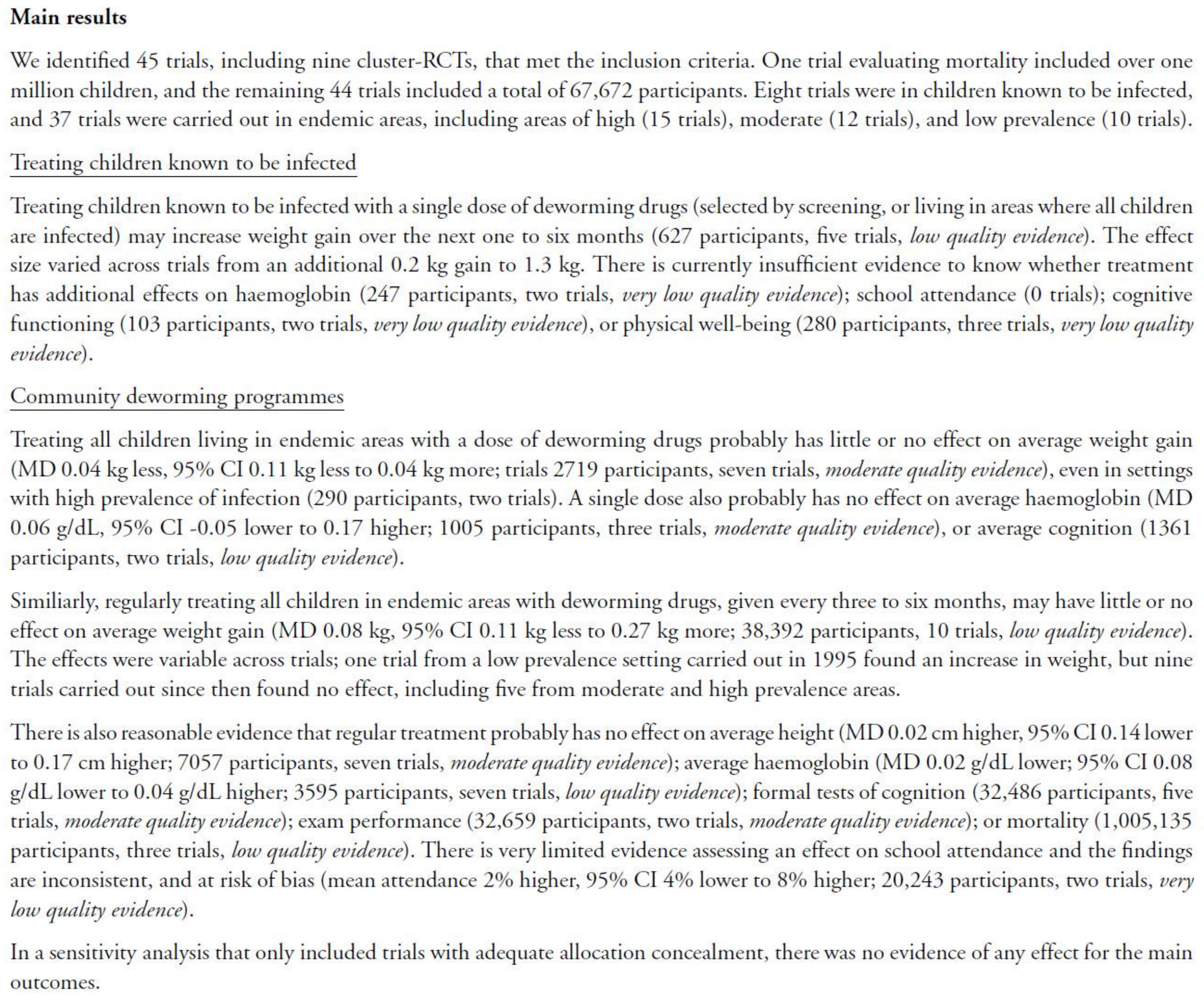
160 pages, 45 studies met criteria, etc.

Worms: the “review”

“Treating children known to have worm infection may have some nutritional benefits for the individual. However, in mass treatment of all children in endemic areas, there is now substantial evidence that this does not improve average nutritional status, haemoglobin, cognition, school performance, or survival.”

Taylor-Robinson, et al., p.2

Worms: the "review"



Main results

We identified 45 trials, including nine cluster-RCTs, that met the inclusion criteria. One trial evaluating mortality included over one million children, and the remaining 44 trials included a total of 67,672 participants. Eight trials were in children known to be infected, and 37 trials were carried out in endemic areas, including areas of high (15 trials), moderate (12 trials), and low prevalence (10 trials).

Treating children known to be infected

Treating children known to be infected with a single dose of deworming drugs (selected by screening, or living in areas where all children are infected) may increase weight gain over the next one to six months (627 participants, five trials, *low quality evidence*). The effect size varied across trials from an additional 0.2 kg gain to 1.3 kg. There is currently insufficient evidence to know whether treatment has additional effects on haemoglobin (247 participants, two trials, *very low quality evidence*); school attendance (0 trials); cognitive functioning (103 participants, two trials, *very low quality evidence*), or physical well-being (280 participants, three trials, *very low quality evidence*).

Community deworming programmes

Treating all children living in endemic areas with a dose of deworming drugs probably has little or no effect on average weight gain (MD 0.04 kg less, 95% CI 0.11 kg less to 0.04 kg more; trials 2719 participants, seven trials, *moderate quality evidence*), even in settings with high prevalence of infection (290 participants, two trials). A single dose also probably has no effect on average haemoglobin (MD 0.06 g/dL, 95% CI -0.05 lower to 0.17 higher; 1005 participants, three trials, *moderate quality evidence*), or average cognition (1361 participants, two trials, *low quality evidence*).

Similiarly, regularly treating all children in endemic areas with deworming drugs, given every three to six months, may have little or no effect on average weight gain (MD 0.08 kg, 95% CI 0.11 kg less to 0.27 kg more; 38,392 participants, 10 trials, *low quality evidence*). The effects were variable across trials; one trial from a low prevalence setting carried out in 1995 found an increase in weight, but nine trials carried out since then found no effect, including five from moderate and high prevalence areas.

There is also reasonable evidence that regular treatment probably has no effect on average height (MD 0.02 cm higher, 95% CI 0.14 lower to 0.17 cm higher; 7057 participants, seven trials, *moderate quality evidence*); average haemoglobin (MD 0.02 g/dL lower; 95% CI 0.08 g/dL lower to 0.04 g/dL higher; 3595 participants, seven trials, *low quality evidence*); formal tests of cognition (32,486 participants, five trials, *moderate quality evidence*); exam performance (32,659 participants, two trials, *moderate quality evidence*); or mortality (1,005,135 participants, three trials, *low quality evidence*). There is very limited evidence assessing an effect on school attendance and the findings are inconsistent, and at risk of bias (mean attendance 2% higher, 95% CI 4% lower to 8% higher; 20,243 participants, two trials, *very low quality evidence*).

In a sensitivity analysis that only included trials with adequate allocation concealment, there was no evidence of any effect for the main outcomes.

Worms: the “review”

“There is also reasonable evidence that regular treatment probably has no effect on ... formal tests of cognition (...five trials...);[or] exam performance (...two trials...); . There is very limited evidence assessing an effect on school attendance ... (two trials, *very low quality evidence*...)**”**

Taylor-Robinson, et al., p.2

Worms: the "review"

"There is also reasonable evidence that regular treatment probably has no effect on ... formal tests of cognition (...five trials...);[or] exam performance (...two trials...); . There is very limited evidence assessing an effect on school attendance ... (two trials, *very low quality evidence*...)"

Taylor-Robinson, et al., p.2

Worms: the “review”

- Cognitive outcomes:
  - Review narrows the evidence to exactly two studies.
    - Miguel and Kremer (2004)
    - Hall , et al (unpublished, 2006)
  - Both studies of school-age children.

Worms: the “review”

Conflating evidence of absence with absence of evidence:

“Treating children known to have worm infection may have some nutritional benefits for the individual. However, in mass treatment of all children in endemic areas, there is now substantial evidence that this does not improve average nutritional status, haemoglobin, cognition, school performance, or survival.”

Taylor-Robinson, et al., p.2

Worms: the “review”

“The replication highlights important coding errors and this resulted in a number of changes to the results: the previously reported effect on anaemia disappeared; the effect on school attendance was similar to the original analysis, although **the effect was seen in both children that received the drug and those that did not**; **and the indirect effects (externalities)** of the intervention on adjacent schools **disappeared** (Aiken 2015). The statistical replication suggested some impact of the complex intervention (deworming and health promotion) on school attendance, but this varied depending on the analysis strategy, and there was a high risk of bias. The replication showed no effect on exam performance (Davey 2015).”

Taylor-Robinson et al, p.10

Externalities – a game of telephone?

- Aiken, et al, pure replication, IJE edition, page 8:

“In corrected re-analysis, the indirect-between-school effect on school attendance had shifted in direction and was less precisely estimated—there was now **little evidence for an effect of this kind in the format of analysis originally employed.** We have not reexamined for evidence of indirect-between-school effect at a distance other than that used in original paper (up to 6km from schools) as this would deviate from our stated pre-analytical plan. We do note that some parameters suggest **effects may be present at distances of up to 3 km.**”

- Aiken, et al, pure replication, IJE edition, abstract:

“after correction of coding errors, there was **little evidence** of an indirect effect on school attendance among children in schools close to intervention schools.”

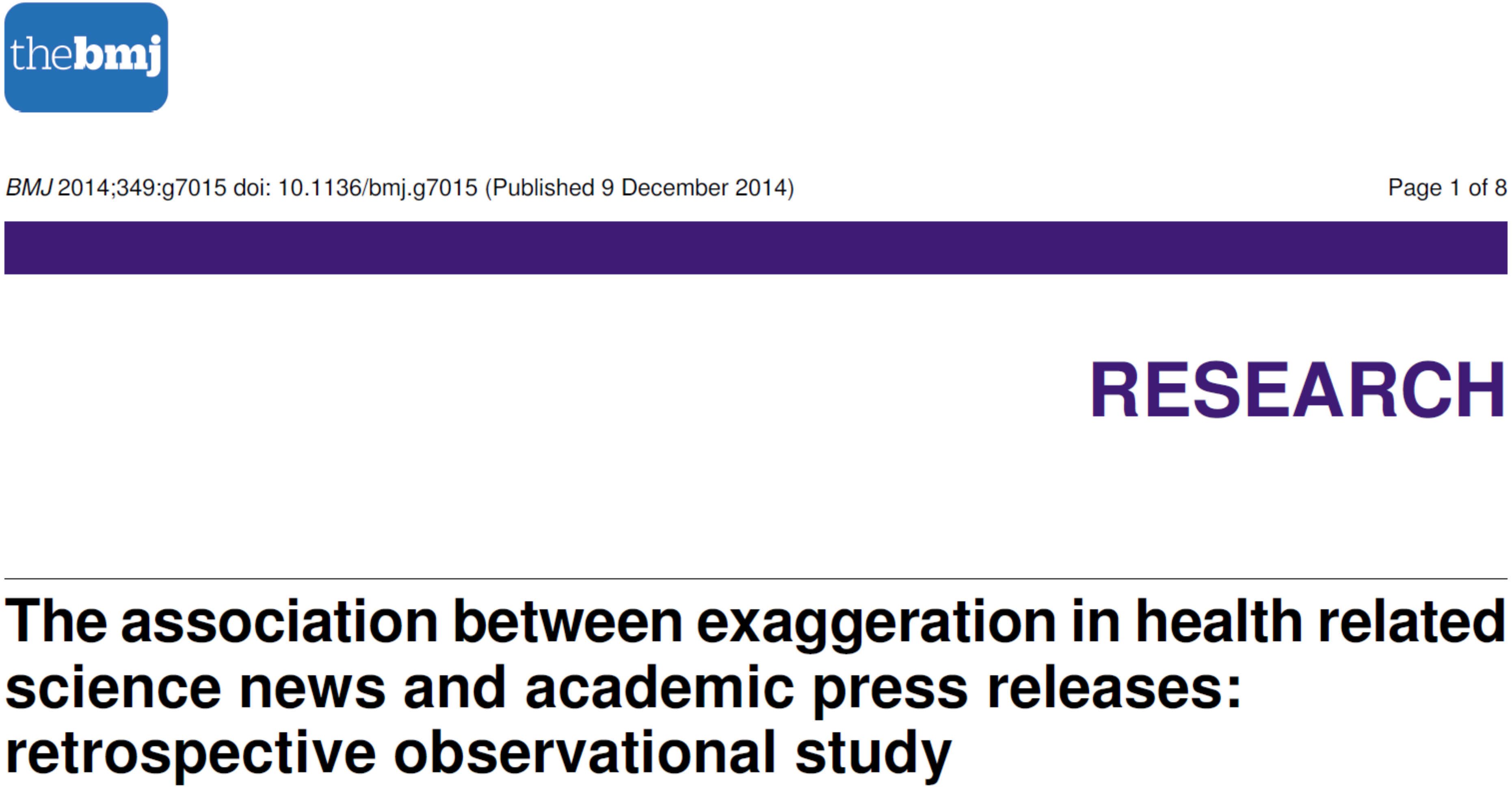
- LSHTM press release

However, the researchers found calculation errors in the original authors’ data which meant there was **no longer evidence** that deworming caused an increase in school attendance among children who attended schools near to the schools where children were treated.

- Taylor-Robinson et al (Cochrane), page 10

“the indirect effects (externalities) of the intervention on adjacent schools **disappeared…”**

The Call Is Coming From Inside The House



thebmj

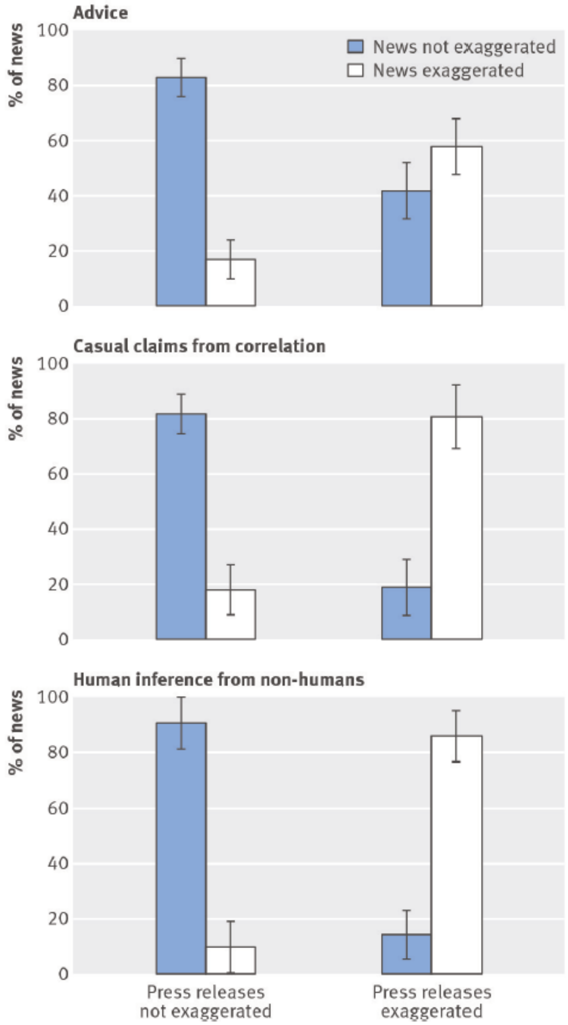
BMJ 2014;349:g7015 doi: 10.1136/bmj.g7015 (Published 9 December 2014)

Page 1 of 8

RESEARCH

The association between exaggeration in health related science news and academic press releases: retrospective observational study

The Call Is Coming From Inside The House



<u>You should</u> drink 8 glasses of water and take 10,000 steps a day

I didn’t omit any variables, did you?

… IN MICE!

Worms: the “review”

- Systematic reviews and meta-analyses are always difficult, and always riddled with judgement calls. In general, I am grateful that people take the time to do this at all, and sympathetic to the challenges.

But critiqued in October 2015 PLoS NTD (and elsewhere):

- Choice of weights (RE/FE) and ages (Croke critique)
- Prevalence – example of a problem with non-worm reviews too:

Worms: the “review”

- Inclusion criteria – what’s missing in this description?

“We included randomized controlled trials (RCTs) and quasi-RCTs comparing deworming drugs for soil-transmitted helminths with placebo or no treatment in children aged 16 years or less, reporting on weight, haemoglobin, and formal tests of intellectual development. We also sought data on school attendance, school performance, and mortality. We included trials that combined health education with deworming programmes.”

Worms: the "review"

- Inclusion criteria – what's missing in this description?

"We included randomized controlled trials (RCTs) and quasi-RCTs comparing deworming drugs for soil-transmitted helminths with placebo or no treatment in children aged 16 years or less, reporting on weight, haemoglobin, and formal tests of intellectual development. We also sought data on school attendance, school performance, and mortality. We included trials that combined health education with deworming programmes."

Worms: the "review"

- Systematic reviews and meta-analyses are always difficult, and always riddled with judgement calls. In general, I am grateful that people take the time to do this at all, and sympathetic to the challenges.

But critiqued in October 2015 PLoS NTD (and elsewhere):

- Choice of weights (RE/FE) and ages (Croke critique)
- Prevalence will change effect size (de Silva critique)
- Duration

Worms: the “review”

- Systematic reviews and meta-analyses are always difficult, and always riddled with judgement calls. In general, I am grateful that people take the time to do this at all, and sympathetic to the challenges.

But critiqued in October 2015 PLoS NTD (and elsewhere):

- Choice of weights (RE/FE) and ages (Croke critique)
- Prevalence will change effect size (de Silva critique)
- Duration (# of studies with follow-up more than 6 years later…)

Worms: the “review”

- Systematic reviews and meta-analyses are always difficult, and always riddled with judgement calls. In general, I am grateful that people take the time to do this at all, and sympathetic to the challenges.

But critiqued in October 2015 PLoS NTD (and elsewhere):

- Choice of weights (RE/FE) and ages (Croke critique)
- Prevalence will change effect size (de Silva critique)
- Duration (# of studies with follow-up more than 6 years later: zero)

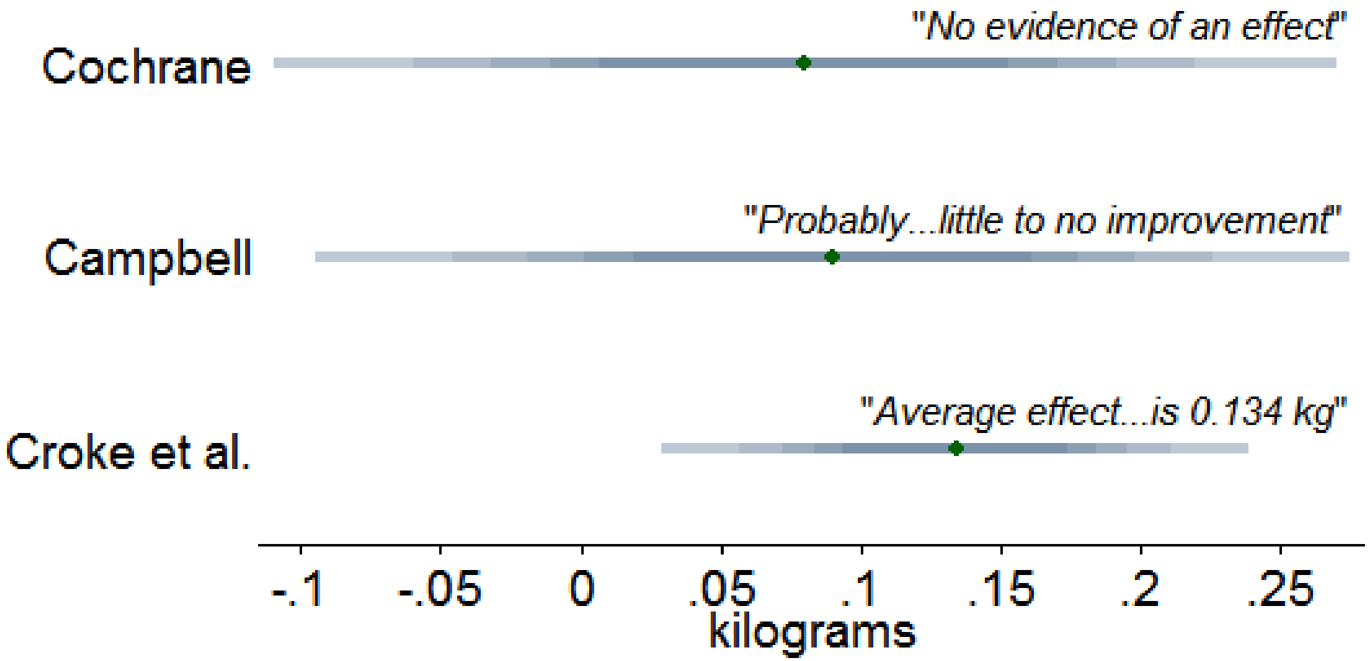
Worms: the “review”

- Systematic reviews and meta-analyses are always difficult, and always riddled with judgement calls. In general, I am grateful that people take the time to do this at all, and sympathetic to the challenges.

But critiqued in October 2015 PLoS NTD (and elsewhere):

- Choice of weights (RE/FE) and ages (Croke critique)
- Prevalence will change effect size (de Silva critique)
- Restricting to short-duration studies, two ways (Montresor critique)
- Conflating absence of evidence with evidence of absence
- Takes the Guardian view of Aiken-Davey (Hicks critique)

Worms: **three** reviews



(source: David Roodman)

Worms: the “review”

“Maybe the Cochrane Collaboration review is chasing something that doesn’t exist.” Angus Deaton, in conversation with Timothy Ogden

Where do we go from here?

- Are there any long term studies of deworming?

Where do we go from here?

- Are there any long term studies of deworming?

Perhaps just four:

Bleakley (published, 2007)

Baird, et al (published, 2016)
Ozier (published, 2018)
Croke and Atun (published, 2019)

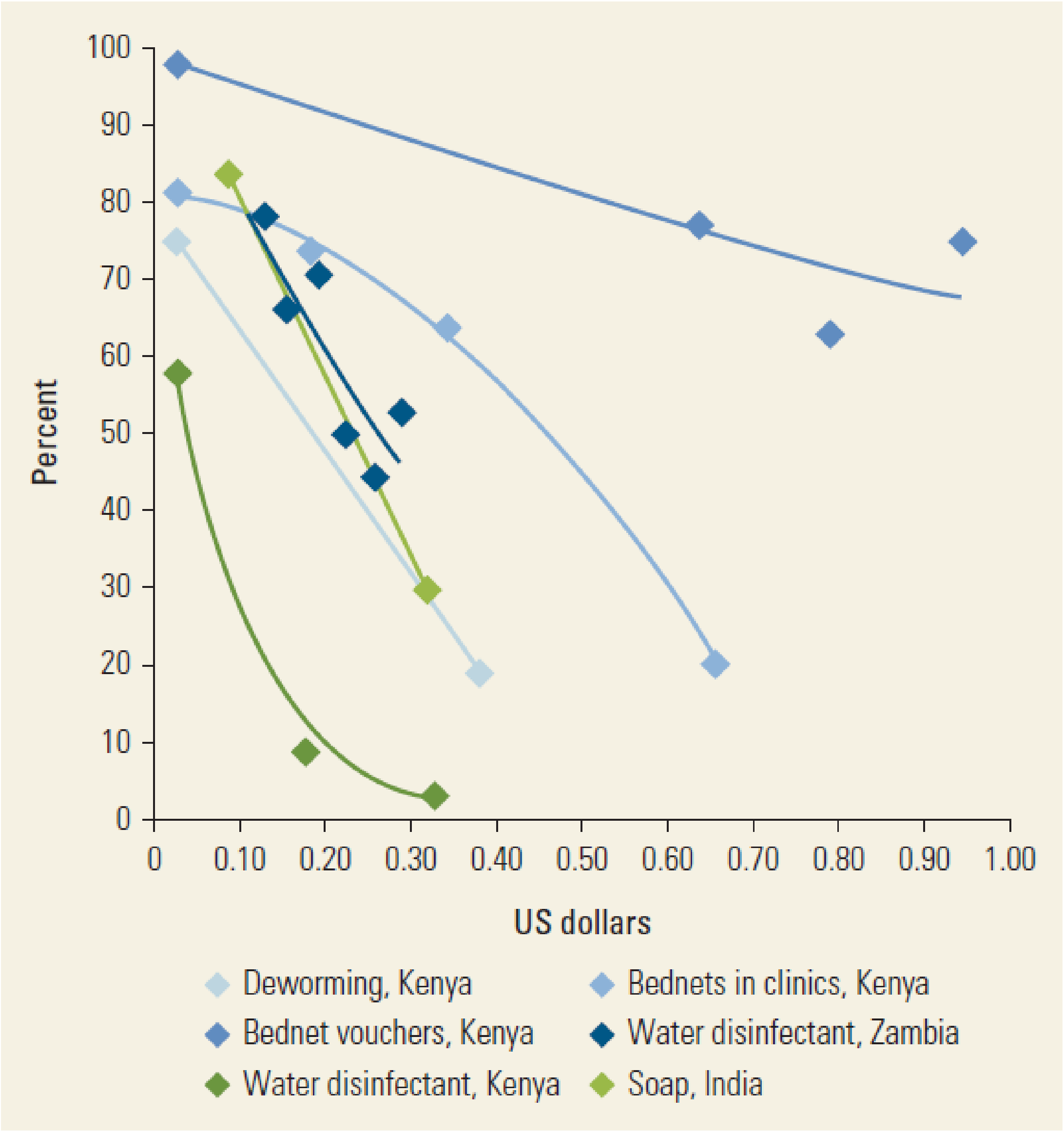
Policy

When is public financing a good idea?

When is an investment cost-effective?

Policy

Figure 29.1 Response of Consumer Demand to Increase in the Price of Health Products

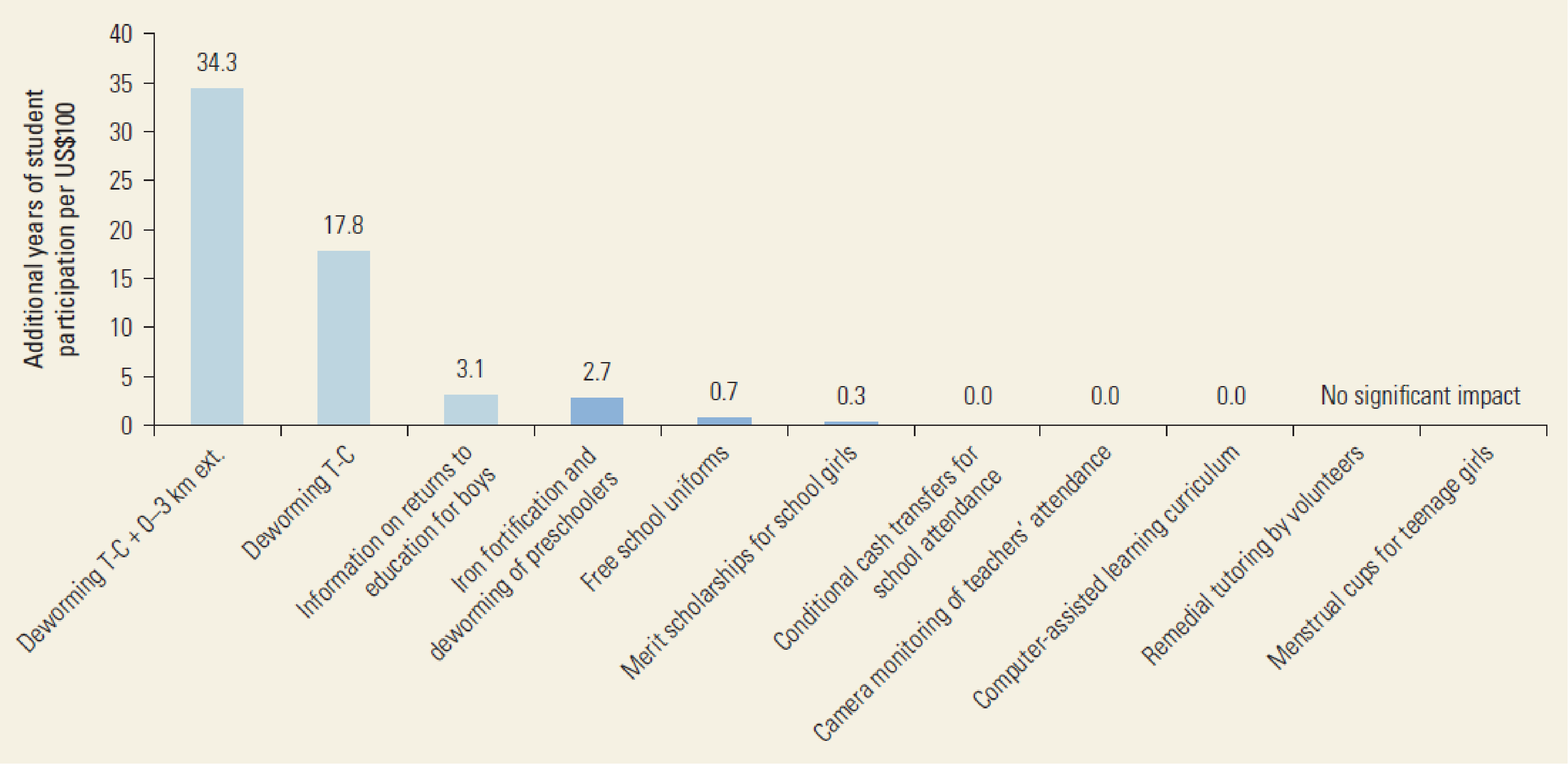


Source: Abdul Latif Jameel Poverty Action Lab 2011.

Ahuja, et al, DCP 2017

Policy

Figure 29.2 Cost-Effectiveness of Development Interventions in Increasing School Attendance



Sources: Hicks, Kremer, and Miguel 2015 based on data from Abdul Latif Jameel Poverty Action Lab.

Note: T–C = the difference between outcomes for those allocated to the deworming treatment group and those allocated to the deworming comparison group; km = kilometers; ext. = externality benefits. Some values are adjusted for inflation but the deworming costs are not. Deworming is costed at US$0.49 per child in Kenya. Some of these programs create benefits beyond school attendance. For example, conditional cash transfers provide income to poor households. The Jameel Poverty Action Lab cost-effectiveness calculations for school participation include conditional cash transfers as program costs.

Ahuja, et al, DCP 2017

In summary:

Miguel and Kremer 2004

- The verification mostly succeeds, but corrects some errors.
- “Headline numbers” (basis for cost-effectiveness) stand up to several approaches.
- The robustness reanalysis also upholds the findings—except when it goes down a road warned against by Deaton, Clemens, Ozler, etc.

Deworming more generally

- Cochrane review’s approach may hobble its own enterprise
- Including or undertaking additional studies would be valuable – to refute, reinforce, or simply refine current thinking.

Policy

- Deworming programs are so inexpensive, it would only take a tiny impact for them to be cost-effective investments; public financing may be the best route when a large part of the benefit accrues to people other than the direct recipient. (see Ahuja et al WBER 2015, Croke et al NBER 2017, Ahuja et al DCP 2017)

In summary:

Press releases, tweets, abstracts, etc.

- Don’t exaggerate.

Replication

- Make our files available.
- Journals play a role: Some suggest, others enforce.
- Just like an RCT, set yourself up so that “a null is publishable.”
    - Camerer, et al. (Nature Human Behavior, Science);
    - Galiani, Gertler, and Romero

Thanks

BGSE Development

Replication and Pre-Analysis Plans (Part 2)

Professors: Pamela Jakiela and Owen Ozier

Excerpts from

“Power to the Plan: Using Pre-Analysis Plans to Learn More from Experiments in Education,”

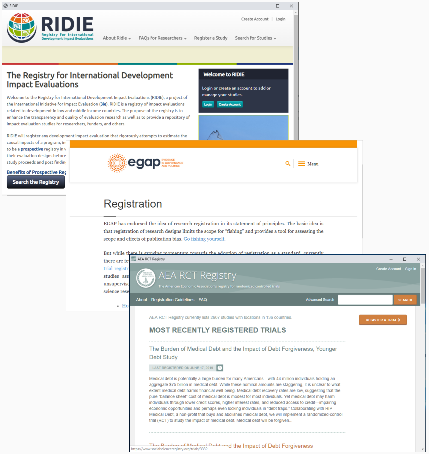
pre-analysis plan October 2018
blog post December 2018
and presentation June 2019

by Clare Leaver, Owen Ozier, Pieter Serneels, and Andrew Zeitlin

Eat your vegetables

A growing view is that pre-analysis plans comprise a necessary, if somewhat painful, approach to mitigating p-hacking in social sciences.

This parallels the use of trial registries, which help circumnavigate the “file drawer” problem in research—the problem of unpublished non-rejections of null hypotheses.



The Registry for International Development Impact Evaluations

Welcome to the Registry for International Development Impact Evaluations (RIDIE), a project of the International Initiative for Impact Evaluation (3ie). RIDIE is a registry of impact evaluations related to development in low and middle income countries. The purpose of the registry is to enhance the transparency and quality of evaluation research as well as to provide a repository of impact evaluation studies for researchers, funders, and others.

RIDIE will register any development impact evaluation that rigorously attempts to estimate the
causal impacts of a program, in
to be a prospective registry in v
their evaluation designs before
study proceeds and post findin

Benefits of Prospective Reg

Login or create an account to add or manage your studies.

Registration

EGAP has endorsed the idea of research registration in its statement of principles. The basic idea is that registration of research designs limits the scope for "fishing" and provides a tool for assessing the scope and effects of publication bias. Go fishing yourself.

But while there is growing momentum towards the adoption of registration as a standard, currently
there are fe
trial registr
studies ass
unsupervise
science rese

- Ho

The American Economic Association's registry for randomized controlled trials

AEA RCT Registry currently lists 2607 studies with locations in 136 countries.

MOST RECENTLY REGISTERED TRIALS

The Burden of Medical Debt and the Impact of Debt Forgiveness, Younger Debt Study

LAST REGISTERED ON JUNE 17, 2019

Medical debt is potentially a large burden for many Americans—with 44 million individuals holding an aggregate $75 billion in medical debt. While these nominal amounts are staggering, it is unclear to what extent medical debt harms financial well-being. Medical debt recovery rates are low, suggesting that the pure "balance sheet" cost of medical debt is modest for most individuals. Yet medical debt may harm individuals through lower credit scores, higher interest rates, and reduced access to credit—impairing economic opportunities and perhaps even locking individuals in "debt traps." Collaborating with RIP Medical Debt, a non-profit that buys and abolishes medical debt, we will implement a randomized-control trial (RCT) to study the impact of medical debt. Medical debt will be forgiven...

The Burden of Medical Debt and the Impact of Debt Forgiveness

https://www.socialscienceregistry.org/trials/3332

Not just work, but guesswork

Part of what bothers researchers is the feeling that they may be leaving important findings on the table.

- Acute to the extent that PAPs force researchers to guess about appropriate measures of a construct.
- Some of this is by design: part of the discomfort reflects the extent to which we subconsciously adapt analyses ex post.
- Some of this is a trade-off between the optimization of statistical power and policy relevance.

A spoonful of sugar

These costs can be at least partially offset by potential gains, which researchers are often leaving on the table.

- Specifically, the PAP offers an opportunity to make analytical decisions that can substantially improve power relative to plain-vanilla analytical strategies.
- One way to do so is through the use of blinded analyses of endline data. This is useful particularly in cases where the generative model has features that have meaningful power implications, but which are hard to guess ex ante.
- We can't take all the guesswork out of writing a PAP, but we can resolve some forms of uncertainty in a way that enhances power.

We provide three examples.

Use case 1: Non-normal errors and the choice of test statistic

Non-normal errors

Consider a typical (ANCOVA) generative model,

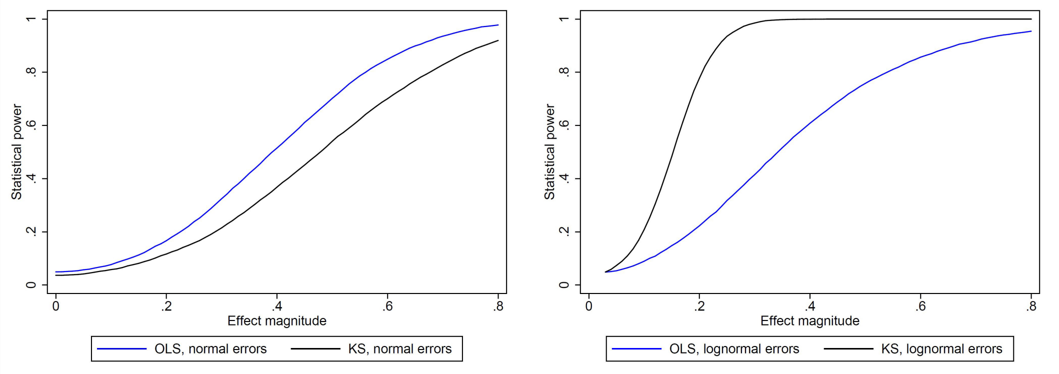
$$y = \beta + \tau T + \rho y_0 + e.$$

with binary treatment T.

The researcher seeks a test statistics that is well powered against alternatives to the null hypothesis H_0: $\tau = 0$.

Is linear regression the best they can do?

Regression coefficients vs KS statistics in simulation



Simulated power in teacher application ‘quality’

Using blinded data from the universe of teacher applications in our study of the recruitment (and other) effects of Pay-for-Performance vs Fixed-Wage contracts in Rwanda, we compare rejection rates for alternative test statistics.

Simulated rejection rates for treatment effects that move a candidate at the median of the application pool by 1, 2, 5, or 10 percentile ranks on the teacher training college exam score:

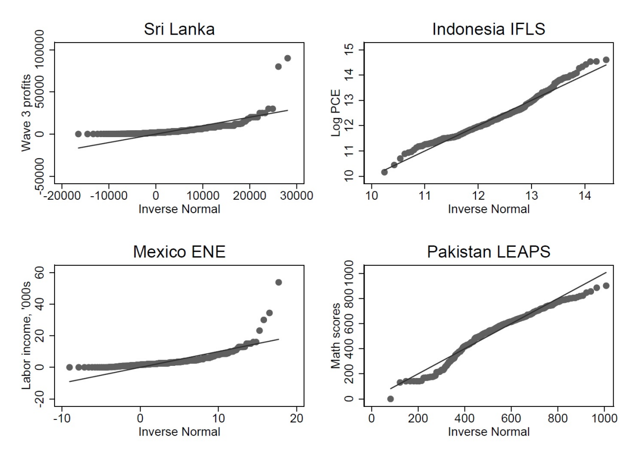
Test statistic	τ_1	τ_2	τ_3	τ_4
T^{KS}	0.45	1.00	1.00	1.00
T^{OLS}	0.11	0.37	0.92	1.00

Note the KS statistic has the advantage in this case that we are also interested in treatment-induced changes in distributions beyond location shifts.

Errors are often far from normal

To illustrate prevalance of non-normal errors in development outcomes, we look at data from Bruhn and McKenzie (2009).

- Departures from normality in many outcomes;
- see also Rachael Meager (2019): microenterprise profits across 7 studies exhibit spike at zero.



Since nothing is entirely free. . .

The downside here is that it may be harder to interpret violations of a 'sharp null': KS statistic, for example, can reject for reasons other than location shifts.

But in a RI context, this is true more generally.

- Rejection of the 'sharp' null that $y_{i0} = y_{i1}$ for all i, based on regression coefficient τ, does *not* imply that $\tau \neq 0$!
- Literature on 'robust' randomization inference highlights asymptotic interpretations of rejections in terms of a non-sharp null (like $\tau = 0$)for, e.g., studentized regression coefficients.

Use case 2: Modeling interdependence

Modeling interdependence—motivation

A common question at the experimental *design* stage is the extent of non-independence among units in treated ‘clusters’.

It is well known that random-effects models can offer efficiency gains at the estimation stage.

RE models are agnostic about the distribution of common shocks, assuming only independence from treatments of interest.

Putting structure on independence

More generally, modeling the distribution of these common disturbances can offer power gains when these models are (approximately) true—but this is hard to know ex ante.

In our Rwanda analysis, we used blinded (single-arm) data under the simulated null of no effect to inform this choice.

- Should we assume normality of common shocks?
- At what level(s) should we model these shocks?

These would be very difficult choices to make guesses about, but blinded data allow doing so. The power gains are surprisingly large relative to, e.g., the financial cost of increases in sample size.

Usefully wrong structure in an RI framework...

From Imbens and Rubin (2015):

> *[A]ny scalar function of the estimated parameters [of models for potential outcomes under control and treatment] is a test statistic that can be used to obtain a p-value for a sharp null hypothesis.*
>
> *Although these test statistics are motivated by statistical models, the validity of an FEP [Fisher exact p-value] based on any one of them does not rely on the validity of these models. In fact, these models are purely descriptive given that the potential outcomes are considered fixed quantities. The reason such models may be useful, however, is that they may provide good descriptive approximations to the sample distribution of the potential outcomes under some alternative hypothesis. If so, the models can suggest a test statistic that is relatively powerful against such alternatives."*

Simulated power for learning outcomes

Our preferred model (LME:RJ) has a standard error of 0.025 on the key coefficient, τ_A^P, reducing the minimum detectable effect by 30 percent from less favorable models (such as RE:RK) and by 17 percent from OLS with district FE.

					Distribution under sharp null				
Model	Sample	FE	RE	$\tilde{z}_0$	τ_A	τ_E	τ_{AE}	τ_A^P	$B \cdot P$
OLS models (fixed-effects for dummy variables)									
OLS:D	All	Districts	·	$\bar{z}_{r-1}$	-0.000 (0.048)	-0.000 (0.053)	0.001 (0.075)	0.000 (0.030)	20 · 200
Random effects models									
RE:RS	All	Districts	Round-School	$\bar{z}_{r-1}$	-0.001 (0.041)	0.000 (0.048)	0.001 (0.061)	-0.000 (0.027)	20 · 200
RE:RJ	All	Districts	Round-Pupil	$\bar{z}_{r-1}$	-0.000 (0.053)	-0.000 (0.058)	0.001 (0.080)	0.001 (0.035)	20 · 200
Linear mixed-effects models									
LME:RS	All	Districts	Round-School	$\bar{z}_{r-1}$	-0.001 (0.041)	0.000 (0.048)	0.001 (0.061)	-0.000 (0.027)	20 · 200
LME:RJ	All	Districts	Round-Pupil	$\bar{z}_{r-1}$	-0.000 (0.039)	0.000 (0.044)	0.000 (0.058)	-0.000 (0.025)	20 · 200

Use case 3: Covariate selection

Covariate selection

Baseline data alone offer little guidance as to how the choice of covariates might absorb residual variation in studied outcomes and improved power.

Machine-learning approaches such as the ‘post-double lasso’ (Belloni, Chernozhukov, Hansen 2014; Chernozhukov et al. 2018) use realized data to make an informed choice about these nuisance parameters.

For searches over functional forms with smaller potential covariate sets—e.g., what is the right functional form for a lag dependent variable—simulations using single-arm endline data seem potentially useful.

Conclusions

Conclusions

Power gains from simulation-based specification choices can provide some compensation for the hand-tying of PAPs.

Intitutionalizing these practices would be helped by. . .

1. Formal mechanisms of blinding;
2. Guidelines on risks (under what circumstances can pooled data reveal information about treatment impacts?)
3. Guidance on cases in which blinded analyses are likely to outperform analyses based on assumed distributions.

Ex ante and ex post simulations can play complementary roles.