Berkeley School of Public Health

Systematic reviews

PHW250B

Why review the literature?

- Summarize the state of knowledge for a given exposure / intervention & disease relationship, including strengths and weaknesses of current body of published literature.
- This may be necessary to prepare for a study of a specific exposure / intervention & disease relationship.
- Reviews can help :
  - Generate new hypotheses
  - Provide context at the beginning of a paper or in a grant application
  - Motivate additional research on a question with a different study design or different exposure / outcome definition

Summary of key points

- Systematic reviews summarize the evidence on a specific research question and are conducted in a fashion such that they can be replicated.
- While often time consuming and cumbersome, systematic reviews provide powerful information that can be used to:
  - Generate new hypotheses
  - Identify gaps in the published literature
  - Prepare for future studies
  - Provide context to your work (grants, papers, etc)
- Systematic reviews should be reported using the PRISMA checklist.

Types of reviews

- **Literature review / narrative review**
  - Typically conducted by an expert on a topic
  - Use a combination of systematic and non-systematic methods to select articles for inclusion
  - Subjective and cannot be replicated
- **Systematic review**
  - Clear and specific research question, inclusion / exclusion criteria, and protocol for finding and selecting articles
  - Can be replicated when conducted properly
  - Time consuming
  - Less subjective than literature / narrative reviews
- **Meta-analysis**
  - Results of individual studies are used to obtain a pooled estimate across multiple studies
  - Purpose is to increase precision of estimates and synthesize findings across populations

Example narrative review

OPEN ACCESS Freely available online

PLoS MEDICINE

Policy Forum

Hygiene, Sanitation, and Water: Forgotten Foundations of Health

Jamie Bartram[1], Sandy Cairncross[2]*

1 Water Institute, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, North Carolina, United States of America, **2** London School of Hygiene & Tropical Medicine, London, United Kingom

This is the introductory article in a four-part *PLoS Medicine* series on water and sanitation.

Globally, around 2.4 million deaths (4.2% of all deaths) [1] could be prevented annually if everyone practised appropriate hygiene and had good, reliable sanitation and drinking water. These deaths are mostly of children in developing countries from diarrhoea and subsequent malnutrition, and from other diseases attributable to malnutrition.

How is an opportunity to prevent so many deaths (and 6.6% of the global burden of disease in terms of disability-burden. Even using the most conservative scenarios, the long-term sequelae due to diarrhoea in early childhood contribute more DALYs than do the deaths [3].

Regrettably, it is no surprise that much ill health is attributable to a lack of HSW. Globally, nearly one in five people (1.1 billion individuals) habitually defecates in the open. Conversely, 61% of the world's population (4.1 billion people) has some form of improved sanitation at home—a basic hygienic latrine or a flush toilet. Between these two extremes, many households rely on dirty, unsafe latrines or shared toilet facilities [4]. Not only can it

quality standards [5]. Reliable safe water at home prevents not only diarrhoea but guinea worm, waterborne arsenicosis, and waterborne outbreaks of diseases such as typhoid, cholera, and cryptosporidiosis.

Much of the impact of water supply on health is mediated through increased use of water in hygiene. For example, hand washing with soap reduces the risk of endemic diarrhoea, and of respiratory and skin infections, while face washing prevents trachoma and other eye infections. A recent systematic review of the literature [6] confirmed that hygiene, particularly hand washing at delivery and postpartum,

- Example of a very general review on a topic rather than a narrow research question.
- "A massive disease burden is associated with deficient hygiene, sanitation, and water supply and is largely preventable with proven, cost-effective interventions."

Bartram J, Cairncross S (2010) Hygiene, Sanitation, and Water: Forgotten Foundations of Health. PLoS Med 7(11): e1000367. doi:10.1371/journal.pmed.1000367

Example systematic review & meta-analysis

Review

Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis

Lancet Infect Dis 2005; 5: 42–52

Lorna Fewtrell, Rachel B Kaufmann, David Kay, Wayne Enanoria, Laurence Haller, and John M Colford Jr

LF and DK are at the Centre for Research into Environment and Health, University of Wales, Aberystwyth, UK; RBK is at the World Bank, Washington DC, and Centers for Disease Control and Prevention, Atlanta, GA, USA; WE and JMCJr are at the University of California, School of Public Health, Berkeley, CA, USA; and LH and JMCJr are at the World Health Organization, Water, Sanitation and Hygiene Unit, Geneva, Switzerland.

Correspondence to: Lorna Fewtrell, Centre for Research into Environment and Health, University of Wales, Aberystwyth, Ceredigion SY23 2DB, UK
Tel +44 (0)1270 250583;

Many studies have reported the results of interventions to reduce illness through improvements in drinking water, sanitation facilities, and hygiene practices in less developed countries. There has, however, been no formal systematic review and meta-analysis comparing the evidence of the relative effectiveness of these interventions. We developed a comprehensive search strategy designed to identify all peer-reviewed articles, in any language, that presented water, sanitation, or hygiene interventions. We examined only those articles with specific measurement of diarrhoea morbidity as a health outcome in non-outbreak conditions. We screened the titles and, where necessary, the abstracts of 2120 publications. 46 studies were judged to contain relevant evidence and were reviewed in detail. Data were extracted from these studies and pooled by meta-analysis to provide summary estimates of the effectiveness of each type of intervention. All of the interventions studied were found to reduce significantly the risks of diarrhoeal illness. Most of the interventions had a similar degree of impact on diarrhoeal illness, with the relative risk estimates from the overall meta-analyses ranging between 0·63 and 0·75. The results generally agree with those from previous reviews, but water quality interventions (point-of-use water treatment) were found to be more effective than previously thought, and multiple interventions (consisting of combined water, sanitation, and hygiene measures) were not more effective than interventions with a single focus. There is some evidence of publication bias in the findings from the hygiene and water treatment interventions.

- Reviewed 2120 abstracts
- 46 studies were eligible
- Pooled relative risk ranged from 0.63 to 0.75.
- “The results generally agree with those from previous reviews, but [...] multiple interventions (consisting of combined water, sanitation, and hygiene measures) were not more effective than interventions with a single focus.”

Lancet Infect Dis 2005; 5: 42–52

Challenges in reviewing the literature

- Large number of potential studies to include
- Inconsistent exposure / outcome definition or terminology makes it difficult to standardize findings across studies
- Different statistical methods make it difficult to compare findings across studies

Systematic review steps

1. Write systematic review protocol
2. Search the literature
3. Review titles, abstracts, and full texts
4. Abstract data from included studies
5. Assess risk of bias
6. Summarize evidence

Step 1: Write systematic review protocol

- Define the study question
- Define the study hypothesis
- Describe the rationale for the study
- Describe the methods for the review, including:
  - Databases that will be searched
  - Exact search queries
  - Study inclusion and exclusion criteria
  - Data that will be abstracted
  - Risk of bias assessment

Step 2: Search the literature

- Search databases such as PubMed, Cochrane Library, Embase, LILACS, Medline, and Pascal Biomed
  - Ideally, select databases that can return fully reproducible results based on your search query
  - Google Scholar is not fully reproducible
- Once initial set of eligible studies is selected, it is common to search those studies' reference list for additional studies that may meet inclusion criteria that were not caught in the initial search

Step 3: Review titles, abstracts, and full texts

- Process
  - If a study title suggests relevance its abstract is reviewed
  - If its abstract suggests relevance its full text is reviewed
  - The full text is used to assess final eligibility.
- Challenges
  - If a large number of studies is identified in the initial search, which is common, this process is very time consuming and requires a team.
  - It is often a good practice to replicate across team members since inclusion decisions can be subjective.

Step 4: Abstract data from included studies

- Time consuming process
- Commonly extracted variables include year of study, study location, exposure/intervention, outcome, study design elements that affect risk of bias (e.g., blinding), confounders controlled for, measure of association, p-value, SE, confidence interval
- It is a good practice to replicate this work with two reviewers to reduce the chance of errors.
- Difficult to decide what to abstract if the number of results is voluminous
  - E.g., a study ran many different statistical models for the same research question
  - Pre-specifying how this situation will be handled in the protocol reduces abstraction time and subjectivity

Example of extracted data

Reference	Intervention	Country (location)	Study quality*	Health outcome	Age group	Measure	Estimate (95% CI)
Khan,[11] 1982	Handwashing with soap	Bangladesh (unstated)	Good	Diarrhoea	All	RR†	0·62 (0·35–1·12)‡
Torún,[12] 1982	Hygiene education	Guatemala (rural)	Poor	Diarrhoea	0–72 months	RR†	0·81 (0·75–0·87)‡
Sircar et al,[13] 1987	Handwashing with soap	India (urban)	Good	Watery diarrhoea	0–60 months	RR†	1·13 (0·79–1·62)
					>5 years	RR†	1·08 (0·86–1·37)
				Dysentery	0–60 months	RR†	0·67 (0·42–1·09)
					>5 years	RR†	0·59 (0·37–0·93)
				Combined outcome	Combined ages	RR†	0·97 (0·82–1·16)‡
Stanton et al,[14] 1988; Stanton and Clemens,[15] 1987	Hygiene education	Bangladesh (urban)	Good	Diarrhoea	0–72 months	IDR	0·78 (0·74–0·83)‡
Alam et al,[16] 1989	Hygiene education (and increased water supply)	Bangladesh (rural)	Good	Diarrhoea	6–23 months	OR	0·27 (0·11–0·66)‡

Lancet Infect Dis 2005; 5: 42–52

Step 5: Assess risk of bias

- Criteria for low, medium, or high risk of bias should be defined in the study protocol. These can include:
  - Sample size
  - Control for confounding
  - Study design (e.g. blinding, matching)
  - Analytic approach
- Publication bias can also be addressed.
  - This occurs when studies with favorable results are more likely to be published than those with null or unfavorable results.
  - “File drawer problem”
  - More on this in the meta-analysis video

Step 6: Summarize evidence

- Summarize size and direction of measure of association across studies
- Highlight any meaningful heterogeneity across studies
- Highlight any gaps in the evidence base
  - E.g., lack of evidence from randomized studies or studies enrolling populations of a certain age group or gender

Reporting systematic reviews

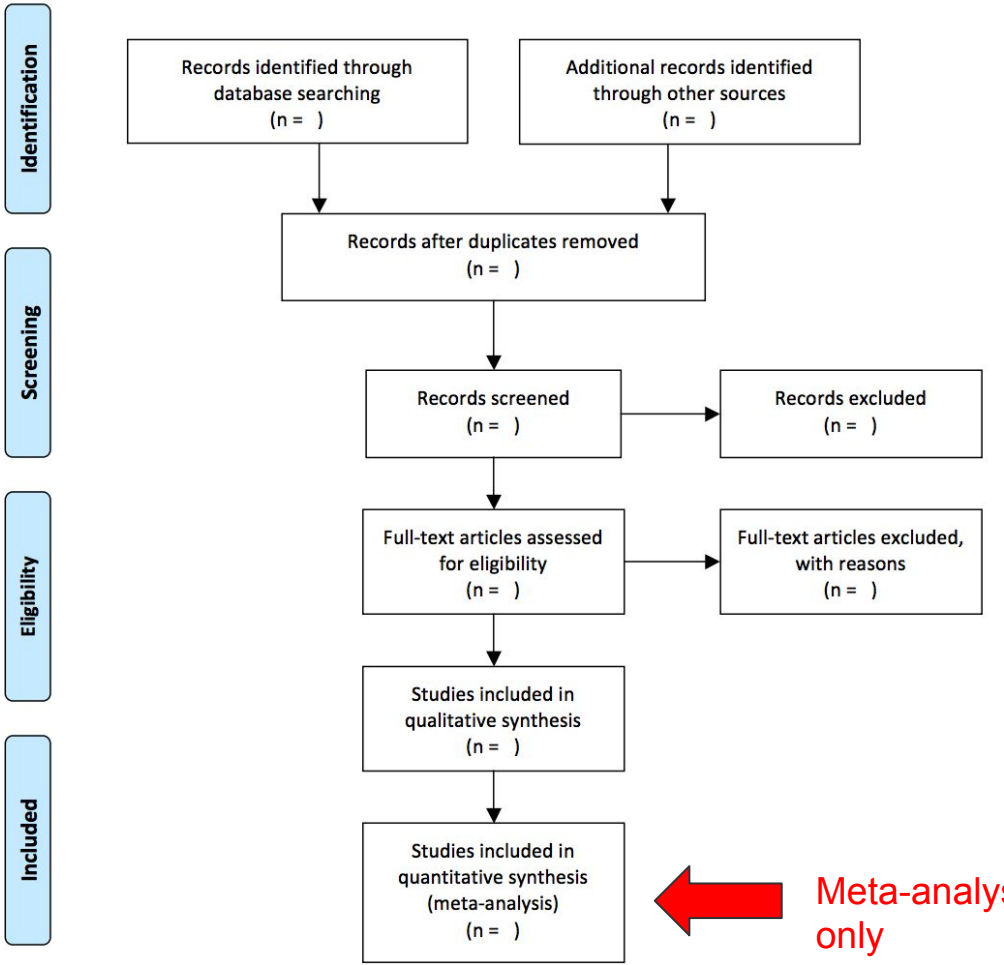
- PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses http://www.prisma-statement.org/
- PRISMA checklist for reporting includes specific guidelines on what to report in a publication of a systematic review

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			

Flow diagram for systematic reviews

- Recommended figure from PRISMA
- Should be included in the publication of every systematic review
- Allows reader to see how many records were assessed at each step



Berkeley School of Public Health

Meta-analyses

PHW250B

Why conduct a meta-analysis?

- Meta-analysis is a statistical approach that combines the effect estimates from multiple studies to obtain a pooled effect estimate with a larger amount of statistical power.
- By combining multiple studies, you achieve greater precision due to the larger sample size. This can help improve precision for estimates, which is useful when estimates are close to the null.
- There is a large body of research using meta-analyses of trials to inform clinical best practices.

Systematic reviews & meta-analyses of trials

Access provided by: UC Berkeley Library | English | Cochrane.org | Sign In

Cochrane Library

Trusted evidence.
Informed decisions.
Better health.

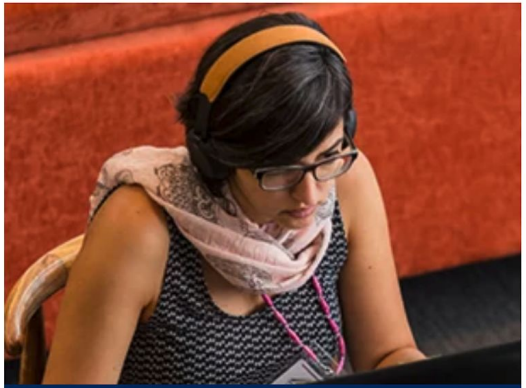
Title Abstract Keyword | Browse | Advanced search

Cochrane Reviews | Trials | Clinical Answers | About | Help

Cochrane Database of Systematic Reviews

The *Cochrane Database of Systematic Reviews* (CDSR) is the leading journal and database for systematic reviews in health care. *CDSR* includes Cochrane Reviews (systematic reviews) and protocols for Cochrane Reviews as well as editorials and supplements.

CDSR (ISSN 1469-493X) is owned and produced by Cochrane, a global, independent network of researchers, professionals, patients, carers, and people interested in health.



Cochrane Clinical Answers

No time to read Cochrane Reviews? Think again - visit Cochrane Clinical Answers

Aims and scope

https://www.cochranelibrary.com/cdsr/about-cdsr

Sources of heterogeneity

- Individual studies' measures of association may differ due to:
  - Population characteristics (age, sex, race)
  - Year of study
  - Location of study
  - Study participants' symptoms / clinical characteristics
  - Variables treated as confounders
  - Study design
  - Sample size
  - Statistical analyses

Meta-analysis steps

1. Write systematic review protocol
2. Search the literature
3. Review titles, abstracts, and full texts
4. Abstract data from included studies
5. Assess risk of bias
6. **Assess heterogeneity of measures of association**
7. **Obtain pooled measure of association**
    - **Conduct any subgroup analyses**
8. **Assess publication bias**

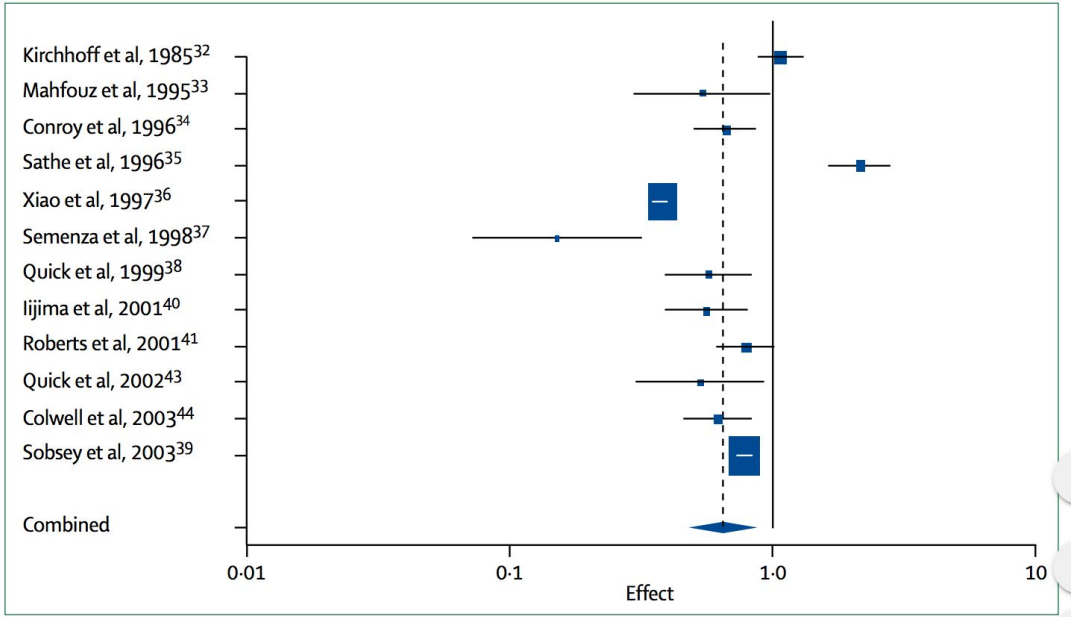
Initial steps are the same as in systematic reviews

Specific to meta-analyses

Step 6: Assess heterogeneity of measures of association

- Examine **forest plot** for heterogeneity
- Each study is one row in the plot
- This plot shows the combined effect estimate, but this is estimated after this initial assessment.
- The dashed line corresponds to the pooled estimate.
- The solid line corresponds to the null.
- Most effect estimates are to the left of the null and have similar value, suggesting little heterogeneity.

Random effects meta-analysis of household treatment water quality interventions



Fewtrell et al. Lancet Infect Dis 2005; 5: 42–52

Step 6: Assess heterogeneity of measures of association

Test for heterogeneity using a Q-statistic (a type of chi-square test)

Null hypothesis: no heterogeneity among original studies

MoA: measure of association

i: indicator for each study

MoA_s: summary measure of association

w_i: weight of ith study

$$Q = \sum_{i=1}^{k} w_i(\mathrm{MoA}_i - \mathrm{MoA}_s)^2$$

$$\mathrm{MoA}_s = \frac{\sum_i w_i \times \mathrm{MoA}_i}{\sum_i w_i}$$

- There are different approaches to weighting
- Larger studies usually have larger weights
- Under the null, the Q statistic follows a chi square distribution with k-1 degrees of freedom, where k is the number of studies
- P-value<0.2: reject the null hypothesis, conclude that heterogeneity is statistically significant

Step 6: Assess heterogeneity of measures of association

Test for heterogeneity using a I^2 statistic

- Assesses the percentage of variability due to heterogeneity rather than chance.

$$I^2 = \left((Q - \mathrm{df})/Q\right) \times 100$$

- Q: defined as in previous slide
- df: degrees of freedom (k-1)
- If I^2 > 50% heterogeneity is substantial

Step 7: Obtain pooled measure of association

- MoA_s: summary or pooled measure of association
- w_i: weight of ith study
- Weights are defined differently depending on the statistical method:
  - **Fixed effects**
    - Inverse variance
    - Mantel-Haenszel
    - Peto
  - **Random effects**
    - DerSimonian Laird

$$\mathrm{MoA}_s = \frac{\sum_i w_i \times \mathrm{MoA}_i}{\sum_i w_i}$$

Step 7: Obtain pooled measure of association

- Choice of statistical method for pooling measures of association depends on whether heterogeneity is present
- **If there is evidence of significant heterogeneity:**
  - It is not usually appropriate to estimate a pooled measure of association
  - You must assess whether the pooled estimate would be meaningful given the heterogeneity.
- If you decide to calculate a pooled estimate, a **random effects model** can be used.
- If there is not evidence of significant heterogeneity, either a **random effects model** or a **fixed effects model** can be used

Step 7: Obtain pooled measure of association

- **Fixed effects models**
  - Assume included studies estimate a common, underlying measure of association
  - Study results only differ because they enrolled a different study population, so there is sampling error.
  - Inferences are made about the only the studies in the meta-analysis
- **Random effects models**
  - Assume included studies were sampled from a hypothetical "population" of studies including the studies in the meta-analysis and other hypothetical studies
  - Individual studies' measures of association vary around a true population effect
  - Inferences are made about the hypothetical "population" of studies

Step 7: Obtain pooled measure of association

- **Choosing between fixed effects and random effects models**
- Partly a decision about heterogeneity
- Partly a decision about framing and level of inference
- Statistical factors
  - Random effects models usually assume that the individual studies' estimates are normally distributed around a true population mean with a specific variance (the variance is often estimated using the DerSimonian and Laird approach).
    - Not possible to validate this assumption
  - Fixed effects models do not make this assumption.

Subgroup analysis

- Sometimes the studies included in a meta-analysis have differing characteristics:
  - Study design (randomized vs. observational)
  - Blinded vs. unblinded trials
  - Categories of countries studied (low vs. high income)
  - Study population age range
- Subgroup analyses of studies with differing characteristics may be of interest in this case.
- Such analyses should be conducted with caution!
- Even if a meta-analysis only includes trials, this does not mean that a subgroup analysis of trials within a meta-analysis can be done with randomization-based inference.
- It is only appropriate to compare the magnitude of the measure of association pooled within subgroups.
- Assessment of statistical significance of differences between subgroups is not appropriate because different subgroups may "contain different amounts of information and thus have different abilities to detect effects, it is extremely misleading simply to compare the statistical significance of the results."

Cochrane Collaboration (https://handbook-5-1.cochrane.org/chapter_9/9_6_3_undertaking_subgroup_analyses.htm)

Example of subgroup analysis in meta-analysis

	Number of studies	Relative risk (95% CI)
Hygiene	11	0·63 (0·52–0·77)
Excluding poor quality studies	8	0·55 (0·40–0·75)
Handwashing	5	0·56 (0·33–0·93)
Education	6	0·72 (0·63–0·83)
Sanitation	2	0·68 (0·53–0·87)
Water supply	6	0·75 (0·62–0·91)
Diarrhoea only	4	1·03 (0·73–1·46)
Household connection	2	0·90 (0·43–1·93)
Standpipe or community connection	3	0·94 (0·65–1·35)
Water quality	15	0·69 (0·53–0·89)
Source treatment only	3	0·89 (0·42–1·90)
Household treatment only	12	0·65 (0·48–0·88)
Household treatment		
• excluding poor quality studies	8	0·61 (0·46–0·81)
• rural location	6	0·61 (0·39–0·94)
• urban/periurban locations	5	0·86 (0·57–1·28)
• urban/periurban excluding Sathe[35]	4	0·74 (0·65–0·85)
Multiple	5	0·67 (0·59–0·76)

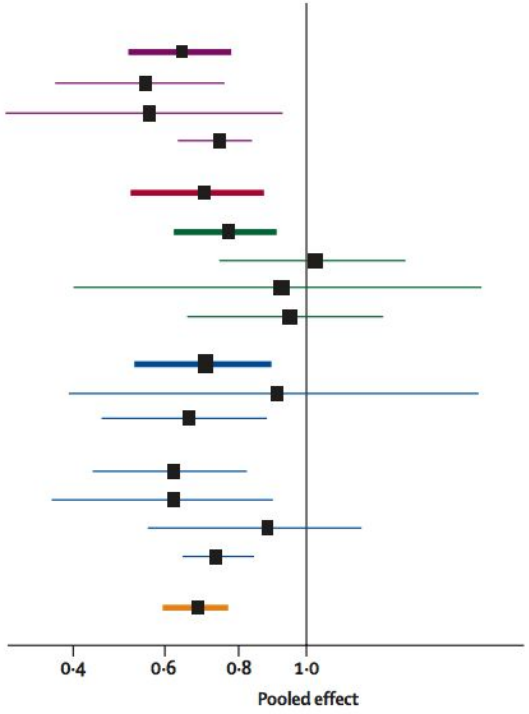


Figure 3: **Summary of meta-analysis results**

Fewtrell et al. Lancet Infect Dis 2005; 5: 42–52

Meta-regression

- Meta-regression can be used when one is interested in assessing variation in effect estimates (across the studies in the meta-analysis)
- Allows for assessment of variation across studies using continuous or categorical variables.
- It is not appropriate to conduct when the number of studies is <10.
- Outcome / dependent variable = effect estimate
- Exposure = characteristics of the study that might affect the magnitude of the effect estimate
- Each unit in the regression = a study's effect estimate stratified within subgroups
- Studies are weighted by their sample size so that larger, more precise studies are more influential.

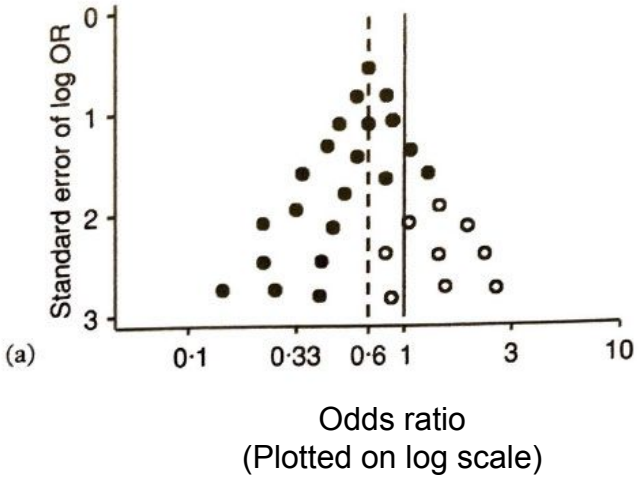
Cochrane Collaboration (https://handbook-5-1.cochrane.org/chapter_9/9_6_4_meta_regression.htm)

Sensitivity analysis

- Sometimes it is of interest to assess whether the findings of a meta-analysis are highly dependent on potentially subjective elements of the study protocol, such as:
  - Inclusion / exclusion criteria
  - Subgroup definition
  - Fixed vs. random effects
- A sensitivity analysis can be done in which the analysis is repeated using alternative decisions (e.g., different inclusion criteria)
- Ideally these analyses would be pre-specified, otherwise there is a natural tendency to perform such analyses only when we see undesirable findings.

Use funnel plots to assess publication bias

- Measure of association on one axis
- Sample size or measure of precision (e.g., standard error) plotted on the other axis
- If no publication bias occurred, we would expect the graph to resemble a funnel.
  - Larger studies with higher precision will be closer to the overall pooled estimate
  - Smaller studies with lower precision will be distributed symmetrically on either side of the average.
- The plot to the right depicts an example with no publication bias.



Egger et al. *Systematic Reviews in Health Care*. BMJ Publishing Group, 2001.

Use funnel plots to assess publication bias

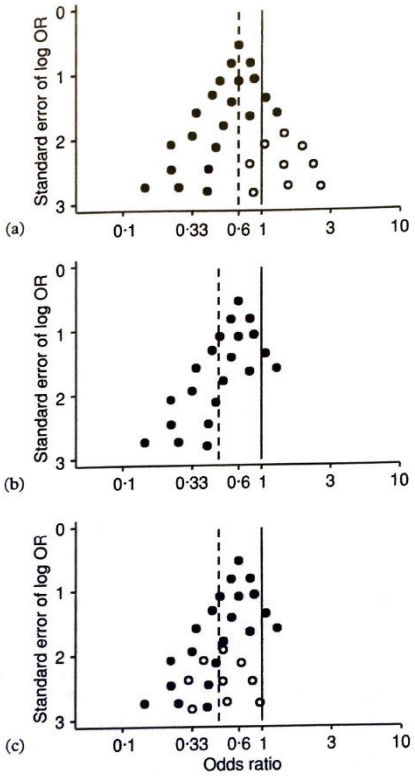
a) Symmetrical plot with **no publication bias**

- **Open circles:** smaller studies with no statistically significant effects

b) Asymmetrical plot **with publication bias.** Smaller studies with no statistically significant effects are missing from the plot.

c) Asymmetrical plot with **bias due to poor quality of smaller studies**

- **Open circles:** smaller studies with poor quality



(Plotted on log scale)

Dashed line: pooled odds ratio

Solid line: null (OR = 1)

Egger et al. *Systematic Reviews in Health Care*. BMJ Publishing Group, 2001.

Challenges in conducting meta-analyses

- Time consuming
- Any bias present in the original studies will carry forward into the meta-analysis
  - Best evidence in meta-analyses comes from those that only enroll trials
  - COCHRANE reviews
- It's not reasonable to pool results if they are very heterogeneous
- For meta-analyses of observational studies we cannot expect that the exposure would have the same association with disease in all study populations

Reporting meta-analyses

- PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses http://www.prisma-statement.org/
- PRISMA checklist for reporting includes specific guidelines on what to report in a publication of a systematic review

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			

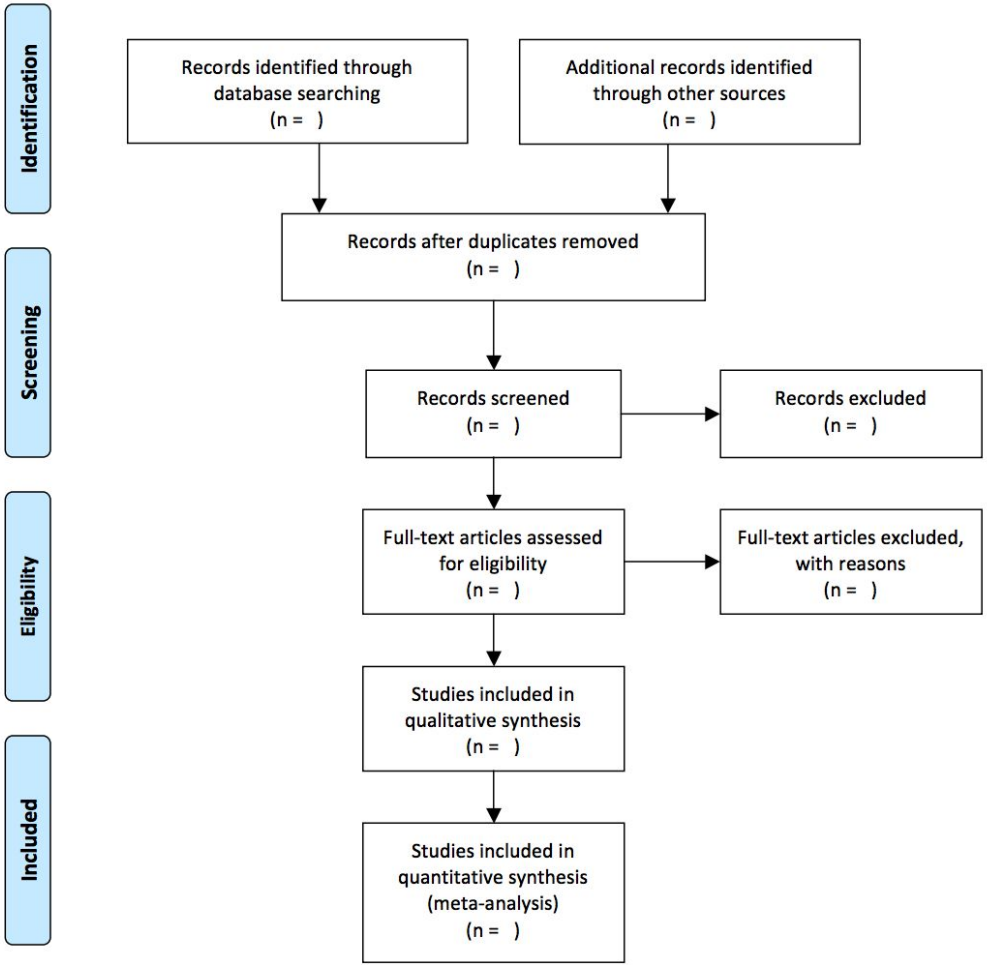
Step 8: Assess publication bias

- To make valid conclusions about a hypothesis based on a systematic review, two criteria must be met:
  - Each study in the review must use unbiased methods
  - Published studies constitute an unbiased sample of a theoretical population of unbiased studies.
- The latter criterion is not met when publication bias is present.
- Why does publication bias occur?
  - Journal editors are often less interested in negative or null results.
  - Researchers may not want to devote the time required to publish results when the results are negative or null.

Szklo & Nieto, 3rd. Ed.

Flow diagram for meta-analysis

- Recommended figure from PRISMA
- Should be included in the publication of every meta-analysis
- Allows reader to see how many records were assessed at each step



Summary of key points

- Meta-analyses include the same initial steps as systematic reviews but also provide a pooled estimate of the measure of association.
- It is important to assess heterogeneity of study findings before conducting a meta-analysis.
- Fixed vs. random effects models ideally should be chosen based on the desired level of inference as well as the level of heterogeneity.
- The COCHRANE collaboration includes a library of meta-analyses of trials and many guidelines for best practices in meta-analysis.
- Meta-analyses should be reported using the PRISMA checklist.

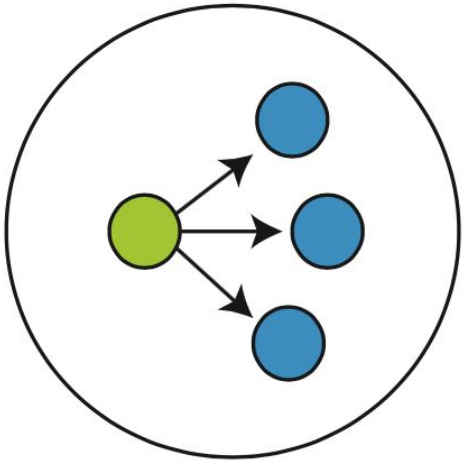
Berkeley School of Public Health

Spillover effects

PHW250B

Definition of spillover effects

- A spillover is the effect of an intervention on people not targeted by an intervention but who were connected to intervention recipients socially, geographically, or by some other means.



Motivation

- In most epidemiologic studies that are not focused on infectious diseases, we assume that individuals are individuals' outcomes are not correlated.
- In infectious disease studies we cannot assume that individuals are independent because of the transmissible nature of disease.
  - The presence of a spillover effect reflects this transmission — an individual who did not receive an intervention was impacted by it.
- In some non-infectious disease studies, it might not be valid to assume individuals' outcomes are not correlated:
  - What if the behaviors of some people affect those of others?
  - What if attitudes of some people affect those of others?

Examples of spillover effects

Infectious disease interventions

Vaccines

Are unvaccinated students who attend schools where free flu shots are offered less likely to get the flu?

Sanitation

Does improving latrines for some households in a community reduce worm infections for their neighbors?

Malaria

Are people who don't own insecticide treated nets living in close proximity to people using nets less likely to become infected?

Examples of spillover effects

Other interventions

Smoking cessation

Are friends of participants in a smoking cessation program more likely to quit?

Obesity

If the majority of people you know are obese, are you more likely to be obese?

Women's peer groups

Are women whose peers participate in peer support groups during pregnancy less likely to experience adverse events during delivery?

Synonyms for spillover effects

- Herd immunity / herd effects
- Indirect effects
- Externalities
- Contagion effects
- Social network effects
- Diffusion

Importance of spillover effects

Bias towards the null

- Ignoring spillovers in the same direction as the treatment effect biases point estimates towards the null.
- If an intervention affects people in the control group, the outcomes of people in the control group will be more similar on average to the outcomes of the people in the intervention group.
- Spillovers are often referred to as “contamination” in the literature on cluster-randomized trials.

Importance of spillover effects

Population level impact and cost-effectiveness

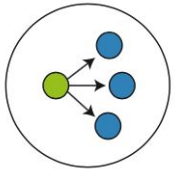
- Measurement of spillovers is needed to accurately estimate the population-level impact and cost-effectiveness of an intervention.
- One of the reasons vaccines are such an effective, commonly used public health intervention are their strong herd effects, which result in high impact and cost-effectiveness.
- Herd immunity: immunity that occurs when a large enough proportion of a population is vaccinated that unvaccinated individuals are protected from infection due to decreased transmission

Spillover measurement across disciplines

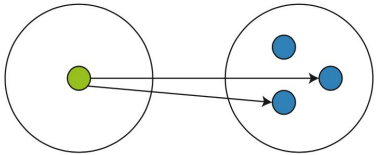
- Rich literature on spillovers related to vaccines and herd immunity
- Many papers using mathematical models to estimate spillover effects, but few empirical studies outside of the vaccine literature
- Growing interest in measuring spillovers in economics and political science covering a wide range of topics
  - Health interventions (deworming, health education, insecticide treated nets, maternal and child health)
  - Conditional cash transfers
  - Women's empowerment programs
  - Information to increase voter turnout

Types of spillover effect parameters

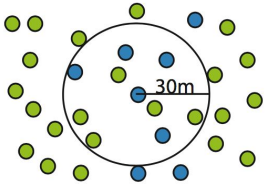
Within-cluster spillover effect



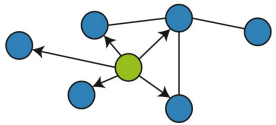
Spillover effect conditional on distance



Spillover effect conditional on treatment density



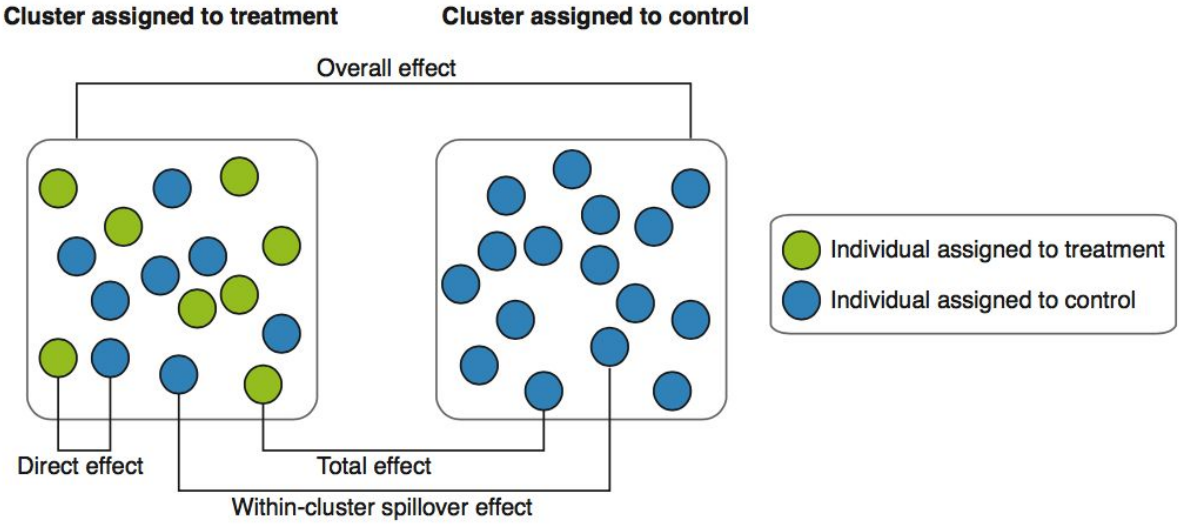
Network spillover effect



Notation used in this video

- Y: outcome / disease
- X: cluster treatment assignment
- T: individual treatment assignment

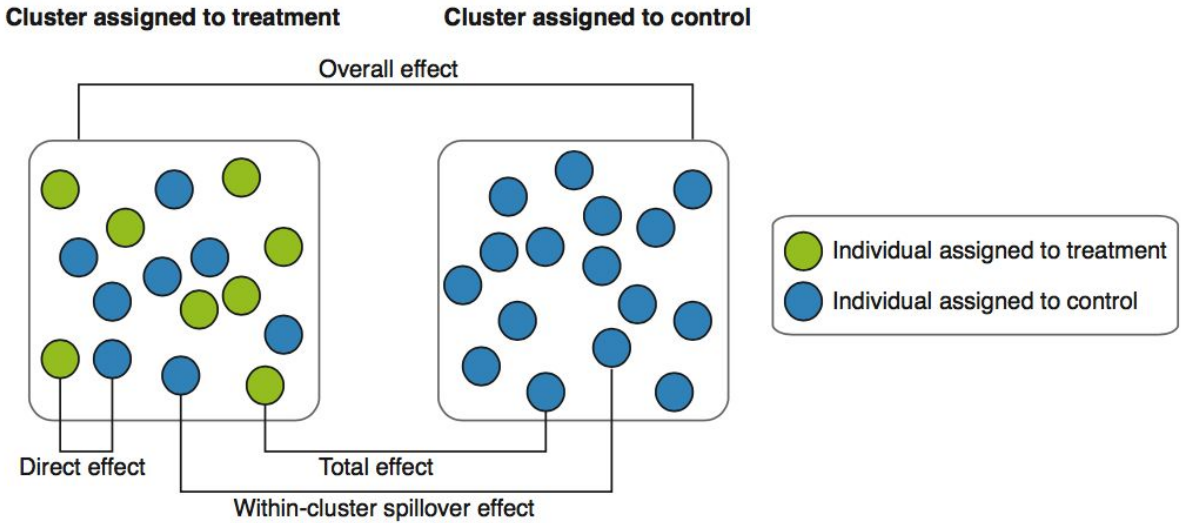
Within-cluster spillover effects



Adapted from Halloran et al., 2010

- Appropriate when spillover effects are expected to only occur within clusters.
- **Example:** spillover effect of a sanitation intervention delivered to some but not all households in a village
- Critical assumption: no spillover effects between clusters

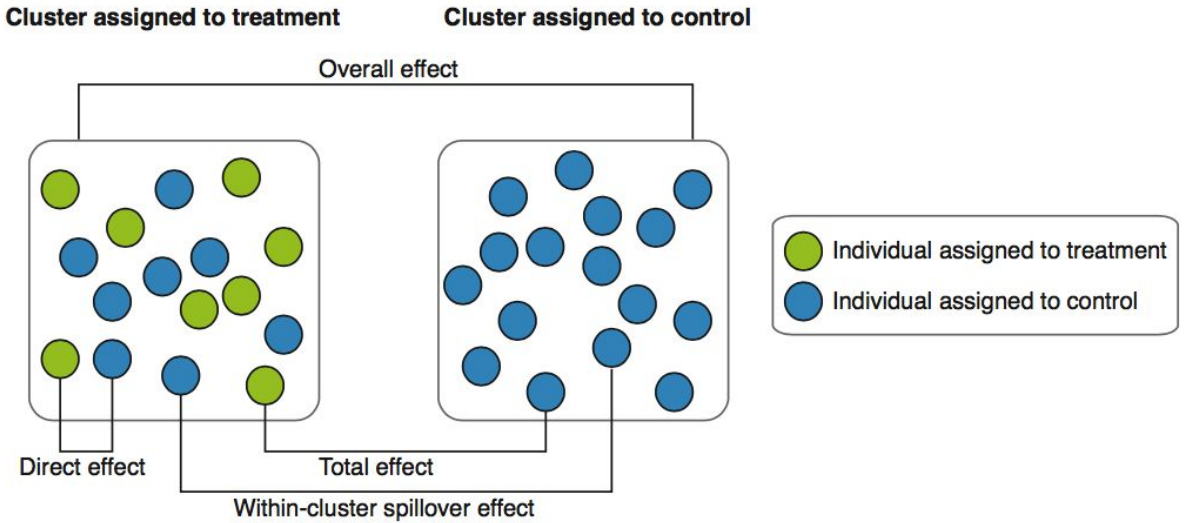
Within-cluster spillover effects



Adapted from Halloran et al., 2010

Overall effect: $E[Y \mid X = 1] - E[Y \mid X = 0]$

Within-cluster spillover effects

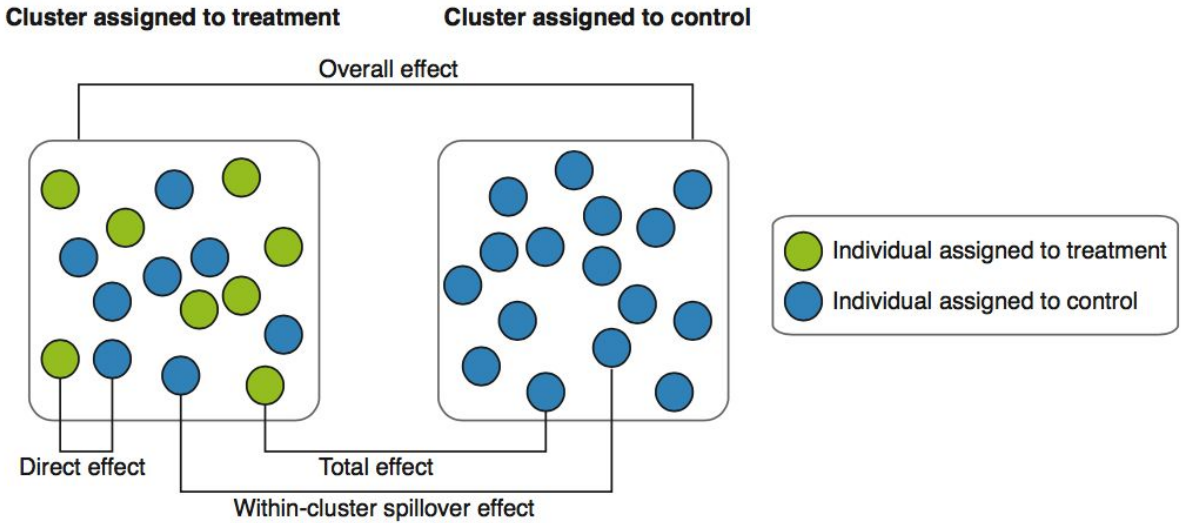


Adapted from Halloran et al., 2010

Overall effect: $E[Y \mid X = 1] - E[Y \mid X = 0]$

Total effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 1, T = 0]$

Within-cluster spillover effects



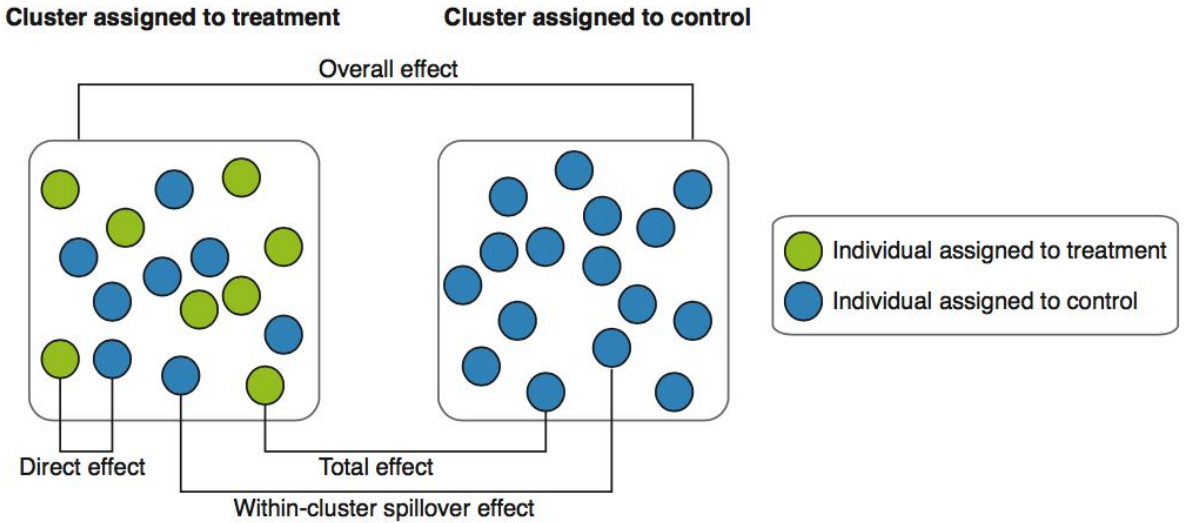
Adapted from Halloran et al., 2010

Overall effect: $E[Y \mid X = 1] - E[Y \mid X = 0]$

Total effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 0, T = 0]$

Direct effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 1, T = 0]$

Within-cluster spillover effects



Adapted from Halloran et al., 2010

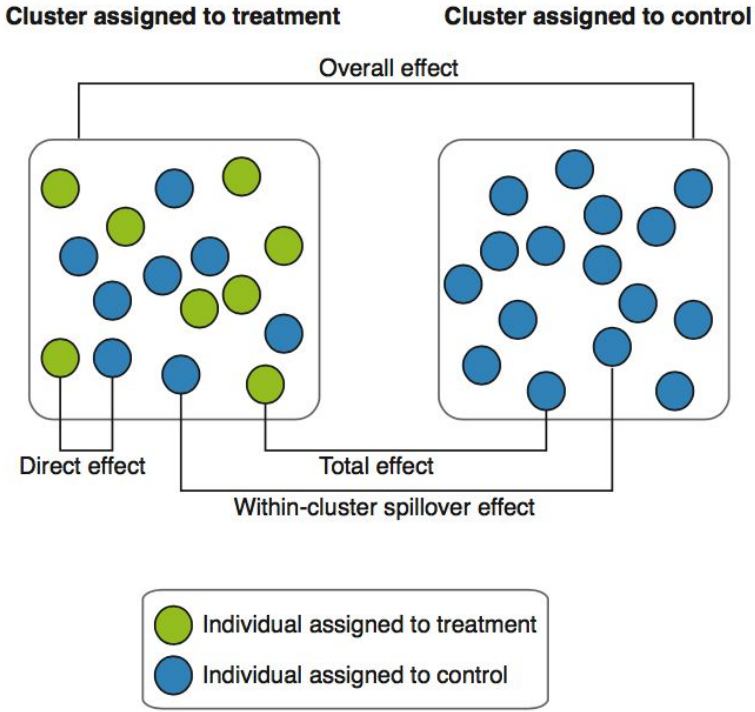
Overall effect: $E[Y \mid X = 1] - E[Y \mid X = 0]$

Total effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 0, T = 0]$

Direct effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 1, T = 0]$

Within-cluster spillover effect: $E[Y \mid X = 1, T = 0] - E[Y \mid X = 0, T = 0]$

Double-randomized trials

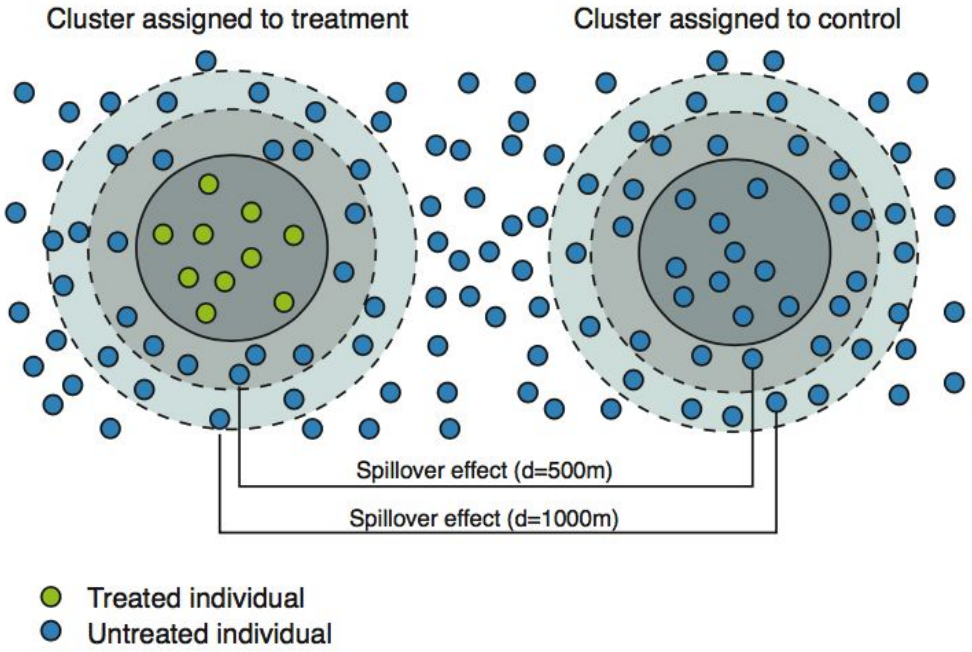


- Double-randomized trials are the most rigorous design for estimating within-cluster spillover effects.
  - Randomize clusters to treatment vs. control
  - Randomize individuals in the treatment arm to treatment vs. control
- Measured and unmeasured confounders are balanced:
  - between study arms
  - between individuals in the treatment arm

Spillover effects and contamination

- When conducting cluster-randomized trials, the goal is to include enough distance between clusters to prevent “contamination”.
- It is important to minimize this type of contamination when estimating within-cluster spillover effects.
- But what if this contamination is of interest?
  - We can estimate spillover effects conditional on distance.

Spillover effect conditional on distance

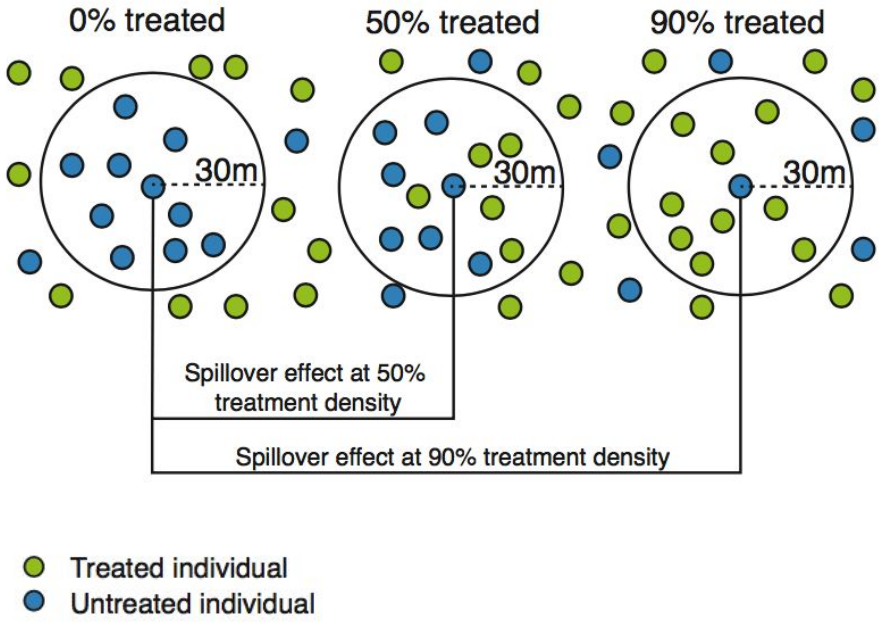


Spillover effect conditional on distance (*D*):

$$E[Y \mid X = 1, T = 0, D = d] - E[Y \mid X = 0, T = 0, D = d]$$

- Appropriate when spillover effects are expected to extend beyond study clusters
- **Example:** Distance from clusters where everyone in the cluster received an insecticide treated bed net to prevent malaria
- Can compare magnitude of spillover effects over a distance gradient
- Important to maintain enough distance between treatment and control clusters that individuals outside of control clusters can serve as a valid counterfactual

Spillover effect conditional on treatment density

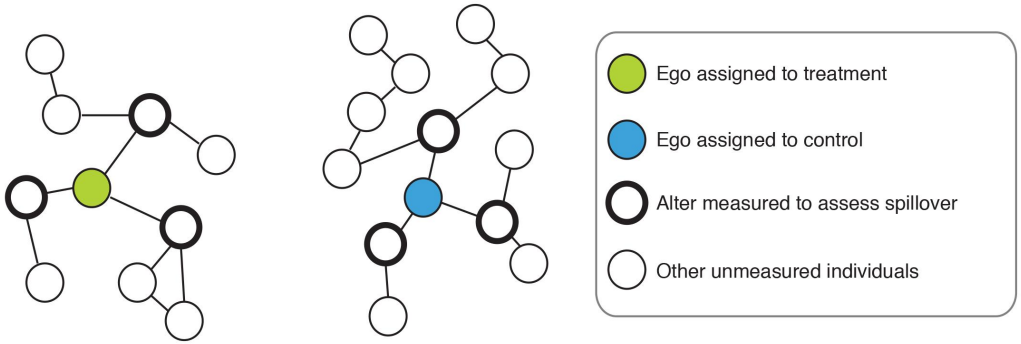


- This example is shown for an individually randomized trial.
- Can also be used with a cluster randomized design if the proportion
- **Example:** Assess spillover effect on malaria associated with percentage of nearby households with an insecticide treated net
- Can compare magnitude of spillover effects over a gradient of treatment density

Spillover effect conditional on treatment density (*P*):

$E[Y \mid T = 0, P = p + \delta] - E[Y \mid T = 0, P = p]$

Network spillover effect



- This example is shown for an individually randomized trial.
- **Example:** Spillover effect of a smoking cessation program among friends of program participants
- Requires information about social or other connectedness, which can be cumbersome to collect.

Network spillover effect:

C = indicator for an alter being connected to an ego

$E[Y \mid T = 1, C = 1] - E[Y \mid T = 0, C = 1]$

Summary of key points

- A spillover is the effect of an intervention on people not targeted by an intervention but who were connected to intervention recipients socially, geographically, or by some other means.
- Alternative names for spillover effects include:
  - Herd immunity / herd effects, indirect effects, externalities, contagion effects, social network effects, diffusion
- In infectious disease studies we cannot assume that individuals are independent because of the transmissible nature of disease. This is also the case in some studies that do not focus on infectious diseases.
- Double-randomized design are the gold standard for estimating within-cluster spillover effects.

Berkeley School of Public Health

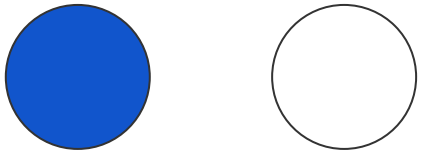
Population intervention effects

PHW250B

Motivation

- Studies often contrast two scenarios:
  - 1) if everyone was treated or exposed
  - 2) of no one was treated or exposed

Relative risk:



Risk if 100% exposed

Risk if 100% unexposed

- However, real-world effects are likely to differ if it is unlikely that an entire population will receive or not receive the intervention or exposure.

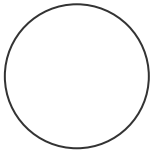
Example

Relative risk

or

Risk difference

R_e = Risk if 100% exposed



R_u = Risk if 100% unexposed

$RR = R_e / R_u$
$RD = R_e - R_u$

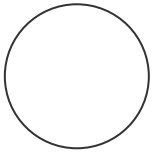
Example

Relative risk

or

Risk difference

R_e = Risk if 100% exposed



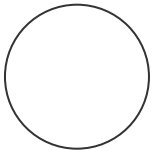
R_u = Risk if 100% unexposed

$RR = R_e / R_u$
$RD = R_e - R_u$

Population attributable fraction

in population in which 75% of people are exposed

R_t = Risk if 75% exposed



R_u = Risk if 100% unexposed

$PAF = (R_t - R_u) / R_t$

- The population attributable fraction is often more realistic because it compares the risk at the current level of exposure in the population (R_t) to the risk among the unexposed.
- But what if it is unrealistic to imagine a scenario in which everyone is unexposed?

Example: SHEWA-B

- SHEWA-B: The Sanitation Hygiene Education and Water Supply in Bangladesh Program
- Implemented by UNICEF and the Government of Bangladesh from 2007-2012
- Targeted 20 million people
- Interim evaluation in 2009 using matched control areas
- Fell short of targets for child illness and health behavior
- Poor results could reflect, poor design, poor implementation, or both

Huda et al., 2012. https://doi.org/10.1016/j.socscimed.2011.10.042
Benjamin-Chung et al., 2017. doi: 10.2105/AJPH.2017.303686

Example: SHEWA-B

- SHEWA-B: The Sanitation Hygiene Education and Water Supply in Bangladesh Program
- Implemented by UNICEF and the Government of Bangladesh from 2007-2012
- Targeted 20 million people
- Interim evaluation in 2009 using matched control areas
- Fell short of targets for child illness and health behavior
- Poor results could reflect, poor design, poor implementation, or both

Ideal scenario:

R_i = Risk when 80% receive program

Observed scenario:

R_o = Risk if 40% receive program

Risk difference for ideal vs. observed scenario: $R_i - R_o$

This is a population intervention effect.

Huda et al., 2012. https://doi.org/10.1016/j.socscimed.2011.10.042
Benjamin-Chung et al., 2017. doi: 10.2105/AJPH.2017.303686

Population intervention effects

- **Population intervention effects** are "causal effects tied to contrasts between the observed population and exposure distributions under realistic interventions"

R_i = Risk when 80% receive program

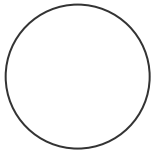
R_o = Risk if 40% receive program

Population intervention effect = $R_i - R_o$

Westreich et al., 2016 doi: 10.2105/AJPH.2016.303226

Comparing different parameters for SHEWA-B evaluation

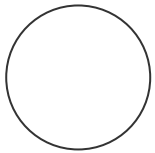
R_e = Risk if 100% exposed



R_u = Risk if 100% unexposed

RD = R_e - R_u

R_t = Risk if 40% exposed



R_u = Risk if 100% unexposed

PAF = R_t - R_u

R_i = Risk when 80% receive program

R_o = Risk if 40% receive program

Population intervention effect = R_i - R_o

Larger effect estimate
Least realistic

Smaller effect estimate
Most realistic

UNICEF never expected to deliver the program to 100% of the target population, so R_e is not realistic.

In this example, R_t was equivalent to R_o— R_t is the risk at the observed level of SHEWA-B coverage. However, it is not of interest to compare R_t to R_u.

This contrast was the most useful to UNICEF because it provided an estimate of the impact the program could have had if it had reached its target level of coverage.

Other examples of population intervention effects

- How much would HIV incidence decline compared to current levels if campaigns for preexposure prophylaxis (PrEP) for HIV reached 85% of their target population?

- What is the difference in risk of binge drinking among neighborhoods with different densities of alcohol outlets?

Westreich et al., 2016. doi: 10.2105/AJPH.2016.303226
Ahern et al., 2016. doi:10.2105/AJPH.2016. 303425

How to estimate population intervention effects

G-computation can be used to estimate population intervention effects.

- Step 1: Estimate the association between the exposure and outcome adjusting for confounders
- Step 2: Use the coefficients from the model to obtain a counterfactual value for each individual using their particular values of each confounder in two scenarios:
  - Counterfactual scenario 1
  - Counterfactual scenario 2 (e.g., observed scenario)
- Step 3: Estimate the measure of disease in the population under the counterfactual scenarios
- Step 4: Take the ratio or difference to obtain the population intervention effect
- Step 5: Use bootstrapping to obtain 95% confidence intervals

Summary of key points

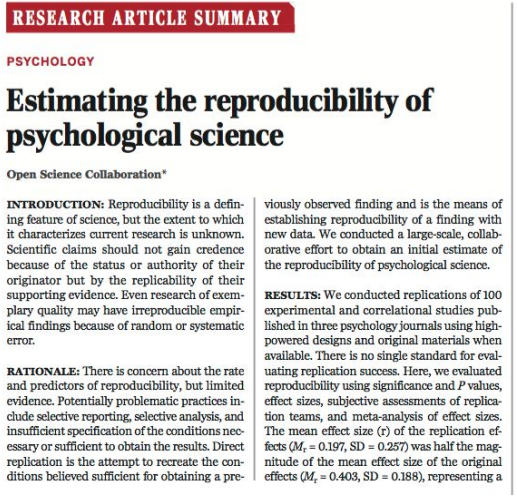
- Studies often contrast two exposure definitions: 1) if everyone was treated and 2) of no one was treated
- However, real-world effects are likely to differ if it is unlikely that an entire population will receive or not receive the intervention or exposure.
- Population intervention effects are parameters that compare two counterfactual scenarios
- These parameters can be used to estimate effects with more realistic and policy-relevant counterfactual scenarios.

Berkeley School of Public Health

Replication, transparency & reproducibility

PHW250B

Many studies' findings cannot be reproduced



RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a

Psychology

Evaluating replicability of laboratory experiments in economics

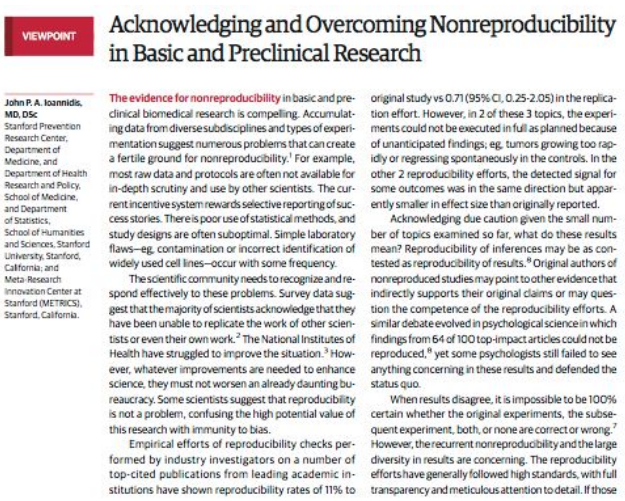
Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10]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

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

The deepest trust in scientific knowledge comes from the ability to replicate empirical findings directly and independently. Although direct replication is widely applauded (*1*), it is rarely carried out in empirical social science. Replication is now more important than ever, because the quality of results has been questioned in many fields, such as medicine (*2*–*5*), neuroscience (*6*), and genetics (*7*, *8*). In economics, concerns about inflated findings in empirical (*9*) and experimental analyses (*10*, *11*) have also been raised. In the social sciences, psychology has been the most active in both self-diagnosing the forces that create "false positives" and conducting direct replications (*12*–*15*). Several high-profile replication failures (*16*, *17*) quickly led to changes in journal publication practices (*18*). The recent Reproducibility Project: Psychology (RPP) replicated 100 original studies published in three top journals in psychology. The vast majority (97) of the original studies reported "positive findings," but in the replications, the RPP only found a significant effect in the same direction for 36% of these studies (*19*).

In this report, we provide insights into the replicability of laboratory experiments in economics. Our sample consists of all 18 between-subject laboratory experimental papers published in the *American Economic Review* and the *Quarterly Journal of Economics* between 2011 and 2014. The most important statistically significant finding,

Economics



VIEWPOINT

Acknowledging and Overcoming Nonreproducibility in Basic and Preclinical Research

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center, Department of Medicine, and Department of Health Research and Policy, School of Medicine, and Department of Statistics, School of Humanities and Sciences, Stanford University, Stanford, California; and Meta-Research Innovation Center at Stanford (METRICS), Stanford, California.

The evidence for nonreproducibility in basic and preclinical biomedical research is compelling. Accumulating data from diverse subdisciplines and types of experimentation suggest numerous problems that can create a fertile ground for nonreproducibility.[1] For example, most raw data and protocols are often not available for in-depth scrutiny and use by other scientists. The current incentive system rewards selective reporting of success stories. There is poor use of statistical methods, and study designs are often suboptimal. Simple laboratory flaws—eg, contamination or incorrect identification of widely used cell lines—occur with some frequency.

The scientific community needs to recognize and respond effectively to these problems. Survey data suggest that the majority of scientists acknowledge that they have been unable to replicate the work of other scientists or even their own work.[2] The National Institutes of Health have struggled to improve the situation.[3] However, whatever improvements are needed to enhance science, they must not worsen an already daunting bureaucracy. Some scientists suggest that reproducibility is not a problem, confusing the high potential value of this research with immunity to bias.

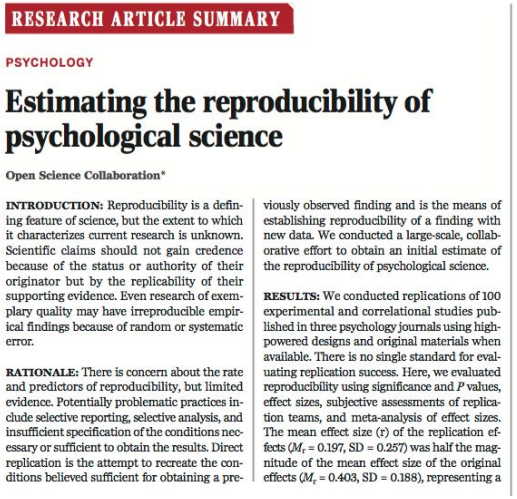
Empirical efforts of reproducibility checks performed by industry investigators on a number of top-cited publications from leading academic institutions have shown reproducibility rates of 11% to original study vs 0.71 (95% CI, 0.25-2.05) in the replication effort. However, in 2 of these 3 topics, the experiments could not be executed in full as planned because of unanticipated findings; eg, tumors growing too rapidly or regressing spontaneously in the controls. In the other 2 reproducibility efforts, the detected signal for some outcomes was in the same direction but apparently smaller in effect size than originally reported.

Acknowledging due caution given the small number of topics examined so far, what do these results mean? Reproducibility of inferences may be as contested as reproducibility of results.[4] Original authors of nonreproduced studies may point to other evidence that indirectly supports their original claims or may question the competence of the reproducibility efforts. A similar debate evolved in psychological science in which findings from 64 of 100 top-impact articles could not be reproduced,[5] yet some psychologists still failed to see anything concerning in these results and defended the status quo.

When results disagree, it is impossible to be 100% certain whether the original experiments, the subsequent experiment, both, or none are correct or wrong.[7] However, the recurrent nonreproducibility and the large diversity in results are concerning. The reproducibility efforts have generally followed high standards, with full transparency and meticulous attention to detail. If those

Basic science

Many studies' findings cannot be reproduced



RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a

Psychology

A study attempted to replicate 100 psychology studies; they only reproduced a little over a third of them.

"If one assumes that the vast majority of the original researchers were honest and diligent, then a large proportion of the problems can be **explained only by unconscious biases.**"

Nuzzo et al., 2015. *Nature* Vol 526.

Threats to reproducibility

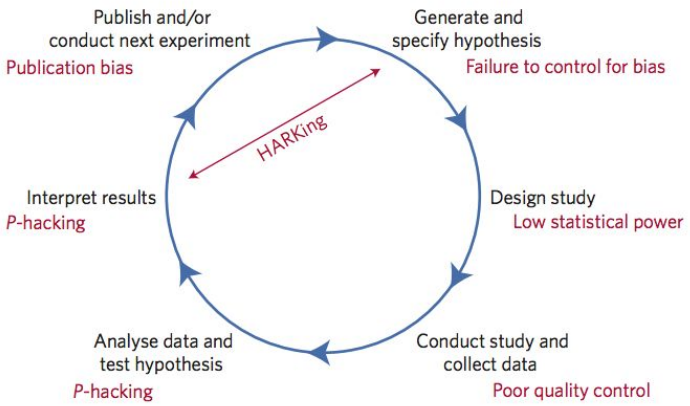


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Failure to control for bias

- **Observational studies:**
  - Unmeasured confounding
  - Measurement error
  - Etc.
- **Randomized trials:**
  - Improper randomization
  - Lack of blinding
  - Lack of allocation concealment
  - Measurement error
  - Etc.

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

Threats to reproducibility

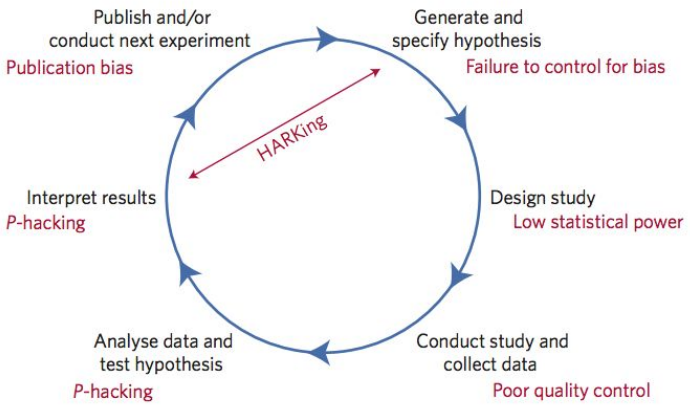


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Low statistical power

- If the sample size is too small, you may fail to detect an effect that is truly there.
- I.e., you may make a Type II error

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

Threats to reproducibility

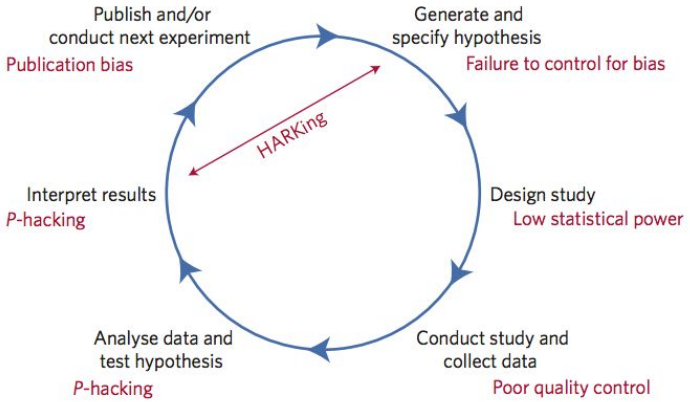


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Poor quality control

- Errors during data collection and data entry can lead to bias and misclassification that are nearly impossible to correct or to identify.
- Robust survey design and staff training is needed to prevent this.

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

Threats to reproducibility

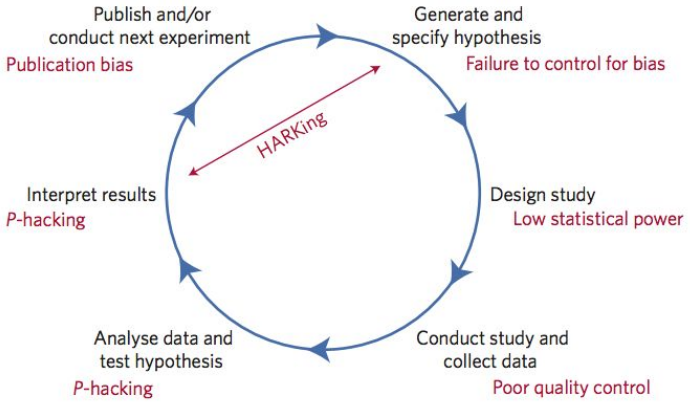


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

P-hacking

- P-hacking occurs when investigators make adjustments to their statistical model repeatedly after viewing the results in order to obtain a p-value < 0.05 or < 0.001
- Statistical significance is based on the probability that a particular result would be observed due only to chance.
- When ignoring the number of models run, eventually it is almost always possible to obtain a p-value < 0.05.
- However, this p-value is confounded by the number of tests and does not necessarily represent the true probability of a Type I error.

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

Threats to reproducibility

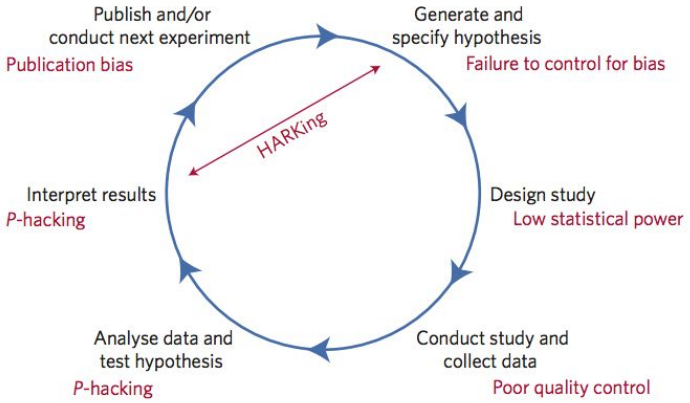


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Publication bias

- Occurs when null or non-desirable findings are not published in the peer reviewed literature.
- If only a small percentage of all null / non-desirable findings are published, those that are published will appear not to be reproducible.
- Failure to publish such studies may lead investigators to conclude that an intervention or exposure has a more desirable impact than it truly does.

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

Confirmation bias is human nature

- It is natural to look for the patterns we expect to find — to confirm our biases.
- When we don't obtain a result we expect to see, we are more likely to carefully vet our code to check for errors.
    - This may mean that errors in results that confirm our expectations go unnoticed.
- Scientists are good at coming up with explanations for unexpected findings.
- How can we improve the scientific process to account for humans' natural confirmation bias?

Nuzzo et al., 2015. *Nature* Vol 526.

The traditional scientific process

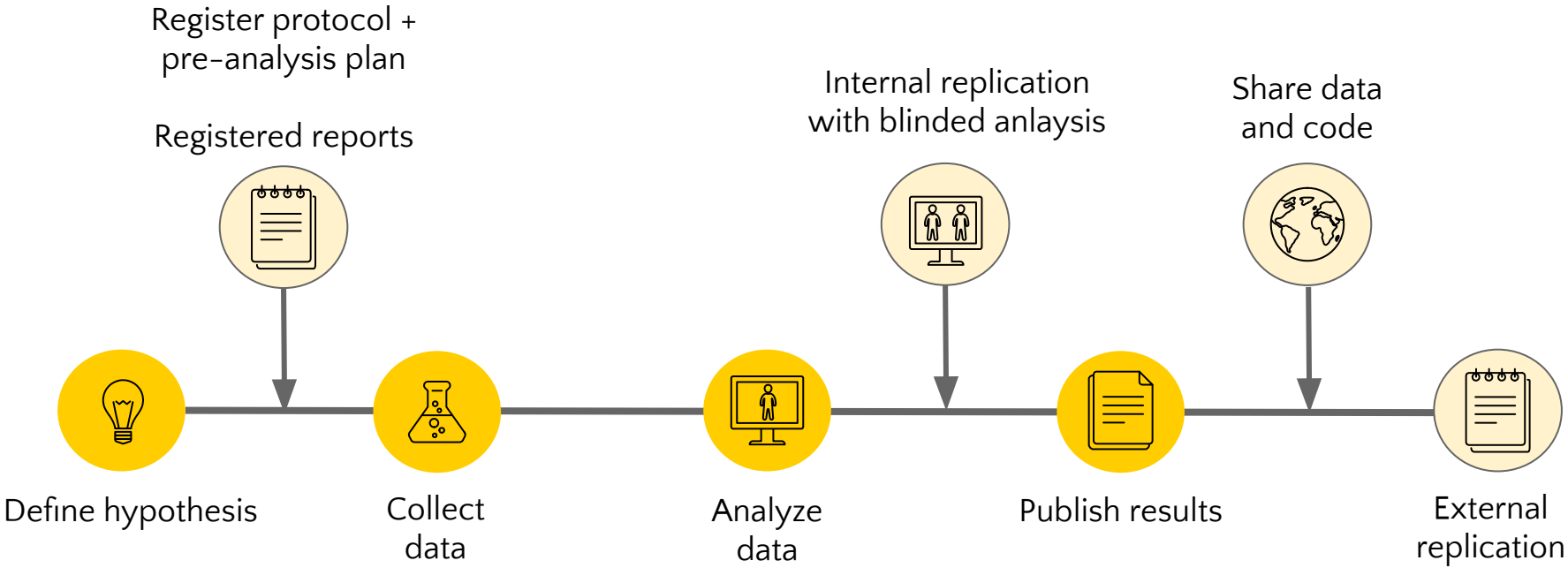
Define hypothesis

Collect data

Analyze data

Publish results

A modern process to increase reproducibility



Study registration

What is it?

- Publish one's study design in a peer-reviewed journal or well-known website:
  - clinicaltrials.gov
  - Open Science Framework (osf.io)
- Registration includes Principal Investigator, hypotheses, study design, outcomes, interventions, study population, etc.
- The is a common practice for clinical trials and is increasing for other types of studies.

Why is it useful?

- Reduces publication bias by creating a repository of studies that are in progress or completed that can be compared to the published literature.
- Helps donors efficiently allocate resources by making them aware of ongoing research.
- May spark collaborations between scientists by making them aware of ongoing research.
- Allows study participants to clearly see the aims and overall design and status of a study.

Pre-analysis plans

What is it?

- Allow for a more detailed description of the variables to be measured and specific statistical analyses to be completed.
- Can publish privately or publicly with a time stamp that allows editors to check that it was published prior to accessing the data.
    - Open Science Framework (osf.io)

Why is it useful?

- It is essentially a form of blinding — requires investigators to plan their analysis prior to seeing the data, so it reduces the influence of confirmation bias on analytic choices.
- Can increase the speed of analysis once data is collected by reducing the number of decisions that must be made once the data is available.

Registered reports

What is it?

- Registered reports allow investigators to write a manuscript summarizing their study aims and designs prior to collecting data.
- A journal will obtain peer review of the manuscript, and if reviews are favorable, they will agree to publish the final manuscript as long as the study adheres to the protocol regardless of the results.

Why is it useful?

- This approach emphasizes the importance of the study design instead of the importance of desirable results.
- It could substantially reduce publication bias.
- This approach is currently being used by a limited number of journals, but interest in it is growing.

Internal replication with blinded analysis

What is it?

- Prior to publication, two or more analysts independently perform data analysis.
- They check their answers and attempt to replicate their results.
- If results are not identical, they identify and resolve discrepancies.
- Ideally this is done while blinding analysts to the true levels of exposure or treatment (ie., using a fake treatment or exposure variable).

Why is it useful?

- Helps identify and resolve errors that can affect final study results prior to publication.
- This means that published studies that are internally replicated are less likely to contain errors, and are thus more reproducible.
- Reduces confirmation bias by blinding analysts to the true values of the treatment or exposure variable.

Data and code sharing

What is it?

- Investigators publish their code (analysis scripts) and study data after completing their primary analyses.
- Platforms for sharing code and data include:
  - Open Science Framework
    osf.io
  - Synapse
    https://www.synapse.org/
  - Github
    https://www.github.com

Why is it useful?

- Allows others to externally replicate their work and identify any potential errors in the published literature.
- Allows for more rapid development of new hypotheses or new studies using existing data.

External replication

What is it?

- After publication, investigators not part of the original study team obtain data from a study and attempt to replicate the published findings.
- This could also be performed with blinding to true treatment or exposure values.

Why is it useful?

- Helps identify and resolve errors in studies that are already published.
- If errors are found, this may lead to a correction or retraction of a published manuscript.

Summary of key points

- Multiple scientific disciplines are in the midst of a “reproducibility crisis”.
- Factors contributing to this crisis include: failure to control for bias, low statistical power, poor quality control, p-hacking, and publication bias
- There is a movement to improve reproducibility through the following tools:
  - Study registration
  - Pre-analysis plans
  - Registered reports
  - Internal and external replication
  - Data and code sharing

WORM WARS

Jade Benjamin-Chung, PhD MPH
Division of Epidemiology & Biostatistics

PH 252C - October 4, 2019

An influential trial

A Replication

A Cochrane Review

A Controversy

THE WORM WARS

The quest for the dream intervention

Chapter 1: Intro to soil-transmitted helminths

Chapter 2: An innovative study

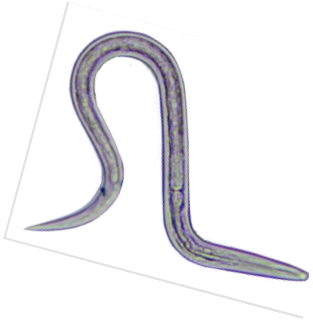
Chapter 3: The Worm Wars

Chapter 4: The moral of the story

Chapter 1

Intro to Soil-transmitted helminths

Soil-transmitted helminths:



Ascaris lumbricoides

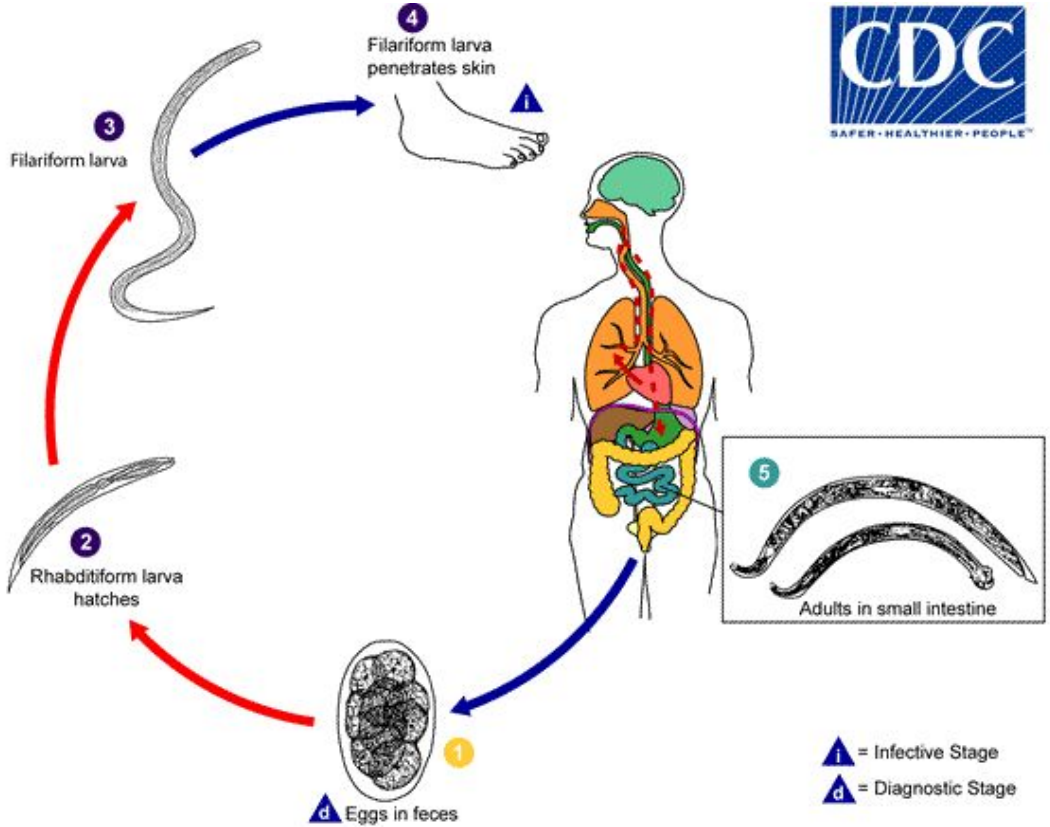
Hookworm

Trichuris trichiura

Share similarities and differences in:

- transmission
- geographic distribution
- symptoms
- treatment efficacy

Hookworm life cycle



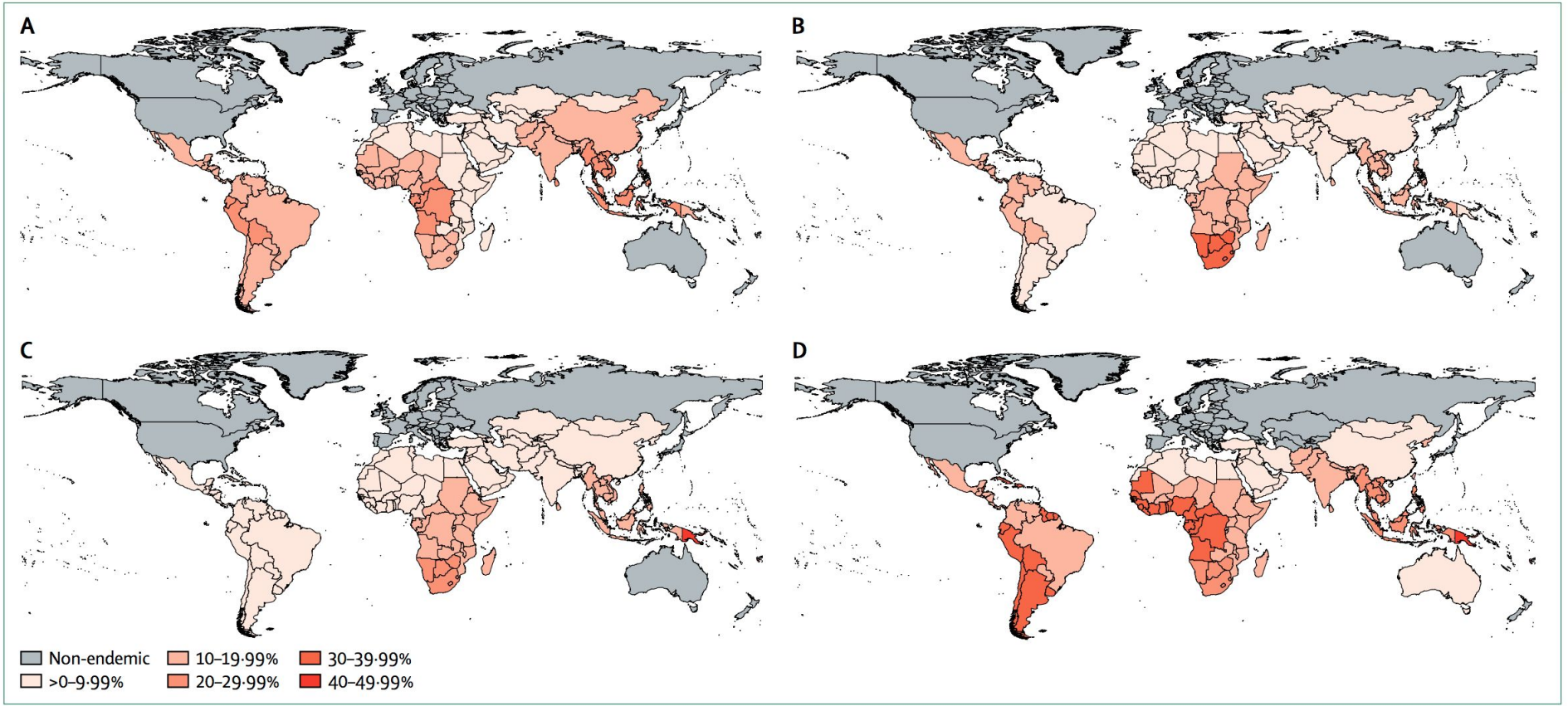


***Figure 1*: Prevalence by global regions of (A) *Ascaris lumbricoides* (for 2010), (B) *Trichuris trichiura* (for 2010), (C) hookworm (*Necator americanus* and *Ancylostoma duodenale*; for 2010), and (D) *Strongyloides stercoralis* (for 2011)**

Data for (A), (B), and (C) from Pullan and colleagues.[2] Data for *S stercoralis* are especially scarce and may be associated with strong publication bias; estimates from data by Schär and colleagues.[3] Data from single community-based studies suggest that *S stercoralis* might be present also in Australia, Israel, and Japan (which is marked as non-endemic on the map).

Jourdan et al., 2017

Symptoms and health impacts of soil-transmitted helminth infections

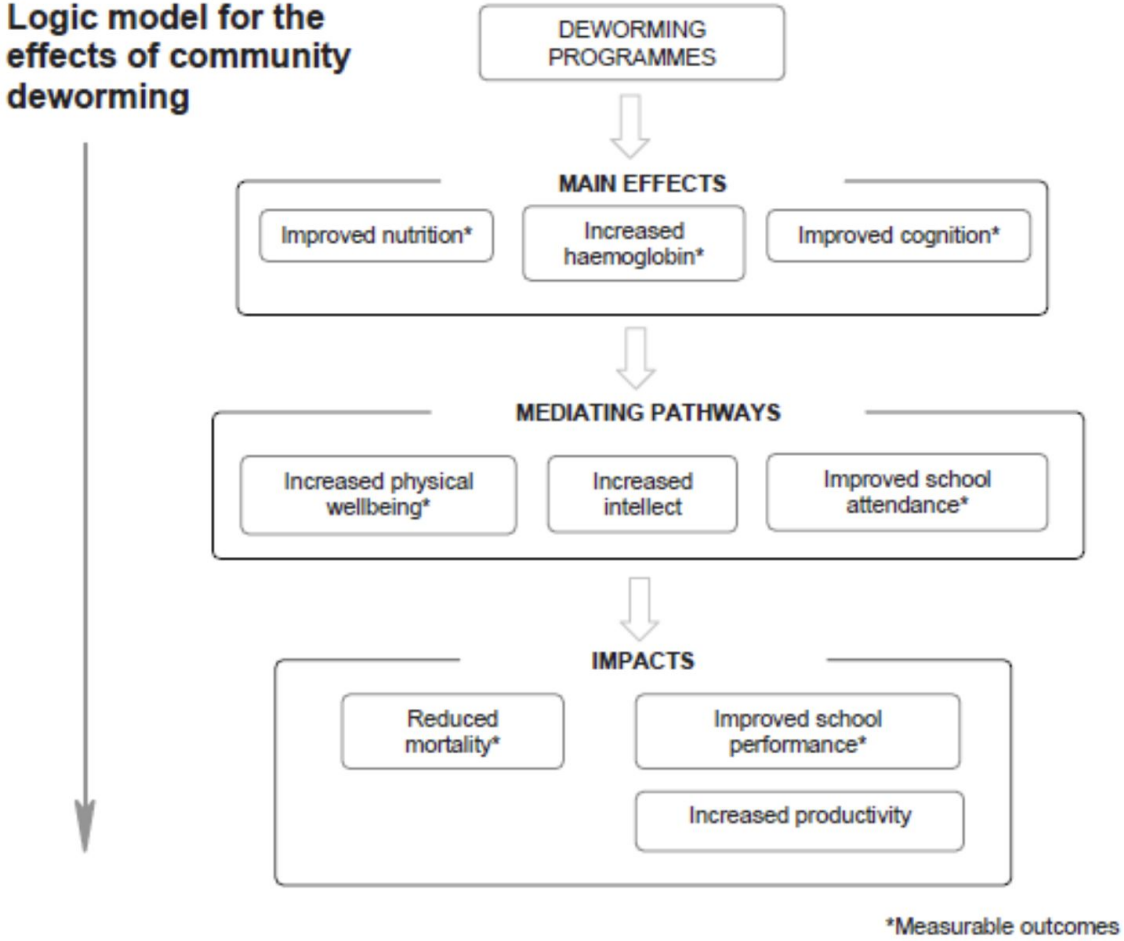
	Symptoms	Health impacts
Ascaris lumbricoides 	Asymptomatic Cough Abdominal discomfort	Altered nutrition and microbiota Intestinal obstruction Anemia Pancreatitis
Hookworm 	Asymptomatic Cough Nausea Vomiting Diarrhea	Severe anemia
Trichuris trichiura 	Asymptomatic Abdominal pain Diarrhea	Dysentery Anemia

Jourdan et al., 2017

Chapter 2

An Innovative Study

School-based deworming



Important! Pathways differ depending on the type of worm infection and the type of deworming

Taylor-Robinson et al., 2012

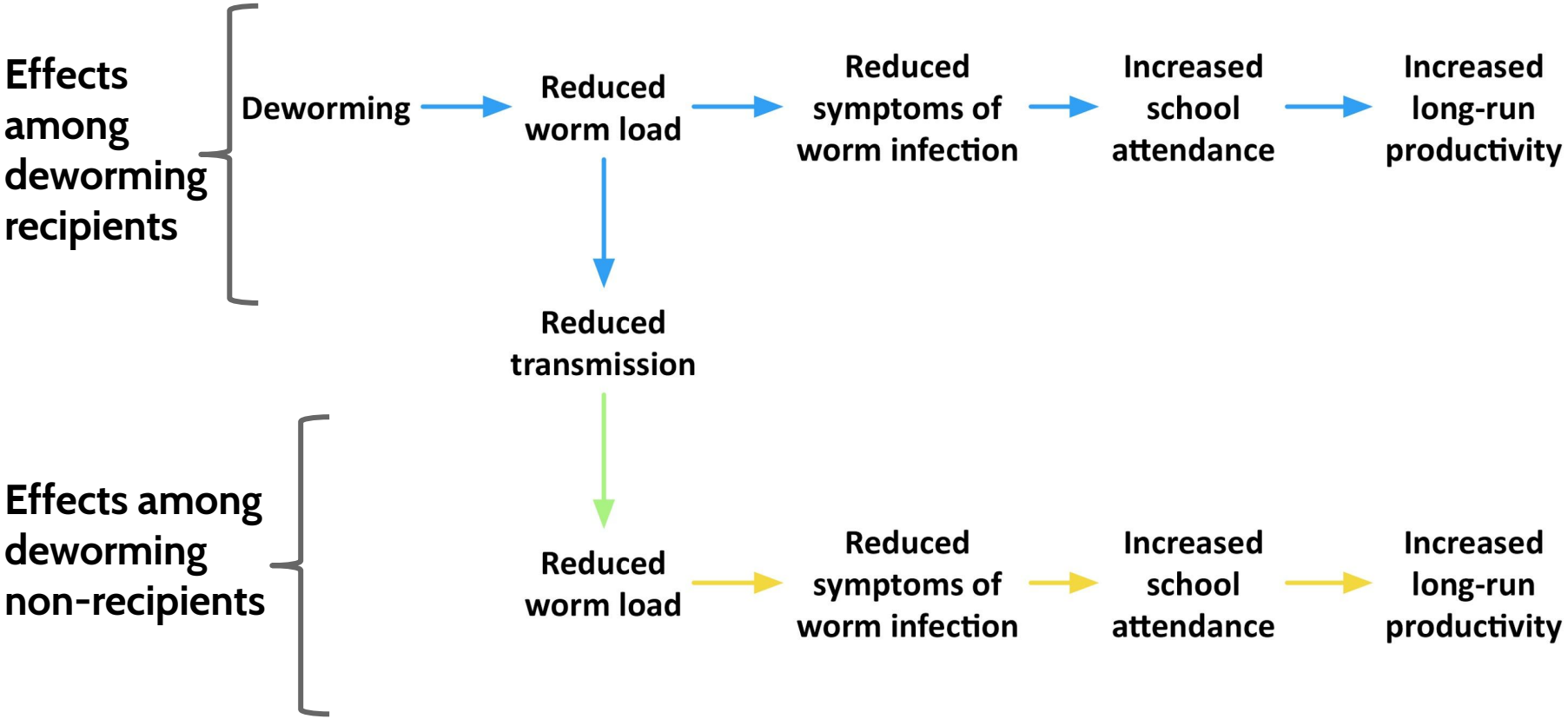
Econometrica, Vol. 72, No. 1 (January, 2004), 159–217

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.


Effects among deworming recipients
Deworming
Reduced worm load
Reduced symptoms of worm infection
Increased school attendance
Increased long-run productivity
Reduced transmission
Effects among deworming non-recipients
Reduced worm load
Reduced symptoms of worm infection
Increased school attendance
Increased long-run productivity

Intervention

- School-based deworming
- Education about worm infections

Outcomes

- Worm infections
- Child growth
- Anemia
- School attendance
- School performance

What kind of Randomization?

Question 1 on your handout.

In January 1998, the seventy-five PSDP schools were randomly divided into three groups of twenty-five schools each: the schools were first stratified by administrative subunit (zone) and by their involvement in other nongovernmental assistance programs, and were then listed alphabetically and every third school was assigned to a given project group.[9] Due to ICS's administrative

What kind of Randomization?

In January 1998, the seventy-five PSDP schools were randomly divided into three groups of twenty-five schools each: the schools were first stratified by administrative subunit (zone) and by their involvement in other nongovernmental assistance programs, and were then listed alphabetically and every third school was assigned to a given project group.[9] Due to ICS's administrative

Quasi-randomization

Question 2 on your handout.

school was assigned to a given project group.[9] Due to ICS's administrative and financial constraints, the health intervention was phased in over several years. Group 1 schools received free deworming treatment in both 1998 and 1999, Group 2 schools in 1999, while Group 3 schools began receiving treatment in 2001. Thus in 1998, Group 1 schools were treatment schools, while Group 2 and Group 3 schools were comparison schools, and in 1999, Group 1 and Group 2 schools were treatment schools and Group 3 schools were comparison schools.

What kind of Randomization?

In January 1998, the seventy-five PSDP schools were randomly divided into three groups of twenty-five schools each: the schools were first stratified by administrative subunit (zone) and by their involvement in other nongovernmental assistance programs, and were then listed alphabetically and every third school was assigned to a given project group.[9] Due to ICS's administrative

Quasi-randomization

Stepped wedge

school was assigned to a given project group.[9] Due to ICS's administrative and financial constraints, the health intervention was phased in over several years. Group 1 schools received free deworming treatment in both 1998 and 1999, Group 2 schools in 1999, while Group 3 schools began receiving treatment in 2001. Thus in 1998, Group 1 schools were treatment schools, while Group 2 and Group 3 schools were comparison schools, and in 1999, Group 1 and Group 2 schools were treatment schools and Group 3 schools were comparison schools.

Stepped wedge randomized trial

Schools	Year 1 (1998)	Year 2 (1999)
Group 1 (n=25)	**Intervention**	**Intervention**
Group 2 (n=25)	**Control**	**Intervention**
Group 3 (n=25)	**Control**	**Control**

Aiken et al., 2015

Was (quasi)-randomization effective?

Question 3 on your handout.

TABLE I

1998 AVERAGE PUPIL AND SCHOOL CHARACTERISTICS, PRE-TREATMENT[a]

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)	Group 1 – Group 3	Group 2 – Group 3
Panel A: Pre-school to Grade 8					
Male	0.53	0.51	0.52	0.01	−0.01
				(0.02)	(0.02)
Proportion girls <13 years, and all boys	0.89	0.89	0.88	0.00	0.01
				(0.01)	(0.01)
Grade progression (= Grade − (Age − 6))	−2.1	−1.9	−2.1	−0.0	0.1
				(0.1)	(0.1)
Year of birth	1986.2	1986.5	1985.8	0.4**	0.8***
				(0.2)	(0.2)
Panel B: Grades 3 to 8					
Attendance recorded in school registers (during the four weeks prior to the pupil survey)	0.973	0.963	0.969	0.003	−0.006
				(0.004)	(0.004)
Access to latrine at home	0.82	0.81	0.82	0.00	−0.01
				(0.03)	(0.03)
Have livestock (cows, goats, pigs, sheep) at home	0.66	0.67	0.66	−0.00	0.01
				(0.03)	(0.03)
Weight-for-age Z-score (low scores denote undernutrition)	−1.39	−1.40	−1.44	0.05	0.04
				(0.05)	(0.05)
Blood in stool (self-reported)	0.26	0.22	0.19	0.07**	0.03
				(0.03)	(0.03)
Sick often (self-reported)	0.10	0.10	0.08	0.02**	0.02**
				(0.01)	(0.01)
Malaria/fever in past week (self-reported)	0.37	0.38	0.40	−0.03	−0.02
				(0.03)	(0.03)
Clean (observed by field workers)	0.60	0.66	0.67	−0.07**	−0.01
				(0.03)	(0.03)
Panel C: School characteristics					
District exam score 1996, grades 5–8[b]	−0.10	0.09	0.01	−0.11	0.08
				(0.12)	(0.12)
Distance to Lake Victoria	10.0	9.9	9.5	0.6	0.5
				(1.9)	(1.9)
Pupil population	392.7	403.8	375.9	16.8	27.9

Was (quasi)-randomization effective?

TABLE I

1998 AVERAGE PUPIL AND SCHOOL CHARACTERISTICS, PRE-TREATMENT[a]

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)	Group 1 – Group 3	Group 2 – Group 3
Panel A: Pre-school to Grade 8					
Male	0.53	0.51	0.52	0.01	−0.01
				(0.02)	(0.02)
Proportion girls <13 years, and all boys	0.89	0.89	0.88	0.00	0.01
				(0.01)	(0.01)
Grade progression (= Grade − (Age − 6))	−2.1	−1.9	−2.1	−0.0	0.1
				(0.1)	(0.1)
Year of birth	1986.2	1986.5	1985.8	0.4**	0.8***
				(0.2)	(0.2)
Panel B: Grades 3 to 8					
Attendance recorded in school registers (during the four weeks prior to the pupil survey)	0.973	0.963	0.969	0.003	−0.006
				(0.004)	(0.004)
Access to latrine at home	0.82	0.81	0.82	0.00	−0.01
				(0.03)	(0.03)
Have livestock (cows, goats, pigs, sheep) at home	0.66	0.67	0.66	−0.00	0.01
				(0.03)	(0.03)
Weight-for-age Z-score (low scores denote undernutrition)	−1.39	−1.40	−1.44	0.05	0.04
				(0.05)	(0.05)
Blood in stool (self-reported)	0.26	0.22	0.19	0.07**	0.03
				(0.03)	(0.03)
Sick often (self-reported)	0.10	0.10	0.08	0.02**	0.02**
				(0.01)	(0.01)
Malaria/fever in past week (self-reported)	0.37	0.38	0.40	−0.03	−0.02
				(0.03)	(0.03)
Clean (observed by field workers)	0.60	0.66	0.67	−0.07**	−0.01
				(0.03)	(0.03)
Panel C: School characteristics					
District exam score 1996, grades 5–8[b]	−0.10	0.09	0.01	−0.11	0.08
				(0.12)	(0.12)
Distance to Lake Victoria	10.0	9.9	9.5	0.6	0.5
				(1.9)	(1.9)
Pupil population	392.7	403.8	375.9	16.8	27.9

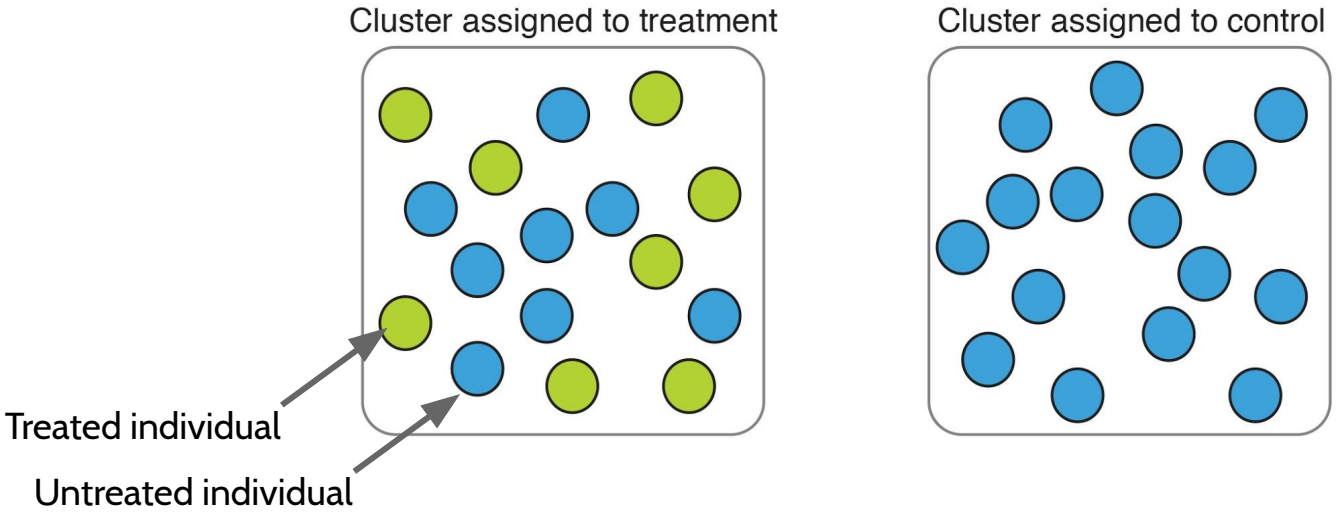
"Group 1 pupils appear to be worse off than Group 2 and 3 pupils along some dimensions, potentially creating a bias against finding significant program effects"

Miguel & Kremer, 2004

Was attrition balanced?

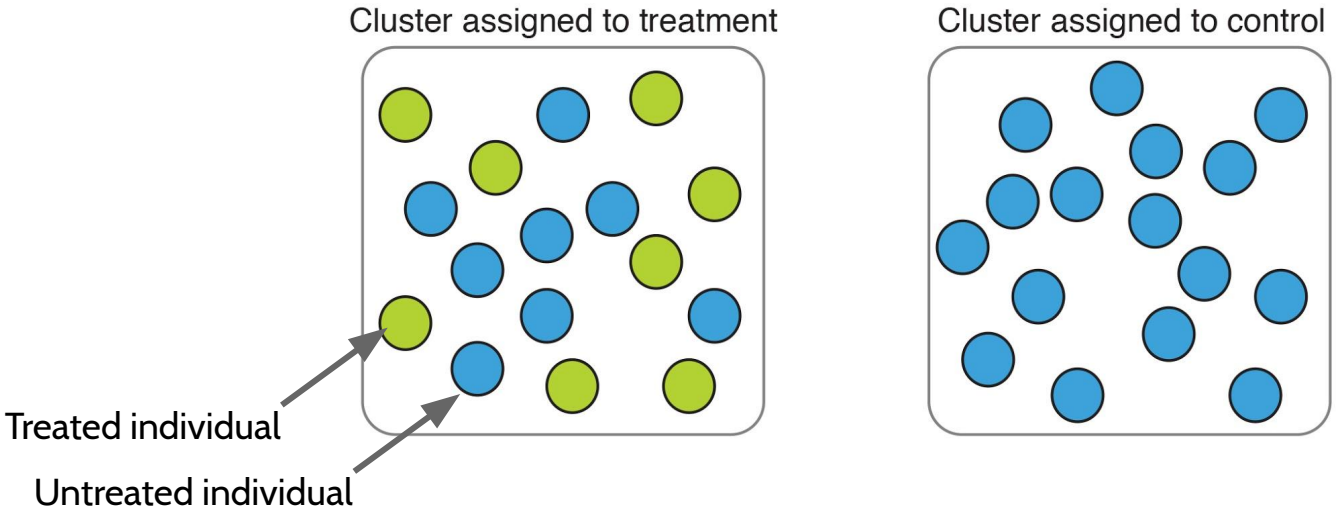
The results could potentially have been affected by differential attrition across treatment and comparison schools, if the additional treatment school pupils who participated in the exam after deworming were below-average performers. The fact that 85 percent of Group 1 pupils took the 1998 ICS exams, compared to 83 percent of Group 2 and Group 3 pupils, suggests that this is a possibility, although the attrition bias is likely to be small.[51] To address

How is a traditional cluster RCT analyzed?



Benjamin-Chung et al., *International Journal of Epidemiology* 2018 doi: 10.1093/ije/dyx201.

How is a traditional cluster RCT analyzed?



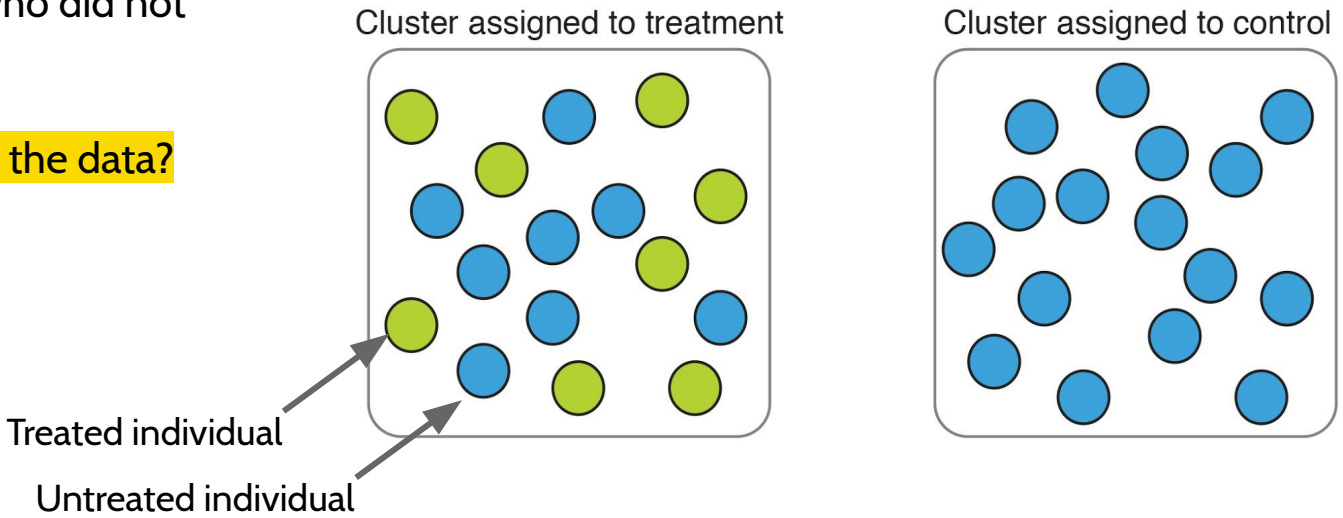
ITT effect = Average in treatment cluster - average in control cluster

Benjamin-Chung et al., *International Journal of Epidemiology* 2018 doi: 10.1093/ije/dyx201.

What if there are indirect effects?

Hypothesis: the intervention benefited children who did not receive it

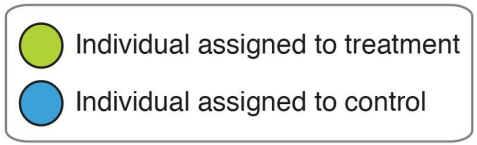
How do you analyze the data?

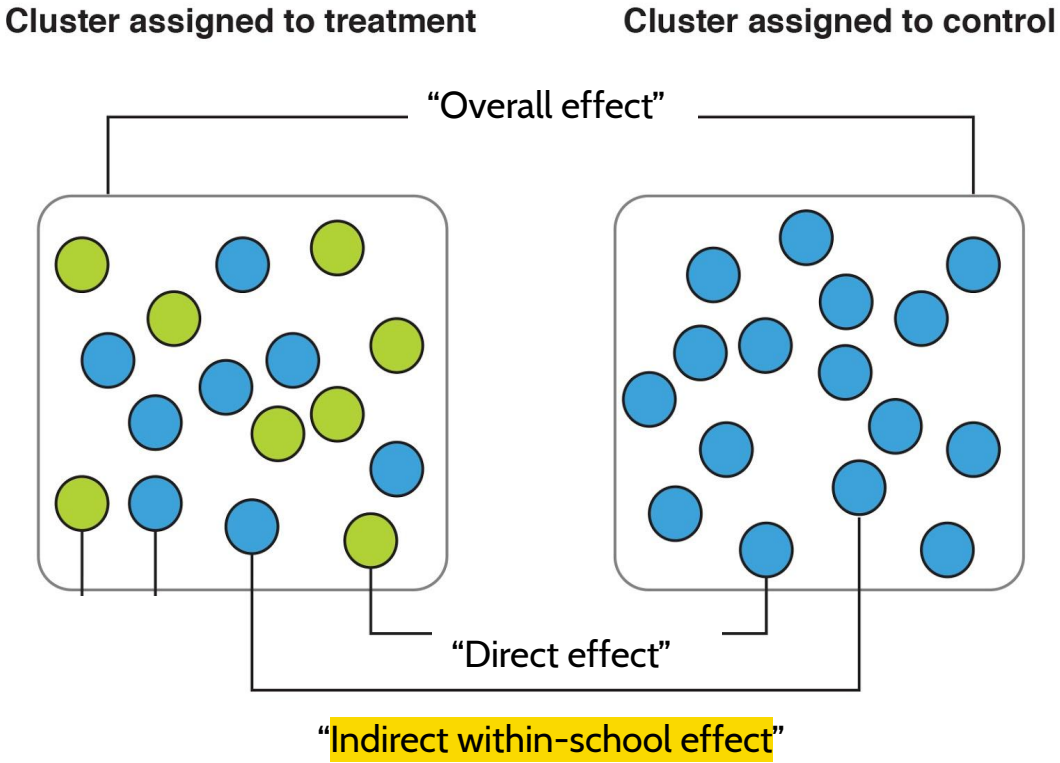


Benjamin-Chung et al., *International Journal of Epidemiology* 2018 doi: 10.1093/ije/dyx201.

Within-school indirect effects

Hypothesis: the intervention benefited children in treatment schools who did not receive it



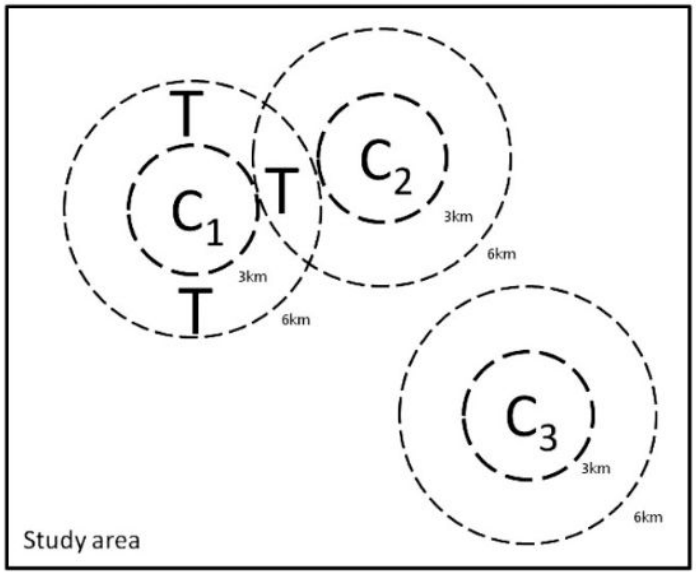


Between-school indirect effects

We first estimate program impacts in treatment schools, as well as cross-school treatment externalities:[24]

$$(1) \qquad Y_{ijt} = a + \beta_1 \cdot T_{1it} + \beta_2 \cdot T_{2it} + X'_{ijt}\delta + \sum_d (\gamma_d \cdot N^T_{dit}) + \sum_d (\phi_d \cdot N_{dit})$$

Hypothesis: the intervention benefited children in control schools who did not receive it



"Indirect between-school effect"

T = treatment school

C_n = control school

Exposure to treatment schools is greatest in school C_1 and least in school C_3

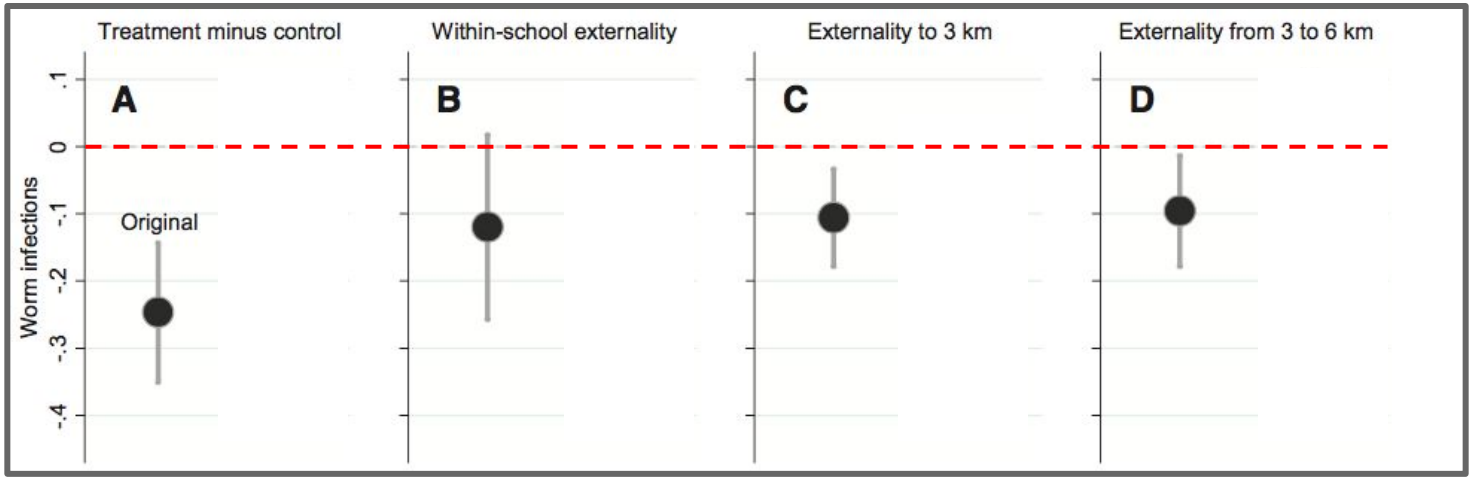
When externalities are present, what is the total impact of the program?

~~Total effect~~
"Naive effect" = Prevalence of worm infections in **treated schools** (regardless of individual treatment) − Prevalence of worm infections in **control schools** (regardless of individual treatment) ?

Total effect = **Direct effect** (on participants) + Within-school **indirect effect** (on non-participants) + Between-school **indirect effect** (on non-participants)

Note 1: This was actually a weighted average, not a simple sum
Note 2: This terminology is not standard

Worm infections - original results

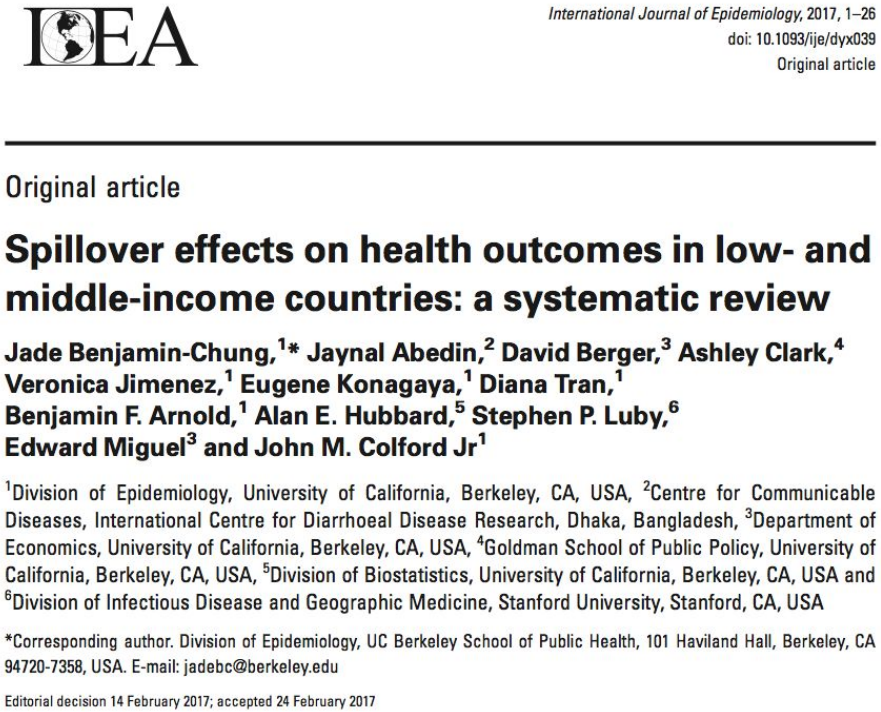


Hicks et al. 2015

"This trial ... helped create an **entire movement** of people doing proper randomized trials ... Before that, the field was in a kind of **dark ages**, blown on the winds of expert opinion and whim."

Goldacre, 2015

The study catalyzed evaluations of spillovers as well



International Journal of Epidemiology, 2017, 1–26
doi: 10.1093/ije/dyx039
Original article

Original article

Spillover effects on health outcomes in low- and middle-income countries: a systematic review

Jade Benjamin-Chung,[1]* Jaynal Abedin,[2] David Berger,[3] Ashley Clark,[4] Veronica Jimenez,[1] Eugene Konagaya,[1] Diana Tran,[1] Benjamin F. Arnold,[1] Alan E. Hubbard,[5] Stephen P. Luby,[6] Edward Miguel[3] and John M. Colford Jr[1]

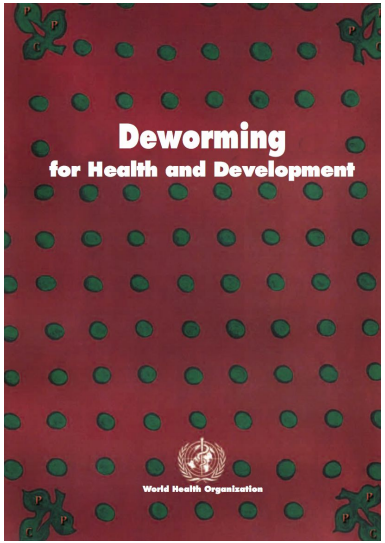
[1]Division of Epidemiology, University of California, Berkeley, CA, USA, [2]Centre for Communicable Diseases, International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh, [3]Department of Economics, University of California, Berkeley, CA, USA, [4]Goldman School of Public Policy, University of California, Berkeley, CA, USA, [5]Division of Biostatistics, University of California, Berkeley, CA, USA and [6]Division of Infectious Disease and Geographic Medicine, Stanford University, Stanford, CA, USA

*Corresponding author. Division of Epidemiology, UC Berkeley School of Public Health, 101 Haviland Hall, Berkeley, CA 94720-7358, USA. E-mail: jadebc@berkeley.edu

Editorial decision 14 February 2017; accepted 24 February 2017

16 out of 54 papers in the review were identified because they cited the Miguel & Kremer paper

Deworming supported by WHO and rapidly expanded through NGOs



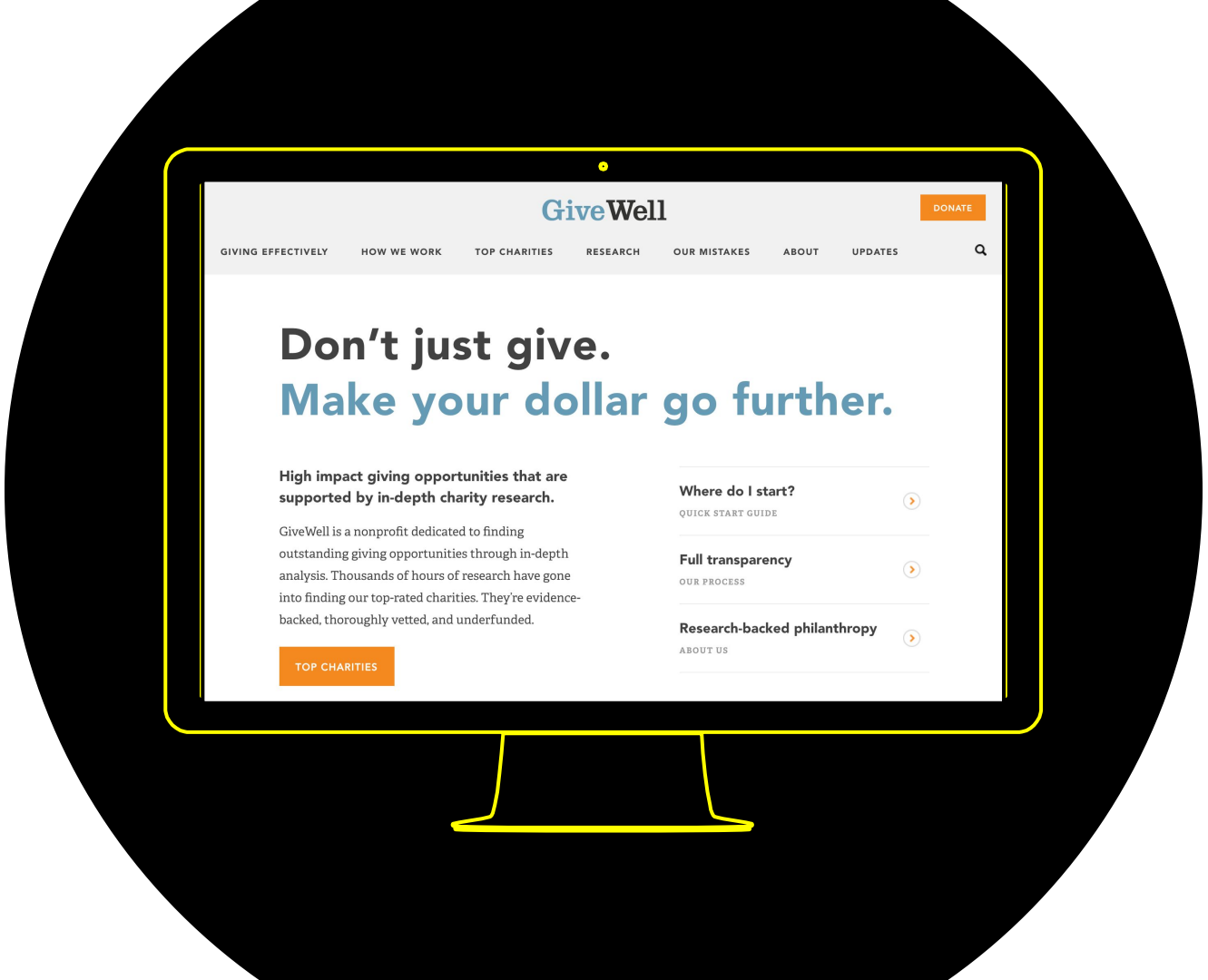
Deworming helps children **"earn their way out of poverty"** through large impacts on cognition and intellectual development.

- World Health Organization

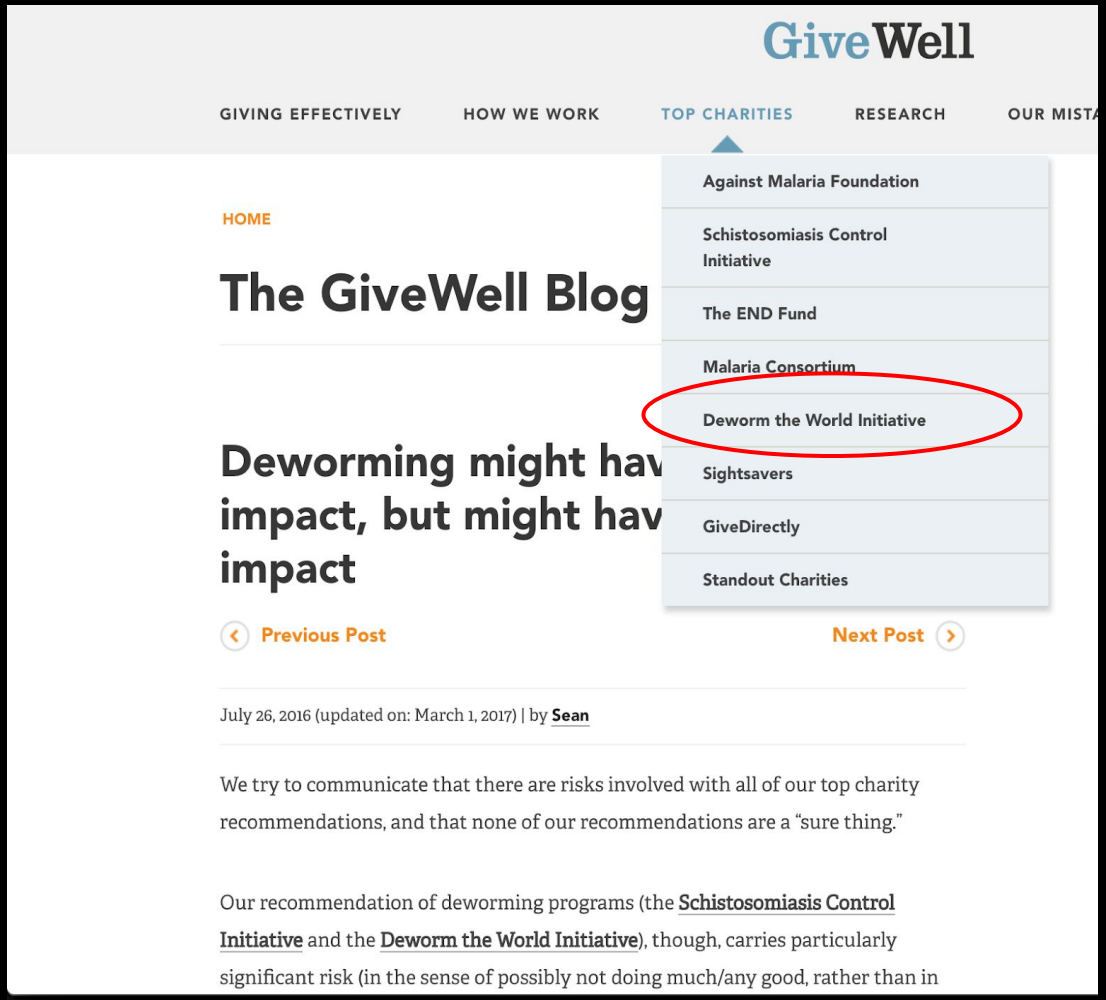
deworm the world

Partnership for Treating and Preventing Intestinal Worms
ChildrenWithoutWorms

WHO 2012



"Unlike charity evaluators that focus solely on financials, assessing administrative or fundraising costs, we conduct in-depth research aiming to determine how much good a given program accomplishes (in terms of lives saved, lives improved, etc.) per dollar spent."

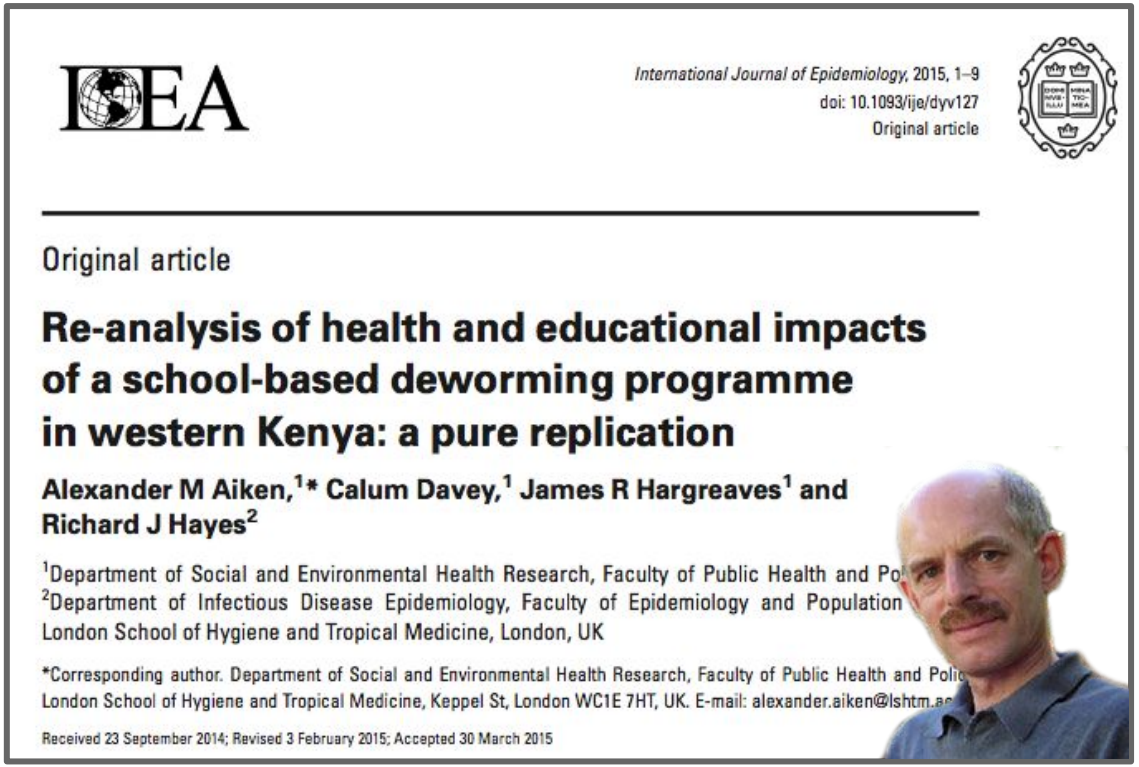


"In our view, the most compelling case for deworming as a cost-effective intervention comes not from its subtle impacts on general health (which appear relatively minor and uncertain) nor from its potential reduction in severe symptoms of disease effects (which we believe to be rare), but from the possibility that deworming children has a subtle, lasting impact on their development, and thus on their ability to be productive and successful throughout life."

Chapter 3

The Worm Wars

Replication Part 1: “Pure replication”



IEA

International Journal of Epidemiology, 2015, 1–9
doi: 10.1093/ije/dyv127
Original article

Original article

Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication

Alexander M Aiken,[1]* Calum Davey,[1] James R Hargreaves[1] and Richard J Hayes[2]

[1]Department of Social and Environmental Health Research, Faculty of Public Health and Po
[2]Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population
London School of Hygiene and Tropical Medicine, London, UK

*Corresponding author. Department of Social and Environmental Health Research, Faculty of Public Health and Poli
London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, UK. E-mail: alexander.aiken@lshtm.a

Received 23 September 2014; Revised 3 February 2015; Accepted 30 March 2015

Objective: Fully replicate the results of the original study using the same raw datasets and information in the original paper

Key errors detected:

- Unclear labeling
- Rounding errors
- Inaccurately reported denominators
- Mislabelled levels of statistical significance
- Coding errors in STATA do files

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)
```

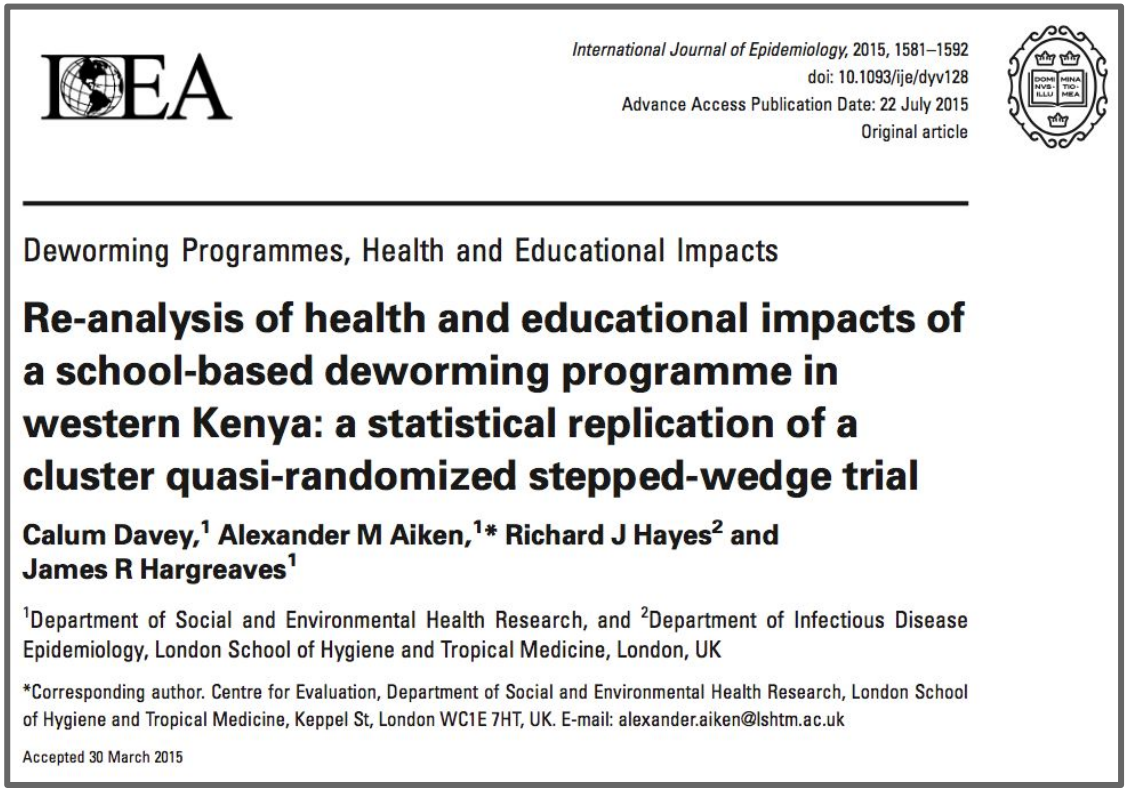
This code was problematic, as it erroneously limited the number of schools that could be included in this matrix calculation to 12, rather than allowing up to 75 as intended.

A second coding error was present that miscalculated local density figures for three of the schools: school number 108 (in Group 1), 109 (in Group 2) and 115 (in Group 3). The code was as follows:

```
if (wgrp[`x', 1] == `i') {;
        matrix S_TEMP1[`i',`j'] = -1;
        matrix S_TEMP2[`i',`j'] = -1;
        matrix S_TEMP3[`i',`j'] = -1;
                }
```

This code was problematic as it erroneously assigned these three schools into a '-1' category, where it ignored their populations when calculating the local densities.

Replication Part 2: Re-analysis



International Journal of Epidemiology, 2015, 1581–1592
doi: 10.1093/ije/dyv128
Advance Access Publication Date: 22 July 2015
Original article

Deworming Programmes, Health and Educational Impacts

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

Calum Davey,[1] Alexander M Aiken,[1]* Richard J Hayes[2] and James R Hargreaves[1]

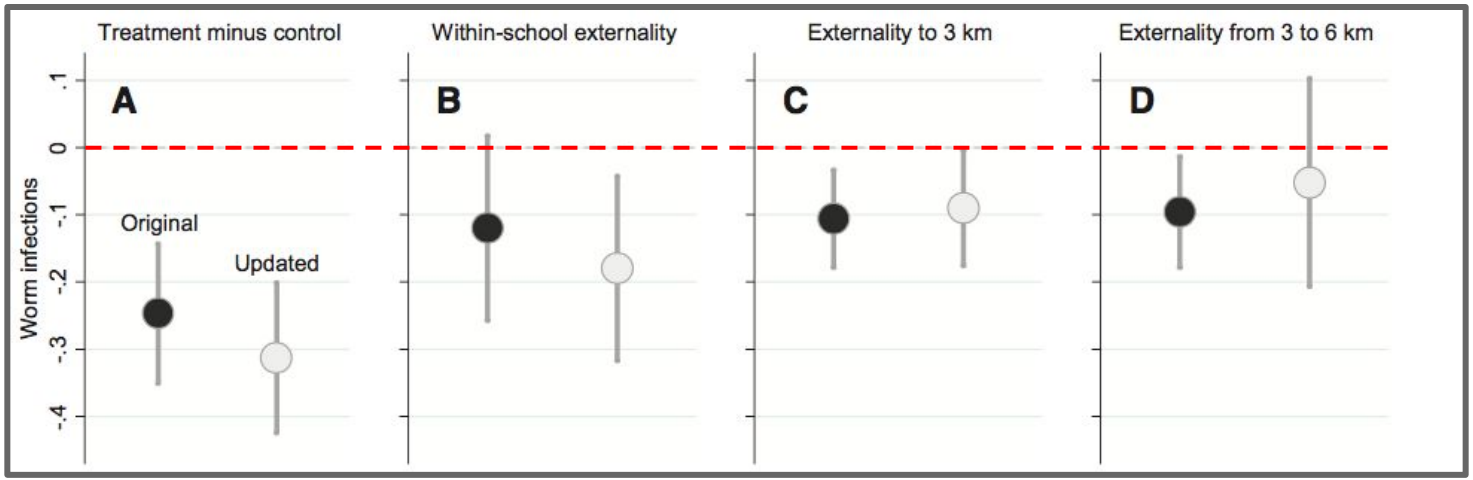
[1]Department of Social and Environmental Health Research, and [2]Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK

*Corresponding author. Centre for Evaluation, Department of Social and Environmental Health Research, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, UK. E-mail: alexander.aiken@lshtm.ac.uk

Accepted 30 March 2015

Objective: To analyze the original data using alternative, pre-specified, methods for data handling and analysis, in line with modern epidemiological approaches

Worm infections – "pure" replication results



Hicks et al. 2015

BuzzFeed NEWS / REPORTING TO YOU BuzzFeed Videos Quizzes Tasty More

Scientists Are Hoarding Data And It's Ruining Medical Research

Major flaws in two massive trials of deworming pills show the importance of sharing data — which most scientists don't do.

Posted on July 22, 2015, at 4:18 p.m.

Ben Goldacre
BuzzFeed Contributor

More

A girl cries while receiving a deworming pill in Managua, Nicaragua.
Hector Retamal / AFP / Getty Images / Via gettyimages.com

David Neal
@DavidPNeal1

Follow

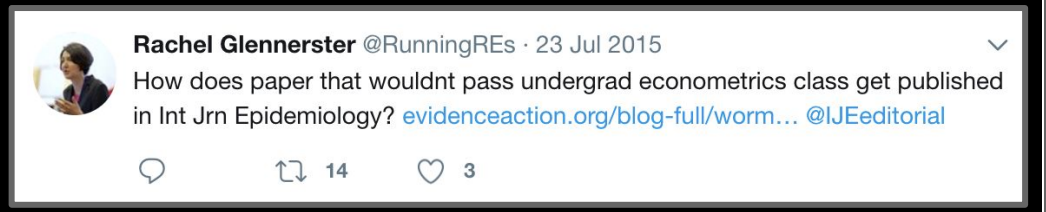
@bengoldacre "I've no position on the Worm Wars." Then why is this article headed by a picture associating deworming with a child crying?

3:21 AM - Jul 23, 2015

2 1

Goldacre, 2015

"DEWORMING WORKS" VS. "DEWORMING DEBUNKED"



https://storify.com/viewfromthecave/tracking-the-wormwars

THE WORLD BANK
IBRD • IDA
Working for a World Free of Poverty

ABOUT DATA RESEARCH LEARNING NEWS PROJECTS & OPERATIONS PUBLICATIONS

Development Impact

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

Bloggers Tags Contact

Most Popular 3 Months

- A new answer to why developing country firms are so small, and how cellphones solve this problem
- Trouble with pre-analysis plans? Try these three weird tricks.
- What a new preschool study tells us about early child education – and about impact evaluation
- How can education systems be better? A round-up of the 2017 RISE conference
- Should we require balance t-tests of baseline observables in randomized experiments?

Latest Bloggers

David McKenzie
Lead Economist, Development Research Group

Worm Wars: A… Kremer's Dewo…

SUBMITTED BY

Share

This post was upda… site from Twitter up… authors' response i… review the papers i… valid as 6 months a… without much chan… have a few new tho… Miguel), which I dis… thank Stefane Helleringer for a nice response he wrote about the definion of ITT in public health: see the back and forth here.

Despite the differences in various methodological and data handling choices, which I discussed below in my original post, it is clear that the interpretation of whether one believes the results of Miguel and Kremer are robust really rests on whether one splits the data or not. Therefore it is important to solely focus on this point and think about which choice is more justified and whether the issue can be dealt with another way. A good starting point is the explanation of DAHH in their pre-analysis plan as to why they decided to split the data into years and analyze it cross-sectionally rather than the difference-in-difference method in the original MK (2004).

The data from a stepped wedge trial can be thought of as a one-way cross-over, and

"Based on what I have seen in the reanalysis study … and the response by [Miguel & Kremer], … I find the findings of the original study more robust than I did before."

Berk Ozler, Economist at the World Bank

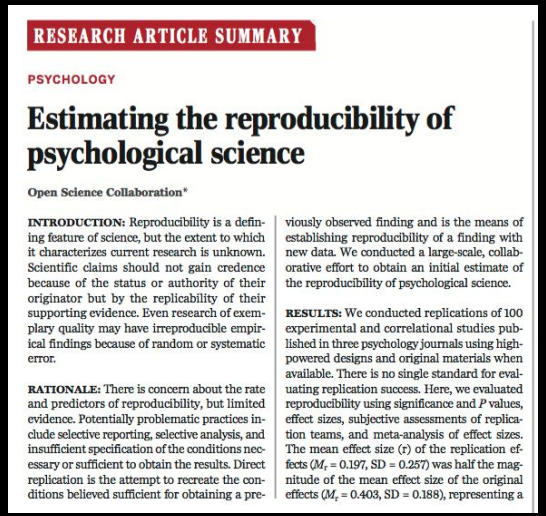
http://blogs.worldbank.org/impactevaluations/warm-wars-review-reanalysis-miguel-and-kremer-s-deworming-study

“They show that you can eliminate the impact of deworming on school attendance — if you torture the data. ...What if the replication, itself, is widely seen as flawed? … This is why an independent peer review process is so important. The editor-in-chief of the IJE is a colleague in the same department as the replication authors. While there is no evidence that the Journal loosened its standards for this paper, it does make one wonder. Frankly, for an issue as sensitive as children’s deworming, the optics would be better had the authors chosen a more independent venue for publishing their work.”

Paul Gertler, Professor of Economics at UC Berkeley

http://blogs.berkeley.edu/2015/08/03/good-science-gone-wrong/

The **WORM WARS** are part of a much bigger phenomenon: the "Reproducibility crisis"



RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

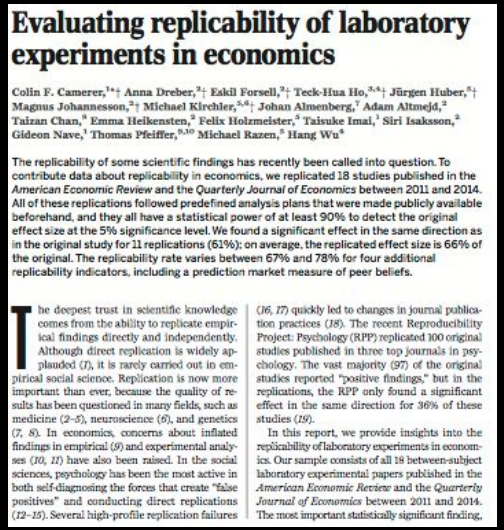
Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a

Psychology



Evaluating replicability of laboratory experiments in economics

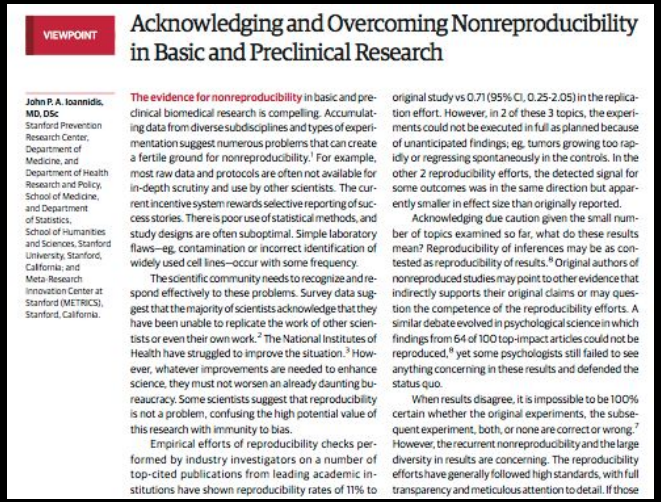
Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10]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

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

The deepest trust in scientific knowledge comes from the ability to replicate empirical findings directly and independently. Although direct replication is widely applauded (*1*), it is rarely carried out in empirical social science. Replication is now more important than ever, because the quality of results has been questioned in many fields, such as medicine (*2*–*5*), neuroscience (*6*), and genetics (*7*, *8*). In economics, concerns about inflated findings in empirical (*9*) and experimental analyses (*10*, *11*) have also been raised. In the social sciences, psychology has been the most active in both self-diagnosing the forces that create "false positives" and conducting direct replications (*12*–*15*). Several high-profile replication failures (*16*, *17*) quickly led to changes in journal publication practices (*18*). The recent Reproducibility Project: Psychology (RPP) replicated 100 original studies published in three top journals in psychology. The vast majority (97) of the original studies reported "positive findings," but in the replications, the RPP only found a significant effect in the same direction for 36% of these studies (*19*).

In this report, we provide insights into the replicability of laboratory experiments in economics. Our sample consists of all 18 between-subject laboratory experimental papers published in the *American Economic Review* and the *Quarterly Journal of Economics* between 2011 and 2014. The most important statistically significant finding,

Economics



VIEWPOINT

Acknowledging and Overcoming Nonreproducibility in Basic and Preclinical Research

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center, Department of Medicine, and Department of Health Research and Policy, School of Medicine, and Department of Statistics, School of Humanities and Sciences, Stanford University, Stanford, California; and Meta-Research Innovation Center at Stanford (METRICS), Stanford, California.

The evidence for nonreproducibility in basic and preclinical biomedical research is compelling. Accumulating data from diverse subdisciplines and types of experimentation suggest numerous problems that can create a fertile ground for nonreproducibility.[1] For example, most raw data and protocols are often not available for in-depth scrutiny and use by other scientists. The current incentive system rewards selective reporting of success stories. There is poor use of statistical methods, and study designs are often suboptimal. Simple laboratory flaws—eg, contamination or incorrect identification of widely used cell lines—occur with some frequency.

The scientific community needs to recognize and respond effectively to these problems. Survey data suggest that the majority of scientists acknowledge that they have been unable to replicate the work of other scientists or even their own work.[2] The National Institutes of Health have struggled to improve the situation.[3] However, whatever improvements are needed to enhance science, they must not worsen an already daunting bureaucracy. Some scientists suggest that reproducibility is not a problem, confusing the high potential value of this research with immunity to bias.

Empirical efforts of reproducibility checks performed by industry investigators on a number of top-cited publications from leading academic institutions have shown reproducibility rates of 11% to original study vs 0.71 (95% CI, 0.25-2.05) in the replication effort. However, in 2 of these 3 topics, the experiments could not be executed in full as planned because of unanticipated findings; eg, tumors growing too rapidly or regressing spontaneously in the controls. In the other 2 reproducibility efforts, the detected signal for some outcomes was in the same direction but apparently smaller in effect size than originally reported.

Acknowledging due caution given the small number of topics examined so far, what do these results mean? Reproducibility of inferences may be as contested as reproducibility of results.[4] Original authors of nonreproduced studies may point to other evidence that indirectly supports their original claims or may question the competence of the reproducibility efforts. A similar debate evolved in psychological science in which findings from 64 of 100 top-impact articles could not be reproduced,[5] yet some psychologists still failed to see anything concerning in these results and defended the status quo.

When results disagree, it is impossible to be 100% certain whether the original experiments, the subsequent experiment, both, or none are correct or wrong.[7] However, the recurrent nonreproducibility and the large diversity in results are concerning. The reproducibility efforts have generally followed high standards, with full transparency and meticulous attention to detail. If those

Basic science

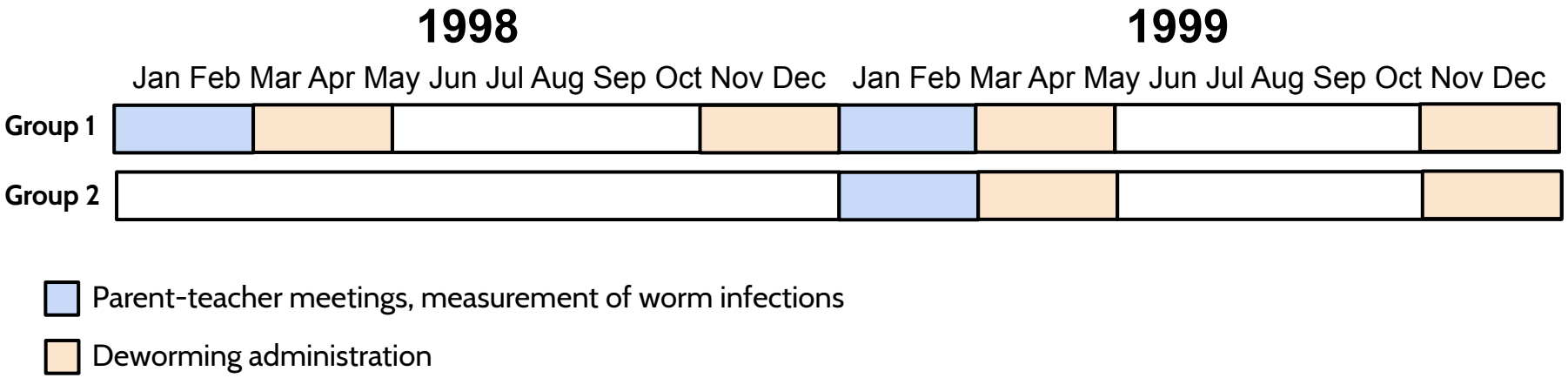
“We found beneficial effects on worm infections similar to or greater than those originally reported, although the **indirect between-school effect was more modest than previously described** and was not statistically significant.” Aiken et al. 2015

"The updated evidence on within-school externalities implies that a key conclusion in Miguel and Kremer—that individually randomized studies underestimate true deworming impacts—remains valid." Hicks, Miguel, & Kremer 2015

Controversy #1: When did treatment officially begin?

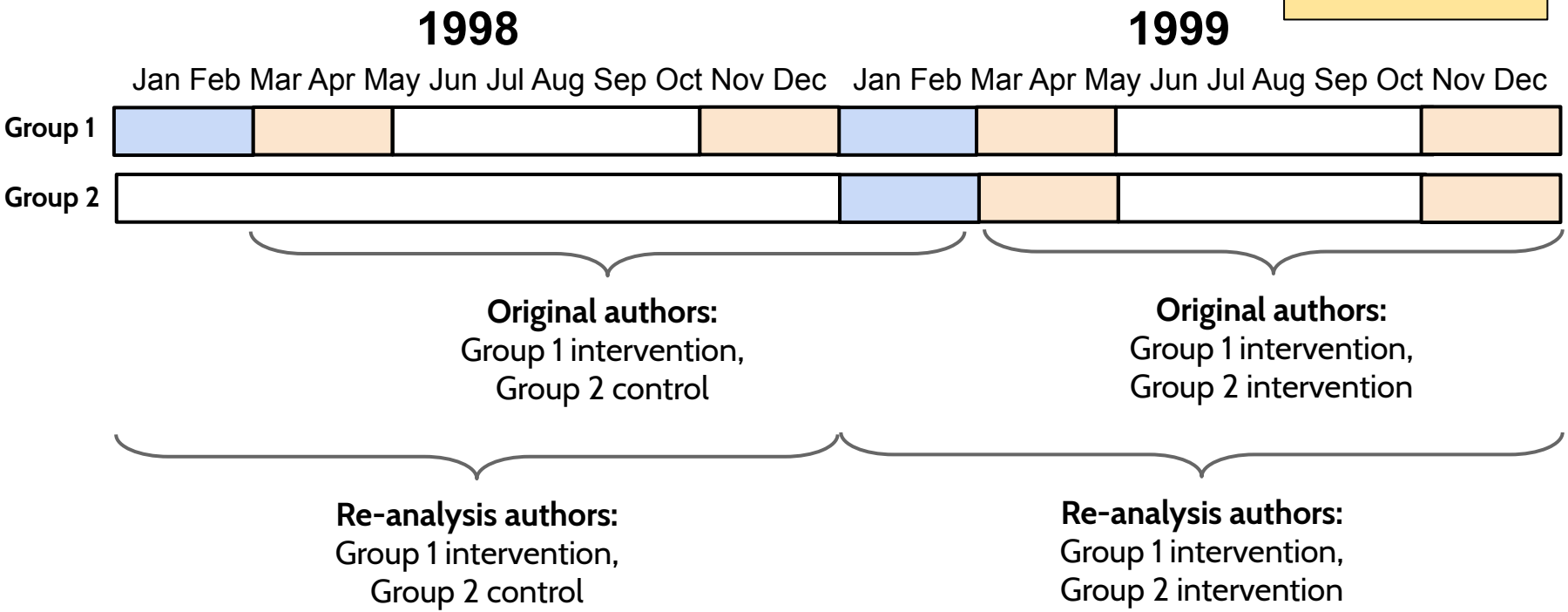
Question 4 on your handout.

For the intention-to-treat analysis, what is the appropriate way to define the intervention with respect to time? Read the excerpts from the original paper and decide for yourself.



Controversy #1: When did treatment officially begin?

Question 4 on your handout.



Original author arguments

"... they misclassify pre-treatment control observations as treatment observations. Group 2 schools began receiving deworming in March 1999. The correct coding of treatment for Group 2 thus begins after March 1999... however, **Davey et al. misclassify the Group 2 observations from early 1999 as treatment observations**. They purport to justify the misclassification of 20% of 1999 observations using an 'intention-to-treat' framework, a framework typically utilized when a population was assigned to treatment, but only some individuals actually received it. Davey et al. **incorrectly apply it to a different situation, in which no individuals were actually treated** (i.e. Group 2 prior to March 1999) nor were any supposed to be treated."

Hicks et al. 2015

Replicator arguments

"[The original authors] describe as an 'error' our decision to include data on school attendance in each year before administration of deworming, saying this decision is unjustified and not in line with our published analysis plan. We disagree. Figure 1 in our analysis plan indicates our understanding that the cross-over point for schools from control to intervention was in line with calendar year. **From the outset, we understood the intervention package (comprising both health education and deworming medications) to have been delivered across full school years**, with effects on school attendance evaluated through assessments in that same school year."

Aiken et al. 2015

Controversy #1: When did treatment officially begin?

....How could it have been prevented?

Controversy #1: When did treatment officially begin?

....How could it have been prevented?

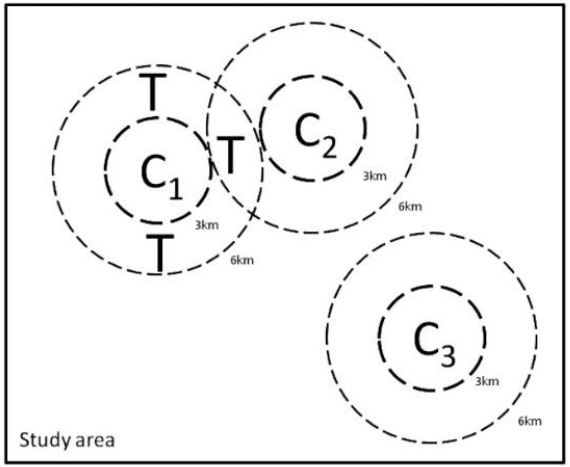
- **Pre-registration** of the study design that clearly defined intervention and control periods
- **Pre-analysis plan** to indicate how the intention-to-treat analysis would be conducted
- **Careful documentation** of any deviation from the original intervention time periods

Controversy #2: What is the right way to calculate the total effect using the updated data?

Question 5 on your handout.

- The original authors defined the total effect as a weighted average of the direct effect and spillover effects.
- The replicators identified a coding error. Once it was resolved, the 3-6km spillover effect was no longer significant because the standard error was much large.

Given this, how would you calculate the total effect?



T = treatment school

C_n = control school

Exposure to treatment schools is greatest in school C_1 and least in school C_3

Original approach (and replicators' approach):

Total effect = **Direct effect** (on participants) + Within-school **indirect effect** (on non-participants) + Between-school **indirect effect** (on non-participants)

Original authors' approach after replication:

Total effect = **Direct effect** (on participants) + Within-school **indirect effect** (on non-participants)

Note: This was actually a weighted average, not a simple sum

Table S1: Worm infection results from Miguel and Kremer (2004), updated and original

	UPDATED			ORIGINAL		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Indicator	-0.347***	-0.333***	-0.313***	-0.347***	-0.311***	-0.247***
	(0.052)	(0.052)	(0.057)	(0.052)	(0.052)	(0.053)
Treatment pupils w/in 3 km		-0.234**	-0.212**		-0.249***	-0.256***
(per 1000 pupils)		(0.097)	(0.104)		(0.085)	(0.087)
Treatment pupils w/in 3 - 6 km			-0.050			-0.140**
(per 1000 pupils)			(0.077)			(0.060)
Total PSDP 'eligible' students w/in 3 km		0.069*	0.046		0.074**	0.109***
(per 1000 pupils)		(0.037)	(0.036)		(0.033)	(0.040)
Total PSDP 'eligible' students w/in 3-6 km			-0.022			0.133**
(per 1000 pupils)			(0.039)			(0.056)
School average of mock score, 1996	-0.208***	-0.216***	-0.188***	-0.208***	-0.220***	-0.093
	(0.055)	(0.052)	(0.073)	(0.055)	(0.048)	(0.068)
Calculated Effects						
Average 0-3 km externality effect		-0.102**	-0.090**		-0.111***	-0.106***
		(0.043)	(0.044)		(0.038)	(0.037)
Average 3-6 km externality effect			-0.052			-0.096**
			(0.079)			(0.042)
Average overall cross-school externality effect		-0.102**	-0.146		-0.111***	-0.212***
		(0.043)	(0.110)		(0.038)	(0.065)
Overall deworming effect	-0.347***	-0.435***	-0.459***	-0.347***	-0.421***	-0.460***
	(0.057)	(0.061)	(0.091)	(0.057)	(0.055)	(0.055)

Note: The sample size in columns (1)-(3) is 2,330, and in (4)-(6) is 2,328. The sample includes pupils in grades 3–8, in 1999 Group 1 and Group 2 schools. Results are from probit estimation, where observations are weighted by total school population. The dependent variable is an indicator for moderate-to-heavy infection. Eligible pupils include girls less than 13 years old and all boys. Additional explanatory variables include indicators for 1998 grade and school SAP participation. Robust standard errors are in parentheses, and disturbance terms are clustered within schools. Stars denote statistical significance at 99 (***), 95 (**), and 90 (*) percent confidence.

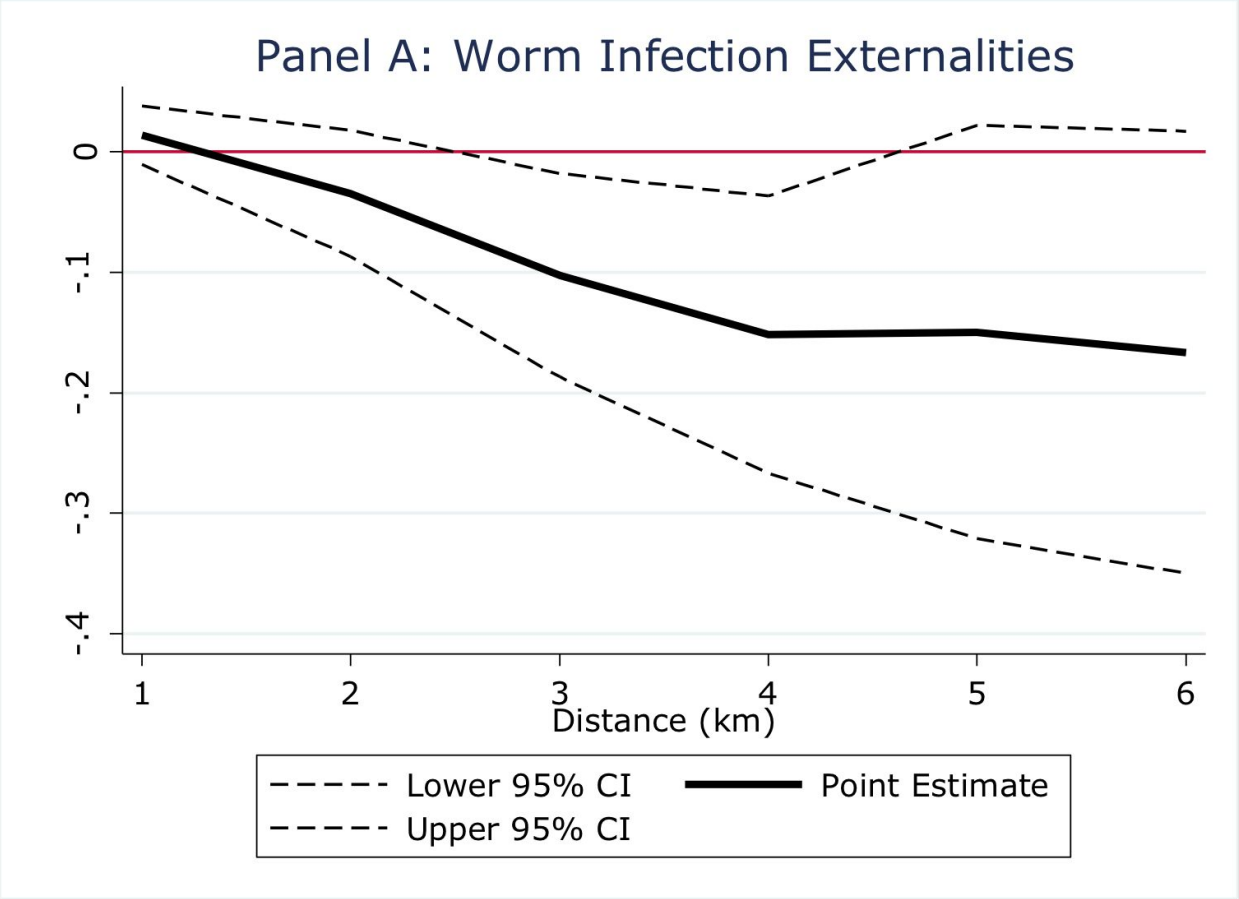
Hicks et al. 2015

Original author arguments

"Given the updated data, a regression specification different from that in the original paper is necessary to precisely estimate the overall externality effect of deworming. While it is natural to first replicate the exact specification used in the original paper, the changes to the data mean that this estimator is no longer appropriate. **More reliable conclusions can be reached by excluding the 3-6 km externality effect from the calculation of overall effects**, since it is adding a tremendous amount of "noise" to the estimate."

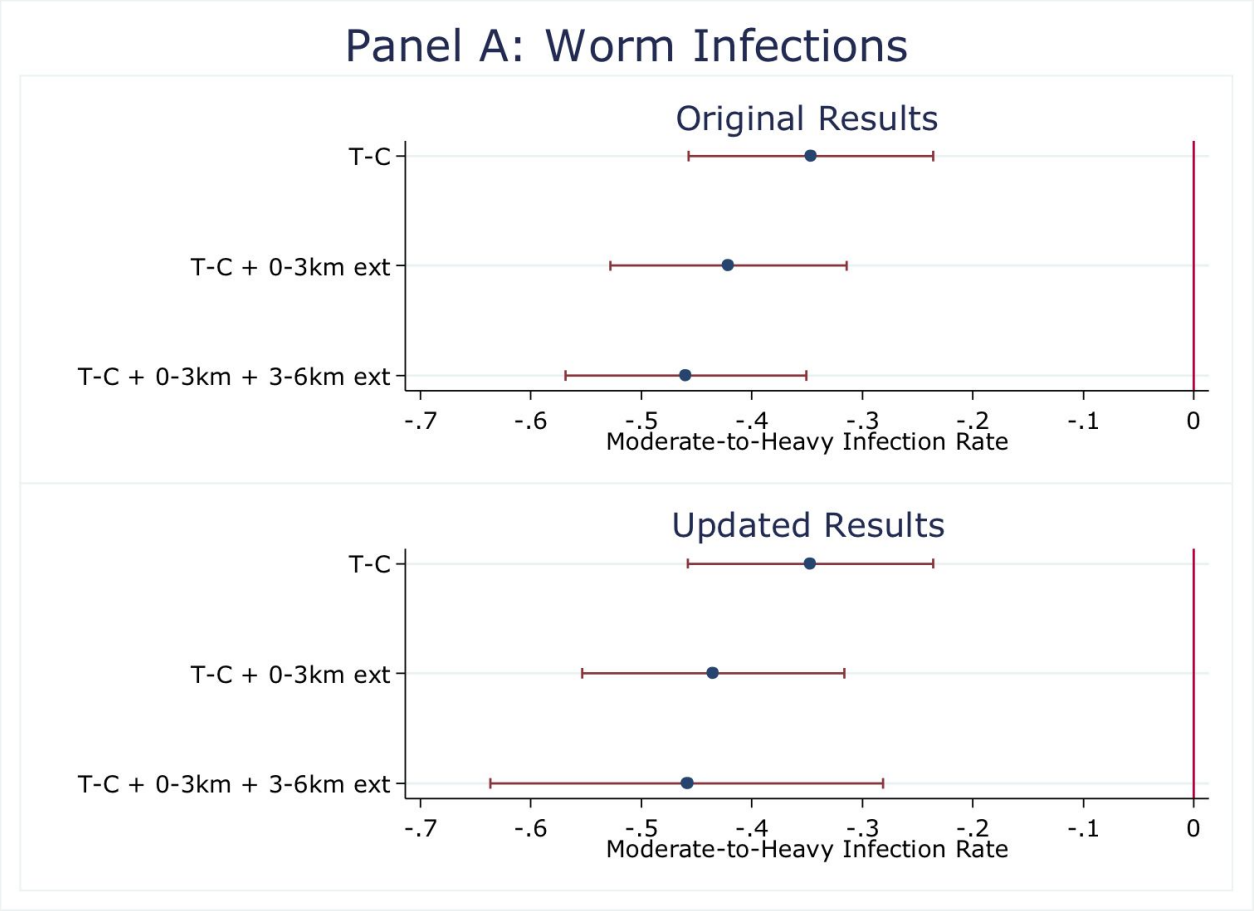
Hicks et al. 2015

Post hoc analysis by original authors



Hicks et al., 2015

Post hoc analysis by original authors



Hicks et al., 2015

Replicator arguments

“In this pure replication, **we did not evaluate the appropriateness of separating effects** into the different categories as described above. Instead, we reproduced the analytical steps to re- determine the results as originally calculated.”

Aiken et al. 2015

Replicator arguments

"We understand their original externality analyses were exploratory and **we value their innovative thinking** in this regard. We simply identified and corrected errors in the original tables, leading us to conclude that there is little evidence, under the authors' original analysis specification, for externalities at the distance (0–6 km) over which these were estimated in [the original paper]. [The original authors] do not dispute this conclusion, but proceed to report further analyses exploring externalities over a range of distances, and contend that we should have followed this approach. On this we must disagree, while hoping that appraisers of the evidence will consider their further analyses. Our view is simply that these were, at the time, innovative hypothesis-generating analyses worthy of further consideration. As with all exploratory analyses, **we caution against over-interpretation of effects seen at specific distances defined after examination of the data**."

Hargreaves et al., 2015

Controversy #2: What is the right way to calculate the total effect using the updated data?

....How could it have been prevented?

Controversy #2: What is the right way to calculate the total effect using the updated data?

....How could it have been prevented?

- Pre-analysis plan that defined the total effect *a priori*
- Internal replication in order to catch errors such as those caught by the external replication team
- Further *post hoc* analyses still permitted but would be clearly labeled as such

Chapter 4

The Moral of the Story

What can we learn from the Worm Wars?

- Different disciplines have different norms around study design, pre-specification, reporting, and analysis that can make it difficult to see eye to eye.

- A lot of pressure was put on the Miguel & Kremer study because policies were based on it alone — basing policies on a larger evidence base means less pressure on single studies (and investigators).

What can we learn from the Worm Wars?

For interventions that have large, long-term potential impacts, are we willing to make policies to scale them up with a weaker evidence base than interventions with smaller potential impacts?

The answer should be no.

But it's human nature to want to say yes.

How can we create incentives and infrastructure for scientists that build in this fact of human nature?

Tools we can use to prevent another “war”

- Internal replication
- External replication?
- Standardization of internal and external replication
- Study registration
- Pre-analysis plans

Berkeley Initiative for Transparency in the Social Sciences

Learn about these tools at BITTS workshops

Catalysts | Education | Leamer-Rosenthal Prizes | SSMART Grants | Research | Resources

Internal & external replication

What is it?

- **Internal replication:** scientists from the original study team independently replicate analyses prior to publication
- **External replication:** scientists not in the original study team independently replicate analyses following publication

Why do it?

- Reduces the chance of errors
- Reduces bias that may occur in the analysis, particularly when coupled with a pre-analysis plan

Study registration

What is it?

- Protocol that describes the study design in detail prior to data collection
- Publish in a peer-reviewed journal (e.g. BMJ Open) or register and time stamp on a public site (e.g. ClinicalTrials.gov, Open Science Framework)

Why do it?

- Clarifies the study design
- Informs other scientists of ongoing research
- Reduces publication bias
- Provides information to study participants
- Supports efficient allocation of research funding

Pre-analysis plans

What is it?

- Protocol that describes the data processing and statistical analysis of a study prior to data cleaning and analysis
- Publish in a peer-reviewed journal (e.g. BMJ Open) or register and/or time stamp on a public site (e.g. ClinicalTrials.gov, Open Science Framework)

Why do it?

- Minimize bias from practices such as data mining
- Facilitate replication
- Can better prepare researchers for data collection/analysis

“Cliff Notes”

Chapter 1: Intro to soil-transmitted helminths

- Soil-transmitted helminths are parasitic worms that live and reproduce in the intestines
- Symptoms and health impacts vary by type of worm
- School-based deworming programs aim to treat children, who have a higher prevalence of infection.

“Cliff Notes”

Chapter 2: An innovative study

- Miguel and Kremer conducted an innovative study that inspired scientists to use randomized trials.
- Their novel finding of between-school spillovers within 0-3 km held up following replication.

“Cliff Notes”

Chapter 3: The Worm Wars

- The replication identified important coding errors and other mistakes that changed some important findings, though many original findings also held up.
- The re-analysis sparked intense debate about the appropriate way to analyze data in such a trial.

“Cliff Notes”

Chapter 4: The moral of the story

- The following tools can help prevent future “wars”: internal replication, external replication, study registration, pre-analysis plans
- Confirmation bias is human nature. As scientists, we should expect it and develop methods to minimize it.
- Openness and transparency methods may be painful in the short term but contribute to better science in the long term.

> Miguel and Kremer had the decency, generosity, strength of character, and intellectual confidence to let someone else peer under the bonnet... One way or another, I can't believe they won't feel bruised by the reanalysis.
>
> And that is where we have gone wrong. It's not just naive to expect that all research will be perfectly free from errors, it's actively harmful.

Goldacre, 2015

Why are replication and transparency important?

"If we are not wrong frequently, we are failing to push on the frontiers of knowledge hard enough. At the same time, because true innovations are rare, **valuing replication will foster efficient filtering** of interesting findings and accelerate knowledge building."

PLOS | BIOLOGY

PERSPECTIVE

Scientists' Reputations Are Based on Getting It Right, Not Being Right

Charles R. Ebersole[1]*, Jordan R. Axt[1], Brian A. Nosek[1,2]

1 University of Virginia, Psychology Department, Charlottesville, Virginia, United States of America, 2 Center for Open Science, Charlottesville, Virginia, United States of America

* cebersole@virginia.edu

Abstract

Replication is vital for increasing precision and accuracy of scientific claims. However, when replications "succeed" or "fail," they could have reputational consequences for the claim's originators. Surveys of United States adults ($N = 4,786$), undergraduates ($N = 428$), and researchers ($N = 313$) showed that reputational assessments of scientists were based more on how they pursue knowledge and respond to replication evidence, not whether the initial results were true. When comparing one scientist that produced boring but certain results with another that produced exciting but uncertain results, opinion favored the former despite researchers' belief in more rewards for the latter. Considering idealized views of scientific practices offers an opportunity to address incentives to reward both innovation and verification.

CrossMark

OPEN ACCESS

Thanks!

Any questions?

You can find me at
jadebc@berkeley.edu

References

Aiken AM, Davey C, Hargreaves JR, Hayes RJ. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *Int J Epidemiol*. 2015 Oct 1;**44**(5):1572–1580.

Davey C, Aiken AM, Hayes RJ, Hargreaves JR. 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. Int J Epidemiol. 2015 Jul 22;dyv128.

Gertler P. 2015. "Good science gone wrong?" http://blogs.berkeley.edu/2015/08/03/good-science-gone-wrong/

Goldacre B. 2015. "Scientists Are Hoarding Data And It's Ruining Medical Research" https://www.buzzfeed.com/bengoldacre/deworming-trials?utm_term=.dy5dWdDAp#.xpN8Q8rwW

Hargreaves JR, Aiken AM, Davey C, Hayes RJ. Authors' Response to: Deworming externalities and school impacts in Kenya. Int J Epidemiol. 2015 Oct 1;44(5):1596–1599.

Hicks JH, Kremer M, Miguel E. Commentary: Deworming externalities and schooling impacts in Kenya: a comment on Aiken et al. (2015) and Davey et al. (2015). Int J Epidemiol. 2015 Oct 1;44(5):1593–1596.

Jourdan PM, Lamberton PHL, Fenwick A, Addiss DG. Soil-transmitted helminth infections. The Lancet [Internet]. [cited 2017 Sep 14]; Available from: http://www.sciencedirect.com/science/article/pii/S014067361731930X

Keiser J, Utzinger J. Efficacy of current drugs against soil-transmitted helminth infections: Systematic review and meta-analysis. JAMA. 2008 Apr 23;299(16):1937–1948.

Miguel E, Kremer M. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Econometrica. 2004 Jan;72(1):159–217.

Ozler B. 2015. "Worm Wars: A Review of the Reanalysis of Miguel and Kremer's Deworming Study" http://blogs.worldbank.org/impactevaluations/warm-wars-review-reanalysis-miguel-and-kremer-s-deworming-study

"Re-analysis opens a big can of #WormWars" https://storify.com/viewfromthecave/tracking-the-wormwars

"Should Deworming Policies in the Developing World be Reconsidered?" http://blogs.plos.org/speakingofmedicine/2012/07/18/should-deworming-policies-in-the-developing-world-be-reconsidered/

Taylor-Robinson DC, Maayan N, Soares-Weiser K, Donegan S, Garner P. Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin and school performance. Cochrane Database Syst Rev. 2012;7:CD000371.

Taylor-Robinson DC, Maayan N, Soares-Weiser K, Donegan S, Garner P. Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance. Cochrane Database of Systematic Reviews [Internet]. John Wiley & Sons, Ltd; 2015 [cited 2017 Sep 14]. Available from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000371.pub6/abstract

World Health Organization. Deworming for health and development [Internet]. Geneva, Switzerland; 2004. Available from: www.searo.who.int/LinkFiles/STH_CDS_CPE_PVC_2005_14.pdf