

Replication Redux: The Reproducibility Crisis and the Case of Deworming

Owen Ozier 

*In 2004, a landmark study showed that an inexpensive medication to treat parasitic worms could improve health and school attendance for millions of children in many developing countries. Eleven years later, a headline in *The Guardian* reported that this treatment, deworming, had been “debunked.” The pronouncement followed an effort to replicate and re-analyze the original study, as well as an update to a systematic review of the effects of deworming. This story made waves amidst discussion of a reproducibility crisis in some of the social sciences. In this paper, I explore what it means to “replicate” and “reanalyze” a study, both in general and in the specific case of deworming. I review the broader replication efforts in economics, then examine the key findings of the original deworming paper in light of the “replication,” “reanalysis,” and “systematic review.” I also discuss the nature of the link between this single paper’s findings, other papers’ findings, and any policy recommendations about deworming. Through this example, I provide a perspective on the ways replication and reanalysis work, the strengths and weaknesses of systematic reviews, and whether there is, in fact, a reproducibility crisis in economics.*

JEL Codes: A14, B41, C18, C38, C59, C80, I10, I15, I18, O15

Keywords: Data access, deworming, health, education, meta-analysis, systematic review, public health, replication, robustness, worms.

Reproducibility

We are in the throes of a “reproducibility crisis” in the sciences, if headlines are to be taken at face value (Maniadis and Tufano 2017). Scholars have expressed concern—in social sciences including psychology and economics, as well as in the natural sciences—that published research findings may not prove to be reproducible, or may not be “robust” (Baker 2016a; Ioannidis, Stanley, and Doucouliagos 2017). Whether a finding is reproducible (or “robust”) may be assessed by trying to replicate

The World Bank Research Observer

© The Author(s) 2020. Published by Oxford University Press on behalf of the International Bank for Reconstruction and Development / THE WORLD BANK. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
doi: 10.1093/wbro/lkaa005

0:1–31

it; but what does it actually mean to *do* such a replication? And how should we respond when a replication seems to cast doubt on a study?

The reproducibility of a result has been central to science for hundreds of years. Robert Boyle wrote about both the importance and the difficulty of reproducibility in 1673.¹ Slightly more recently, in a seminal 1935 book, R. A. Fisher not only quoted Boyle on the topic, but wrote that as users of tests of statistical significance, “we ... admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the ‘one chance in a million’ will undoubtedly occur, ... however surprised we may be that it should occur to us. In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result,” (Fisher 1935, 13–14).

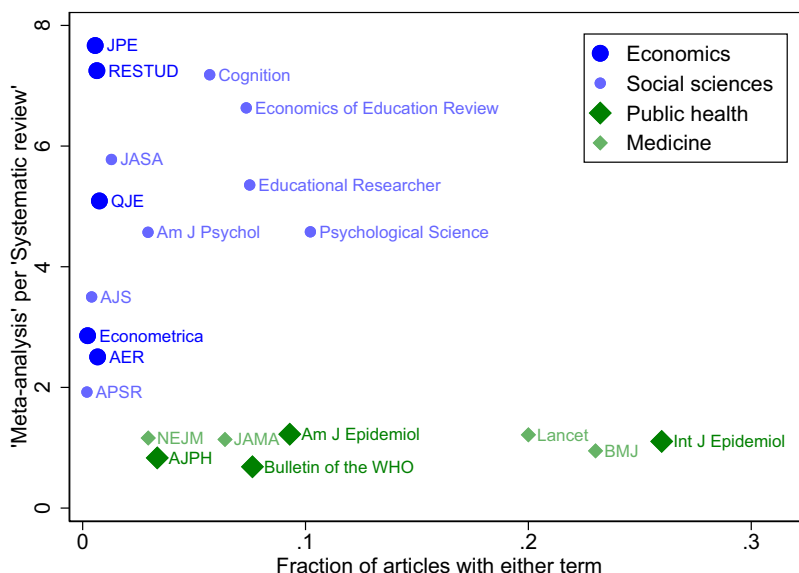
Fisher refers to “a reliable method of procedure” as that which renders a phenomenon “demonstrable.” This is reproducibility in the sense of defining an experimental procedure, which, if followed, will yield a predictable result. For an experiment in chemistry or physics, reproducibility might be easy to imagine (though sometimes still challenging to achieve in practice). In the case of an empirical study in economics or epidemiology, however, the definition is harder to pin down. Reproducing a “procedure” might mean doing an entire study again at a site with similar conditions, or simply re-running the code or calculations to make sure others would arrive at the same result given the existing data and research design.² At the other extreme, reproducibility animates our thinking on the aggregation of evidence across studies: if a policy generates similar results in multiple studies, contexts, and so forth, then it is clearly reproducible in this sense, and may be seen as a policy with well-understood impacts.³

Aggregating Results across Studies

Bringing together evidence across studies first requires a search of the literature; this is one of the early steps in what is sometimes called a “systematic review.” With a set of studies in hand, the next step in the review may (or may not) be a statistical approach to analyzing the published findings, usually called “meta-analysis” or a variant thereof. Different disciplines emphasize different aspects of this process, and afford the process different levels of scholarly prominence, as shown in figure 1.

Several patterns are evident in figure 1. First, in the field of economics, these kinds of aggregations across studies are mentioned much less commonly than in medicine: among the journals shown in the figure, a mention of “systematic review” or “meta-analysis” occurs only in roughly 0.6 percent of articles in economics journals; the

Figure 1. The Terms “Meta-Analysis” and “Systematic Review” across Disciplines



Note: The vertical axis of figure 1 shows the ratio of the number of articles using the term “meta-analysis” to the number of articles using the term “systematic review.” The horizontal axis shows the fraction of articles that use either of these terms. All data gathered from Google Scholar, February, 2019. Abbreviations: AER = American Economic Review; AJPH = American Journal of Public Health; AJS = American Journal of Sociology; APSR = American Political Science Review; BMJ = British Medical Journal; JAMA = Journal of the American Medical Association; JASA = Journal of the American Statistical Association; JPE = Journal of Political Economy; NEJM = New England Journal of Medicine; QJE = Quarterly Journal of Economics; RESTUD = Review of Economic Studies. Additional journals in these fields would corroborate the pattern shown in the figure, but clutter the figure, and are not shown. Social Forces and Sociology of Education are both very close to the statistics of the American Economic Review; the Journal of Politics is very close to the American Political Science Review; the American Journal of Political Science is very close to the Quarterly Journal of Economics; the Annals of Internal Medicine and the Journal of Pediatrics are very close to both the New England Journal of Medicine and the American Journal of Public Health; and so on.

comparable figure for public health is 7.1 percent, while for medicine it is 16.7 percent (almost 30 times the rate seen in economics). Other social sciences, a sampling of which are shown, fall somewhere in between, with these terms appearing in 4.6 percent of articles. The second pattern in the figure is that economics places much greater emphasis on the statistics of a “meta-analysis” than on the formal procedures of “systematic review,” with the former term occurring more than three times as often, on average, than the latter. In the other social sciences, that pattern is even more extreme: meta-analysis is mentioned six times more often than systematic review. In medicine and public health, there is relatively more emphasis on “systematic reviews:” the ratio is reliably one to one.

Though economists appear to discuss meta-analyses relatively infrequently, they have offered suggestions on how to go about them. A 2001 article in the *Journal of Economic Perspectives* described in four paragraphs how to systematically trawl the literature, then spent many more pages providing guidance on, and examples of, helpful statistical approaches (Stanley 2001). Despite its brevity, the article points to well-known reviews of the effects of minimum wages, tax policies, and the returns to education, among other topics.

In the field of health, the Cochrane Collaboration is one of the more prominent organizations curating and aggregating evidence; it has offered guidelines on systematic reviews since the 1990s (Petticrew and Roberts 2008). Within the 265-page 2006 edition of its handbook, it offered 14 pages on locating studies, 12 pages on assessing their quality, six pages on eliciting information from the studies, and 70 pages on analysis. Further discussion on other topics included whether and how to analyze subpopulations, and so on (Higgins and Green 2006). Thus, there is ample guidance on how to conduct a review and analyze resulting data. What can go wrong?

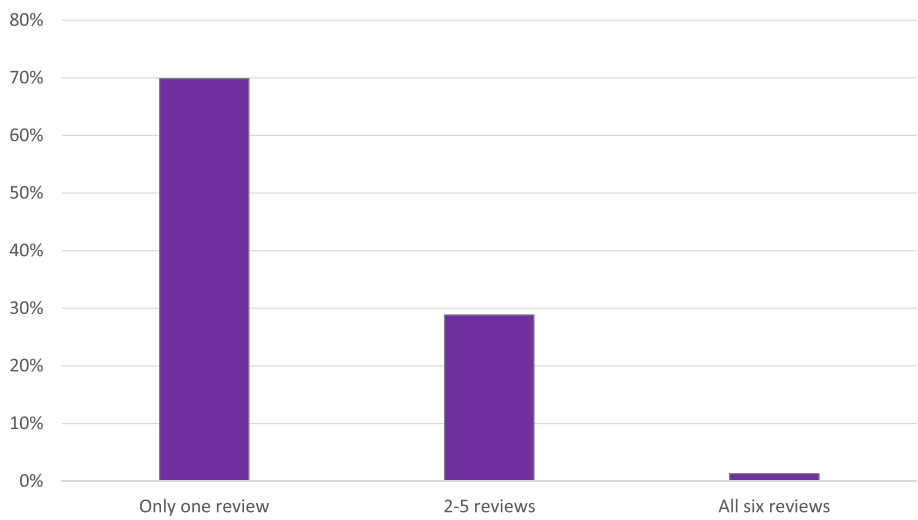
The point of a systematic review is that one study alone has some chance of being erroneous or unique to a specific context, so one learns more by drawing together evidence across studies. Doing so is not easy, though.

As an example of the subtleties involved in any such review, however, consider that a reviewer would not want to double-count evidence. This may be relatively straightforward in the case of medical trials: in principle, every trial is entirely distinct (in terms of population and so forth) from every other. The Cochrane 2006 handbook never uses the phrase “double counting.” But when interventions take place at a larger scale, more than one study may examine the same phenomenon, in the same place and time, even using overlapping—possibly identical—underlying data; determining how to aggregate studies statistically without double-counting observations could be a difficult task (a concern raised by Goldfarb and Stekler 2002, for example).

A second subtlety is that while a large group of randomized trials may be somewhat straightforward to assess and interpret, quasi-experimental studies (such as difference-in-difference, regression discontinuity, and other designs) provide estimates whose consistency depend on differing assumptions, some of which (depending on context) may be more credible than others. Study design criteria can therefore play a role in determining whether a study is included in any particular review. This may be an area where the approach to reviews that is common to medicine and public health may serve medicine better than public health: in public health, a larger share of the evidence may come from quasi-experimental designs than is common in the medical literature, potentially necessitating a wider net or more elaborate statistical work when aggregating studies.

A crucial step for any systematic review is that of defining a category of study well enough that there are both enough studies to meaningfully aggregate, and yet that the topics of the studies remain sufficiently similar: “whether experiments can be

Figure 2. Across Six Reviews, Number of Studies Appearing In. . .



Source: Adapted from Evans and Popova (2016).

pooled to provide cumulative evidence depends further on which features of a study or results are considered scientifically equivalent enough to pool,” as Goodman, Fanelli, and Ioannidis (2016) wrote. To provide some perspective on the challenges inherent in systematic reviews, consider the pattern, depicted in figure 2, based on Evans and Popova (2016). These authors examined six reviews of education interventions in developing countries to understand why the reviews had come to differing conclusions. The set of studies ultimately included in a review is, of course, central to the review’s conclusions. The pattern that Evans and Popova (2016) uncover is stark (shown in figure 2). Despite the similarity of their goals at the outset, the six reviews defined their meta-analytic inclusion criteria slightly differently, so that of 229 studies included in *any* of the reviews, most underlying studies were included in just *one* of the six reviews. Of the 229 underlying studies, only three of them (roughly 1 percent, shown in figure 2) appeared in all six systematic reviews. Seeing this, it is hardly surprising that the reviews arrived at different conclusions; they reviewed different papers!

Some general points to take away from this discussion are that there is more than one approach to assembling evidence across studies, and that the array of decisions required along the way can easily influence a review’s conclusions. In relation to the example provided by Evans and Popova, however, a more specific lesson emerges: it is reasonable to assess the robustness of a study by comparing it to other “similar” studies, but only if particular care is taken to determine what is in fact “similar,” and if—having done this—a reasonable number of studies are left to examine.

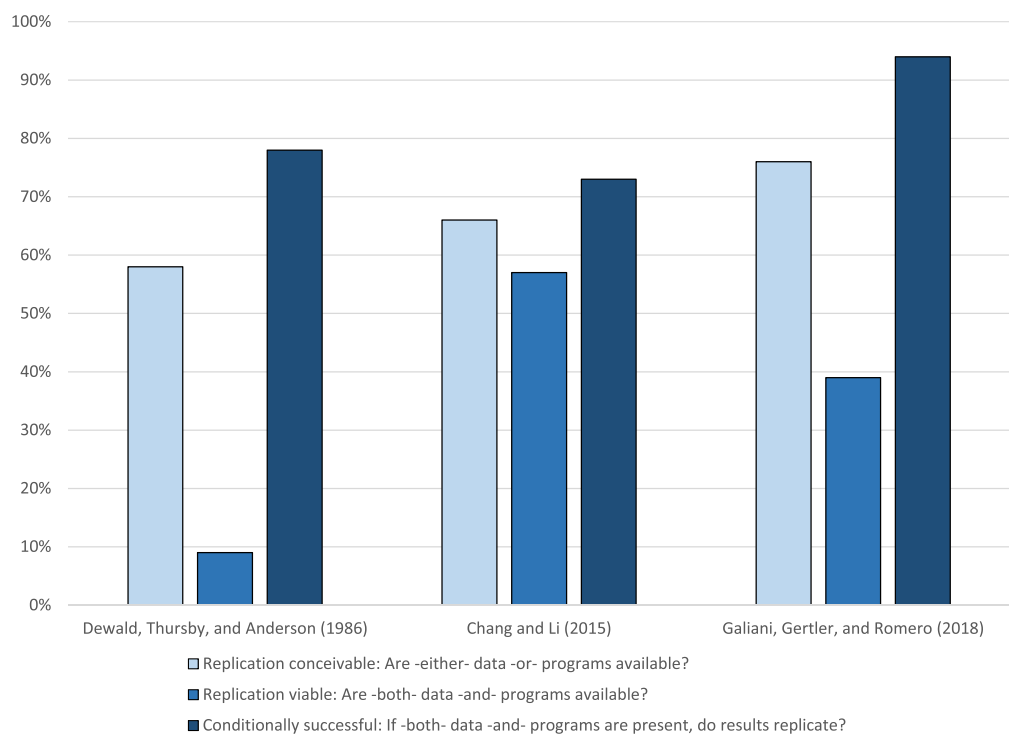
Comparison with other studies is not the only measure of “reproducibility,” however. The rise of computing power has created the opportunity, as well as the need, for another kind of reproducibility: checking prior studies for computational mistakes.

Over the last several decades, the social sciences have seen a rapid increase in the availability of “replication” data. That is, once researchers publish papers, they share their data and computer programs so that others may re-run the analysis for themselves, and potentially even expand on it. This been an important development, in part, because well-known attempts to replicate papers en masse have done famously poorly. Mind you, this is only an attempt to re-run the calculations in existing papers; something that seems almost mechanical. Yet within economics, a 1986 study showed that only 15 percent (8 of 54 sets) of replication files were complete enough to permit replication; more than 30 years later, a 2018 study showed that only 14 percent of studies supplied the materials needed for replication. These numbers, side-by-side, make it seem that almost nothing is replicating, and that almost nothing has changed. The headlines, however, obscure several underlying changes, shown in [figure 3](#).

When is a replication even conceivable? 30 years ago, [Dewald, Thursby, and Anderson \(1986\)](#) were only able to obtain something resembling replication data for just over half of the economics papers for which they sought data; that figure has risen to between 60 and 80 percent; in other words, the rate of data unavailability has nearly been cut in half.⁴ Getting usable data and code in hand to make a *viable* replication attempt was very hard in those days: of 54 received data sets, only 8 satisfied a basic usability criterion (yielding the 15 percent number from the 1986 study), driving the chances of even being able to try a replication below 10 percent. As [Figure 3](#) shows, that viability number has risen considerably, though how high depends on which sample of journals is examined.⁵ [Galiani, Gertler, and Romero \(2018\)](#) found a roughly 39 percent viability rate, while [Chang and Li \(2015\)](#) found a 58 percent rate (38 viable data sets from 67 attempts).⁶ Thus, by this comparison, viability of replication has risen by as much as a factor of six since the 1980s. Conditional on a viable combination of data sets and programs, findings can be *successfully* replicated 70 or more percent of the time, with that fraction rising in the most recent study to above 90 percent.

One can still find ways of taking a dim view of the current situation, of course. If one wishes, one can count as “not successfully replicating” any case in which a typographical, data, or programming error is uncovered that changes numbers somewhere in the paper but does not change the overall qualitative findings. Imposing this requirement would bring Dewald, Thursby, and Anderson’s 78 percent conditional success rate down to 22 percent. Alternatively, one could add the requirements that (a) graphics (figures) be replicable in addition to tables, (b) that this be possible

Figure 3. How Replicable are Studies in Economics?



without any help from the original authors, (c) that this take less than half a day’s work, and (d) that the replication program works not only from the estimation data set but also all the way from any raw data that were originally gathered. Having imposed this higher standard, [Galiani, Gertler, and Romero \(2018\)](#) drive the net success rate down to their headline number of 14 percent.

Replication Terminology

Having seen changes in the patterns of replicability over the past three decades, it is worth asking whether clear names can be given to different kinds of replication. Earlier, I referred to what Fisher called a “demonstrable” experiment (one for which an experimenter is able to reliably conduct a procedure that produces a predictable result) as “reproducible,” while I have followed recent years’ efforts in the social sciences in referring to the successful checking of computer programs as “replication.”⁷ To confuse matters, not everyone uses those same terms. In the field of biostatistics, [Leek and Peng \(2015\)](#) suggest using these terms in almost exactly the opposite way.⁸ In economics, [Clemens \(2017\)](#) spends a thoughtful 17 pages on

which way the terms have been used, and what a good way forward might be.⁹ The bottom line is that what [figure 3](#) describes as “replication” is sometimes called “pure replication” ([Hamermesh 2007](#)), or “verification” ([Clemens 2017](#)). Within the social sciences, this kind of code-checking is a reasonably well-defined exercise. However, there is little agreement on how to categorize all the other kinds of data reproducibility exercises that go further: analyzing the same data in new ways, or testing the analysis for additional robustness. The ambiguity about what “replication” and “reproducibility” mean, even within a single discipline, could easily contribute to either misunderstandings in the media or conflict among experts.

Case in Point: [Miguel and Kremer \(2004\)](#)

With an understanding of the possible meanings of reproducibility established, but before offering any recommendations, it is instructive to examine a recent, prominent case. In introducing this case, I first recapitulate the arc of the original underlying paper; I discuss its relationship to the prior and subsequent literature; I then proceed to describe the specifics of a replication, reanalysis, and review.

Worms

Nearly two billion people around the world are infected by intestinal worms ([World Health Organization 2017](#)). These species of parasitic worm inhabit the human digestive tract; they spread by expelling their eggs via the excrement of infected people. Without good sanitation, these microscopic eggs can find their way, unnoticed, onto the skin (or food) of another person. Once someone ingests an egg, the reinfection cycle continues. Some of these parasites’ life cycles are more exotic and complex, but they have in common that poor sanitation facilities and hygiene practices allow infections to spread locally. The medication to treat the worms has few side effects and is remarkably cheap. Many of the people most infected by worms, it should be noted, are children.

The Original Study and the “Worm Wars”

In 2004, Miguel and Kremer published a study showing that an inexpensive deworming medication improved health and school attendance in Kenya. [Miguel and Kremer \(2004\)](#) went on to show that these effects could previously have been hidden from view by a subtlety in the design of many randomized trials: many studies had not accounted for the ways that worm infections can spread from one person to another, as I discuss further below. With a randomized trial designed to overcome this obstacle, the effects on health and school attendance were easy to see. The World Health

Organization, the international donor community, and country governments all supported policies of deworming.¹⁰ When, in 2015, *The Guardian* headline described deworming as “debunked,” dozens of blog posts, journal articles, and stories in the popular media sprung forth to debate deworming (Evans 2015). The back-and-forth involved technical experts from a range of fields, hundreds of pages of analysis and critique, and considerable misunderstanding. Though seeing science happening in “real time” may have been thrilling for some, it was disorienting and time-consuming for many. What opinion you left with might depend on what opinion you started with, which discipline’s training you received, who you trusted, or which article you happened to read last before losing interest. This outburst, colloquially known as the “Worm Wars,” raised practical as well as philosophical questions.

What did the original paper find, and how exactly had it not been discovered before? In light of the definitions of reproducibility and replication discussed earlier, what did it practically mean to “replicate” and “reanalyze” the study, when no new deworming trial had taken place? What conclusions should we then draw from these “replication” and “reanalysis” studies? What role did systematic reviews play, and where does this all leave deworming?

Cluster Randomization

Most randomized trials of deworming prior to that of Miguel and Kremer assigned “treatment” at the individual level: that is, within some group, such as a village or school, exactly which children were given medications during the trial was randomized. The problem with this research design is reinfection. Children living close to one another can infect one another. Why does this matter? Because if one child takes a drug (say, mebendazole) that kills the parasitic worms living in her gut, she might be worm-free. But if her siblings or neighbors did not receive this treatment, the infected ones will continue to excrete worm eggs into the environment; before long, the dewormed child may be re-infected. Thus, the “treatment” group in such a study may not be entirely free of worms. The complication doesn’t end there, however. When many of the children in a neighborhood are treated for worm infections, even those who do not receive medication could benefit from the reduced reinfection rate. The “control” group effectively gets some treatment as well. This may be thought of as “crossover” or “contamination” in the design of a clinical trial, brought about by what could be termed an “indirect effect,” “spillover,” or “externality.” The upshot is that in an individually-randomized deworming trial, the “treatment” and “control” groups may not be what the researcher meant for them to be, so any differences that deworming medication could bring about—in health, in schooling, or in anything else—could be substantially muted when simply comparing these two groups.

Starting in 1998, for a period of several years, Miguel and Kremer did things differently. Rather than randomizing at the individual level, they assigned treatment at the

level of the school: if the school was assigned to be “treated,” everyone who came to school on deworming day received medication. If the school was assigned to be “control,” nobody received such medication (at least not until a few years later). This has advantages and drawbacks. A great advantage, in light of the reinfection dynamics described above, is that within the immediate vicinity of the school, the reinfection spillovers do not dampen the intensity of effects in the study. In treatment areas, children who receive deworming medication gain both the direct benefit of treatment and the indirect benefit of living around others who are less likely to spread worm eggs into the environment. In control areas, neither the direct nor indirect effect is present—at least, if the schools are far enough away from one another. A drawback of this “cluster-randomized” design is that the statistics must eventually be adjusted for the inherent correlations in outcomes among children living in the same neighborhood and attending the same school, so the study must be quite large in order to precisely measure effects; Miguel and Kremer’s study thus involved tens of thousands of pupils.

Miguel and Kremer found that deworming reduced worm infections (as one would hope), improved self-reported health status, and improved school attendance: in the simplest analysis, the likelihood of attending school on a given day was increased by 5.1 percentage points, against a background of a 20 to 30 percent absenteeism rate.¹¹ This finding was notable, since some previous studies had not found effects on student absenteeism, though perhaps previous non-findings had been due to study design issues of the kind described above.

Because Miguel and Kremer had randomly varied which schools were dewormed, there was not only random variation in whether a child’s school was dewormed, there was also random variation in how many nearby schools’ students were also dewormed. This variation allowed Miguel and Kremer to estimate the spillovers themselves, up to some distance limit: with a large enough distance, there would be no such variation since all schools in the study area would eventually be included.

In the original paper, there seemed to be enough variation to precisely and separately estimate several kinds of reinfection-related spillovers: spillovers within schools (from students who were present to take the medication on the day of deworming to those who were ineligible or absent that day); spillovers from dewormed schools within 3 kilometers; and spillovers from schools between 3 and 6 kilometers away. Taking just the 3-kilometer spillovers into account (calculated by adding the direct and indirect effects together, weighting the spillovers by the number of students in the area), the overall effect of deworming was about an 8.1 percentage point improvement in school attendance. The spillovers from 3–6 kilometers away seemed to be beneficial for health outcomes but were not significantly different from zero for school attendance. In fact, though statistically indistinguishable from zero, the estimate of these long-distance spillovers on school attendance was negative, so incorporating them into the overall calculation reduced the overall calculated effect of deworming to a 7.5 percentage point improvement in attendance. This figure,

however, was still clearly statistically nonzero, so despite the imprecision in the calculation, reporting the smaller of these precisely measured effects could be seen as a cautious choice. Thus was born the oft-cited 7.5-percentage-point improvement in school attendance that has appeared in policy briefs and textbooks for years (e.g., J-PAL 2012, Glennerster and Takavarasha 2013, De Janvry and Sadoulet 2015).

The paper was influential: it quickly gathered hundreds of citations and played a role in policy discussions. Since Miguel and Kremer had found evidence that an inexpensive drug could improve a range of important outcomes, it was actionable. Given its actionable nature, it seemed like the kind of finding whose reproducibility should be confirmed. Of course, Miguel and Kremer had taken most of a decade to prepare for, conduct, analyze, and publish their work. Was someone supposed to do that all over again? The wide, sometimes contradictory, range of standards and definitions as to what constitutes “reproducibility” or “replication” discussed above played a role in the communication around the “Worm Wars.” But, with the framework of Clemens (2017) and others in mind, the contours of what happened next are easily understood.

Replication and Reanalysis of Miguel and Kremer (2004)

In 2013, the International Initiative for Impact Evaluation (3ie) commissioned a replication study of Miguel and Kremer’s work; by then, Miguel and Kremer had already prepared their data set so that it could be made accessible to the public.¹² In early 2013, a replication plan was made public; by the end of 2014, Miguel and Kremer had posted their public data set online, and two kinds of “replication” had been completed.

The first thing to notice about this replication effort was its ambition. The effort not only included a “pure replication,” or “verification” type replication, in which the goal was to follow the original Miguel and Kremer analytical approach to check whether the published results hold up to an effort at re-calculation; it also included analyses with different handling of the raw data, which the replication team describes, extending the terms of Hamermesh (2007), as “internal scientific replication” and “internal statistical replication.”¹³ Perhaps most notably of all, adding a layer of challenge: the effort was not undertaken by economists, but by epidemiologists. (How many of you, reading this, have ever undertaken a replication of a study in a different discipline than the one in which you were trained?) This creates an additional terminology challenge at the outset, which the authors grapple with head-on: the first page of the replication plan sets out a glossary of terms, comparing them across disciplines: an economist’s “externality” becomes the medical “indirect benefit;” “systematic errors in data” become “bias;” and so forth (Aiken et al. 2013).¹⁴

While terminology may have been an obstacle, the interdisciplinary nature of this replication effort, and of the ensuing “worm wars” more generally, yielded

as a byproduct a perspective on the way economics and other disciplines conduct their scientific inquiry, and how that has changed over time. Miguel and Kremer did not file a trial registry or pre-analysis plan, for example. In the 1990s, however, when their study began, norms were somewhat different. To take an example from medicine: what would become known as the “International Standard Randomised Controlled Trial Number” registry did not come into existence until 2000 ([ISRCTN 2018](#)). Now, of course, many in the discipline of economics have started considering whether, when, and how it makes sense to register experiments and to specify analytical designs in advance ([Coffman and Niederle 2015](#); [Olken 2015](#); [Anderson and Magruder 2017](#); [Fafchamps and Labonne 2017](#)).¹⁵ Here is another reflection on scientific norms: unlike in an efficacy trial of a new medication, it is difficult to imagine how double-blinding would be possible for many experiments in the social sciences. While it may be possible to obscure the treatment status of study participants so that field data collectors may be blind to it in some cases, treatment and comparison groups generally know what they are receiving. Yet a third epistemological issue exists: in the social sciences, studies regularly sample from a large population of interest, making CONSORT-compliant diagrams—a standard for communicating randomized trials in medicine—more complicated to depict ([Schultz et al. 2010](#)).

The back-and-forth between the replication authors and the study authors, while sometimes reading as a tense disagreement, is for the most part a polite and gracious scholarly exchange.¹⁶ Aiken, Davey, and colleagues thank Miguel, Kremer, and colleagues for their openness with data and assistance with aspects of the replication; Hicks, Miguel, and Kremer, in turn, thank the replication team for identifying issues in the analysis that could be resolved in the public replication data. So why the apparent conflict of the “worm wars?” Below, I go into some detail describing the two replications, their findings, and how those findings were communicated.

Pure (verification) Replication

The replication report’s abstract makes the pattern of results sound complicated and nuanced: “We noted various discrepancies between the published results and those from this reanalysis. These ranged in importance from minor (for example, rounding errors) to moderate (for example, inaccurately labelled significance) to major (for example, coding errors),” ([Aiken et al. 2014](#)). Indeed, mirroring the replication report, the original study authors themselves discuss a wide range of underlying data problems in the data manual that accompanies their replication files ([Miguel and Kremer 2014](#)).¹⁷ To understand what this means, we can divide the findings from the original paper that the replication focused on into two main categories: those that appeared directly in the regression tables in the original paper, and other numbers from the text that were calculated on the basis of several estimated quantities. [Table 1](#)

Table 1. Replication of Key Coefficient Estimates

	Original	Revised
Naïve effect, reduced worm infection	−0.25 (0.05) ***	−0.31 (0.06) ***
Within-school externality on worm infection	−0.12 (0.07) *	−0.18 (0.07) **
Within-school externality on attendance	+0.056 (0.02) ***	+0.056 (0.02) ***

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

(above) shows how the replication unfolded for three of the prominent findings in the first category.

The first finding in [table 1](#) is that when schools were randomly assigned to begin deworming treatment, pupils there experienced a rate of moderate-to-heavy worm infection that was 25 percentage points lower than the rates at nearby schools. When that finding was re-visited in the replication effort, the pattern grew slightly stronger: the best estimate was now a 31 percentage point reduction in such infections. The second finding in [table 1](#) is that, within schools receiving deworming treatment, children who for various reasons did not receive deworming medication still benefited: the original paper showed that they had a rate of moderate-to-heavy infection 12 percentage points lower than they would have without the intervention; in the revised calculations with all data and programming errors resolved, this estimate also grew stronger (and more statistically significant). Within treated schools, the effect on the school attendance of untreated children was an increase of 5.6 percentage points: unchanged. In this portion of the replication, Miguel and Kremer’s results, if anything, grow stronger.

Two Characters and the Headline Number

The replication seemed less clear-cut for two of the prominent findings in the second category: numbers calculated from several estimated quantities. One of the worm wars’ more important disagreements surrounding one of these numbers is the 7.5-percentage-point improvement in school attendance often mentioned from the original 2004 study. When Miguel and Kremer released their public data files, they corrected several mistakes in their original data construction, described in detail in Miguel and Kremer’s replication manual.¹⁸ A single two-character programming error proved to be pivotal, however, in the part of a program that counted the

number of pupils within a specified radius of a given school. The loop that counted the number of pupils nearby stopped too soon. This matters because Miguel and Kremer estimate that a child experiences benefits of deworming even when the dewormed children attend a school a few kilometers away.

Aiken et al. (2014) relate the relevant code excerpt, provided to them by Miguel and Kremer, as including the number **12** as the maximum number of schools to be tallied where the maximum should instead be **75**, the total number of schools in the study. This does not affect the calculation of the number of pupils within three kilometers, since there are never more than 12 schools within this short distance, but it does affect the calculation of the number of pupils within six kilometers.¹⁹

This has several implications. In specifications that do not involve estimation of spillovers beyond 3 kilometers, not much changes. Direct effects are slightly more statistically significant, and slightly more pronounced, in the corrected data: a previously-estimated 5.36-percentage-point increase in school attendance caused by the direct effect of deworming is corrected to 5.78 percentage points, for example (appendix table A1.1, panel A). Spillover effects within 3 kilometers are nearly unchanged, revising a previously-reported 2.78-percentage-point benefit to a corrected 2.70-percentage-point benefit (appendix table A1.1, panel C). Summing these two effects up to 3 km, the picture did not change: the total benefit had been 8.14 percentage points, but was now 8.48 percentage points.

The spillovers from 3–6 kilometers, however, had previously been fairly precisely estimated to be nearly zero. That is, in the 2004 analysis, whether to add the 3–6 km spillovers to a cumulative estimate of direct effects and within-3-km effects was inconsequential: it neither changed the magnitude nor the standard error very much. This inclusion reduced the estimate from 8.14 percentage points to 7.47 percentage points, rounding to the familiar 7.5-percentage-point effect that is so often quoted. However, in the corrected data, the coefficient on 3–6 km spillovers was more negative (though still not statistically significant), and now that the number of children in this area was being correctly counted, there were more than twice as many of them, driving up the associated standard error by more than a factor of two. Now, in the corrected data, whether to add in the large, negative, and imprecisely estimated 3–6 km spillover estimate was consequential: it would take the 8.48-percentage-point effect, reduce it, and make it imprecise enough that it could no longer be distinguished from zero. Thus, following exactly the original steps but with corrected data, the 7.5 percentage point change in school attendance, dropping to an imprecise 3.9 percentage points, seems not to hold up.

The twist is that there is no particular reason that 6 km is the right radius to check. In the 2004 analysis, it seemed that there was no meaningful difference in either point estimate or precision, whether the 3–6 km spillovers were included or not. However, the step including these spillovers now appears to be a problematic one. The distant spillovers are too imprecisely estimated to add in, so should not have been included in

the first place. Even in the original analysis, there must have been insufficient variation in treatment at still greater distances to estimate spillovers from afar, so the line had to be drawn somewhere. Where should it be drawn now, and why? Hicks et al. (2015) point out that adding imprecisely estimated quantities to precisely estimated ones yields imprecision, so even if adding in the 3–6 km spillover yields an unbiased estimator of the true total effect, the expected distance from such an estimate to the truth (the “mean squared error”) is larger than it would be, had one used the slightly biased but lower variance estimator involving only the direct effect and 0–3 km spillovers. This reasoning requires knowing the magnitude of the bias, which Hicks et al. (2015), take from the relative magnitude of the estimates of direct and 3–6 km spillover effects on worm infections: the 3–6 km spillovers appear to cause relatively little change in worm infection, thereby making it likely that they also cause very little change in school attendance. By this reasoning, the line should be drawn at 3 km.

There is a simpler way of looking at this. Without the suggestion from Hicks et al. (2015), of a rule meant to minimize mean squared error, one can take both the original distance thresholds as equally reasonable options for effect summation. Having two statistics to choose from, however, one must confront the problem that usually arises when picking whichever of two specifications yields a more statistically significant effect: “p-hacking.”

“P-hacking” is a term describing a problematic research practice. Quite often, a fixed statistical significance threshold (or “p-value”) is seen as desirable; this “p-value” is usually seen as the chance that a study’s finding is a “false positive”—that is, that such a strong pattern would have been found by chance (driven by measurement error or sampling variation perhaps), had there been no underlying pattern to detect. A common threshold value to consider has been 0.05, perhaps ever since R. A. Fisher wrote, “We shall not often be astray if we draw a conventional line at .05,” (Fisher 1934). This may be a fine threshold to consider if there is one only possible statistical test to consider in a given study or in relation to a given hypothesis. But if a researcher is free to try many similar variants of an analysis until that threshold is met, reporting (or being able to publish) only the analysis that meets the threshold, then with enough persistence, a researcher can bend nearly any data set to meet the threshold; spurious results begin to appear; and resulting p-values are “hacked,” in the words of Simmons, Nelson, and Simonsohn (2011 and 2013), and Simonsohn, Simmons, and Nelson (2014).²⁰

There are well-known antidotes to this problem. One option is that before a study is started, a researcher could pre-commit to a specific analysis, as trial registries and pre-analysis plans allow, thereby avoiding the temptation to search among different specifications (Olken 2015). As discussed earlier, such plans were uncommon in the social sciences (and less formalized even in medicine) at the time of the original deworming study. However, even without such pre-analysis planning, there is another solution that is available after the fact, as long as the number of statistical tests

is known. In the present example, there are only two possible tests. The antidote? Correction for multiple hypothesis testing.²¹ In the present case, if we are free to choose between these two specifications, the Dunn/Bonferroni style of correction simply involves multiplying the p-value by two; other corrections that take into consideration correlations between outcomes generally require smaller corrections.²² Allowing for both the test of the coefficient 0.0387 (yielding a conventionally insignificant p-value of around 0.2) and the test yielding of the coefficient 0.0848 (yielding a conventionally very significant p-value of less than 0.00001) to be considered, then, is there a significant result—after correcting for performing two tests—that agrees qualitatively with the original paper? There is. The p-value associated with the 8.48 percentage point increase in school attendance, when doubled, is still significant by the conventional standards of both economics and epidemiology.

This is the first place where drawing a conclusion from the replication exercise proved to be divisive. Miguel and Kremer's own replication files include all these calculations, old and new. Following the corrected data construction, but without any adjustment to analytical decision-making in light of the data, one could conclude that the 7.5-percentage-point finding drops to insignificance. Allowing for such adjustments, either via a minimum mean-squared-error approach or a multiple-testing-adjustment approach, one would conclude that the 7.5-percentage-point finding is now an 8.5-percentage-point finding, and is stronger than before.

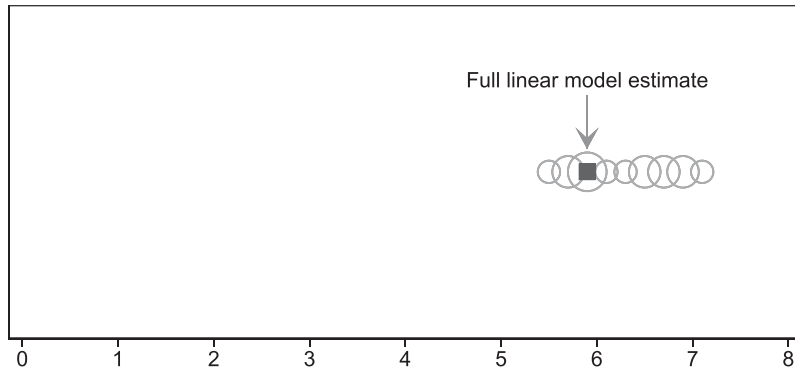
Either way, everyone agrees that the study shows (and the corrected data can be replicated to show) an effect of deworming treatment on direct recipients of medication, on children in the same school, and on children within three kilometers; this is true both in terms of moderate-to-heavy worm infections and in terms of school attendance. In fact, in the published version of the pure (verification) replication, the replication authors wrote, "We do note that some parameters suggest effects may be present at distances of up to 3 km", ([Aiken et al. 2015](#)).

So why the worm wars?

Alternative Analysis (Robustness Reanalysis)

The second exercise undertaken in the replication effort was to re-organize the presentation of Miguel and Kremer's experiment, and to re-analyze the data from Miguel and Kremer's study, in a way more consistent with epidemiological reporting, so that it might be more easily assessed for incorporation in systematic reviews. This part of the project, reported on in [Davey et al. \(2014\)](#), makes several analytical decisions that differ from those in the original paper. These are detailed in [Hicks et al. \(2015\)](#) and in a blog post by [Özler \(2015\)](#). One early decision was to ignore the possibilities of cross-school externalities, focusing on naïve treatment effects. This may drive estimates toward zero, but because the naïve effect was clearly present in the original study, this does not seem to do much harm to the possibility of a treatment effect.

Figure 4. Coefficient Estimates: Percentage Point Impact on School Attendance



Source: Adapted from Hicks et al. 2015, figure 2, panel A.

Beyond that, five decisions follow: covariates, eligibility, treatment definition, weights, and splitting the sample.

The first decision was whether or not to control for additional covariates (as Miguel and Kremer had). A second decision was whether to restrict attention to pupils eligible to receive deworming medications (excluding girls over the age of 13 who were not eligible for deworming treatment under the medical recommendations of that time), or to include the full set of pupils (as Miguel and Kremer had). A third decision concerned how to handle the dates of deworming: Davey et al. (2014) state that their “interpretation of the study” was that treatment had been intended to start at the exact beginning of each calendar year, leading them to consider attendance observations done in a January or February prior to a March deworming round as “dewormed,” when deworming had not yet arrived, in the spirit of an “intention to treat” analysis, if this was indeed the intention.²³ Hicks, et al. (2015) dispute that this was ever the intention; Miguel and Kremer consider observations prior to the first deworming treatment as “not dewormed.” A fourth decision concerned the weighting of observations: Davey, et al. (2014) note that attendance checks are not perfectly evenly distributed across schools and pupils. To address this issue, Hicks, et al. (2015) suggest the alternative of weighting the analysis equally by pupil rather than by attendance observation (as Miguel and Kremer had).

Analyzing the data through each of the 16 combinations of possible approaches that result from the four decisions above produces a range of point estimates near the original: roughly a 6-percentage-point effect on attendance. As shown in figure 4 (adapted from Hicks et al. 2015), the distribution of these 16 estimates is not spread far from the original estimate (larger circles indicate multiple estimates near that value). All resulting estimates are positive, all fall between 5.5 and 7.5 percentage

points, and all are statistically significant with a p-value less than 0.001 (Hicks et al. 2015).

Davey et al. (2014) opt for a different approach to the fourth decision, weighting the analysis by school rather than by pupil, leading some study participants (in smaller schools) to have seven times the analytical importance of others. Hicks et al. (2015) and Davey et al. (2014) disagree on the right approach here, and Özler (2015) and others have discussed the meaning behind the different weights, but a typical consequence of using widely varying weights is the reduction of statistical power: that is, the study is effectively driven by a smaller number of observations. There is a fifth decision that Davey et al. (2014) face: whether to use all years of data simultaneously (as Miguel and Kremer had), or whether to examine a single year of data at a time. Noting that the point estimate from a specification pooling across the two years is not equal to an average of the two separate within-year estimates, they opt for the latter approach, splitting the sample into two separate tests, which also has the natural consequence of reducing the number of observations in each test. This again means further reducing statistical power—the chance of finding an effect if it is there. When all five of these decisions are made together, the result (the impact of deworming on school attendance) is finally no longer statistically significant.²⁴ Of this battery of robustness tests, Özler (2015) wrote: “if anything, I find the findings of the original study more robust than I did before. ... a number of unconventional ways of handling the data and conducting the analysis are jointly required to obtain results that are qualitatively different than the original study.”

The replication authors themselves explained much of this in their paper. In the subsequently published paper’s abstract, they wrote, “In year-stratified cluster-summary analysis, there was no clear evidence for improvement in ... school attendance. In year-stratified regression models, there was some evidence of improvement in school attendance ... When both years were combined, there was strong evidence of an effect on attendance ...” (Davey et al. 2015).

Thus, though the re-analysis found that the study could be split into separate studies too small to provide definitive evidence, all the data combined to show the effect originally reported.

Aggregating Deworming Findings (“Systematic Review”)

Why spend too much energy focusing on a single study, such as Miguel and Kremer (2004), when the robustness of a finding across studies is probably more informative? The worm wars still might not have taken place, had it not been for the simultaneous release of an update to the Cochrane Collaboration’s systematic review on deworming (Taylor-Robinson et al. 2015).

As described earlier, systematic reviews are challenging to undertake, and the guidelines for conducting them are extensive. In the present case, the authors of the 160-page 2015 Cochrane Collaboration systematic review of deworming clearly worked very hard to cover the subject at hand. Yet shortly after its publication, a range of critiques appeared in print. One critique pointed out that different levels of worm prevalence prior to treatment should change effect sizes, so the aggregation of low-prevalence and high-prevalence studies is both underpowered and pursues an average whose meaning may not be relevant to any particular context (de Silva et al. 2015). Another critiqued the review for being restricted to short-duration studies, thus missing any long-term effects (Montresor et al. 2015). Whatever the critiques, the authors of the systematic review are constrained both by the definitions they choose, and by the literature they are trying to summarize. For the effects of deworming on formal tests of cognition, they are able to find only five studies. For effects on school attendance, only two.

How do they treat Miguel and Kremer (2004)? Recall that the text of Aiken et al. (2015) stated that, upon re-examination, “effects may be present at distances of up to 3 km.” However, in their *abstract*, Aiken et al. (2015) write that “For school attendance, re-analysis showed benefits similar to those originally found in intervention schools for both children who did and those who did not receive deworming drugs. However, . . . there was little evidence of an indirect effect on school attendance among schools close to intervention schools.” So, though Aiken et al. (2015) are thorough in their text, their abstract shortens the message to a more negative one in terms of spillovers. The associated Cochrane Review text (Taylor-Robinson et al. 2015, p. 10) reads: “The indirect effects of the intervention on adjacent schools disappeared.” A press release from Aiken and Davey’s home institution then appeared simultaneously with the Cochrane Review in July of 2015 (LSHTM 2015), proclaiming that “deworming children may not improve school attendance.”

In summary: two very detailed and nuanced replication reports appeared, with commentary from the original study authors. The abstracts of two very detailed replication efforts took a relatively negative view of the replication’s outcome, despite the very different interpretation that the original study authors and some consumers of the replication came to. The Cochrane review took the pessimism to heart, a press release led with the negative angle, *The Guardian* headline followed, and with it, the “worm wars.”

Perhaps the over-simplified message of the press release is a cautionary tale about producing inaccurate summaries of research; perhaps the incredible ambition of this replication effort was simply too much; or perhaps simply with modern trial registries and widespread data availability requirements, this series of misunderstandings would not be likely to happen again.

Lessons

We have seen how the replication, reanalysis, and review that considered the work of [Miguel and Kremer \(2004\)](#) made substantial progress in pointing out the ways analysis may or may not be robust, but what can others take from this, in their efforts toward transparency and reproducibility? And where, again, does this leave deworming?

How Should Replication Work?

Citing several examples including this one, [Clemens \(2017\)](#) wrote about the broader category into which this situation falls, regarding what separates a replication from a robustness check: “Confusion in the meaning of replication harms research... anyone can find ‘plausible’ ways to change someone else’s regression so that coefficient estimates change... A well-known form of this problem is that it is a simple matter to change any result into a null result by running modified versions of the same test that are underpowered by construction.”

Indeed, in this case, changing the definition of treatment, splitting the sample, and applying uneven weights does reduce the power of the study by construction.²⁵ What possible benefit is there in approaching a study in this way? Two recent examples provide some insight. [Galiani, Gertler, and Romero \(2018\)](#) point out (via a survey of editors) that academic journals are part of this: while all editors surveyed indicated interest in publishing a replication study that overturned the results of the original, only one-quarter were interested in replications that upheld the original result.²⁶ Replication teams, sometimes perceived as setting out simply to overturn others’ work, have even been called “replication bullies” for this ([Meyer and Chabris 2014](#)). This may seem pernicious, but journals are intended to print new findings; a successful re-running of someone’s code is just not very much new information (particularly since most of these efforts are successful given code availability, as in [figure 3](#)), but the discovery of a problem in re-running an analysis is new information. The problem is not limited to editors’ survey responses: Gertler, as chair of 3ie’s board of directors, saw many of the 3ie replications unfold firsthand. He and his coauthors report that of 27 replication studies 3ie commissioned, 20 upheld the original results, while 7 reported being “unable to fully replicate the results in the original article.” Galiani, Gertler, and Romero note: “The only replication published in a peer-reviewed journal claimed to refute the results of the original paper,” citing the present case of [Davey et al. \(2015\)](#).²⁷

Before making any recommendation, however, Galiani, Gertler, and Romero have led by example, following in a long tradition. Replication efforts that focus on a single paper face a problematic set of incentives surrounding what makes an interesting publication. Replication efforts that group together a large number of papers have

more options, particularly in relation to how that work is distilled in short summaries and media reports: the fraction of a category of papers that proves to be replicable, by some standard, is of interest to readers, no matter whether it is high or low. This kind of multi-paper exercise can also be combined with other types of analysis. [Galiani, Gertler, and Romero \(2018\)](#) examined hundreds of papers, combined their analysis with a survey of journal editors, and published their analysis in *Nature*. [Camerer et al. \(2016\)](#), for example, examined 18 prominent lab-experimental economics papers for their results' reproducibility using a new set of experimental participants, and combined it with predictions of replicability. Eleven (61 percent) of 18 studies' results were replicated successfully; the paper describing the effort was published in *Science*. [Camerer et al. \(2018\)](#), examined 21 experiments across the social sciences, found that 13 (62 percent) replicated in a general sense, but discovered that the replicated effects were generally smaller than the originals; this analysis was published in *Nature Human Behaviour*.

[Humphreys \(2015\)](#) observes that slight changes in levels of statistical significance in a single re-visited study should not be able to radically shift the academic or policy community's beliefs (or the findings of an analysis across many studies). If we are doing science right, the reasoning goes, the stakes associated with a single replication should not be that high: pushing some individual study's p-value just past a salient or popular threshold so that the finding is "insignificant" should not really change our beliefs that much—unless it is the only study of its kind in the world, and little else informs our beliefs about the effect in question. Putting a related idea into practice, coupling a measure of expert beliefs with a replication project is exactly what [Camerer et al. \(2016\)](#) did.

Galiani, Gertler, and Romero's main recommendation focuses on the way journals structure incentives. If journals require some combination of public data availability and data checks in the review process, much of the replicability "problem" might disappear. This sort of requirement is empirically important: [Chang and Li \(2015\)](#) report that they are able to obtain replication data more often for papers published in journals with a data availability requirement than for those published in journals without such a requirement.

However, not all disciplines and journals are approaching this the same way. The data availability and replication policy that would have mandated that Miguel and Kremer's data be available for others to explore two years sooner is now in place at *Econometrica*, as it is across most top journals in economics, shown in Galiani, Gertler, and Romero's "Data checked?" figure. Many political science journals require this as well, though this is currently much less common in sociology or psychology. Notably, in relation to the present case, the journal which published the replication of the deworming study, the *International Journal of Epidemiology*, presently has no replication or data availability policy ([International Journal of Epidemiology 2019](#)).

What about Deworming?

Since the Cochrane Review, two other reviews have appeared on the topic of deworming: that of the Campbell Collaboration ([Welch et al. 2017](#)), and that of [Croke et al. \(2016\)](#). The biggest difference across these studies, in the area that they all cover (the outcome of weight gain) is precision. Croke, et al. take pains to obtain raw data from authors whose studies are not immediately amenable to meta-analysis, thereby including more studies; methodologically, they also weight the studies in a way that does not allow small-sample-size, imprecise studies to drown out larger-scale, more precise ones. Croke, et al., also focus on populations with high enough worm prevalence for the World Health Organization to suggest mass deworming in those settings. The result is that Croke, et al., find a positive effect of deworming on weight, where neither of the other two efforts does. The other two efforts' estimates are much less precise, however: they cannot reject that the quantity that Croke et al. estimate is equal to the one they estimate, but they cannot reject the null effect either. So, deworming has substantial benefits—for example, on the growth of children—but whether a systematic review detects them depends on whether it requires there to be worm infections to treat, how many papers it includes, and how detailed the subsequent statistical work is.²⁸

Besides precision, one notable gap in the Cochrane Collaboration's review is of long-term studies. The number of studies gathering outcome data six years or more after deworming and included in the 2015 review? Zero. Not because there are no such studies, but because they did not satisfy certain details of the inclusion criteria. The following three published studies on this topic all find large benefits of deworming: [Bleakley \(2007\)](#), who finds wide-ranging benefits of deworming in the U.S. south; [Baird et al. \(2016\)](#), who find the participants of the [Miguel and Kremer \(2004\)](#) study earning more money as adults; and my own work, [Ozier \(2018\)](#), in which I follow the younger siblings and neighbors of the children in the Miguel and Kremer study, and find that children exposed to deworming spillovers in early childhood grow up to perform much better on cognitive tests a decade later. The exclusion of these long-term studies, as well as the consequential statistical work by [Croke et al. \(2016\)](#), both illustrate how the medical approach to systematic review may hobble our understanding in matters of public health.

In terms of policy, one must weigh expected costs against expected benefits. If there is any uncertainty around our beliefs about a program's effect, then “the costs of proceeding when there is still substantial doubt as to the outcome needs to be weighed against the cost of missing an intervention that may be valuable,” as one anonymous referee of this manuscript has pointed out. In the case of deworming, one must further consider the likelihood that public (or NGO) financing of deworming medications influences uptake. [Ahuja et al. \(2017\)](#) illustrate that even under pessimistic interpretations of the re-analysis of [Miguel and Kremer \(2004\)](#), the incredibly low

cost of deworming relative to other interventions still makes it highly cost-effective in high-worm-prevalence settings.²⁹ Ahuja et al. (2017) also point out that preventative health investments are precisely the ones that are least likely to be invested in by the poor; Kremer and Glennerster (2011) discuss how up-front costs for health technologies—deworming medication, bed nets to protect against malaria, clean water, and so forth—sharply reduce take-up, even when the cost itself is very low.

The “worm wars” teach us that scientific inquiry is difficult, and that those of us producing and communicating our findings do make mistakes along the way. Deworming remains lower-cost than almost any other intervention, and our best estimates still suggest that it has lasting benefits. Because it requires up-front investments and has substantial positive externalities, people are likely to underinvest in it as individuals. The World Health Organization and other coordinating entities can continue to play a role in ensuring that deworming benefits do reach people, however. As for the replication effort? After all is said and done, most of the original study findings hold up, but the externality benefits probably do not reach as far as 6 km. That just does not make a very exciting headline.

Notes

Owen Ozier is an Associate Professor at Williams College, on leave from the Development Research Group of the World Bank. Email: oo3@williams.edu. The author thanks Joan Hamory Hicks, Pam Jakiela, Ricardo Maertens, Lance Ozier, Linda Ozier, Berk Özler, and especially Adam Wagstaff for helpful comments on earlier drafts, as well as to David Evans, Alan Fuchs, and Jeff Tanner for providing opportunities to present this work. Also, in the interest of full disclosure, the author is grateful to both Michael Kremer and Ted Miguel for hiring him to work in Kenya 16 years ago, and for their guidance since then, though they have not commented on this manuscript. The views expressed are the author’s, and do not represent those of the World Bank, its Executive Directors, or the governments they represent.

1. Of reproducing a result, Boyle (1673) wrote: “It is much more difficult than most men can imagine, to make an accurate Experiment [sic],” (100).

2. Famously, with no agreement on the research design or “identification strategy,” and no agreed-upon source of experimental variation, different researchers may use the same observational data set to come to different conclusions. In a well-known case, 29 research teams all used the same observational dataset to arrive at somewhat differing results; only one of the teams is recorded even mentioning the phrase “identification strategy,” however (Silberzahn et al. 2018).

3. Whether a finding is similar across contexts is related to the question of whether a particular study is “externally valid,” and to whether a finding *ought* to be similar across contexts, issues discussed by Allcott (2015), Peters, Langbein, and Roberts (2018), and others. Reflecting on this, Gene Glass wrote, “Where ten studies might suffice to resolve a matter in biology, ten studies on computer assisted instruction or reading may fail to show the same pattern of results twice,” (Glass 1976).

4. Dewald, Thursby, and Anderson (1986) found that replication datasets were twice as likely to be available if the concerned papers were in the process of being published at the time of the replication attempt, compared to papers that had been published years before the attempt.

5. Datasets and programs failed to combine to produce viable and successful replications for a list of reasons that is, by turns, tragic, comic, incredible, and instructive. Dewald, Thursby, and Anderson (1986) describe a common situation: a regularly updated government dataset is used in analysis, but

neither a copy of the relevant vintage of the public dataset nor the precise date when it was obtained is included in the replication files, thus preventing would-be replicators from knowing whether the dataset they obtain is the same as what the original study authors had used. [Chang and Li \(2015\)](#) mention, as the 1986 paper also did, the occasional problem of confidential datasets and unavailable software packages. Alongside uncommented computer programs and unintuitive variable names, [McCullough, McGeary, and Harrison \(2006\)](#) describe cases of forgotten subroutines, cases in which “the person responsible for archiving the data and code stopped doing this part of his job,” a case of an ASCII data file in which “we are supposed to guess the names of the variables,” and one case in which the original study author had included the pessimistic caveat that the program supplied to replicators was “not necessarily the one that produced the results reported in the paper.” The author was right: it was not.

6. [Chang and Li \(2015\)](#) also noted that journal requirements mattered: replication a data set was twice as likely to be available when the journal required it, compared to when the journal did not.

7. [Dewald, Thursby, and Anderson \(1986\)](#), [McCullough, McGeary, and Harrison \(2006\)](#), [Chang and Li \(2015\)](#), and [Galiani, Gertler, and Romero \(2018\)](#) all describe re-running programs on data as “replication.” [Clemens \(2017\)](#) points out that, in the context of the *American Economic Review*, “the term replication unambiguously means using the original data and code to get exactly the same results as appear in the paper.”

8. [Leek and Peng \(2015\)](#) begin their article: “Reproducibility—the ability to recompute results—and replicability—the chances other experimenters will achieve a consistent result—are two foundational characteristics of successful scientific research.” [Baker \(2016b\)](#) suggests that, in the sciences, there is not yet wide agreement regarding the precise use of terminology around reproducibility. [Barba \(2018\)](#) documents the diverse use of the terms “reproducible” and “replication” across a range of funding agencies and scientific disciplines.

9. [Goodman, Fanelli, and Ioannidis \(2016\)](#) recognize that “reproducibility,” “replicability,” and “repeatability” are all synonymous enough in common English that to impose special technical meanings upon them in any scientific discipline may be fruitless. These authors propose using the terms “methods reproducibility,” “results reproducibility,” and “inferential reproducibility,” but quickly point out the many unsettled boundaries that these proposed terms leave unresolved. [Hamermesh \(2007\)](#) describes, on one end of the spectrum, “pure replication,” which amounts to “checking on others’ published papers using their data,” and “scientific replication,” which can go further in any of a few ways. [Reed \(2017\)](#) and [Clemens \(2017\)](#) each point out that there are at least two dimensions to the characteristics of these terms, when applied to new analysis following an earlier study. One dimension is whether the new analysis involves the same population and/or dataset as the original study. The other dimension is whether the analytical approach is the same as before, and thus whether the parameters being estimated really ought to be the same as before.

10. Note that the WHO already supported community-level deworming treatment prior to the publication of the 2004 study. See, for example, [WHO \(1987\)](#), table 3.

11. See [Miguel and Kremer \(2004\)](#), table IX, column 1.

12. Though the journal that had published the paper, *Econometrica*, had no stated replication policy when MK submitted or published their paper, by 2005 it had begun to require (of new submissions) that data be made available when possible. Miguel and Kremer began making their data public in 2007.

13. [Davey et al. 2014](#), write, “The analysis in this report comprises an ‘internal statistical replication’ and an ‘internal scientific replication’. We use the term ‘internal statistical replication’ to mean a reanalysis of the study’s original hypotheses using different handling of the same raw data (e.g., different variable constructs, different data handling). We use the term ‘internal scientific replication’ to mean the introduction of a (different) explicit causal framework to guide analysis and interpretation of the statistical results, similar to the ‘theory of change’ process (Vogel 2012). We have used the qualifier ‘internal’ to differentiate the statistical and scientific replication analyses in this report from replication work involving the collection of new data. [Hamermesh \(2007\)](#) [sic] uses these terms without the “internal” qualifier to describe what we would describe as “external replication”, which uses new samples or data on different populations.”

14. The communication styles are remarkably different across disciplines. As one example, “bias,” in the replication reports, may be intended to mean anything that could alter findings, in the sense used by [Moher et al. \(1995\)](#): “to minimize bias . . . the quality assessor should not know the (masked) identity of the trial’s author(s). . . .” This is much broader than the specific meaning of “bias” described in a standard statistics text: see, for example, [Casella and Berger \(2002\)](#), p.330. [Humphreys \(2015\)](#), a political scientist commenting on the “worm wars,” also commented on difficulty understanding the replication’s use of the term “bias” from his own disciplinary perspective. As a further example of the difference in communication styles between disciplines: the abstract to the alternative analysis includes a word, “coprimary,” which appears with some regularity in every one of *JAMA*, *BMJ*, *Lancet*, and *New England Journal of Medicine*, but has never appeared in the title, abstract, or text of any paper published in any of the so-called top five journals in economics (*AER*, *Econometrica*, *JPE*, *QJE*, *Review of Economic Studies*). On the other hand, the word “externality” appears in the abstract of the original [Miguel and Kremer \(2004\)](#) article—just as it occurs hundreds of times in each of the most prominent journals in economics—but the word “externality” almost never appears in medical journals.

15. Norms are changing in economics as in other disciplines: the American Economics Association began offering trial registration in 2012, for example ([American Economic Association 2018](#)).

16. Much of the exchange is documented in the documents both available online at 3ie’s website and published subsequently in the *International Journal of Epidemiology*; these materials are cited in the bibliography at the end of this paper.

17. As a historical comparison, five of the seven papers that a 1986 effort was able to replicate still did not replicate exactly: there were discrepancies in some calculations, and programming errors were found along the way, described then as “some minor, some serious,” ([Dewald, Thursby, and Anderson 1986](#), p. 594).

18. A variety of minor problems are described. This includes rounding errors in which 0.787 became 0.78 rather than 0.79; cases of sequential rounding resulting in errors, for example, in which 0.7745 first became 0.775 and then 0.78 rather than becoming 0.77 as it should have; and simple cases of annotating coefficients with the level of statistical significance in the published paper. MK also mention incorrect reported numbers of observations: 1,467 was listed where there should have been 1,466, and so forth. One variable having to do with worm infection incorrectly mapped the underlying egg counts to the thresholds associated with “moderate-to-heavy” infection, though the original and corrected variables only differ in a small fraction of cases.

19. For those interested, Appendix A shows the numbers underlying this discussion in considerable detail.

20. As the insidiousness of “p-hacking” became well-known, some well-known papers were retracted; the popular press article by [Rosenberg and Wong \(2018\)](#) provides one immediate example. In economics, the pervasiveness of this pattern (whether perpetrated by authors, referees, editors, or a combination thereof) varies with study characteristics. Randomized trials display relatively less of this problem, for example, as [Brodeur et al. \(2016\)](#) show in their figure 6.

21. An in-depth discussion of the worm wars that also emphasizes this point is provided by [Humphreys 2015](#).

22. The intuition behind this correction is that, whatever vanishingly small probability there is of estimating a very statistically significant effect when one really isn’t there, the odds of finding it in either of two separate attempts are about twice that, if the tests are not correlated with one another. More recently developed (and more statistically powerful) corrections are also possible, but this is perhaps the simplest and most conservative. Olive Jean Dunn published on this correction in 1957 and 1958. Her work builds on, and refers to, Carlo Emilio Bonferroni’s work in the 1930s, but to the best of my knowledge, his work did not include the correction itself. [Stigler \(1980\)](#) suggests that, in statistics, discoveries are named after someone other than their discoverer; but not having been able to quickly discern who actually proposed this correction, I mention them both.

23. Given no evidence that exact calendar year timing was ever intended, this assertion of what [Hicks et al. \(2015\)](#) describe as an “incorrect” definition of treatment has been termed a “very unusual”

choice by the replication team (Özler 2015). Intention-to-treat analyses may, in some settings, be usefully employed as safeguards against manipulative deviations from a protocol, but their application here does not seem likely to have achieved that purpose (Blattman, HELLERINGER, and ÖZLER 2015).

24. See Davey et al. (2014), table 4, as well as Hicks et al. (2015) appendix table S7.

25. To “underpower,” or to “reduce power,” in this context, means making estimates less precise, and thus less likely to be statistically different from zero even when an effect is present.

26. Prior to Glass (1976), Galiani, Gertler, and Romero (2018) quoted Lewis M. Branscomb as writing, “when professional advancement and peer recognition are so heavily oriented toward original discovery, and research funding is largely restricted to original . . . research, it is hard to motivate a scientist to write scholarly reviews.” Dewald, Thursby, and Anderson (1986) also paraphrased Thomas S. Kuhn in a similar vein: “Thomas Kuhn (1970) emphasized that replication – however valuable in the search for knowledge – does not fit within the ‘puzzle-solving’ paradigm which defines the reward structure in scientific research.”

27. After the present paper had been drafted and revised for publication, several more of the 3ie-funded replications were published in a special issue of the *Journal of Development Studies* edited by Annette N. Brown and Benjamin D. K. Wood, directors of 3ie’s replication programme (Brown and Wood 2019).

28. An excellent visualization of these confidence intervals is provided by Roodman (2017).

29. Note that the full list of authors of this study is Ahuja, Baird, Hicks, Kremer, and Miguel.

References

- AEA. 2018. “AEA RCT Registry.” Accessed March 16, 2018. <https://www.socialscisearch.org/site/about>.
- Ahuja, A., S. Baird, J. H. Hicks, M. Kremer, and E. Miguel. 2017. “Economics of Mass Deworming Programs.” In *Disease Control Priorities (third edition): Volume 8, Child and Adolescent Health and Development*, edited by D. Bundy, N. de Silva, S. Horton, D. T. Jamison, G. Patton. Washington, DC: World Bank.
- Aiken, A. M., C. Davey, R. J. Hayes, and J. Hargreaves. 2013. Deworming Schoolchildren in Kenya - Replication Plan. International Institute Impact Evaluation (3ie). Accessed March 5, 2018. http://www.3ieimpact.org/media/filer_public/2013/05/14/aiken_replication_plan_final.pdf.
- Aiken, A. M., C. Davey, J. Hargreaves, and R. J. Hayes. 2014. “Reanalysis of Health and Educational Impacts of a School-based Deworming Program in Western Kenya: Part 1, Pure Replication.” 3ie Replication Paper 3, part 1. Washington, DC: International Initiative for Impact Evaluation (3ie). Accessed March 9, 2018. http://www.3ieimpact.org/media/filer_public/2015/01/07/3ie_rps3_worms_replication_1.pdf.
- . 2015. “Reanalysis of Health and Educational Impacts of a School-based Deworming Program in Western Kenya: A Pure Replication.” *International Journal of Epidemiology* 44 (5): 1572–80.
- Allcott, H. 2015. “Site Selection Bias in Program Evaluation.” *Quarterly Journal of Economics* 130 (3): 1117–65.
- Anderson, M. L., and J. Magruder. 2017. “Split Sample Strategies for Avoiding False Discoveries.” (mimeo, No. w23544). National Bureau of Economic Research.
- Baird, S., J. H. Hicks, M. Kremer, and E. Miguel. 2016. “Worms at Work: Long-run Impacts of a Child Health Investment.” *Quarterly Journal of Economics* 131 (4): 1637–80.
- Baker, M. 2016a. “Is there a Reproducibility Crisis? A Nature Survey Lifts the Lid on how Researchers View the ‘crisis’ Rocking Science and What They Think Will Help.” *Nature* 533 (26): 353–66.
- . 2016b. “Muddled Meanings Hamper Efforts to Fix Reproducibility Crisis.” *Nature News*. doi:10.1038/nature.2016.20076.

- Barba, L. A. 2018. "Terminologies for Reproducible Research." *arXiv preprint* arXiv:1802.03311.
- Blattman, C., S. Helleringer, and B. Özler. 2015. Comments on "Dear Journalists and Policy-makers: What You Need to Know about the Worm Wars." *Chris Blattman blog*. Accessed April 9 2018. <https://chrisblattman.com/2015/07/23/dear-journalists-and-policy-makers-what-you-need-to-know-about-the-worm-wars/#comment-192758>.
- Bleakley, H. 2007. "Disease and Development: Evidence from Hookworm Eradication in the American South." *The Quarterly Journal of Economics* 122 (1): 73–117.
- Boyle, R. 1673. *Certain Physiological Essays And Other Tracts: Written at Distant Times, and on Several Occasions By the Honourable Robert Boyle.... Wherein Some of the Tracts are Enlarged by Experiments, and the Work Is Increased by the Addition of a Discourse about the Absolute Rest in Bodies*.
- Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg. 2016. "Star Wars: the Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.
- Brown, A. N., and B. D. K. Wood. 2019. "Replication Studies of Development Impact Evaluations." *Journal of Development Studies* 55 (5): 917–25.
- Camerer, C. F., A. Dreber, E. Forsell, T. H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, and E. Heikensten. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–6.
- Camerer, C. F., A. Dreber, F. Holzmeister, T. H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, and A. Altmejd. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015." *Nature Human Behaviour* 2: 637–44.
- Casella, G., and R. L. Berger. 2002. *Statistical Inference* (second edition). Pacific Grove, CA: Duxbury.
- Chang, A. C., and P. Li. 2015. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not.'" *Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington, D.C.* Accessed March 7, 2018. <http://dx.doi.org/10.17016/FEDS.2015.083>.
- Clemens, M. A. 2017. "The Meaning of Failed Replications: A Review and Proposal." *Journal of Economic Surveys* 31 (1): 326–42.
- Coffman, L., and M. Niederle. 2015. "Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29 (3): 81–98.
- Croke, K., J. H. Hicks, E. Hsu, M. Kremer, and E. Miguel. 2016. "Does Mass Deworming Affect Child Nutrition? Meta-analysis, Cost-effectiveness, and Statistical Power." *NBER Working Paper No. w22382*, National Bureau of Economic Research.
- Davey, C., A. M. Aiken, R. J. Hayes, and J. Hargreaves. 2014. "Reanalysis of Health and Educational Impacts of a School-based Deworming Program in Western Kenya: Part 2, Alternative Analyses." 3ie Replication Paper 3, part 2. Washington, DC: International Initiative for Impact Evaluation (3ie). Accessed March 9, 2019. http://www.3ieimpact.org/media/filer_public/2015/01/07/rps_3_part_2_top_copy_reduced_size_1_7_15-top.pdf.
- . 2015. "Reanalysis of Health and Educational Impacts of a School-based Deworming Program in Western Kenya: A Statistical Replication of a Cluster Quasi-randomized Stepped-wedge Trial." *International Journal of Epidemiology* 44 (5): 1581–92.
- De Janvry, A., and E. Sadoulet. 2015. *Development Economics: Theory and Practice*. Routledge.
- de Silva, N., B. Ahmed, M. Casapia, H. J. de Silva, J. Gyapong, M. Malecela, and A. Pathmeswaran. 2015. "Cochrane Reviews on Deworming and the Right to a Health, Worm-Free Life." *PLOS Neglected Tropical Diseases* 9 (10): e0004203.
- Dewald, W., J. Thursby, and R. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *American Economic Review* 76 (4): 587–603.

- Evans, D. K. 2015. "Worm Wars: The Anthology" *Development impact blog*. Accessed March 9 2018. <https://blogs.worldbank.org/impactevaluations/worm-wars-anthology>.
- Evans, D. K., and A. Popova. 2016. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Research Observer* 31 (2): 242–70.
- Fafchamps, M., and J. Labonne. 2017. "Using Split Samples to Improve Inference on Causal Effects." *Political Analysis* 25 (4): 465–82.
- Fisher, R. A. 1934. *Statistical Methods for Research Workers* (fifth edition). Edinburgh: Oliver and Boyd.
- . 1935. *The Design of Experiments*. Edinburgh; London: Oliver and Boyd.
- Galiani, S., P. Gertler, and M. Romero. 2018. "How to Make Replication the Norm." *Nature* 554 (7693): 417–9.
- Glass, G. V. 1976. Primary, Secondary, and Meta-analysis of Research. *Educational Researcher* 5 (10): 3–8.
- Glennerster, R., and K. Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press.
- Goldfarb, R. S., and H. O. Stekler. 2002. Meta-analysis. *The Journal of Economic Perspectives* 16 (3): 225–6.
- Goodman, S. N., D. Fanelli, and J. P. A. Ioannidis. 2016. "What Does Research Reproducibility Mean?" *Science Translational Medicine* 8 (341ps12): 1–6.
- Hamermesh, D. S. 2007. "Replication in Economics." *Canadian Journal of Economics/Revue canadienne d'économie* 40 (3): 715–33.
- Hicks, J. H., M. Kremer, and E. Miguel. 2015. "Commentary: Deworming Externalities and Schooling Impacts in Kenya: A Comment on Aiken et al. (2015) and Davey et al. (2015)." *International Journal of Epidemiology* 44 (5): 1593–6.
- Higgins, J. P. T., and Green S., eds. 2006. *Cochrane Handbook for Systematic Reviews of Interventions* 4.2.6 Accessed February 15, 2019. <https://training.cochrane.org/handbook>.
- Humphreys, M. 2015. *What Has Been Learned from the Deworming Replications: A Nonpartisan View*. Accessed March 9, 2018. <http://www.columbia.edu/~mh2245/w/worms.html>.
- International Journal of Epidemiology. 2019. "Instructions to Authors." Accessed March 9, 2018. https://academic.oup.com/ije/pages/Instructions_To_Authors.
- Ioannidis, J. P. A., T. D. Stanley, and H. Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127 (October): F236–65.
- ISRCTN. 2018. "ISRCTN - About." Accessed March 14, 2018. <https://www.isrctn.com/page/about>.
- J-PAL. 2012. "Deworming: A Best Buy for Development." *J-PAL Policy Bulletin*. March. Accessed March 9, 2018. <https://www.povertyactionlab.org/sites/default/files/publications/2012.3.22-Deworming.pdf>.
- Kremer, M., and R. Glennerster. 2011. Improving Health in Developing Countries: Evidence from Randomized Evaluations. In *Handbook of Health Economics* 2: 201–315. Elsevier.
- Leek, J. T., and R. D. Peng. 2015. "Opinion: Reproducible Research can Still be Wrong: Adopting a Prevention Approach." *Proceedings of the National Academy of Sciences* 112 (6): 1645–1646.
- LSHTM. 2015. "Educational Benefits of Deworming Children Questioned by Re-analysis of Flagship Study." Accessed March 9, 2018. https://www.lshtm.ac.uk/newsevents/news/2015/educational_benefits_of_deworming_children_questioned.html.
- Maniadis, Z., and F. Tufano. 2017. "The Research Reproducibility Crisis and Economics of Science." *The Economic Journal* 127 (October): F200–8.
- McCullough, B. D., K. A. McGeary, and T. D. Harrison. 2006. "Lessons from the JMCB Archive." *Journal of Money, Credit, and Banking* 38 (4): 1093–107.

- Meyer, M. N., and C. Chabris. 2014. "Why Psychologists' Food Fight Matters." *Slate*, July 31. Accessed: February 16, 2019. <https://slate.com/technology/2014/07/replication-controversy-in-psychology-bullying-file-drawer-effect-blog-posts-repligate.html>.
- Miguel, E., and M. Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159–217.
- . 2014. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Guide to Replication of Miguel and Kremer (2004)." Accessed March 6, 2018. http://emiguel.econ.berkeley.edu/assets/miguel_research/46/PSDP-REP__2014-11.pdf.
- Moher, D., A. R. Jadad, G. Nichol, M. Penman, P. Tugwell, and S. Walsh. 1995. Assessing the Quality of Randomized Controlled Trials: An Annotated Bibliography of Scales and Checklists. *Controlled Clinical Trials* 16 (1): 62–73.
- Montresor, A., D. Addiss, M. Albonico, S. M. Ali, S. K. Ault, A.-F. Gabrielli, A. Garba, E. Gasimov, T. Gyorkos, M. A. Jamsheed, B. Levecke, P. Mbabazi, D. Mupfashoni, L. Savioli, J. Vercruyssen, and A. Yajima. 2015. "Methodological Bias Can Lead the Cochrane Collaboration to Irrelevance in Public Health Decision-Making." *PLOS Neglected Tropical Diseases* 9 (10): e0004165.
- Olken, B. 2015. "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.
- Ozier, O. 2018. "Exploiting Externalities to Estimate the Long-term Effects of Early Childhood Deworming." *American Economic Journal: Applied Economics* 10 (3): 235–62.
- Özler, B. 2015. "Worm Wars: A Review of the Reanalysis of Miguel and Kremer's Deworming Study." *Development impact blog*. Accessed March 9, 2018. <http://blogs.worldbank.org/impactevaluations/warm-wars-review-reanalysis-miguel-and-kremer-s-deworming-study>.
- Peters, J., J. Langbein, and G. Roberts. 2018. "Generalization in the Tropics—Development Policy, Randomized Controlled Trials, and External Validity." *World Bank Research Observer* 33 (1): 34–64.
- Petticrew, M., and H. Roberts. 2008. *Systematic Reviews in the Social Sciences: A Practical Guide*. John Wiley & Sons.
- Reed, W. R. 2017. "Replication in Labor Economics." *IZA World of Labor*. doi:10.15185/izawol.413.
- Roodman, D. 2017. "How Thin the Reed? Generalizing from 'Worms at Work'." *GiveWell blog*. Accessed March 16, 2018. <https://blog.givewell.org/2017/01/04/how-thin-the-reed-generalizing-from-worms-at-work/>.
- Rosenberg, E., and H. Wong. 2018. "This Ivy League Food Scientist Was a Media Darling. He Just Submitted His Resignation, the School Says." *Washington Post*, Washington DC, 20 September. Accessed: February 11, 2019. <https://www.washingtonpost.com/health/2018/09/20/this-ivy-league-food-scientist-was-media-darling-now-his-studies-are-being-retracted/>.
- Schultz, K. F., D. G. Altman, and D. Moher. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMJ* 340: c332.
- Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š., F. Bai, C. Bannard, and E. Bonnier et al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1 (3): 337–56.
- Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. "False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66.
- . 2013. "Life after p-hacking." *Meeting of the Society for Personality and Social Psychology*, New Orleans, LA, 17–19 January. Accessed February 16, 2019. <https://ssrn.com/abstract=2205186>.
- Simonsohn, U., J. P. Simmons, and L. D. Nelson. 2014. "Anchoring is Not a False-Positive: Maniadis, Tufano, and List's (2014) 'Failure-to-Replicate' is Actually Entirely Consistent with the Original." Accessed February 15, 2019. <https://ssrn.com/abstract=2351926>.

- Stigler, S. M. 1980. (Merton Festschrift Volume, Ed. T. Gieryn), ed. "Stigler's Law of Eponymy." *Transactions of the New York Academy of Sciences*, Ser. 2, 39: 147–58.
- Stanley, T. D. 2001. "Wheat from Chaff: Meta-Analysis as Quantitative Literature Review." *Journal of Economic Perspectives* 15 (3): 131–50.
- Taylor-Robinson, D. C., N. Maayan, K. Soares-Weiser, D. Donegan, and P. Garner. 2015. "Deworming Drugs for Soil-transmitted Intestinal Worms in Children: Effects on Nutritional Indicators, Haemoglobin, and School Performance (Review)." *Cochrane Database of Systematic Reviews* 7.
- Welch, V. A., E. Ghogomu, A. Hossain, S. Awasthi, Z. A. Bhutta, C. Cumberbatch, R. Fletcher, J. McGowan, S. Krishnaratne, E. Kristjansson, and S. Sohani. 2017. "Mass Deworming to Improve Developmental Health and Wellbeing of Children in Low-Income and Middle-Income Countries: A Systematic Review and Network Meta-Analysis." *The Lancet Global Health* 5 (1): e40–50.
- World Health Organization. 1987. *Prevention and Control of Intestinal Parasitic Infections*. Report of the WHO Scientific Group. WHO Technical Report Series: 749. Geneva: WHO.
- . 2017. *Guideline: Preventive Chemotherapy to Control Soil-Transmitted Helminth Infections in at-risk Population Groups*.

Appendix A1: Recalculating the Headline Number

Panel A of [table A1.1](#) provides excerpted coefficient estimates from a regression of school attendance on a set of right-hand-side variables describing who was dewormed where. Columns (2) and 4 provide the original and revised estimates for a simple specification involving the effect of deworming a child, herself, and the externality effect of deworming each additional thousand children within 3 km. The coefficients do not change appreciably in magnitude, but the direct effect is slightly more precisely estimated (and the spillover effect slightly less precisely estimated) in the revised calculations. Columns (1) and (3) provide the original and revised estimates when the specification also includes the externality effect of deworming each additional 1,000 children between 3 km and 6 km away. The revised calculations include an estimated effect of more distant (3 km–6 km) deworming that is negative and twice as large as before. It remains statistically indistinguishable from zero, consistent with the lack of any known mechanism by which deworming children far away would harm the health of those nearby. But its change in size, from -0.01268 in column (1) to -0.02429 in column (3), is important in what follows.

Panel B of [table A1.1](#) provides the mean number of treated children in two areas: 0 km–3 km, and 3 km–6 km. Again, the 0–3 km row hardly changes, but the 3–6 km row shows more than a doubling in the number of children in this area, from just over 700 to just over 1,600. This has consequences in Panel C of [Table A1.1](#). In column (1), taking into account both 0–3 km and 3–6 km spillovers, Miguel and Kremer had originally found an imprecisely estimated 2.0 percentage point spillover total, leading to a 7.47 percentage point total effect—the “7.5 percentage point increase in school participation” that is well-known from this study. This is the more conservative of the two figures that might have been presented from that version of the data: column (2) shows that without the more distant spillovers, the total effect would have

Table A1.1. Estimated and Calculated Effects on School Participation

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

Note: Calculations above are author's original calculations based on data and replication files provided by Miguel and Kremer (2014), following and expanding upon some parts of Miguel and Kremer (2014). Table B2. Standard errors appear in parentheses: Asterisk * denotes significance at the 10 percent level, ** = 5 percent level; and *** = 1 percent level.

been estimated to be 8.14 percentage points. However, with the corrected data set, the imprecise 2.0 percentage point increase in attendance caused by spillovers at both 0–3 km and 3–6 km distances (standard error 1.35 percentage points) changes sign to become a much more imprecise 1.66 percentage point decrease in attendance (standard error 3.00 percentage points). Adding in the direct effect, the total effect estimate is an imprecisely estimated 3.87 percentage point increase in school attendance. With the corrected data set, column (4) shows that without the more distant (and now less precisely estimated) 3–6 km spillovers, the total effect is estimated to be an 8.48 percentage point increase in school attendance, which is statistically different from zero.