Mapping the Worm Wars: What the Public Should Take Away from the Scientific Debate about Mass Deworming

7/30/15 | Michael Clemens and Justin Sandefur

Global Development: Views from the Center | RCTs, Aid Effectiveness, Evaluation

Leave a comment

A quarter of all people on earth, horrendously, have worms living inside their bodies. That includes 600 *million* school-age children. They get usually get worms from contact with things contaminated by the excrement of infected people, which means that these are diseases inflicted on poor people who lack basic sanitation.

There is a large campaign for "deworming," the mass treatment of children across poor countries to either prevent or reverse worm infections. That campaign is something of a poster child for the effective altruism movement, which seeks to channel philanthropy toward things that have scientific evidence showing they work.

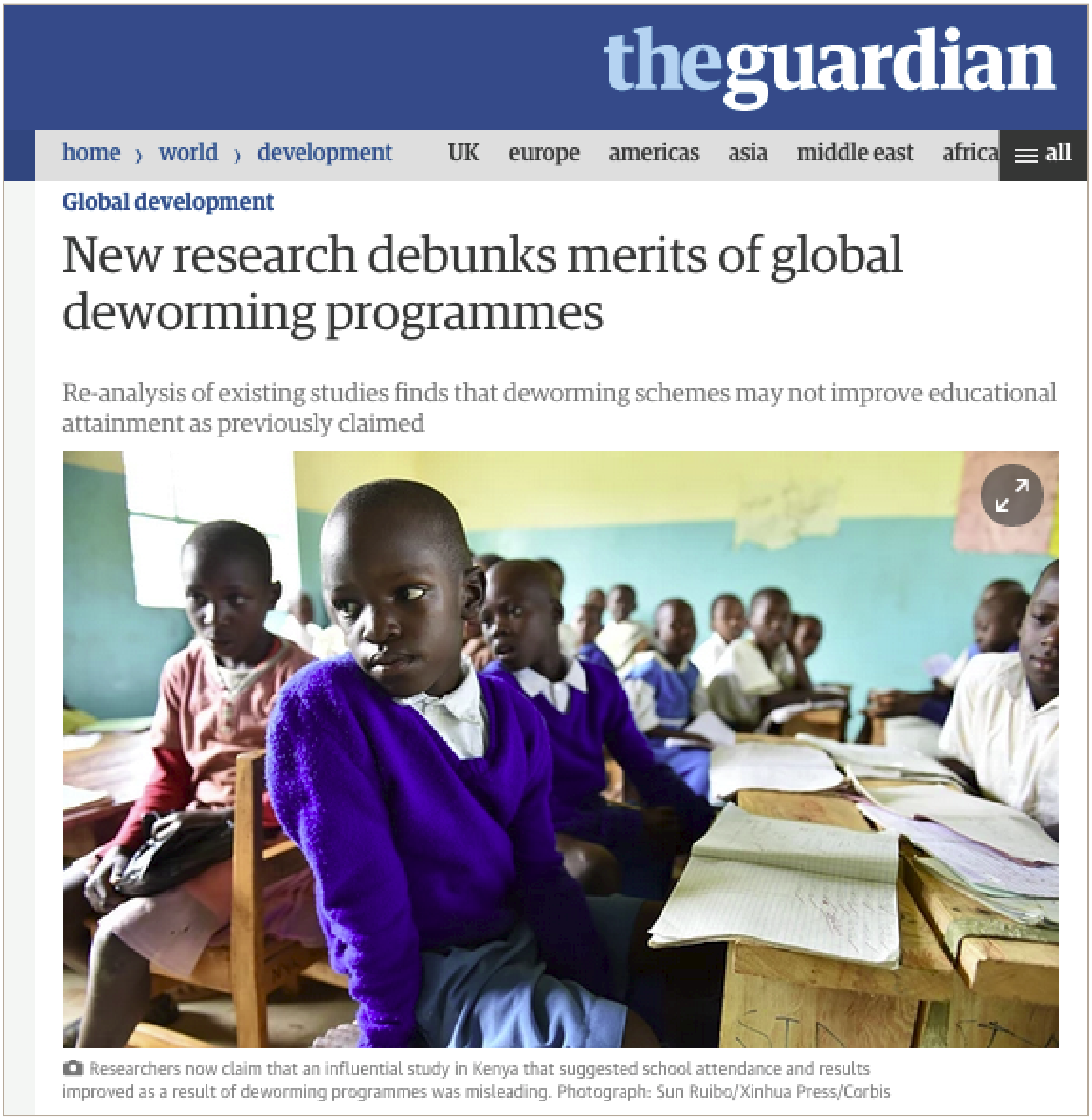
So it was a big deal when various media outlets declared last week that the evidence to support mass deworming had been "debunked." The stories reported on critiques of a famous randomized experiment conducted in Kenya in the late 1990s, and a new summary of evidence from a scientific body known as the Cochrane Collaboration.

The unique feature of that original study from Kenya, by Michael Kremer and Ted Miguel, was that it found important effects of mass deworming on school attendance — not just for children that got deworming pills, but for other kids who didn't get the pills, both in the same school and in nearby schools. That's plausible because the more infected kids take the pills, the less infected excrement is in the nearby environment, so it could be that fewer nearby kids who *didn't* take pills get infected in the first place. This "spillover" effect to other kids would tend to magnify the pills' impact, so Kremer and Miguel concluded that mass deworming is "far cheaper than alternative ways of boosting school participation."

That's the core of the debate. The debate is not about whether children sick with worms should get treated (everyone says yes), but whether the *mass* treatment of *all* kids — including those not known to be infected, just in case — is a cost-effective way to raise school attendance.

The healthiest parts of the debate have been about the need for transparency, data sharing, and more replication in science, as Ben Goldacre emphasized in his *Buzzfeed* piece that kicked off a firestorm in social media. We endorse those points. Goldacre is a longtime champion of transparency in science.

Here, we're going to focus here on the narrower question of the evidence for mass deworming specifically, which is where some journalists — like the *Guardian* headline writer below — have gotten things quite wrong. If you're pressed for time, scroll down to the bottom for the takeaway.


theguardian
home › world › development UK europe americas asia middle east africa ≡ all

Global development

New research debunks merits of global deworming programmes

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

Researchers now claim that an influential study in Kenya that suggested school attendance and results improved as a result of deworming programmes was misleading. Photograph: Sun Ruibo/Xinhua Press/Corbis

Cutting to the chase: new information about the original deworming study qualifies its findings, but certainly does not "debunk," "overturn," or negate its findings. Donors should remain open to and encourage new evidence but should not shift their priorities on deworming in response to this episode.

Here's what happened. The respected *International Journal of Epidemiology* just published two separate critique papers that revisited the original Kremer and Miguel results — and reached very different conclusions with the same data. The key to understanding what this means is understanding why there were two separate critique papers in the first place. That requires a detour, but we'll get back to the story soon.

Replication Tests and Robustness Tests Are Very Different

The two critique papers were supported by a replication project at a research funding body called the International Initiative for Impact Evaluation or 3ie (which our employer helped create). 3ie's "replication window" offers small grants for researchers to revisit influential papers in international development to see how solid their evidence base is. The Kremer and Miguel deworming paper was one of the first nominees, and a team from the London School of Hygiene and Tropical Medicine set off to revisit Miguel and Kremer's analysis, starting from the original data and computer code. They produced the two critiques. This division into two papers was required by 3ie, for reasons that will become clear.

The first critique paper contains a "pure replication"; the second critique paper contains "alternative analyses." The goal of the "pure replication" is simply to see if the computer code and dataset match the findings reported in the paper. That's a check against errors and fraud. If a finding in a paper fails a test like that, it's appropriate to say that the finding "could not be replicated" or "failed a replication test."

The goal of the second, "alternative analyses" paper by Davey et al. is to try ways of analyzing the data that the original authors did not. The authors did some sensible things in the re-analysis that weren't common practice in economics in the 1990s when Miguel and Kremer were writing: they pre-registered their analysis plan to minimize the risk of fishing for a particular result, and they attempted to follow the guidelines for transparent reporting of randomized trials issued by CONSORT and endorsed by many leading medical journals. Additional analysis like this is valuable because the authors of any original study might have omitted important analyses. Adhering to contemporary (2015) standards for the clinical trials also makes it easier to put this trial in the context of other deworming trials.

However, if "alternative analyses" give a different result, it would not be right to say that the original result "could not be replicated." The results of any scientific paper ever written can be changed with alternative analyses of the same data, a well-known problem in social science. For example, an experiment that had a detectable effect on an entire high school as a whole might not have a detectable effect on each class looked at separately; slicing up the analysis can change the signal-to-noise ratio in the data. In this example it would be right to say that the result is *not robust* to separating by class. But it would be wrong to say that the effect on the whole high school "can't be replicated," because there could be good arguments for doing it each way. Neither is clearly an error, and each result could be right on its own terms.

That is, the two critiques are doing two separate, totally different tests.

The first is a test for indisputable errors or fraud. The second is a test for different results from legitimately disputable choices. If the first kind of test gives different results, this often brings stinging opprobrium on a scientist; if the second kind of test gives different results, this is just the normal progression of science, as colleagues try different things and learn more.

One of us (Clemens) has argued that these two kinds of tests need unmistakably different names, and shown that many researchers recognize the need for this distinction. But right now we don't have clear terms to separate them, so the public discussion ends up calling both kinds of studies "replication" tests, and any different results from either kind of test get described as a failed replication. That's very bad, because science needs a lot more studies that check previous results, as Ben Goldacre says, and it harms that enterprise to make researchers facilitating fruitful robustness checks on their work feel that they'll end up smeared by the undeserved association with incompetence or fraud.

So Did the Original Deworming Paper Fail a Replication Test?

Okay, back to the story. For the deworming study, both of the critiques report results that are substantially different from the original study's results. But only the first study (by Alexander Aiken et al.) even has the potential to "debunk" or "overturn" the original result. The second study (by Calum Davey et al.) does not, and its results cannot be described as a failed replication.

The second, "alternative analyses" study could help us learn important things about the world by showing what happens when researchers try different ways of approaching the data. For example, the Davey et al. study explores what happens when different years of data are analyzed separately, rather than considering all years at once. But unless it's beyond reasonable dispute that the original analysis *should* have done what the critique does, the critique can't claim to be uncovering errors or "debunking" the original. It can claim to be checking robustness to alternative analyses, period. There are highly qualified people disputing the choices in the second, Davey et al. paper (such as here and here). We don't take any position on that dispute here; we just note that the choices in the "alternative analyses" paper are apparently not beyond dispute. The choices have advantages and disadvantages.

What we want to be clear about is that no "alternative analyses" can be said to objectively uncover *errors* in the original if the choices they alter are disputable choices. If a choice is objectively an error, it means that no reasonable person could make that choice.

That's why everything below is about the first study, by Aitken et al. If that study were to find that the analysis performed for the original paper does not match what's in the paper, those discrepancies are rightly called errors. If those errors were to greatly change the result, it would be right to say that Aitken et al. "failed to replicate" the result.

Bottom line: we see no reason to doubt the Aitken et al. study is correct on its own terms, and is executed competently in a good faith effort to test individual findings from the original paper. The Aitken et al. paper fails to replicate some of the statistical estimates in the original paper exactly as they were executed in that paper. It would be completely wrong, however, to say that the Aitken et al. study "debunks," "overturns," or otherwise negates the original study. The effect of primary interest that Kremer and Miguel claimed is clearly present in the data even after the corrections by Aitken et al., albeit at a somewhat different magnitude and for a slightly different set of schools than originally reported.

Here's an analogy that aptly describes this situation: Suppose a chemistry lab claimed that when it mixed two chemicals, the mixture rose in temperature by 60 degrees. Later, a replication team reviewed the original calculations, found an error, and observed that the increase in temperature was only 40 degrees. It would be strictly correct for the replication team to announce, "We fail to replicate the original finding of 60 degrees." That's a true statement by itself, and it doesn't fall within the strict purview of a pure replication to do additional tests to see whether the mix rose by 30 degrees, or 40 degrees, or whatever. But it in this situation it would be excessive to claim that replication "debunks the finding of a rise in temperature," because the temperature certainly did rise, by a somewhat different amount. This is basically what's happened with the deworming replication, as we'll explain.

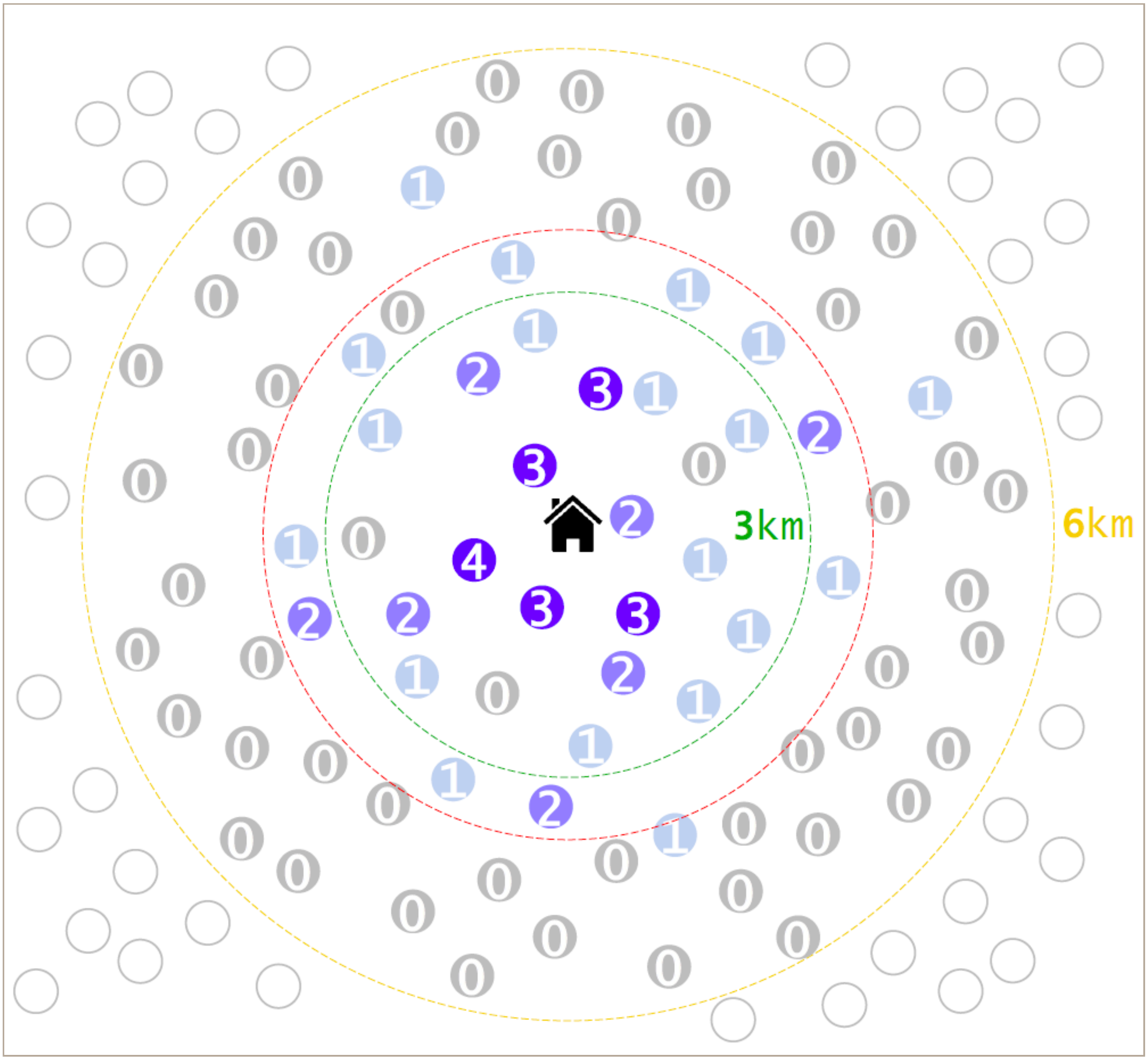
The Technical Stuff: Why the Original Deworming Result is Not "Debunked"

The original Miguel and Kremer paper found that deworming had a positive spillover effect on school attendance for children in nearby schools of +2%. The most contentious part of the Aiken et al. replication is their conclusion that "after correction for coding errors, there was little evidence of an indirect effect on school attendance among children in schools close to intervention schools." They get an estimate of –1.7%. In their reply, Hicks, Kremer, and Miguel (2015) declare that "we interpret the evidence from the reanalysis as strongly supporting the findings of positive deworming treatment externalities and school participation impacts."

How can two teams of very sophisticated researchers look at the same data and reach opposite conclusions? We're going to take you back to high school geometry, where you learned that the area of a circle is proportional to the circle's radius squared ($A = \pi r^2$). So if you double a circle's radius, you quadruple the area of the circle. That informs how you would analyze the spillover effects of a treatment within a certain radius of a treated school.

Suppose the spillover effects looked like they do in this schematic example:

Figure 1. Hypothetical benefits of deworming for neighboring schools



The school that got treated is the black house in the center. Each circle around the black house is some other school that didn't get treated. The number on each of those other schools is the spillover effect from treatment at the school in the center. For example, the number could be the percentage increase in school attendance at each *untreated* school due to spillover effects from the treated school.

Looking at the map, in this schematic example, it's obvious that there is a spillover effect from treatment. You don't need any statistics to tell you that. Schools near the treated school have big increases in attendance, schools far away don't. It's obviously very unlikely that's this pattern is just coincidence.

We can use this example to see how correcting a mistake in the original deworming paper could make it look like there are no spillover effects, even when there are.

The original paper arbitrarily set two concentric circles around each treated school: It measured short-range spillovers inside radius 3km (in green). In the schematic picture above, using the made-up numbers there, the average spillover effect inside the green circle is 1.6. Suppose that, due to statistical noise, we can only detect an effect above 1; so this short-range effect is easy to detect.

The original deworming paper also measured long-range spillovers for *some* of the schools between 3km and 6km away. Why not measure long-range spillovers for *all* the schools between 3km and 6km? This would give excess "weight" to the schools furthest away, where spillover effects would tend to be smallest. (You can see that in the picture: for example, there are a lot more schools around 5km than around 2km. As we talked about above, the outer ring between green and gold has triple the area of the inner green circle.) Giving excess weight to schools with the lowest spillovers could conceal the spillovers amidst the statistical noise.

Thus the analysis underlying original paper measured long-range spillovers only for a subset of the *closest* schools between the green and gold circles. The picture above shows that schematically: it's roughly like considering long-range spillovers only for the schools between the green and *red* circles. In this example, the average spillover for the 11 schools in that narrow band is 1.1.

Here's where the mistake happened in the original paper: the write-up of this analysis in the original paper said that it did the equivalent, in our schematic example, of measuring long-range spillovers for *all* 55 schools between the green and gold circles. That's not what the original statistical analysis actually did, but it's what the write-up said it had done. If you do that for our schematic example, the average effect in the 3km to 6km is only 0.25. That's below our detectable threshold of 1, so we can't distinguish it from zero. Furthermore, in this example, the average spillover effect at *all* 76 schools inside 6km is just 0.6 — a statistical goose egg.

How would you report a correction to this mistake? There are two ways you could do it, ways that would give opposite impressions of the true spillover effects.

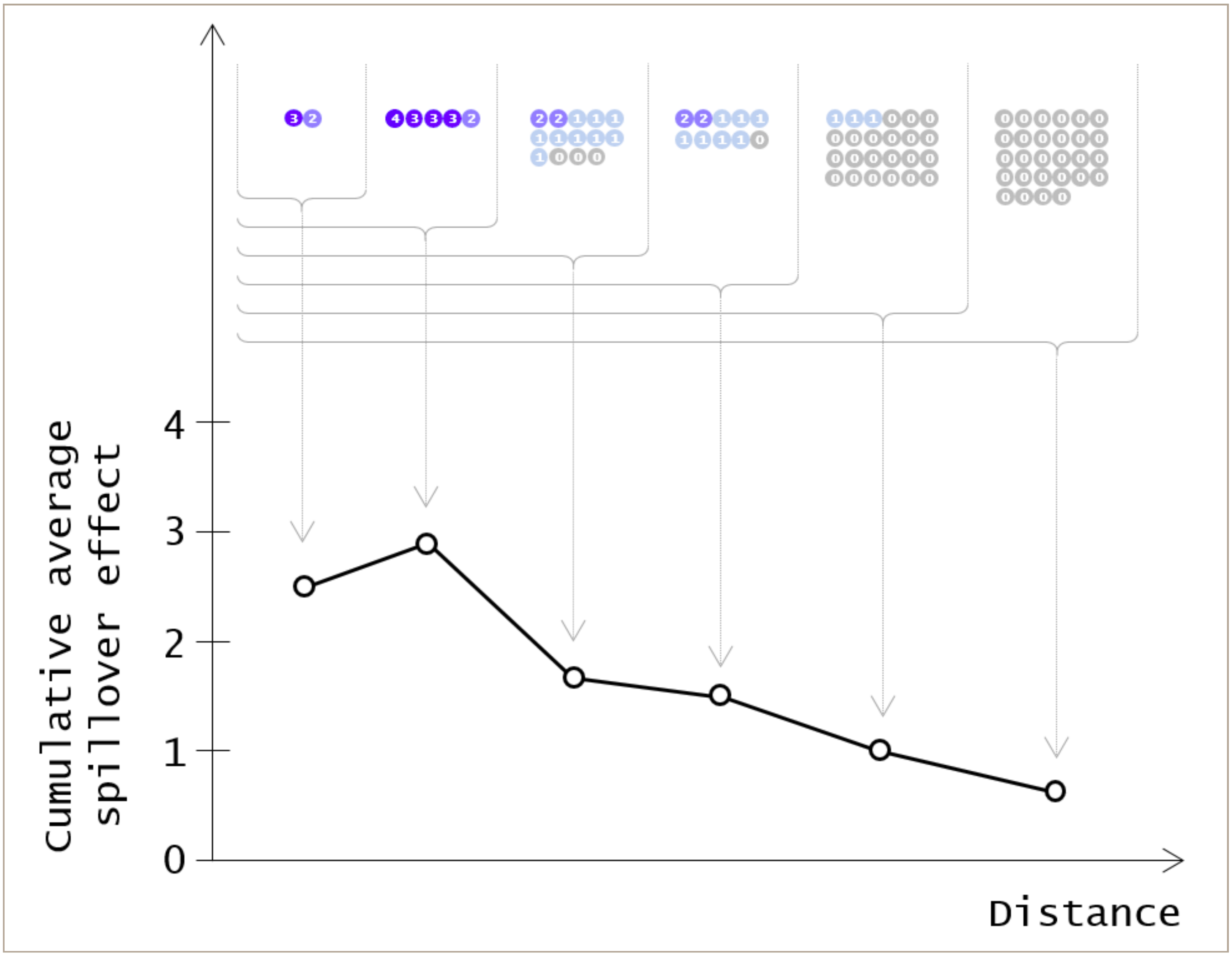
You could simply state that when you correct the error, the average spillover effect on all 76 schools in the correct 6km radius is 0.6, which is indistinguishable from zero. That's an accurate statement in isolation. This is essentially all that is done in the tables of the published version of the replication paper. On that basis you could conclude, as that paper does, that "there was little evidence of an indirect [spillover] effect on school attendance among children in schools close to intervention schools." Strictly on its own terms, that is correct. That's the average value in all the circles in that picture.

But wait a minute. Look back at our schematic picture. It's obvious that there *is* a spillover effect. So something's incomplete and unsatisfying about that portrayal. First of all, the average spillover inside the 3km green circle is 1.6, which in this example we can distinguish from zero. So it's certainly not right to say there is "little evidence" of a spillover effect "close to" the treatment schools.

Second, there's nothing magical about 3km and 6km. Neither the original study nor the replication had any theoretical reason to think that spillovers happen at 2km but don't happen at 4km, nor did they have any reason to think that spillovers *should* happen with 6km but not within other ranges.

So how could you report this correction differently, in a way that shows the obvious spillover effect? Using the same hypothetical data from the figure above, you could show this:

Figure 2. Hypothetical example continued -- benefits of deworming for neighboring schools

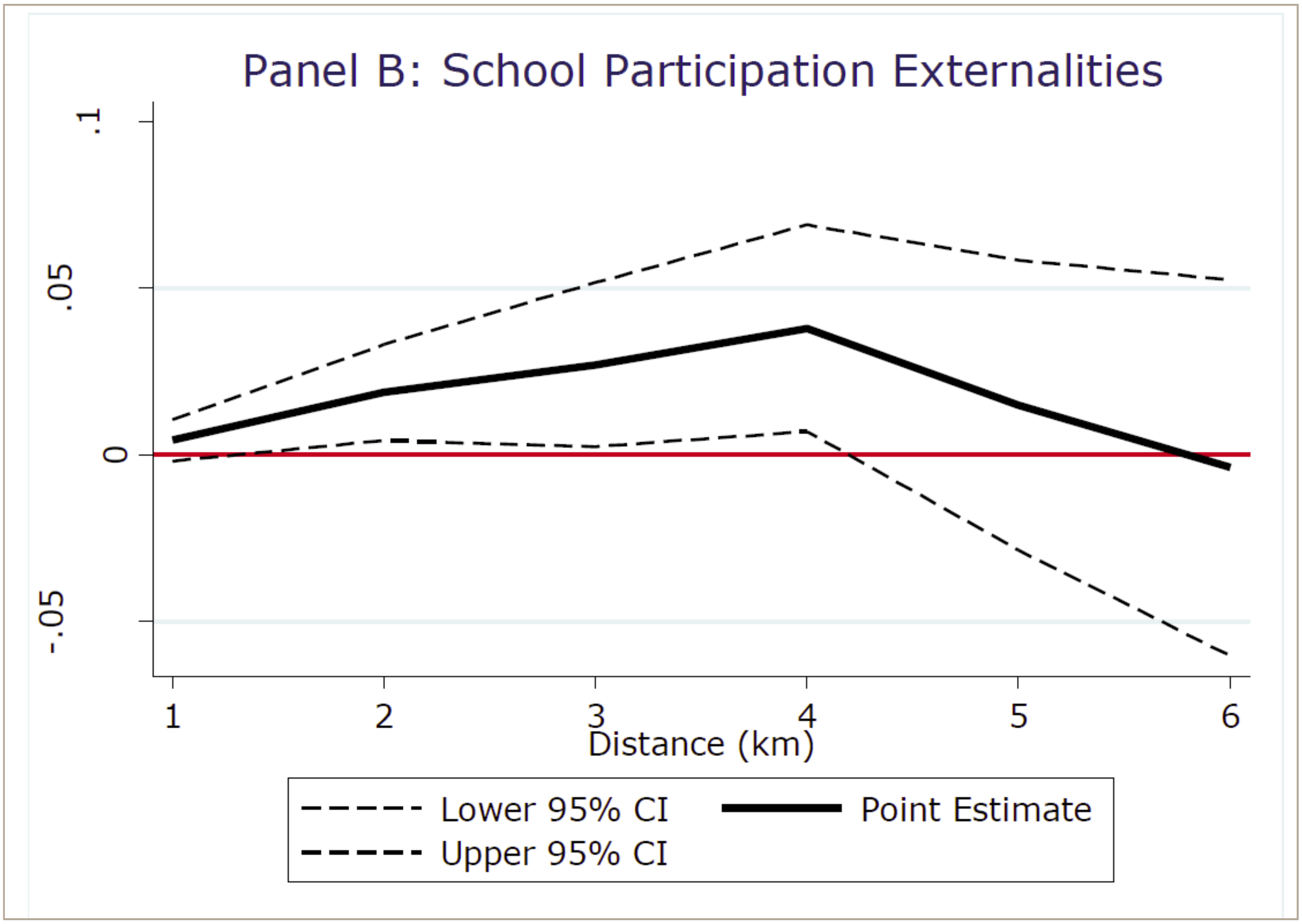


This picture shows, again for our schematic example, the average cumulative spillover effect out to various distances from the treated school: all the schools out to 1km away, all the schools out to 2km, all the schools out to 3km, and so on.

Here, there's a big spillover effect nearby the treated school. That effect peters out as you expand the radius. In this example, it gets undetectable (falls below 1) once you consider all the schools within 5km, because the overall average starts to include so many faraway, unaffected schools.

The authors of the original paper made the corresponding graph for the real data in their original reply to the critique (the graph isn't in the journal-article version). It uses the real data incorporating all the corrections made by Aitken et al., and looks like this:

Figure 3. Actual benefits of deworming on neighboring schools from Hicks, Miguel, and Kremer 2015



Just like in the schematic example before, the cumulative average spillover effect within a given radius first rises and then starts to fall as the radius expands. When the radius gets to 5km, the average spillover effect across *all* schools inside that radius becomes indistinguishable from zero.

This figure makes it clear that in the real, corrected data, there certainly is a detectable spillover effect: within 2km of the treated school, and within 3km, and within 4km. This is strictly incompatible with interpreting the Aitken et al. corrections as negating, debunking, or overturning the original study's finding of spillover effects.

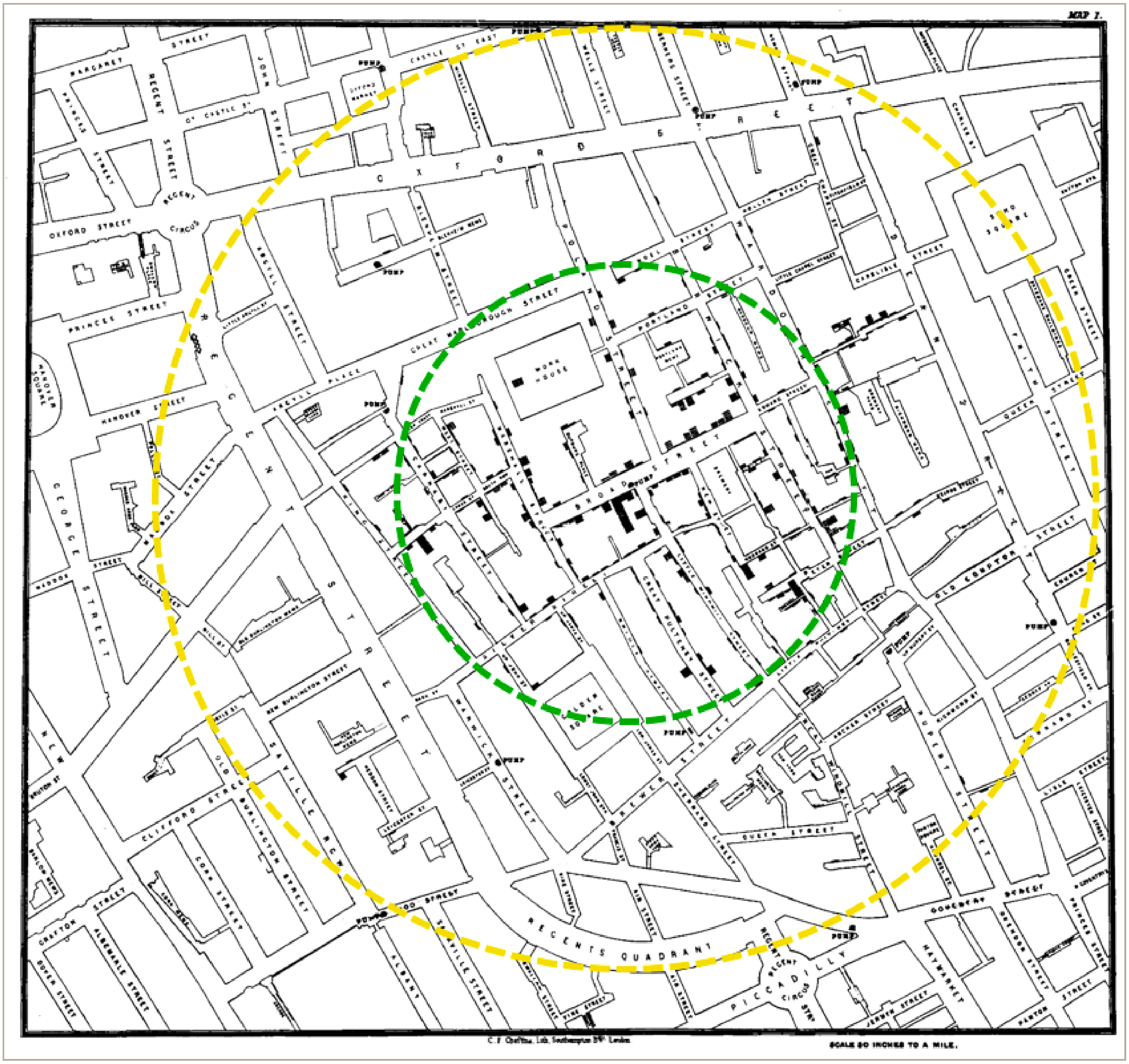
An accurate summary of the Aitken et al. result would be "The statistical analysis underlying the original study found there to be positive spillover effects extending 4km away from treated schools. The write-up of the original study incorrectly reported that those spillover effects extended to 6km. Aitken et al. correct errors in the original study's computer code. They do successfully replicate the spillover effects at 4km, but fail to replicate the spillover effects at 6km claimed by the original study."

In other words, the Aitken et al. finding is correct on its own terms.

The corrected results also *confirm* the existence of spillover effects detected by the original paper, with a somewhat different magnitude. It is wrong for others reading Aitken et al.'s work to report that it debunks the original finding of spillovers.

If the corrected data show "little evidence" of spillovers in the Kremer and Miguel study, the same logic could imply the perverse conclusion that John Snow had little evidence of cholera near London's Broad Street pump in 1854. Snow revolutionized public health by proving the link between contaminated water and cholera. He did it by showing that infections during one outbreak were higher close to a single water pump, near which cholera-infected diapers had been washed. Here is Snow's legendary map, with the infected pump at the center of the circles we've drawn on it, and black rectangles showing cholera cases:

Figure 4. Jon Snow's iconic map of cholera cases around London's Broad Street pump in 1854



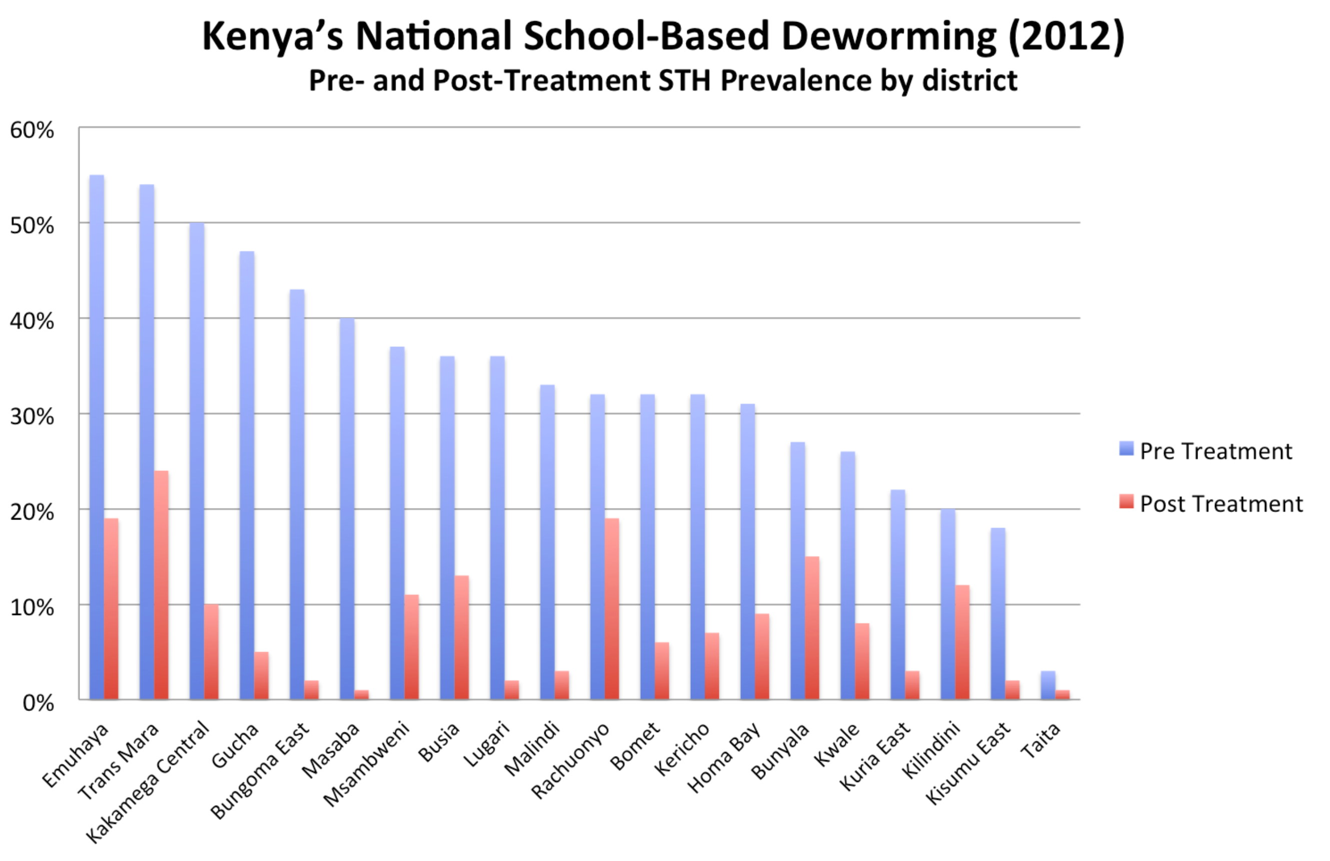
If you were to count up the density of cholera infections inside the green circle, you'd find it to be above average, incriminating the pump. But if you counted the density inside the entire gold circle, which has four times the area, the density would be much lower. It might be hard to statistically detect any difference between the density of infections inside that big circle and the density in other parts of London. If you reported exclusively that finding, you'd state that Snow had been debunked and move on. But that would not be a correct and responsible interpretation of the evidence; it would be an artifact of where you placed the circle. If you step back and simply look at the map, it's quite obvious that there are spillovers of cholera from that pump.

Well Then, Should Donors Invest in Mass Deworming?

Contrary to some of the more sensational media coverage, we feel the policy case for mass deworming is largely unchanged after the latest kerfuffle.

Before we dive into the details of the policy debate, it's useful to step back and examine the track record of one of the major policy initiatives built on the back of the Miguel and Kremer paper — Kenya's National School-Based Deworming program, which administered deworming tablets to nearly 6 million children across more than 13,000 schools in 2012–13. Monitoring and evaluation results from the Ministries of Education and Health suggest that in the 20 districts where the program was active in its first year, worm loads declined from an average of 33% in Jan-Apr to just 9% after treatment in Sep-Oct. (Thanks to Willa Friedman for pointing us to these data.)

Regardless of whatever further impacts deworming may have on health and education, this is a huge welfare gain for over a million children who no longer have these parasites living inside them. There are clearly questions of causality here — seasonal effects and other interventions may be contributing to this decline. But we take this as important prima facie evidence of the Kenyan government's willingness and ability to implement an effective deworming program at scale.



Source: Authors' chart based on "Kenya National School Based Deworming Year 1 (2012-2013) Results", Kenyan Ministry of Education, Science, and Technology, and Ministry of Health (p. 7).
Note: Pre-treatment worm loads were measured in 200 schools in Jan-Apr 2012, and post-treatment measurement was conducted in 60 schools in Sep-Oct 2012.

Returning to the issue at hand: Note that the key policy question up for debate in the worm wars is *not* whether deworming treatments are effective at reducing worm loads — everyone agrees deworming treatments reduce worm levels and are good for infected kids — but whether deworming brings enough additional health and education benefits to justify mass, preemptive treatment of entire populations of children. Essentially this is a debate over whether deworming should be rolled out preemptively like vaccination or withheld until you're sick, like antibiotics.

Everyone also agrees that no big policy question like this should hinge on a single study, and should reflect the totality of the evidence. The sticking point is over which evidence to admit into that totality.

Prior to the latest round of debate, the World Health Organization deemed mass deworming to be cost effective, and the nonprofit GiveWell ranked two deworming efforts — the Schistosomiasis Control Initiative and Deworm the World — as two of its (top four) most effective charities in terms of lives saved per dollar spent. In a comprehensive review of the current debate GiveWell has reiterated its earlier finding that donors should fund mass deworming. GiveWell has no reputational or financial stake in the outcome of this debate, and we find their assessment independent and rigorous. (Both GiveWell and our employer receive support from the philanthropy Good Ventures.)

Proponents of mass deworming rest their case mostly on the long-term educational benefits of deworming found in studies that track children over many years. Roughly a decade after the Miguel and Kremer trial, the authors returned to Western Kenya and found positive effects on the employment and income of kids who had been dewormed (Baird et al. 2011). Another follow-up analysis found significant gains in test scores for younger siblings of the original treatment group -- kids who were more likely to have been worm-free during the critical developmental period of infancy (Ozier 2014). Across the border in Uganda, researchers also found big gains in test scores among children who had benefitted from a separate RCT years earlier (Croke 2014). And that evidence coincided with historical evidence on the schooling benefits from the Rockefeller Foundation's big push to eradicate hookworm in the US south before World War I (Bleakley 2007). The researchers involved with this work are some of the most able on earth.

Skeptics don't contest this evidence, but have set criteria by which this evidence does not get a seat at the table. A systematic review of evidence by the Cochrane Collaboration concluded that mass deworming had no impact on nutrition, health, or learning outcomes — a finding reached by omitting all the studies cited above.

For the narrow goals of the Cochrane review, there’s good reason to omit the US South study, as it’s not a randomized trial. (Though policy makers could still consider the evidence, appropriately caveated.) The rationale given by the Cochrane review for excluding the Kenyan and Uganda studies is that the control group eventually received deworming treatment, and other deworming programs now exist in both countries. As Alexander Berger of GiveWell noted, by this criterion it seems that no long-term study will ever be admissible, because deworming programs are proliferating everywhere. It doesn't require fancy econometrics skills to recognize that if the control group in a long-term study eventually benefits from a different deworming campaign, this will only lead researchers to *under*-estimate the impact of the original deworming. Ignoring positive findings because they might underestimate effects, and then concluding there is no effect, is puzzling. Conservative criteria can be a good thing, but it is also possible to be too conservative.

We conclude that GiveWell struck a good balance here (they call spillover effects an “externality”):

> We continue to believe that it is extremely valuable and important for authors to share their data and code, and we appreciate that Miguel and Kremer did so in this case. We’re also glad to see the record corrected regarding the 3-6km externality terms in Miguel and Kremer 2004. But our overall impression is that this is a case in which the replication process has brought more heat than light. We hope that the research community can develop stronger norms supporting data sharing and replication in the future.

Donors should proceed by encouraging and supporting ongoing, iterative evaluation and learning about the long-term effects of deworming. They should also limit their diet of overblown journalism.

Conclusion / TL;DR

1. The latest academic debate over a famous paper on deworming in Kenya (Miguel and Kremer 2004) questions whether deworming in one school had benefits for student attendance for neighboring schools. The original

authors made an error, defining “neighboring” in the underlying code more narrowly than it was defined in the text. Upon replication, it appears there are spillover benefits up to about 4km, but not out to 6km. We see this as a useful correction, made in good faith by a competent replication team, and the result verifies the original paper’s finding that effects spill over to untreated schools.

2. The policy case for mass deworming rests largely on the long-term benefits to children's cognitive development. But a new report summarizing deworming evidence by the Cochrane Collaboration rules out consideration of almost all long-term impact studies. This exclusion seems hard to justify. Aggregators of the evidence should explore criteria that allow consideration, to some degree, of long-term studies. The existing long-term studies, while fewer in number than one might wish, appear to justify investment in deworming. Ongoing long-term follow-up is needed to continue learning about the effects of mass deworming.

Disclosure: The authors have institutional links both with the authors of the critique studies and with the authors of the original study. Clemens and Sandefur’s employer, the Center for Global Development, created the working group that founded 3ie, the organization that encouraged and funded the critique studies; Clemens and Sandefur also have been involved in consultations with 3ie on the design of the “replication window” that supported the critiques. In addition, the authors of the original study are affiliates of the Center for Global Development: Michael Kremer as non-resident fellow, and Ted Miguel as a member of the Advisory Group.

RELATED POSTS:

The Final Word on Microcredit?
End the Evaluation Wars: A Plea to Shift from the Abstract to the Specific
An Homage to the Randomistas on the Occasion of the J-PAL 10th Anniversary: Development as a Faith-Based Activity
The R-Word Is Not Dirty
RCTs in Development, Lessons from the Hype Cycle

Authors:

Michael Clemens

VIEW PROFILE

Justin Sandefur

VIEW PROFILE

Comments