GiveWell

DONATE

GIVING EFFECTIVELY HOW WE WORK TOP CHARITIES RESEARCH OUR MISTAKES ABOUT UPDATES

HOME

The GiveWell Blog

Why I mostly believe in Worms

Previous Post

December 6, 2016 (updated on: December 7, 2016) | by **David Roodman**

The following statements are true:

- "**GiveWell** is a nonprofit dedicated to finding outstanding giving opportunities through in-depth analysis. Thousands of hours of research have gone into finding our top-rated charities."
- GiveWell **recommends** four deworming charities as having outstanding expected value. Why? Hundreds of millions of kids harbor parasitic worms in their guts[1]. Treatment is safe, effective, and cheap, so much so that where the worms are common, the World Health Organization **recommends** administering pills once or twice a year to *all* children without i[ncurring the cost of determining who is infected.

Recent Blog Posts

Want to stay updated on GiveWell's research? Read our blog or follow us by email, Facebook, Twitter, or RSS.

Why I mostly believe in Worms

Our updated top charities for giving season 2016

New Incentives update

Updates on AMF's transparency and monitoring

- Two respected organizations, **Cochrane** and the **Campbell** Collaboration, have systematically reviewed the relevant studies and found little reliable evidence that mass deworming does good.

That list reads like a logic puzzle. GiveWell relies on evidence. GiveWell recommends mass-deworming charities. The evidence says mass deworming doesn't work. How is that possible? Most studies of mass deworming track impact over a few years. The handful that look longer term find big benefits, including one in Kenya that reports **higher earnings in adulthood**. So great is that benefit that even when GiveWell **discounts it by some** *99%* out of doubts about generalizability, deworming charities look like promising bets.

Still, as my colleagues have written, the evidence on deworming is **complicated** and **ambiguous**. And GiveWell **takes seriously** the questions raised by the Cochrane and Campbell evidence reviews. Maybe the best discount is not 99% but 100%. That would make all the difference for our assessment. This is why, starting in October, I delved into deworming. In this post and the next, I will share what I learned.

In brief, my confidence rose in that Kenya study's finding of higher earnings in adulthood. I will explain why below. My confidence fell in the generalizability of that finding to other settings, as discussed in the next post.

As with all the recommendations we make, our calculations may be wrong. But I believe they are reasonable and quite possibly conservative. And notice that they do not imply that the odds are 1 in 100 that deworming does great good everywhere and 99 in 100 that it does no good anywhere. It can instead imply that kids receiving mass deworming today need it less than those in the Kenya study, because today's children

September 2016 open thread

Would other organizations have funded AMF's bednet distributions if AMF hadn't?

Deworming might have huge impact, but might have close to zero impact

Update on GiveWell's web traffic / money moved: Q1 2016

Weighing organizational strength vs. estimated cost-effectiveness

Why we don't currently recommend charities focused on vaccine distribution

FOLLOW US:

have fewer worms or because they are healthy enough in other respects to thrive despite the worms.

Unsurprisingly, I do not know whether 99% overshoots or undershoots. I wish we had more research on the long-term impacts of deworming in other settings, so that we could generalize with more nuance and confidence.

In this post, I will first orient you with some conceptual and historical background. Then I'll think through two concerns about the evidence base we're standing on: that the long-term studies lack design features that would add credibility; and that the key experiment in Kenya was not randomized, as that term is generally understood.

Background

Conclusions vs. decisions

There's a deeper explanation for the paradox that opens this post. Back in 1955, the great statistician John Tukey gave an after-dinner talk called "**Conclusions vs Decisions**," in which he meditated on the distinction between judging what is true—or might be true with some probability—and deciding what to do with such information. Modern gurus of evidence synthesis retain that distinction. The Cochrane Handbook, which guides the Cochrane and Campbell deworming reviews, is **emphatic**: "Authors of Cochrane reviews should not make recommendations." Indeed, researchers arguing today over the impact of mass deworming are mostly arguing about *conclusions*. Does treatment for worms help? How much and under what circumstances? How confident are we in our answers? We at GiveWell—and you, if you're considering our **charity recommendations**—have to make *decisions*.

The guidelines for the GRADE system for rating the quality of studies nicely illustrates how reaching conclusions, as hard and

Recent Comments

CATHERINE (GIVEWELL) COMMENTED ON
Our updated top charities for giving season 2016

JULIA WISE COMMENTED ON
Why I mostly believe in Worms

DAVID ROODMAN COMMENTED ON
Why I mostly believe in Worms

AMANDA GLASSMAN COMMENTED ON
Why I mostly believe in Worms

DAVID ROODMAN COMMENTED ON
Why I mostly believe in Worms

JAKE GOLDSTON COMMENTED ON
Why I mostly believe in Worms

DAVID ROODMAN COMMENTED ON

complicated as it is, still leaves you several logical steps short of choosing action. Under the heading, "**A particular quality of evidence does not necessarily imply a particular strength of recommendation**," we read:

> *For instance, consider the decision to administer aspirin or acetaminophen to children with chicken pox. Observational studies have observed an association between aspirin administration and Reye's syndrome. Because aspirin and acetaminophen are similar in their analgesic and antipyretic effects, the low-quality evidence regarding the potential harms of aspirin does not preclude a strong recommendation for acetaminophen.*
>
> *Similarly, high-quality evidence does not necessarily imply strong recommendations. For example, faced with a first deep venous thrombosis (DVT) with no obvious provoking factor patients must, after the first months of anticoagulation, decide whether to continue taking warfarin long term. High-quality randomized controlled trials show that continuous warfarin will decrease the risk of recurrent thrombosis but at the cost of increased risk of bleeding and inconvenience preferences. Because patients with varying values and preferences are likely to make different choices, guideline panels addressing whether patients should continue or terminate warfarin may—despite the high-quality evidence—offer a weak recommendation.*

I think some of the recent deworming debate has nearly equated the empirical question of whether mass deworming "works" with the practical question of whether it should be done. More than many participants in the conversation, GiveWell has seriously analyzed the logical terrain between the two questions, with a **explicit decision framework** that allows and forces us to estimate a dozen relevant parameters. We have found the decision process no more straightforward than the

Why I mostly believe in Worms

HOWARD WHITE COMMENTED ON
Why I mostly believe in Worms

CHRISTIAN SMITH COMMENTED ON
Our updated top charities for giving season 2016

COLIN RUST COMMENTED ON
September 2016 open thread

Full archive

All posts

Archives

2016

December

November

October

research process appears to be. You can argue with how GiveWell has **made its calls** (and we hope you will, with specificity), and such argument will probably further expose the trickiness of going from conclusion to decision.

The rest of this post is about the "conclusions" side of the Tukey dichotomy. But having spent time with our **spreadsheet** helped me approach the research with a more discerning eye, for example, by sensitizing me to the crucial question of how to generalize from the few studies we have.

The research on the long-term impacts of deworming

Two studies form the spine of GiveWell's support for deworming. Ted Miguel and Michael Kremer's seminal **Worms** paper reported that after school-based mass deworming in southern Busia county, Kenya, in the late 1990s, kids came to school more. And there were "spillovers": even kids at the treated schools who didn't take the pills saw gains, as did kids at nearby schools that didn't get deworming. However, children did not better on standardized tests. In all treatment schools, children were given albendazole for soil-transmitted worms—hookworm, roundworm, whipworm. In addition, where warranted, treatment schools received praziquantel for schistosomiasis, which is transmitted through contact with water and was common near Lake Victoria and the rivers that feed it.

Worms at Work, the sequel written with Sarah Baird and Joan Hamory Hicks, tracked down the (former) kids 10 years later. It found that the average 2.4 years of extra deworming given to treatment group children led to **15% higher non-agricultural earnings[2]**, while hours devoted to farm work **did not change**. The earnings gain appeared concentrated in wages (as distinct from self-employment income), which rose 31%.**[3]** That's a huge benefit for a few dollars of deworming, especially if it accrued for years, and is what drives **GiveWell's recommendations of deworming charities**.

Four more studies track impacts of mass deworming over the long run:

- In 2009–10, Owen Ozier surveyed children in Busia who were too young to have participated in the Kenya experiment, since they were not in school yet, but who might have benefited through the deworming of their school-age siblings and neighbors. (If your big sister and her friends don't have worms, you're less likely to get them too.) Ozier **found** that kids born right around the time of the experiment scored higher on cognitive tests years later.

- The Worms team confidentially shared initial results from the latest follow-up on the original experiment, based on surveys fields in 2011–14. Many of those former schoolchildren now have children of their own. The results shared are limited and preliminary, and I advised my colleagues to wait before updating their views based on this research.

- Kevin Croke **followed up** on a **deworming experiment** that took place across the border in Uganda in 2000–03. (GiveWell summary **here**.) Dispensing albendazole (for soil-transmitted worms) boosted children's scores on **basic tests of numeracy and literacy** administered years later, in 2010 and 2011. I am exploring and discussing the findings with Kevin Croke, and don't have anything to report yet.

- In a remarkable act of historical scholarship, Hoyt Bleakley **tracked the impacts** of the hookworm eradication campaign initiated by the Rockefeller Foundation in the American South a century ago.

> Though not a randomized experiment, his analysis indicates that children who benefited from the campaign went on to earn more in adulthood.

These studies have increased GiveWell's confidence in generalizing from Worms at Work—but perhaps only a little. Two of the four follow-up on the original Worms experiment, so they do not constitute fully independent checks. One other is not experimental. For now, the case for mass deworming largely stands or falls with the Worms and Worms at Work studies. So I will focus on them.

Worm Wars

A few years ago, the International Initiative for Impact Evaluation (3ie) **funded British epidemiologists Alexander Aiken and Calum Davey** to replicate Worms. (I **served on 3ie's board** around this time.) With coauthors, the researchers first exactly **replicated the study** using the original data and computer code. Then they **analyzed the data afresh** with their **preferred methods**. The deeply critical write-ups appeared in the *International Journal of Epidemiology* in the summer of 2015. The next day, Cochrane (which our Open Philanthropy Project **has funded**) updated its review of the deworming literature, finding "**quite substantial evidence that deworming programmes do not show benefit**." And so, on the dreary plains of academia, did the great **worm wars** begin.

I read through the blogospheric explosion of debate.**[4]** Much of it is secondary for GiveWell, because it is about the reported bump-up in school attendance after deworming. That matters less to us than the long-term impact on earnings. Getting kids to school is only a means to other ends—**at best**. Similarly, much debate centers on those spillovers: all sides agree that the original Worms paper overestimated their geographic reach. But that is not so important when assessing charities that aim to deworm *all* (school-age) children in a region rather than a subset as in the experiment.

I think GiveWell should focus on these three criticisms aired in the debate:

- The Worms experiment and the long-term follow-ups lack certain design features that are common in epidemiology, with good reason, yet are rare in economics. For example, the kids in the study were not "blinded" through use of placebos to whether they were in a treatment or control group. Maybe they behaved differently merely because they knew they were being treated and observed.
- The Worms experiment wasn't randomized, as that term is usually meant.
- Against the handful of promising (if imperfect) long-term studies are several dozen short-term studies, which in aggregate find **little or no benefit** for outcomes such as survival, height, weight, hemoglobin, cognition, and school performance. The surer we are that the short-term impacts are small, the harder it is to believe that the long-term impacts are big.

I will discuss the first two criticisms in this post and the third in the next.

"High risk of bias": Addressing the critique from epidemiology

Perhaps the most alarming charge against Worms and its brethren has been that they are at "high risk of bias" (**Cochrane**, **Campbell**, **Aiken et al.**, **Davey et al.**). This phrase comes out of a method in epidemiology for assessing the reliability of studies. It is worth understanding exactly what it means.

Within development economics, Worms is seminal because when it **circulated in draft in 1999**, it launched the field experimentation movement. But it is not as if development economists invented randomized trials. Long before the "randomistas" appeared, epidemiologists were running experiments to evaluate countless drugs, devices, and therapies in countries rich and poor. Through this experience, they developed norms about how to run an experiment to minimize misleading results. Some are codified in the **Cochrane Handbook**, the bible of *meta-analysis*, which is the process of systematically synthesizing the available evidence on such questions as whether breast cancer screening saves lives.

The norms make sense. An experimental study is more reliable when there is:

- Good **sequence generation**: The experiment is randomized.
- **Sequence concealment**: No one knows before subjects enter the study who will be assigned to treatment and who to control. This prevents, for example, cancer patients from dropping out of a trial of a new chemotherapy when they or their doctors learn they've been put in the control group.
- **Blinding**: During the experiment, assignment remains hidden from subjects, nurses, and others who deliver or sustain treatment, so that they cannot adjust their behavior or survey responses, consciously or otherwise. Sometimes this requires giving people in the control group fake treatment (placebos).
- **Double-blinding**: The people who measure outcomes—who take blood pressure, or count the kids showing up for school—are also kept in the dark about who is treatment and who is control.

- Minimized **incomplete outcome data** (in economics, "attrition"): If some patients on an experimental drug fare so poorly that they miss follow-up appointments and drop out of a study, they could make the retained patients look misleadingly well-off.
- No **selective outcome reporting**: Impacts on *all* outcomes measured are reported—for otherwise we should become suspicious of omissions. Are the researchers hiding contrary findings, or **mining for statistically significant impacts**? One way researchers can reduce selective reporting and the appearance thereof is to pre-register their analytical plans on a website outside their control.

Especially when gathering studies for meta-analysis, epidemiologists prize these features, as well as clear reporting of their presence or absence.

Yet most of those features are scarce in economics research. Partly that is because economics is not medicine: in a housing experiment, to **paraphrase Macartan Humphreys**, an agency can't give you a placebo housing voucher that leaves you sleeping in your car without your realizing it. Partly it is because these desirable features come with trade-offs: the flexibility to test un-registered hypotheses can let you find new facts; sometimes the hospital that would implement your experiment has its own views on how things should be done. And partly the gap between ideal and reality is a sign that economists can and should do better.

I can imagine that, if becoming an epidemiologist involves studying examples of how the absence of such design features can mislead—even kill—people, then this batch of unblinded, un-pre-registered, and even un-randomized deworming studies out of economics might look passing strange.[5] So might GiveWell's reliance upon them.

The scary but vague term of art, "high risk of bias," captures such worries. The term arises from the **Cochrane Handbook**, which, as I've mentioned, is the authoritative guide for the process of systematically synthesizing available research on a health-related question. The Handbook, like meta-analysis in general, strives for an approach that is mechanical in its objectivity. Studies are to be sifted, sorted, and assessed on observable traits, such as whether they are blinded. In providing guidance to such work, the Handbook **distinguishes** credibility from quality. "Quality" could encompass such traits as whether proper ethical review was obtained. Since Cochrane focuses on credibility, the handbook authors excluded "quality" from their nomenclature for study design issues. They settled on "risk of bias" as a core term, it being the logical antithesis of credibility.

Meanwhile, while some epidemiologists have devised scoring systems to measure risk of bias—plus 1 point for blinding, minus 2 for lack of pre-registration, etc.—the Cochrane Handbook **says** that such scoring is "is not supported by empirical evidence." So, out of a sort of humility, the Handbook recommends something simpler: run down a checklist of design features, and for each one, **just judge whether a study has it or not**. If it does, label it as having "low risk of bias" in that domain. Otherwise, mark it "high risk of bias." If you can't tell, call it "unclear risk of bias."

Thus, when a study earns the "high risk of bias" label, that means that it lacks certain design features that all concerned agree are desirable. Full stop.

So while the Handbook's checklist brings healthy objectivity to evidence synthesis, it also brings limitations, especially in our context:

- Those unversed in statistics, including many decision-makers, may not appreciate that "bias" carries a **technical meeting** that is less pejorative than the everyday one. It doesn't mean "prejudiced." It means "gives an answer different from the true answer, on average." So, especially in debates that extend outside of academia, its use tends to sow confusion and inflame emotions.

- The binaristic label "high risk of bias" may be humble in origins, but it does not come off as humble in use. At least to non-experts the pronouncement, "the study is at high risk of bias," seem confident. But how big is the potential bias and how great the risk? More precisely, what is the probability distribution for the bias? No one knows.

- While useful when distilling knowledge from reams of research, the objectivity of the checklist comes at a price in superficiality. And the trade-off becomes less warranted when examining five studies instead of 50. As members of the Worms team **point out**, some Cochrane-based criticisms of their work make less sense on closer inspection. For example, the lack of blinding in Worms "cannot explain why untreated pupils in a treatment school experienced sharply reduced worm infections." As we will see, by probing beneath the surface of a study—engaging with its specifics, examining its data and code—one can learn much that can enhance or degrade credibility.

- The checklist is incomplete. E.g., **with an assist from Ben Bernanke**, economics is getting better at transparency. Perhaps we should brand all studies for which data and code have not been publicly shared as being at "high risk of bias" for opacity. The controversy that ensued after

> the 3ie-funded replication of Worms generated a lot of heat, but light too. There were **points of agreement**. New analysts brought new insights. Speaking personally, exploring the public data and code for **Worms** and **Worms at Work** ultimately raised my trust in those studies, as I will explain. If it had done opposite, that too would have raised my confidence in whatever conclusion I extracted. Arguably, Worms is now the most credible deworming study, for no other has **survived** such scrutiny.

So what is a decisionmaker to do with a report of "high risk of bias"? If the choice is between relying on "low risk" studies and "high risk" studies, **all else equal**, then the choice is clear: favor the "low risk" studies. But what if all the studies before you contain "high risk of bias"?

That question may seem to lead us to an analytical cul-de-sac. But some researchers have pushed through it, with *meta-epidemiology*. A **1995 article** (hat tip: Paul Garner) drew together 250 studies from 33 meta-analyses of certain interventions relating to pregnancy, labor, and delivery. They asked: do studies lacking blinding or other good features report bigger impacts? The answers were "yes" for sequence concealment and double-blinding and "not so much" for randomization and attrition. More studies have been done like that. And researchers have even aggregated *those*, which I suppose is meta-meta-epidemiology. (OK, not really.) **One example** **cited by the Cochrane Handbook** finds that lack of sequence concealment is associated with an **average impact exaggeration of 10%**, and, separately, that lack of double-blinding is associated with **exaggeration by 22%**.**[6]**

To operationalize "high risk of bias," we might discount the reported long-term benefits from deworming by such factors. No one knows if those discounts would be right. But they would make GiveWell's ~99% discount—which can compensate for 100-fold (10000%) exaggeration—look conservative.

The epidemiological perspective should alert economists to ways they can improve. And it has helped GiveWell appreciate limitations in deworming studies. But the healthy challenge from epidemiologists has not undermined the long-term deworming evidence as completely as it may at first appear.

Why I pretty much trust the Worms experiment

*Here are Stata **do** and **log** files for the quantitative assertions below that are based on publicly available data.*

I happened to attend a conference on "**What Works in Development**" at the Brookings Institution in 2008. As economists enjoyed a free lunch, the speaker, Angus Deaton, launched a **broadside** against the randomization movement. He made many points. Some were so deep I still haven't fully grasped them. I remember best two less profound things he said. He suggested that Abhijit Banerjee and Esther Duflo flip a coin and jump out of an airplane, the lucky one with a parachute, in order to perform a much-needed **randomized controlled trial of this injury-prevention technology**. And he pointed out that the poster child of the randomization movement, Miguel and Kremer's Worms, wasn't actually randomized—at least not as most people understood that term.

It appears that that the **charity** that carried out the deworming for Miguel and Kremer would not allow schools to be assigned to treatment or control via rolls of a die or the computer equivalent. Instead, Deaton said, the 75 schools were listed alphabetically. Then they were assigned cyclically to three groups: the first school went to group 1, the second to group 2, the third to group 3, the fourth to group 1, and so on. Group 1 started receiving deworming treatment in 1998; group 2 in 1999;

and group 3, the control, not until after the experiment ended in 2000. During the Q&A that day at Brookings, Michael Kremer politely argued that he could think of no good theory for why this assignment system would generate false results—why it would cause, say, group 1 students to attend school more for some reason other than deworming.**[7]** I think Deaton replied by citing the example of a study that was widely thought to be well randomized until someone showed that it wasn't.**[8]** His point was that unless an experiment is randomized, you just can't sure be that no causal demons lurk within.

This exchange came to mind when I began reading about deworming. As I say, GiveWell is less interested in whether treatment for worms raised school attendance in the short run than whether it raised earnings in the long run. But those long-term results, in Worms at Work, depend on the same experiment for credibility. In contrast with the meta-analytic response to this concern, which is to affix the label "high risk of bias for sequence generation" and move on, I dug into the study's data. What I attacked hardest was the premise that before the experiment began, the three school groups were statistically similar, or "balanced."

Mostly the premise won.

Yes, there are reasons to doubt the Worms experiment…

If I were the prosecutor in *Statistical balance police v. Miguel and Kremer*, I'd point out that:

- Deaton had it wrong: schools were not alphabetized. It was worse than that, in principle. The 75 schools were sorted alphabetically by division and zone (units of local geography in Kenya) and within zones by *enrollment*. Thus, you could say, a study famous for finding more kids in school *after* deworming formed its treatment groups on how many kids were in school *before* deworming. That is not ideal. In the worst case, the 75 schools would have been situated in 25 zones, each with three schools. The cyclic algorithm would then have always put the smallest school in group 1, the middle in group 2, and the largest in group 3. And if the groups started out differing in size, they would probably have differed in other respects too, spoiling credibility. (In defense of Deaton, I should say that the authors' description of the cyclical procedure changed between **2007** and **2014**.)

- Worms **reports** that the experimental groups *did* start out different in some respects, with statistical significance: "Treatment schools were initially somewhat worse off. Group 1 pupils had significantly more self-reported blood in stool (a symptom of schistosomiasis infection), reported being sick more often than Group 3 pupils, and were not as clean as Group 2 and Group 3 pupils (as observed by NGO field workers)." Now, in checking balance, **Table I of Worms** makes 42 comparisons: group 1 vs. group 3 and group 2 vs. group 3 for 21 variables. Even if balance were perfect, when imposing a p = 0.05 significance threshold, one should expect about 5% of the tests to show up as significant, or about two of 42. In the event, five show up that way. I confirmed with formal tests that these differences were unexpected in aggregate if the groups were balanced.

- Moreover, the groups differed before the experiment in a way not previously reported: in school attendance. Again, this looks very bad, at least on the surface, since attendance is a major focus of Worms. According to school registers, attendance in grades 3–8 in early 1998 averaged **97.3%, 96.3%, and 96.9%** in groups 1, 2, and 3 respectively. Notice that group 3's rate put it between the two others. This explains why, when Worms separately compares groups 1 and 2 to 3, it does not find terribly significant differences (p = 0.4, 0.12). But the distance from group 1 to 2—which is not checked—is more significant (p = 0.02), as is that from group 1 to 2 and 3 averaged together (p = 0.06). In the first year of the experiment, only group 1 was treated. So if it started out with higher attendance, can we confidently attribute the higher attendance over the following year to deworming?

Miguel and Kremer point out that school registers, from which those attendance rates come, "are not considered reliable in Kenya." Indeed, at about 97%, the rates converge rather implausibly toward perfection. This is why the researchers measured attendance by independently sending enumerators on surprise visits to schools. They found attendance around 68–76% in the 1998 control group schools (bottom of **Table VI**). So should we worry about a tiny imbalance in nearly meaningless school-reported attendance? Perhaps so. I find that at the beginning of the experiment the school- and researcher-reported attendance correlated positively. Each 1% increase in a school's self-reported attendance—equivalent to moving from group 2 to group 1—predicted a 3% increase in researcher-recorded attendance (p = 0.008), making the starting difference superficially capable of explaining roughly half the **direct impact found in Worms**.

…but there are reasons to trust the Worms experiment too

To start with, in response to the points above:

- Joan Hamory Hicks, who manages much of the ongoing Worms follow-up project, sent me the spreadsheet used to assign the 75 schools to the three groups back in 1997.

Its contents do not approximate the worst case I described, with three schools in each zone. There are eight zones, and their school counts range from four to 15. Thus, cyclical assignment did introduce substantial arbitrariness with respect to initial school enrollment. In some zones the first and smallest school went into group 1, in others group 2, in others group 3.

- As for the documented imbalances, such as kids in group 1 schools being sick more often, Worms **points out** that these should probably make the study conservative: the groups that ultimately fared better started out worse off.

- The Worms team began collecting attendance data in all three groups, in early 1998 before the first deworming visits took place. Those more-accurate numbers do not suggest imbalance across the three groups (p = 0.43). And the correlation of school-recorded attendance, which is not balanced, and researcher-recorded attendance, which is, is not especially dispositive. If you looked across a representative 75 New York City schools at two arbitrarily chosen variables were—say, fraction of students who qualify for free meals and average class size—they could easily be correlated too. Finally, when I modify a basic Miguel and Kremer attendance regression (**Table IX, col. 1**) to control for the imbalanced school-recorded attendance variable, it hardly perturbs the results (except by restricting the sample because of missing observations for this variable). If initial treatment-control differences in school-recorded attendance were a major factor in the celebrated impact estimates, we would expect that controlling for the former would affect the latter

In addition, three observations more powerfully bolster the Worms experiment.

First, I managed to identify the 75 schools and link them to a **public database of primary schools in Kenya.** (In email, Ted Miguel expressed concern for the privacy of the study subjects, so I will not explain how I did this nor share the school-level information I gained thereby, except the elevations discussed just below.) This gave me fresh school-level variables on which to test the balance of the Worms experiment, such as institution type (religious, central government, etc.) and precise latitude and longitude. I found **little suggestion** of imbalance on the new variables as a group (p= 0.7, 0.2 for overall differences between group 1 or 2 and group 3; p = 0.54 for a difference between groups 1 and 2 together and group 3, which is the split in Worms at Work). Then, with a Python program I wrote, I used the geo-coordinates of the schools to query **Google** for their elevations in meters above sea level. The hypothesis that the groups differed on elevation is **rejected** at p = 0.36, meaning once more that a hypothesis of balance on a new variable is not strongly rejected. And if we aggregate groups 1 and 2 into a single treatment group as in Worms at Work, p = 0.97.

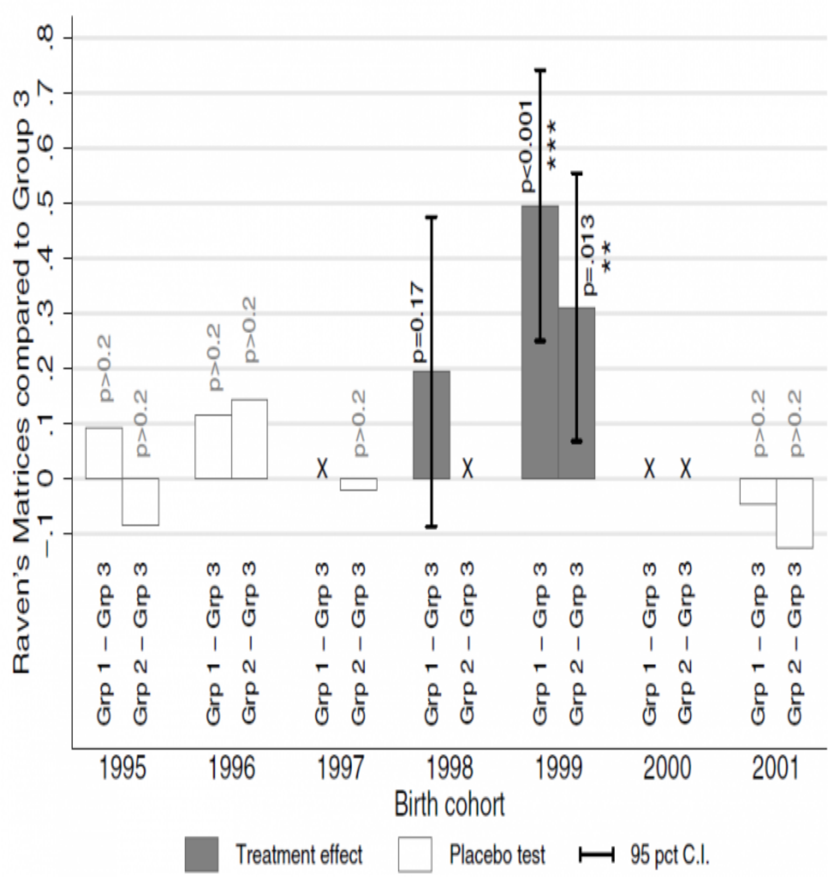
Second, after the Worms experiment finished in 2000—and all 75 schools were receiving deworming—Miguel and Kremer launched a **second, truly randomized experiment in the same setting.** With respect to earnings in early adulthood (our main interest), the new experiment generates similar, if less precise, results. The experiment took on a hot topic of 2001: whether to charge poor people for basic services such as schooling and health care, in order to make service provision more financially sustainable as well as more accountable to clients. The researchers took the 50 group 1 and group 2 schools from the first experiment and randomly split them into two new groups. In the new control group, children continued to receive deworming for free. In the new treatment group, for the duration of 2001, families were charged 30 shillings ($0.40) for albendazole, for soil-transmitted worms, and another 70 shillings ($0.90) for praziquantel, where warranted for schistosomiasis. In response to the "user fees," take-up of deworming medication fell 80% in the treatment group (which therefore, ironically, received less treatment). In effect, a second and less impeachable deworming experiment had begun.

Like the original, this new experiment sent ripples into the data that the Worms team collected as it tracked the former schoolchildren into adulthood. Because the user fee trial affected a smaller group—50 instead of 75 schools—for a shorter time—one year instead of an average 2.4 in the original experiment—it did not generate deworming impact estimates of the same precision. This is probably why Worms at Work gives those impact estimates less space than the ones derived from the original experiment.

But they are there. And they tend to corroborate the main results. The regression that has anchored GiveWell's cost-effectiveness analysis puts the impact of the first experiment's 2.4 years of deworming on later wage earnings at +31% (p = 0.002). If you run **the publicly available code on the publicly available data**, you discover that the same regression estimates that being in the treatment arm of the second experiment cut wage earnings by 14% (albeit with less confidence: p = 0.08). The hypothesis that the two implied rates of impact are equal—31% per 2.4 years and 14% per 80% x 1 year—fits the data (p = 0.44). More generally, Worms at Work states that among 30 outcomes checked, in domains ranging from labor to health to education, the estimated long-term impacts of the two experiments agree in sign in 23 cases. The odds of that happening by chance alone are 1 in 383.**[9]**

The third source of reinforcement for the Worms experiment is **Owen Ozier's follow-up**. In 2009 and 2010, he and his assistants surveyed 2400 children in the Worms study area who were born between about 1995 and 2001. I say "about" because their birth dates were estimated by asking them how many years old they were, and if a child said in August 2009 that she was eight, that meant that she was born in 2000 or 2001. By design, the survey covered children who were too young to have been in school during the original Worms experiment, but who might have benefited indirectly, through the deworming of their older siblings and neighbors. The survey included several cognitive tests, among them **Raven's Matrices**, which are best understood by looking at an **example**.

This graph from the Ozier working paper shows the **impact of Miguel and Kremer's 1998–2000 deworming experiment on Raven's Matrix scores of younger children, by approximate year of birth**:



To understand the graph, look at the right end first. The white bars extending slightly below zero say that among children born in 2001 (or maybe really 2002) those linked by siblings and neighbors to group 1 or group 2 schools scored slightly lower than those linked to group 3 schools—but not with any statistical significance. The effective lack of difference is easy to explain since by 2001, schools in all three groups were or had been receiving deworming. (Though there was that user fee experiment in 2001….) For children in the 2000 birth cohort, no comparisons are made, because of the ambiguity over whether those linked to group 3 were born in 2000, when group 3 didn't receive deworming, or 2001, when it did. Moving to 1999, we find more statistically significant cognitive benefits for kids linked to the group 1 and 2 schools, which indeed received deworming in 1999–2000. Something similar goes for 1998. Pushing farther back, to children born before the experiment, we again find little impact, even though a few years after birth some would have had deworming-treated siblings and neighbors and some not. This suggests that the knock-on benefit for younger children was largely to confined to their first year of life.

The evidence that health problems in infancy can take a long-term toll is interesting in itself. But it matters for us in another way too. Suppose you think that because the Worms experiment's quasi-randomization failed to achieve balance, initial cross-group differences in some factor, visible or hidden, generated the Worms at Work results. Then, essentially, you must explain why that factor caused long-term gains in cognitive scores only among kids born during the experiment. If, say, children at group 1 schools were less poor at the onset of the experiment, creating the illusion of impact, we'd expect the kids at those schools to be less poor a few years before and after too.

It's not impossible to meet this challenge. I conjectured that the Worms groups were imbalanced on elevation, which differentially exposed them to the destructive flooding caused by the **strong 1997–98 El Nino**. But my theory foundered on the lack of convincing evidence of imbalance on elevation, which I described above.

At any rate, the relevant question is not whether it is *possible* to construct a story for how poor randomization could falsely generate all the short- and long-term impacts found from the Worms experiment. It is how plausible those explanations would be. The more strained the alternative theories, the more credible does the straightforward explanation become, that giving kids deworming pills measurably helped them.

One caveat: GiveWell has not obtained Ozier's data and code, so we have not vetted this study as much as we have Worms and Worms at Work.

Summary

I came to this investigation with some reason to doubt Worms and found more when I arrived. But in the end, the defenses persuade me more than the attacks. I find that:

- The charge of "high risk of bias" is legitimate but vague.
- Under a barrage of tests, the statistical balance of the experiment mostly survives.
- The original experiment is corroborated by a second, randomized one.
- There is evidence that long-term cognitive benefits are confined to children born right around the time of the experiment, a pattern that is hard to explain except as impacts of deworming.

In addition, I plan to present some fresh findings in my next post that, like Ozier's, seem to make alternative theories harder to fashion.

When there are both reasons to doubt and reasons to trust an experiment, the right response is not to shrug one's shoulders, or give each point pro and con a vote, or zoom out and ponder whether to side with economists or epidemiologists. *The right response is to ask: what is the most plausible theory that is compatible with the entire sweep of the evidence?* For me, an important criterion for plausibility is Occam's razor: simplicity.

As I see it now, the explanation that best blends simplicity and compatibility-with-evidence runs this way: the imbalances in the Worms experiment are real but small, are unlikely to explain the results, and if anything make those results conservative; thus, the reported impacts are indeed largely impacts. If one instead assumes the Worms results are artifacts of flawed experimental design, execution, and analysis, then one has to construct a complicated theory for why, e.g. the user fee experiment produces similar results, and why the benefits for non-school-age children appear confined to those born in the treatment groups around the time of differential treatment.

I hope that anyone who disagrees will prove me wrong by constructing an alternative yet simple theory that explains the evidence before us.

I'm less confident when it comes to *generalizing* from these experiments. Worms, Worms at Work, and Ozier tell us something about what happened after kids in one time and place were treated for intestinal helminths. What do those studies tell us about the effectiveness of deworming campaigns today, from Liberia to India? I'll explore that next.

Notes

[1]The WHO **estimates** that 2 billion people carry soil-transmitted "geohelminths," including hookworm, roundworm, and whipworm. Separately, it **reports** that 258 million people needed treatment for schistosomiasis which is transmitted by contact with fresh water. Children are disproportionately affected because of their play patterns and poorer hygiene.

[2] **Baird et al. (2016)**, Table IV, Panel A, row 3, estimates a 112-shilling increase over a control-group mean of 749/month. Panel B, row 1, suggest that the effect is concentrated in wage earnings.

[3] **Baird et al. (2016)**, Table IV, Panel A, row 1, col. 1, reports 0.269. Exponentiating that gives a 31% increase.

[4] For an overview, I recommend Tim Harford's **graceful take**. To dig in more, see the **Worms authors' reply** and the posts by **Berk Ozler, Chris Blattman**, and my former colleagues **Michael Clemens and Justin Sandefur**. To really delve, read **Macartan Humphreys** and the **reply thereto**.

[5] For literature on the impacts of these study design features on results, see the first 10 references of **Schulz et al. 1995**.

[6] Figures obtained by dividing the "total" point estimates from the linked figures into 1. The study expresses higher benefits as lower risk estimates, in the sense that risk of bad outcomes is reduced.

[7] The **Baird et al. (2016) appendix** defends the "list randomization" procedure more fully.

[8] Deaton may have mentioned **Angrist (1990)** and **Heckman's critique of it**. But I believe the lesson there is not about imperfect quasi-randomization but local average treatment effects.

[9] For the cumulative distribution function of the binomial distribution, F(30,7,0.5) = .00261.

Posted in **Uncategorized** | **Permalink**Tagged **deworming**, **geohelminths**, **schistosomiasis** | **7 Comments**

Previous Post

Comments

Howard White on **December 6, 2016 at 1:41 pm** said:

David
Thanks for the detailed exposition. It is of course very useful to test the robustness of study findings in this way. But the Campbell and Cochrane conclusions of no impact do not coming from doubting the findings of the Kenya study. They come from combining it in meta-analysis with a large number of other studies undertaken in other countries which find no impact.
It is not clear why you or Givewell would base a global recommendation on this one study, especially when there are so many other studies (mostly RCTs) giving contrary findings. Why is the focus so much on just one study in a huge literature?
We had hoped that the Campbell review would throw light on why there were positive impacts in Kenya, but not in most other cases. Unfortunately the analysis was unable to do that. So there may be some settings in which mass deworming is appropriate, though the evidence does not allow us to say what they are at present.
I await your external validity post. But it clearly makes no sense to extrapolate findings of long run impact from a context in which there was also a short run impact to estimate a long run impact for contexts in which we have studies showing no short run impact.

With regards
Howard

David Roodman on **December 6, 2016 at 3:13 pm** said:

Hi Howard.

Good questions, thoughtfully put. Since, as you suggest, I will get to these questions in my next post (how well, you'll need to judge), I won't attempt a complete response here.

But one thought. The evidence we have seems to say: short-term impacts are clear in trials that target infected children (Cochrane reviewed them but I think Campbell didn't); short-term impacts are zero or small in trials of mass deworming; long-term impacts look pretty substantial, at least in Busia 1998–99. If we take all of it at face value, then the simplest explanation is that the short-term effects are modest—too modest to be easily detected in mass deworming trials—and the long-term effects much greater on average. This is not an impossible state of the world. And if it is the real state of the world then there's a strong argument for mass deworming.

Now, I'm not suggesting that we should just take the evidence at face value, else I wouldn't be writing two long blog posts about it. But I think the argument I made is not crazy either, for the simple reason that it fits all the data.

If the key for you is a belief that one should never implement a program on a global scale based on just one encouraging trial—also not a crazy idea—I'd be grateful for elaboration. Why in the case of deworming should we believe that doing nothing (yet) would be better in expectation than plunging ahead?

–David

Jake Goldston on **December 6, 2016 at 3:23 pm** said:

It should be noted that the worms study took place in an area of Kenya with incredibly high rates of geophagy (73% of children according to the study cited in Miguel & Kremer), which seems like a real threat to external validity since most people don't eat dirt every day.

David Roodman on **December 6, 2016 at 4:13 pm** said:

Jake, are you aware of any information about how that compares to other parts of Africa, or South Asia?

Amanda Glassman on **December 6, 2016 at 4:41 pm** said:

WHO data on prevalence are not very good either, worth being more critical on that as well?

David Roodman on **December 6, 2016 at 5:00 pm** said:

Hi Amanda. GiveWell has obtained **some data on prevalence and intensity**, which is fed into the cost-effectiveness calculations. So if "the" prevalence is half what it was in Worms at baseline, that generates a 50% discount. But I don't know how good the data are, and I'm not certain whether they pertain to the areas where the charities work or to countries as a whole.
–David

Julia Wise on **December 7, 2016 at 11:53 am** said:

Fascinating about the geophagy, Jake. The study you refer to is about schoolchildren (**https://www.ncbi.nlm.nih.gov/pubmed/9270730**)
My guess as a parent is that most babies and toddlers do eat dirt every day. Not sure how that would affect worm infection rates at school age, or the effectiveness of re-treating schoolchilren. Perhaps in areas where children eat dirt in the first two years but not much after, a single treatment at age 5 is needed but subsequent treatments aren't? Total guesswork here.

New Comment

Your email address will not be published. Required fields are marked *

Comment *

□ Notify me of followup comments via e-mail

Name *

Email *

Website

Syariah BANCHORY

Type the text

Privacy & Terms

POST COMMENT

HOME CONTACT STAY UPDATED FAQ FOR CHARITIES SITE MAP
OPEN PHILANTHROPY PROJECT

FOLLOW US:

SUBSCRIBE TO EMAIL UPDATES:

EMAIL ADDRESS Submit