What Has Been Learned from the Deworming Replications: A Nonpartisan View

Macartan Humphreys, Columbia University, 18 August 2015

For those in a hurry, follow the orange text.

Summary: A heated discussion on the value of mass deworming campaigns followed the release by a team of epidemiologists (Aiken et al, 2015; Davey et al 2015) of a replication analysis of an influential study on the educational benefits of deworming in Western Kenya (Miguel and Kremer 2004). Despite strong critiques of a seminal paper, and many rounds of responses, neither side appeared to change their views much on the quality of the evidence in the paper. Here I go back to the data and the arguments on both sides. My conclusion is that the replication *has* raised (or in some cases, highlighted) important questions both over the strength of evidence for spillovers and for the strength of the direct effects of deworming on school attendance – at least insofar as these pass through a worms mechanism. There should have been learning here. I point to structural factors that may contribute to the polarization of this discussion, make it hard for authors to acknowledge errors, and inhibit learning from this kind of replication exercise.

1 Background to the Worm Replications

Last month there was another battle in an ongoing dispute between economists and epidemiologists over the merits of mass deworming. In brief, economists claim there is clear evidence that cheap deworming interventions have large effects on welfare via increased education and ultimately job opportunities. It's a best buy development intervention. Epidemiologists claim that although worms are widespread and can cause illnesses sometimes, the evidence of important links to health is weak and knock-on effects of deworming to education seem implausible. As stated by Garner (http://www.cochrane.org/news/educational-benefits-deworming-children-questioned-re-analysis-flagship-study) "the belief that deworming will impact substantially on economic development seems delusional when you look at the results of reliable controlled trials."

Aside: Framing this debate as one between economists and epidemiologists captures some of the dynamic of what has unfortunately been called the "worm wars" but it *is* a caricature. The dispute is not just between economists and epidemiologists. For an earlier round of this see this discussion here (http://blogs.plos.org/speakingofmedicine/2012/07/18/should-deworming-policies-in-the-developing-world-be-reconsidered/), involving health scientists on both sides. Note also that the WHO advocates (http://www.who.int/elena/titles/deworming/en/) deworming campaigns.

So. Deworming: good for educational outcomes or not?

On their side, epidemiologists point to 45 studies that are jointly analyzed in Cochrane reports. Among these they see few high quality studies on school attendance in particular, with a recent report (http://www.cochrane.org/CD000371/INFECTN_deworming-school-children-developing-countries) concluding that they "do not know if there is an effect on school attendance (very low quality evidence)." Indeed they also see surprisingly few health benefits. One randomized control trial (http://www.thelancet.com/journals/lancet/article/PIIS0140-6736%2812%2962126-6/abstract) included one million Indian students and found little evidence of impact on health outcomes. Much bigger than all other trials combined; such results raise questions for them about the possibility of strong downstream effects. Economists question the relevance of this result and other studies in the Cochrane review.

On their side, the chief weapon in the economists' arsenal has for some time been a paper from 2004 on a study of deworming in West Kenya by Ted Miguel and Michael Kremer, two leading development economists that have had an enormous impact on the quality of research in their field. In this paper, Miguel and Kremer (henceforth MK) claimed to show strong effects of deworming on school attendance not just for kids in treated schools but also for the kids in untreated schools nearby. More recently a set of new papers focusing on longer term impacts, some building on this study, have been added to this arsenal. In addition, on their side, economists have a few things that do not depend on the evidence at all: determination, sway, and the moral high ground. After all, who could be against deworming kids?

The most recent battle played out quickly:

- It began on 22 July 2015 when a 3ie (http://www.3ieimpact.org)-funded replication and reanalysis of the critical 2004 study came out as two articles in the International Journal of Epidemiology (http://ije.oxfordjournals.org/content/early/2015/07/21/ije.dyv127.abstract). The studies by Alex Aiken, Calum Davey and colleagues found (or, rediscovered—more on that below), myriad errors in the 2004 study, left the replicators in doubt about the quality of the design and data, and in doubt about the credibility of some of the core results.
- The media moved quickly, making hay of these results, claiming that epidemiologists had debunked the case for worming, even though they had only cast doubts on one study. The Guardian (http://www.theguardian.com/global-development-professionals-network/2015/aug/05/explainer-wormwars-deworming-science-kenya) was one of the worst offenders.
- Soon after there was a blister of blogs and a torrent of tweets defending MK and the case for deworming, oftentimes maligning the epidemiologists and in some cases stating that confidence in the findings was strengthened not weakened by the replication. In perhaps the low point, Gertler, a prominent development economist, essentially accused the epidemiologists of cheating in their analysis (http://blogs.berkeley.edu/2015/08/03/good-science-gone-wrong/); another economist said that the epidemiologists' work would not pass an undergraduate econometrics exam (https://twitter.com/RunningREs/status/624279860651888640). The drama was palpable. The subject was important, the lines sharply drawn. From the outside at least, it all looked surprisingly partisan.

A take home, articulated by The Guardian (http://www.theguardian.com/global-development-professionals-network/2015/aug/05/explainer-wormwars-deworming-science-kenya), was that there is no real news for deworming (they quote CGD (http://www.cgdev.org/blog/mapping-worm-wars-what-public-should-take-away-scientific-debate-about-mass-deworming): "the policy case for mass deworming is largely unchanged") and we have learned about how science is both messy and self-correcting. On to the next topic.

In my less optimistic read, the replication *has* raised important questions over the claims of the paper, if the policy conclusions are unchanged it is only because there is no agreement on the policy conclusions, and my optimism in the self-correcting nature of science has taken a beating.

To make these points I will try to stay close to the data. MK have made this very easy by posting data and extensive documentation at dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28038). This kind of transparency is still unusual. Here I will try to keep up on the transparency end by writing this document in R markdown (http://rmarkdown.rstudio.com/), a free interface which allows me to do analysis and writing at the same time. My code is integrated with the .Rmd (https://www.dropbox.com/s/m9xm4jb41fuzafr/worms.Rmd?dl=0) version of this comment (here (https://www.dropbox.com/s/m9xm4jb41fuzafr/worms.Rmd?dl=0)). No more waiting for the code to be cleaned up after the paper is produced. I will talk first about the study that this battle was fought over. Then I will talk about what if anything we have learned from this for the broader case for deworming. Then I will talk about science and how poorly the system has worked here.

Note I call this a nonpartisan view because I am in neither discipline and do not work on the topic. I have also reached out and discussed a number of these points with authors on both sides of this discussion. With that said I am not completely detached as I am a coauthor on an article (http://www.sciencemag.org/content/343/6166/30.summary) on transparency for which Miguel (also briefly a former classmate!) was the lead author. My initial sympathies, as well as the way I think about experiments and experimental estimates, lie closer to economists, even if many of my conclusions do not.

2 The paper I: Overview

2.1 The key analyses

The main claims of the 2004 paper are that a field experiment shows that deworming has (a) strong effects on worm infection in treated students (b) strong effects on infection in nearby untreated (and treated) students (c) strong effects on school attendance in treated students and (d) strong effects on school attendance for nearby untreated students. However, (e), there is little evidence for short-term effects on school performance. Finally, (f) the way this works is through the medication and through health. Claim (f) is implicit in the policy prescriptions emanating from the research.

The design of the field experiment is a little complicated since there were three sets of schools and two years. Group 1 had treatment in both years; Group 2 had treatment in the second year only; and Group 3 never had treatment. Below (following a chunk of code) is the raw data, showing the unweighted average attendance for each of these groups (I am doing this and everything else, in R (https://www.r-project.org/), not STATA):

```
# Get the data
T9U               <- read.dta(paste(datapath, "MK.T9.U.dta", sep=""))
T9U$years_treated <- T9U$t1+2*T9U$t2
T9U$groups        <- 3 - 2*ave(T9U$t1+T9U$t2,T9U$pupid)
# Cross tab
present           <- xtabs(prs ~ yr+groups, data = aggregate(prs ~ yr+groups,T9U,mean))
row.names(present) <- c("Year 1", "Year 2"); colnames(present)  <- c("Group 1", "Group 2", "Group 3")
knitr::kable(round(100*present, 1), caption="Raw Attendance Data.  Group 1 is treated in years 1 and 2; Group 2 in year 2 only; Group 3 is not treated either year")
```

Raw Attendance Data. Group 1 is treated in years 1 and 2; Group 2 in year 2 only; Group 3 is not treated either year

	Group 1	Group 2	Group 3
Year 1	82.8	72.7	75.1
Year 2	70.4	70.9	65.3

You will notice that with this kind of design there are really four distinct experiments taking place at the same time. These comparisons are implied by the design (but they do not all jump out from regression tables used by MK):

- Effect a1: the effect of a single year of treatment in year 1: compare Group 1 against Groups 2 and 3 (pooled) in the first year
- Effect a2: the effect of a single year of treatment in year 2: compare Group 2 against Group 3 in the

second year
- Effect b: the effect of two years of treatment: compare Group 1 and Group 3 in the second year, and
- Effect c: the effect of a *second* year of treatment: compare Groups 1 and Group 2 in the second year.

Note that effect *c* is the same as the difference between *b* and effect *a2*. That is, it is the difference between two years of treatment and one year of treatment delivered *in the second year*. Now for efficiency one might think of years as blocks and estimate the one year treatment effect as the average of *a1* and *a2*. Call this effect *a*. I think that is reasonable to do this kind of pooling here (though Cochrane (http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000371.pub6/abstract) does not because of lack of independence), but note that in that case there is no reason to expect that *c* is the difference between *a* and *b*.

In addition MK[1] (and Ozler (http://blogs.worldbank.org/impactevaluations/worm-wars-review-reanalysis-miguel-and-kremer-s-deworming-study) after them) argue for a comparison of Group 1 both years against Group 3 both years. This makes good use of the data but note that it means taking an average of the effects of two different treatments (one year of treatment and two years of treatment).

The implied effects for *a*, *b*, and *c* from the raw data look like this:

```
dir.effects(present)
```

```
##  One year effect:  Two year effect:  2nd year effect:
##              7.24              5.04             -0.54
```

MK argue however that there are plausibly important externalities, in which case we cannot read the treatment effects straight from this table of average values. Instead they employ a model that seeks to account for spillovers, and that includes various controls, weights and so on. (The weights here seem to be used to put more weight on individuals for whom more attendance data was collected, presumably to improve efficiency). The spillover variables include measures of proximity to other students treated by the project within 0-3km and 3-6km bands. Note that when those measures were created there were some calculation errors made by MK which they and the replication team agree about.

The below regression model uses the corrected data as provided by MK and their original model.

```
# Table 9 Col 3 Replication
# First a little rescaling gets more manageable coefficients
T9U$pop_3km_updated <- T9U$pop_3km_updated/1000;  T9U$pop_36k_updated <- T9U$pop_36k_updated/1000
# Now run model T.9
T.9 <-  lm(prs~t1+t2+
          elg98+p1+mk96_s+ Y98sap1 +Y98sap2 +Y98sap3 +Y98sap4 + sap1 + sap2+ sap3+ sap4 +
          Istd1+  Istd2 +  Istd3+ Istd4+  Istd5 +  Istd6 +  Istd7 +Istd8 +Istd9 + Isem1 +
          Isem2 + Isem3 +  pop_3km_updated+popT_3km_updated+pop_36k_updated+ popT_36k_updated,
          weight = obs, data = T9U, na.action="na.exclude" )
T.9.miss <- is.na(predict(T.9))
# Adjustment for clusters
T.9.se <- cl(T9U, T.9, T9U$sch98v1)[c(2:3, 27,29),1:4]
rownames(T.9.se) <- c("First time treated", "Second time treated", "Treated pop within 3km", "Tre
ated pop 3 - 6 km")
# Display
knitr::kable(round(T.9.se,3), caption="Key estimates from the model: direct effects and spillover
effects.")
```

Key estimates from the model: direct effects and spillover effects.

	Estimate	Std. Error	t value	Pr(>|t|)
First time treated	0.062	0.014	4.507	0.000
Second time treated	0.033	0.021	1.568	0.117
Treated pop within 3km	0.039	0.022	1.807	0.071
Treated pop 3 - 6 km	-0.024	0.015	-1.656	0.098

The results reproduced here correspond to MK’s corrected model, with a small difference in one final digit due I believe to rounding error (here we get the exact computer printout delivered without need for human transcription and human rounding).

The fitted values for different treatment effect combinations (all other values held at their (unweighted) means) is given below:

```
knitr::kable(round(100*present.T.9, 1), caption="Fitted values from model; all else at mean.  Gro
up 1 is treated in years 1 and 2; Group 2 in year 2 only; Group 3 neither year.")
```

Fitted values from model; all else at mean. Group 1 is treated in years 1 and 2; Group 2 in year 2 only; Group 3 neither year.

	Group 1	Group 2	Group 3
Year 1	76.3	70.1	70.1
Year 2	73.4	76.3	70.1

You can see from the fitted values table that the model is constraining in a couple of ways (ie there are only 3 distinct numbers in the 6 cells); for example it does not allow for a time trend or for the spillover effects to work differently in different periods. In effect this means that the one year effect is constrained to be the same in year 1 and year 2, and the effect of the second year is the difference in effects for a two year treatment and the period-averaged one period treatments. However if we fit a “saturated” model that more closely captures the experimental comparisons as taking place *within* year we end up getting very similar treatment effects (fitted values below).

```
knitr::kable(round(100*present.T.9.S, 1), caption="Fitted values from (partially) saturated mode
l.  Group 1 is treated in years 1 and 2; Group 2 in year 2 only; Group 3 neither year.")
```

Fitted values from (partially) saturated model. Group 1 is treated in years 1 and 2; Group 2 in year 2 only; Group 3 neither year.

	Group 1	Group 2	Group 3
Year 1	81.6	74.8	76.3
Year 2	67.6	70.6	63.9

Returning to the model in the paper, the first estimated effect is the marginal effect of one year of treatment on a treated unit, using the corrected data. The implied percentage point increase in attendance is 6.24, or equivalently the percentage increase is 8.9 (or more dramatically the percentage *decline* in absenteeism is 20.9). As with the raw data you can see that a second year of treatment is much less powerful than a first year (though note that the difference between the two effects is itself not significant, $p = 0.17$. To be clear this does not just mean that the marginal gain from the second year is lower, it means that places with two years of treatment are worse off than places with one year. The difference between the two is also negative here (by construction it has to be here since it is the difference between the two year and one year effects). Interestingly using the model, the effects are *smaller* than we see with the raw data. This is contrary to what you might expect from a spillovers logic but then again the model includes lots of things, weights, controls and so on.

Note that although this one year effect is very similar to MK's reported 7.5 percentage point drop in absenteeism (or 25% drop), the most cited number from this study, that's measuring something different. I will get to that number below. Note also that the one year effect looks strongest in year 1 results (for schools in Group 1, year 1).

All sides agree on a sizable marginal direct effect of the program on absenteeism (as long as the original estimation strategy is used on the corrected data). However there are disagreements about whether this is the right estimation strategy and about how to interpret the number.

2.2 Where the disagreement is and isn't

More generally here are the points of agreement and disagreement between the original team and the replication team.

2.2.1 Errors and issues that both sides agree about

There are at least six important points where MK and the replication team are in agreement.

1. **Headline effect miscalculated:** The headline effect of the original paper is a 7.5 point drop in absenteeism (or 25% drop). This number is still being quoted (http://blogs.berkeley.edu/2015/08/03/good-science-gone-wrong/). The effect represents the difference in outcomes for a typical treated unit relative to a hypothetical situation where *no* units are treated (I discuss that calculation of this more below). This is different from the marginal effect since it includes externalities to treated groups. **In fact**, correcting the error and using the same estimator as in the original article you get an overall effect just below 4 points, and not significant. Share 0.73 of the original effect was due to the direct effect. If you alter the model a bit you can get the effect back up to 8.5 (http://escholarship.org/uc/item/8db127cm). More on that below.

2. **Weaker evidence on spillovers (infections):** The paper stated that there were significant spillover effects (at the 5% level) on infections at the 0-3km range and the 3-6km range and that the overall spillover effect between 0-3km and 3km-6km was about -0.23 and statistically significant (MK05, p187).
In fact using the corrected data and original strategy, only the first estimate is significant, it is about 20% smaller (-0.26 versus -0.21) and the uncertainty around it (in terms of its standard error) has increased. The other two quantities are not significant.

3. **Weaker evidence on spillovers (attendance):** The paper stated that there were significant spillover effects at the 5% level on attendance at the 0-3km range (but not the 3-6km range) and that the overall spillover effects between 0-3km and 3-6km was about 0.2 (and statistically significant).
In fact using the corrected data and original strategy, none of these three quantities are significant at 5% and the overall estimated effect is -0.02 not +0.02. Again one regain significance on the 0-3km effects but it requires changing the estimation strategy.

4. **Estimates of direct effects largely unaffected by the calculation error:** The *direct* marginal effects on infections are similar in the original and in the corrected tables both in terms of magnitude and uncertainty. Indeed in some specifications they are slightly larger. These correspond to a 0.31 reduction in infection rates for children in treated schools and, for attendance, a 0.062 increase in the first year of treatment (regression estimates), which declines to 0.03 for children in the second year of treatment (regression estimates).
5. **Weaker second year effects:** The estimated *second-year of treatment* effect – i.e. comparing children who have been taking treatment for two years compared to none, was marked as significant at the 10% level in the original article.
   In fact this second year effect is not significant even at 10%.[2]
6. **Anemia link not significant:** The paper erroneously stated that there was a relation between deworming and anemia. As noted by MK this was a substantively weak relationship in the first place, though it was marked as significant at 5%. **In fact** this relationship is not significant even at 10%. This means that in the corrected tables the only relations between treatment and health significant at 5% are on the self-reported measures, not the anthropometric measures.

These are not small points of agreement. To be clear, although you mightn't guess it from the tone of exchanges, **all sides agree that errors were made and that these affect a substantial number of the estimates provided by the paper, including the headline claims**. Interestingly, the original research team discovered and wrote up many of these in 2007, so to a large extent what this replication has done is *drawn attention* to known errors.

2.2.2 Other concerns raised by the replication team

Here are two other important concerns raised by the team that I elaborate on below:

1. This is not really a randomized trial and moreover the way it was done created potential **imbalance in the handling of treatment and control groups** which could lead to biased estimates, perhaps creating a spurious estimate of effects on attendance.
2. The **explanation** does not hold up. In the absence of health effects, the presumed relation between worm infections and attendance seems implausible.

These two concerns are not the ones at the center of recent discussions, though I think they might be the most important ones (MK did however provide a response to 1. especially, here (http://www.3ieimpact.org/media/filer_public/2015/01/07/rps3_worms-3ie-pure-response_2014-12-22-part_2.pdf)).

2.2.3 Other concerns that have me puzzling

Four other things have left me puzzling:

1. **Quantities of interest** I am a little puzzled by the headline quantity of interest and think there are more natural ones to employ both for understanding direct and total effects. The results on some of these are more worrying however, notably the effect of having two years of treatment is weaker than the effect of having one year of treatment, and the implied effect of a *second* year of treatment is actually negative.
2. **Specification selection** MK defend the headline results by changing the analysis, and in my read, changing both the estimand and the estimator. I can understand the temptation to do that, but given the work by Miguel and other colleagues (http://qje.oxfordjournals.org/content/127/4/1755.abstract) on the dangers of *post hoc* selection of estimation strategies I cannot understand why this is not accompanied at least by a warning to downweight the strength of the evidence.

3. **Spillovers to untreated schools.** The claim that the treatment improved outcomes for children in **untreated schools** does not seem to be supported by most of the tables.
4. **Communication of errors.** I have been teaching this paper for years, and I did not know that the original research team had found out about these errors many years ago; to date the journal has published no *errata* and it seems that in the exchanges with the Cochrane team these errors were not mentioned.[3]

2.2.4 Other points of disagreement between MK and replicators

The replicators suggest that the main effects are not robust and lose significance with various modifications of the analysis. I think there is reasonable justification for a lot of these analyses but in my read they don't make all that much of a difference. The coefficient on direct effects is robust to many of the alternative estimation strategies that Davey et al examined. To illustrate robustness a bit the table below runs the model with no controls or weights and focusing only on Group 1 and Group 3 and *counting only one observation per pupil* (the average attendance across both years). Things hold up pretty well both in terms of effect sizes and the stats.

```
D4      <- aggregate(prs ~ groups + sch98v1 + year +pupid,T9U[T9U$groups!=2,],mean)
M4      <- lm(D4$prs ~ D4$groups==1)
T.4.se  <- cl(D4, M4, D4$sch98v1)
knitr::kable(round(round(T.4.se[1:2,],3),2), caption="A very bare bones model: Groups 1 and 3 onl
y, one observation per student; no controls or weights.")
```

A very bare bones model: Groups 1 and 3 only, one observation per student; no controls or weights.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.70	0.02	31.96	0.00
D4$groups == 1TRUE	0.06	0.03	2.21	0.03

2.3 What's gone and what remains?

My overall reading is this:

Many of the claims of the paper are weakened. The link to health outcomes (especially anemia) is weakened. The evidence of externalities is weakened. The evidence for externalities on the untreated schools *specifically* was never clearly there.

What remains is a tightly estimated coefficient on the direct effect linking being in a treated school and attending classes. However there are still at least three reasons to worry about this main effect: (a) there are missing links in the preferred mechanism: there is little evidence for direct effects on the intermediate health outcomes, and the estimated effect of treatment seems to operate even after controlling for worms (b) it seems possible that other features of the design gave rise to spurious estimates, in particular the uneven treatment of treatment and control and (c) the effect is only clearly present for the first year of this strategy anyhow. In the second year it is much smaller and not significant. All three of these claims are consistent with the possibility that this attendance effect, if real, is not an effect of the medication itself.

Also remaining is evidence that worm medication gets rid of worms, but no one ever seemed to doubt that.

3 The Paper II: Five takes on the analysis

There are different things to look out for in every paper, but I think five key things that very often warrant particular attention are the following:

1. **The identification**: On what basis does the paper claim to have estimated a true *causal* effect? For an RCT the identification is generally justified by appeal to the virtues of random assignment to treatment.
2. **The measurement**: How do you measure the inputs and outputs? Are the measures reliable? Validated? Properly constructed?
3. **The interpretation**: How do you map from the quantities that you estimate to the quantities you care about? For example from a coefficient in a regression table to a claim about effects in some population.
4. **The explanation**: Though not essential for causal inference, compelling accounts tend to have plausible stories attached to them so that we can be convinced of the logic and we can reason about where else we might expect such effects to operate.
5. **The communication**: How were the results communicated? Did the right things get highlighted?

In my reading of the replication and reanalysis, and re-reading of the paper, there are possibly significant concerns regarding all five dimensions. Let’s take each in turn.

3.1 Identification

Figuring out what causes what is one of the hardest challenges of social science. So many correlations suggest spurious causal relations. Did taking my umbrella out cause it to rain? More subtly, did wearing rainboots cause me to take my umbrella out? Social scientists have come up with a few strategies to figure out causal relations with confidence, but the most important is the use of randomization. Miguel and Kremer and their colleagues have had a huge impact on development economics and beyond by advancing the use of randomization to assess causal effects. The basic insight of randomization is a simple one: If units are randomly assigned to a treatment (and not self-selected in some way) then the only thing different between them and people not in treatment (“in expectation”) is the treatment and so later differences we see between them and people not getting the treatment must be due to the treatment.[4]

That’s all fine, but it assumes three things. First that there really is randomization. Second that there are no spillovers in the sense that a given unit’s treatment status does not affect another unit’s outcomes (or, alternatively, that all spillovers are fully understood). Technically this is called the SUTVA (https://en.wikipedia.org/wiki/Rubin_causal_model) assumption. Third, that the allocation of the treatment works only through the treatment itself and not through *how* it is allocated (technically this is an exclusion restriction).[5]

The MK study is taught as a casebook randomized control trial. How well does it do on these three criteria?

3.1.1 Randomization

As highlighted by the replication team, there was not, strictly, randomization of the direct treatment. Thus to claim the benefits of randomization you need to make an *as-if* random assumption. In this case, given the method that was used, I expect many people would be willing to make that assumption. However, more subtly, there was also no randomization of the “spillover treatment.” Here however an *as-if* random assumption is very hard to justify since the probability of assignment to spillovers (if we assume spillovers work the way MK do) is demonstrably a function of how many neighbors one has. As I pointed out before (http://www.columbia.edu/~mh2245/w/Worms_Again.pdf), regression with interactions does not restore *as-if* randomization. It’ll work only if you get the model right. An implication is that the claim of unbiased estimates of the spillover effect depends here on the model and not just the design.

3.1.2 SUTVA

If there are spillovers then estimates that do not properly account for these will be biased. For instance if I give a medication randomly to one of two sick people and they both get better because of it, then the medication will have helped two people, but in comparing treatment and control I will think it helped none. MK are keenly conscious of this; they criticize others for not taking it into account, and they argue that many of their own estimates are *under*estimates because of this. They may be right but if you are depending on a model for inference you can be off in very surprising ways. In an example I provided before (http://www.columbia.edu/~mh2245/w/Worms_Again.pdf) you can see how linear regression can pull a positive effect from a zero when it does not take account of assignment propensities properly even with interactions included. The bottom line is that whereas in the absence of spillovers, randomization can get you an unbiased estimate of the average treatment effect, in the presence of spillovers, the randomization can only justify claims to unbiasedness *conditional* on the particular model of spillovers (http://arxiv.org/pdf/1305.6156.pdf) being correct (and correctly estimated).

3.1.3 Exclusion restriction

Perhaps the biggest concern is around the exclusion restriction. I think this one is bothering the epidemiologists a lot. When health researchers administer treatments they are very worried about various kinds of spurious effects: placebo effects, Hawthorne effects, and so on. They use placebos and they blind patients to treatment status because they want to make sure that the only thing that is being done differently by the assignment to treatment is the provision of the treatment itself. They don’t want to pick up the effect of increased contact with subjects, or of subjects trying to please them and taking actions that might complement or counter a treatment.

In social science experiments we often do not have placebos because we cannot. You cannot give a family Placebo housing or placebo lower tax rates. And as MK point out in a response, social scientists often care about the social effects. But you can give placebo medication. In principle this study could have treated treatment and control subjects in the same way – except for treatment. But it did not. Perhaps ethical or practical reasons justify this decision. (Also, in correspondence MK argue that a Placebo would not blind to treatment status well since subjects observe the dispelling or worms, and at the cluster level this would be observable by agents). Either way, it means in practice there were differences beyond treatment between treatment and control subjects. The replication team made many of the points (http://ije.oxfordjournals.org/content/early/2015/07/21/ije.dyv128.abstract) (p11) that follow quite clearly:

> Allocation to the intervention arm could therefore plausibly have affected school attendance through behavioural pathways affected by the educational component of the intervention, the placebo effect of the drug treatment or the Hawthorne effect.

Let’s talk through some of the threats from differential treatment of groups.

1. First there were various difference both in the definitions of the subject population and in how they were treated. For the analysis of worms it turns out that the subjects were sampled using different procedures. Baseline data was gathered on Group 1 treatment students only. This produces a difference in engagement beyond the treatment. Then the researchers sought to follow up on these but failed to relocate, or gather data on, a large group of them. The Group 2 control students in contrast, were sampled afresh. This implies that the groups themselves are not readily comparable.

These concerns are particularly important for the infection results, and not, I believe, for the attendance results. There are perhaps two more important sources of imbalance for the attendance results.

2. Second, in fact the treatment was a bundled treatment. In addition to the medication there were also "regular public health lectures, wall charts, and the training of teachers … on worm prevention" (MK05, 169); and the intervention also "promoted hand-washing, wearing shoes, and avoiding contact with fresh water" (Data User's Guide, p7 (http://emiguel.econ.berkeley.edu/assets/miguel_research/46/PSDP-DUG_2014-11.pdf)). The question then is: could these separate components have had positive effects independent of the deworming? The CDC (http://www.cdc.gov/handwashing/why-handwashing.html) counts hand washing, for example, as "one of the most important steps" to avoid sickness. They list various mechanisms including "Salmonella, E. coli O157, and norovirus that cause diarrhea," as well as adenovirus and hand-foot-mouth disease." Could it be that all this evidence on school absenteeism is really evidence for the effectiveness of these other components, and not deworming at all? Note that MK do check for evidence of cleanliness, as observed by enumerators, and find no evidence of difference between treatment and control schools (their Table V). But no evidence for an effect is not the same as evidence for no effect. So there is ambiguity here.

3. These extra treatments sound like they add up to a lot of contact with kids in the treatment schools that you do not have in the control schools as well as possibly greater knowledge about the experiment in the treatment schools (consent was sought in year 1 only for the treated). The question is: when you have so much contact with these schools and are providing so many services and benefits, is it possible that when you then ask them about attendance, up to six times over the year, that they start making sure people are there for it? This is exactly the kind of concern that epidemiologists have when they worry about blinding and it is why they use placebos and try to ensure balance in terms of contact with research teams and so on etc.

It's hard to make much progress with this issue in the data, though the replication team looked into it a little. It ends up feeling a bit speculative. If I were to investigate demand effects I might look at the year 1 Group 1 schools where so much of the action seems to happen. Many of these first contact schools are reporting almost no absences (see below for the number with attendance above 90%). This pattern is consistent with a half dozen schools working to boost attendance to please the research team. This is of course highly speculative; but if I were going to investigate further I would try to learn more about the connections between these schools and the NGO implementing treatment.

```
school.unweighted <- summaryBy(groups + prs~ sch98v1 + yr, FUN=c(mean), na.rm=TRUE, data=T9U)
school.unweighted$few.absent <- school.unweighted$prs >= .9
xtabs(few.absent ~ yr+groups.mean, data = school.unweighted)
```

```
##    groups.mean
## yr   1  2  3
##   1 10  4  6
##   2  0  1  0
```

3.2 Measurement

The study is marked by real innovations in measurement; including the careful collection of attendance data at multiple points in time for over 50,000 students. This was done via surprise visits rather than relying on class records.

In addition to the innovations there are some documented concerns with the measurement in this study: the replication team had difficulties understanding many features of the measurement, including sampling schemes; and they had concerns over unusually high levels of missing data. They had additional worries over weighting schemes. Lack of complete documentation is not terribly surprising given how long ago the research took place.

The measurement error that seems particularly consequential, and emphasized in the replication, was the miscalculation of the number of neighboring students in treatment in the 3-6km window. Correcting for the mistake results in various changes to estimates as described above. MK suggest a strategy to fix this and some, including CGD ((http://www.cgdev.org/blog/mapping-worm-wars-what-public-should-take-away-scientific-debate-about-mass-deworming)), see this as reasonable. In my read this involves a change of strategy in order to produce significant results (MK basically describe it as such). More on this below. Even after doing that the estimated externality effects for infections are smaller in magnitude than what was estimated before and they come with larger *p* values. All this, together with the identification concerns, leads me to downweight my confidence in this finding.

3.3 Inference

Once models are run there comes the task of interpreting the coefficients. What inferences can be made? There has been a lot of reasonable concern about what inferences you can make from West Kenya to the rest of the world. Few will doubt that I think. Here I focus on simpler issues regarding the inferences for the study population itself.

3.3.1 Adding up the benefits: where the 7.5 number came from and what it means

There is some trickiness in adding up the benefits from an intervention with spillovers like this.

The headline result is that ``[t]he program reduced school absenteeism in treatment schools by one-quarter'' It took me a while to figure out what the headline result meant and I had it wrong in my discussion here (http://www.columbia.edu/~mh2245/w/Worms_Again.pdf).[6] Here is how to think about it. Say you had a process of the form:

$$y = a + bT + cn^T + dn + \dots$$

where y is the outcome (absenteeism say), T indicates whether or not you are treated, n^T indicates the number of neighbors you have that are treated and n is the total number of neighbors you have.

We often then wonder what the marginal effect of a change might be: what's the gain from assigning one more person to treatment, say. That is easy enough here if we believe the model. Imagine one extra person, Molly, gets treated. What are the gains? The gain for Molly is just b. We can call this the "*marginal private benefits for a treated unit*" (note, there is also possibly an extremely small indirect effect for Molly which arises from the indirect benefits to her that circle back from the people that indirectly benefit from her treatment, but those effects are ignored here). Then there are also the indirect benefits to all the other people from a marginal increase in the number of treated people close to Molly. These all get benefit c. How many people are there like that? There are n (not n^T). So the total gains are $b + cn$, for a unit with n neighbors. We can call these "*the marginal social benefits*."

MK however calculate something between b and $b + cn$, they are interested in $b + c\overline{n^T}$, where $\overline{n^T}$ is the average number of treated neighbors a unit has. What is this quantity? This is the typical benefit to a treated unit from the *entire* intervention, assuming that is, that treated units also have $\overline{n^T}$ neighbors on average. Call it the "*total program effect for a treated unit*."

It's a curious quantity. If you believe the model, the marginal effect depends on the geography, but not on treatment saturation. But the total program effect for the treated does depend on saturation (as well as geography), which here is a function of experimental design only. Interestingly, if you are interested in the effect on the treated *and* are interested in mass deworming, and believed the linear model, then $b + cn$ would be both the marginal effect *and* the total effect on the treated (from a *mass* program). So the $b + \overline{cn^T}$ quantity gets at an estimand you might not care too much about (the effect of one third / two thirds treated) but also puts pressure on the model since you have to extrapolate to an imaginary control with no intervention at all (a type of “uniformity trial”).

It was also hard to figure out this number because the paper states that number comes from regression 3 of Table 9 but I think that’s not right; I believe it is generated by estimating a model where the first year and second year effects are replaced with an “any year” variable that forces a kind of common effect over first and second year of treatment.

In summary, though it gets cited a lot, I expect there is a lot of confusion over what the headline quantity mean. It is not the marginal social benefit of treatment. If one did focus on the marginal social benefit this would in fact increase the estimated benefits by a factor of two or three.[7]

3.3.2 Who are the spillovers good for?

The model above, which is similar to the core model used by MK, assumes that spillover effects are the same for treated and untreated individuals. If you thought that spillovers worked differently for people in treatment and control, or if you were interested in effects specifically for either of these groups, then you might estimate a model with *interactions* like this:

$$y = a + bT + cn^T + dTn^T + \dots$$

In that case the spillover gains from treatment for a neighbor (for whom n^T has increased) is c if they are not themselves treated and $c + d$ if they are. This kind of model can produce evidence that a treatment produced benefits *specifically* for the untreated; that would register as a strong coefficient on c.

Although the abstract and, recent discussions (http://blogs.berkeley.edu/2015/08/03/good-science-gone-wrong/), describe the punchline as being about the effects on untreated schools, suggesting something like the interactions equation above, the main results (Table 9 of the main article) do not have these interactions. If you do them it looks like the effects are mostly happening for the treated, not the untreated (see Table below). Similarly if you look at the infection results[8], evidence does not seem to be there for action on the *untreated* specifically.[9]

```
knitr::kable(round(T.9INT.se,2), caption="Interactions of direct and indirect effects for the att
endance model.")
```

Interactions of direct and indirect effects for the attendance model.

	Estimate	Std. Error	t value	Pr(>|t|)
0-3km effect for control	0.02	0.02	0.92	0.36
Additional 0-3 km effect for treatment	0.05	0.02	1.93	0.05
3-6km effect for control	-0.03	0.02	-1.96	0.05
Additional 3-6 km effect for treatment	0.02	0.01	1.67	0.09

Given MK’s logic it might make some sense if the gains are in fact strongest for the treated since the externality works not through removing an infection but from preventing a *reinfection*.

If that's right this suggests something of a re-interpretation of results but not necessarily a weakening of them. Interestingly, it might increase support for a deworming program since participants, as a group, internalize more of the benefits.

3.3.3 Multiple Comparisons

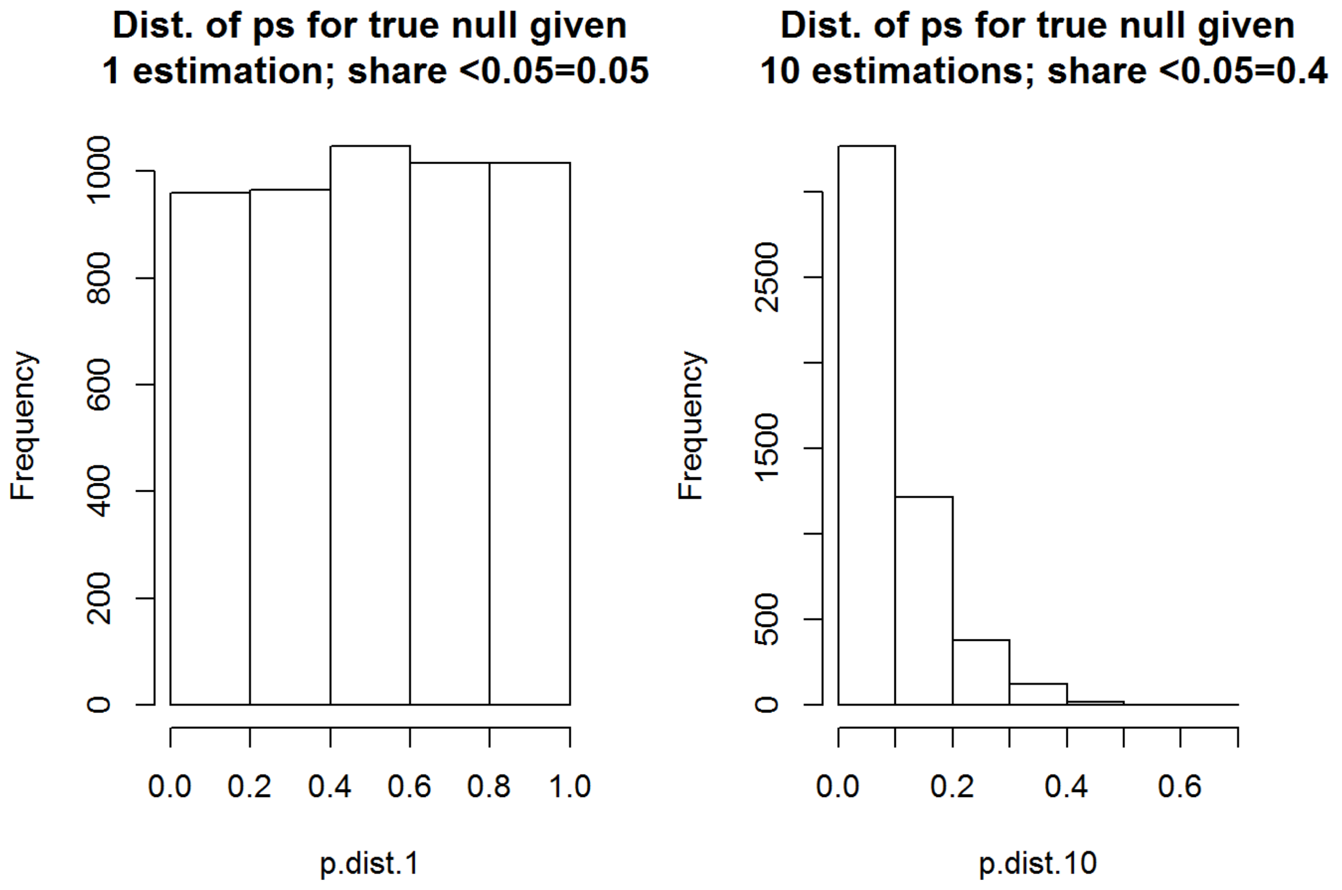
The original worms team seem not to take the worry about the calculation error too seriously because they see a principled reason to alter their analysis in a way that recovers a strong effect. Epidemiologists might worry that this is a form of unprincipled data mining: essentially if you reject an analysis just because it produces estimates with high variance then you are selecting for results that are more likely to be significant even if the null of no effect is true. Interestingly on the opposite side the epidemiologists were accused of this kind of fishing even though their analysis, unlike the original, followed a preanalysis plan (with deviations transparently declared). Of course when it was written there would have been no expectation that the original MK analysis would have a predefined analysis plan.

Even still, the puzzling thing for me was the change of strategy by the authors now. They had clearly thought about this issue, but I could not understand their answer. They argue that they prefer the new specifications over the old ones because they minimize the mean squared error – that is, the loss in precision outweigh the gains in bias reduction from the old analysis; it's better to go for the more precise estimate.

Here's a thought experiment that shows why you might be uneasy with this argument. Imagine that there were no concerns at all about bias and you could choose between multiple specifications that, though each unbiased, varied in precision (but with similar estimates). If you then selected the more precise one, you would be likely to find a significant effect even if there were none, and, your reported p values would be incorrect.

Here's a quick simulation that makes the point. Imagine no effect of X on Y but you have access to 10 distinct and entirely informationless measures of X. You decide to use the one that minimizes estimated error. The figure shows what kind of *p* values you might expect to get.[10] The bottom line is that the probability of getting $p \leq 0.5$ is much larger than 0.05 even if there is no true relationship. (Note, this simulation is designed to make the general point, though in in the MK reanalysis they had fewer degrees of freedom than assumed in this simulation).

```
# Say there are 10 completely noisy measures of some value X
# You want to examine the effect of X on Y. Say the true effect is 0.
set.seed(20150812)
ssr.p   <- function(x,y=NULL) {M <- lm(y~x)          # Get ssr and p
                        c(mean(M$residuals^2),
                        coef(summary(M))[2,4])}
p.sim <- function(m=10, n = 1000){
  y <- rnorm(n)
  X <- matrix(rnorm(m*n), n,m)
  q <- t(apply(X, 2, ssr.p, y = y))
  min.ssr <- q[,1]==min(q[,1])
  q[min.ssr, 2]
  }
p.dist.1  <- replicate(5000, p.sim(m = 1))
p.dist.10 <- replicate(5000, p.sim(m = 10))
{par(mfrow = c(1,2))
hist(p.dist.1, main = paste("Dist. of ps for true null given", "\n 1 estimation; share <0.05=",
  round(mean(p.dist.1 <= 0.05),2), sep=""), breaks=5)
hist(p.dist.10, main = paste("Dist. of ps for true null given", "\n 10 estimations; share <0.0
5=",
  round(mean(p.dist.10 <= 0.05),2), sep=""), breaks = 5)
}
```



To be clear it makes sense to use estimators that minimize mean squared error; the problem arises when your estimates of the MSE are a function of the same data that you then use the estimator on. Even if you planned ex ante to select estimators in this way, you still have this problem, though it becomes more clearly a multiple comparisons problem and not a selective reporting problem. In such cases it is increasingly common practice to use multiple comparison corrections (http://egap.org/resources/guides/10-things-you-need-to-know-about-multiple-comparisons/).

In exchanges, MK suggested an approach to multiple comparisons corrections suggesting one could do a False Discovery Rate (FDR) or a Bonferroni correction on different estimates they provide. They suggest doing this on three estimates of total effects (ie estimates that a) include no spillover terms, b) include the 0-3km terms only, and c) include both the 0-3km and the 3-6km terms). If you do this the first two survive fine. In my read this confirms the robustness of the direct effect, which is important, but my worry has been around the robustness of the spillover effects. The suggested test could yield significant effects even if there were no spillover effects. However it is easy enough to do the same thing on estimates of the “average overall cross school externality effect” using the original model with and without the 3-6km terms. Then however things don’t look so strong. Things are so close to the wire that small changes in specification push you above or below significance too easily (in a sense this is a rut that the discussion of this piece needs to get out of; more on that below).[11]

```
T.9.2 <-  lm(prs~t1+t2+
          elg98+p1+mk96_s+ Y98sap1 +Y98sap2 +Y98sap3 +Y98sap4 + sap1 + sap2+ sap3+ sap4 +
          Istd1+  Istd2 +  Istd3+ Istd4+  Istd5 +  Istd6 +  Istd7 +Istd8 +Istd9 + Isem1 +
          Isem2 + Isem3 +  pop_3km_updated+popT_3km_updated,
          weight = obs, data = T9U, na.action="na.exclude" )
round(cl(T9U, T.9.2, T9U$sch98v1)[27,],3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##      0.046      0.022      2.123      0.034
```

This *p* value, together with an estimate of the *p* value from the original model imply the following corrected values under different approaches to correction for multiple comparisons:

```
p=c(cl(T9U, T.9.2, T9U$sch98v1)[27,4], 2*(1- pnorm(0.017/0.030)))
rbind(p=round(p,3),
bonferroni = round(p.adjust(p, "bonferroni"),3),
holm = round(p.adjust(p, "holm"),3),
BH = round(p.adjust(p, "BH"),3))
```

```
##             [,1]  [,2]
## p          0.034 0.571
## bonferroni 0.068 1.000
## holm       0.068 0.571
## BH         0.068 0.571
```

3.4 Explanation: The Mechanisms

The *explanation* for the effects on attendance suggested by the Worms paper is simple: worms in your gut make you sick and this leads you to miss school. The explanation seems obvious but still it doesn’t seem to convince epidemiologists. Why not?

The key thing is that although they could see evidence for a reduction in worms they could not see the changes in health outcomes that you would expect to see to explain behavioral change. The Cochrane reviews, examining a wide variety of studies, but excluding the economics paper because of doubts about various features of design and measurement and analysis (as well as odder concerns about baseline data and provision of benefits to controls), concluded "in mass treatment of all children in endemic areas, there is now substantial evidence that this does not improve average nutritional status, haemoglobin, cognition, school performance, or survival." (Note: some surprise to see a positive statement about null effects here.)

If deworming doesn't improve health much, how could it have knock-on effects to education? Understanding the mechanisms becomes particularly important if there are worries about the exclusion restriction, Hawthorne effects etc, as there are here. Simply put, if there is no evidence for the right mechanism then isn't it reasonable to think that the result could just be due to Hawthorne or other spurious effects? I already noted above possible rival explanations (Hawthorne effects etc). In addition there are a couple of things we can do to get a handle on the plausibility of the proposed mechanism though. One is a thought experiment. The other a kind of mediation analysis.

3.4.1 A thought experiment: How great would the effect of worms on attendance have to be to explain this large treatment effect?

Let's ask: how strong would the effect of worms on attendance have to be to make sense of the attendance result? Is it within the sort of ranges that epidemiologists might believe?

Luckily, we have the elements to do a rough calculation. Say that we believe the following two relations from the MK data for Groups 1 and 2: deworming students cuts worm infections from `56.2%` to `24.9%` ; deworming students cuts absenteeism from `29.9%` to `23.7%` (based on direct effect of one year of treatment).

Then, let's assume an average probability for going to school if you do and do not have worms, *and that this probability itself does not depend upon treatment*. Then if we work those probabilities out we find that for everything to add up, kids with worms would have to have a `38.6` % probability of being absent compared to a baseline absenteeism rate of `18.7` % for kids without (note, in correspondence HMK suggest that the effect on infections here might be an underestimate).[12] The table below summarizes the types of implied absences.

```
knitr::kable(round(profile,1), caption="Student types implied by key MK results under assumption
that treatment had effects on attendance only through the effect on worms.")
```

Student types implied by key MK results under assumption that treatment had effects on attendance only through the effect on worms.

	Control	Treatment
Absent with worms and because of worms	11.2	5.0
Absent with worms but not because of them	10.5	4.7
Present, with worms	34.5	15.3
Absent without worms	8.2	14.1
Present without worms	35.6	61.1
Total	100.0	100.0

So to make sense of the results, and to maintain an account that this treatment worked through the worms, you need to think worms would more than double a kid's chances of missing school in West Kenya and that `37.4` % of all absenteeism would be due to worms. I expect that to the epidemiologists that sounds implausible. Their evidence of the link between worms and poor health suggests it is weak in general and it also seems weak in this study, outside the self-reported measures. Could worms really increase absenteeism by 20 percentage points?

We can do a kind of reality check of these numbers against the data (MK also do this in Model 6 of the original Table 9). Remember for all those kids in Groups 1 and 2 with worms data we also have their past attendance. Did those with worms attend much less than those that did not? Not really. There is a significant relation between worms and attendance, but the estimated effect is just 2-3 percentage points, not 19.9. This is of course a non-experimental estimate, but it is still much more in line with the modest expectations of epidemiologists (though see MK's arguments on p 197-8 of the original article for why this number may be too low and for their IV estimate).[13]

3.4.2 Is a mediation analysis possible?

Besides this sort of orders-of-magnitude check, there are more formal ways to assess mechanisms. In general, identifying mechanisms is a really hard problem and recent work recommends worrying about this first at the design stage, and not just at the analysis stage. The kind of assumption that you need to justify common approaches to estimate mechanisms is that conditional on treatment, having worms or not is essentially random (http://imai.princeton.edu/projects/mechanisms.html) (with a lower probability for treated people, obviously). If in contrast people that are treated that have worms are different in other ways from children that are treated without worms (e.g. they are from poorer families), then any relation identified between worms and attendance may be spurious; for example the correlation may reflect the underlying household poverty, and not the effects of worms at all.

Conscious of the strength of these assumptions there are various ways to study mechanisms. By far the most popular (with a stunning 53,000 citations), and also among the simplest, is the (recently maligned (http://psycnet.apa.org/journals/psp/98/4/550/)) Baron-Kenny approach.[14] The Baron-Kenny (http://psycnet.apa.org/psycinfo/1987-13085-001) approach assumes simple linear structures, similar to those assumed in MK as well as an assumption on the independence of errors.[15] It works like this. First look at the relation between treatment and attendance and note the estimated effect; then do the same thing but this time control for the supposed mediator, worms. If after controlling for the worms the relationship between treatment and attendance disappears then you know that the effect passes through the worm infection; in a sense after accounting for the actual worms there is nothing more to be explained by treatment.

In fact if you do this with this data you see that controlling for worms has hardly *any* effect on the relation between treatment and attendance. Moreover the worms (ICS) effect itself is no longer significant at 5%. It's as if the effect is entirely passing through other mechanisms.[16]

```
knitr::kable(round(Mediation,3), caption="Baron-Kenny style mediation analysis. Three models repo
rted, the interest is in how the coefficient for treatment declines when we control for infection
s (ICS). We see here that the estimate in Column 5 is almost the same as the  estimate in Column
1.")
```

Baron-Kenny style mediation analysis. Three models reported, the interest is in how the coefficient for treatment declines when we control for infections (ICS). We see here that the estimate in Column 5 is almost the same as the estimate in Column 1.

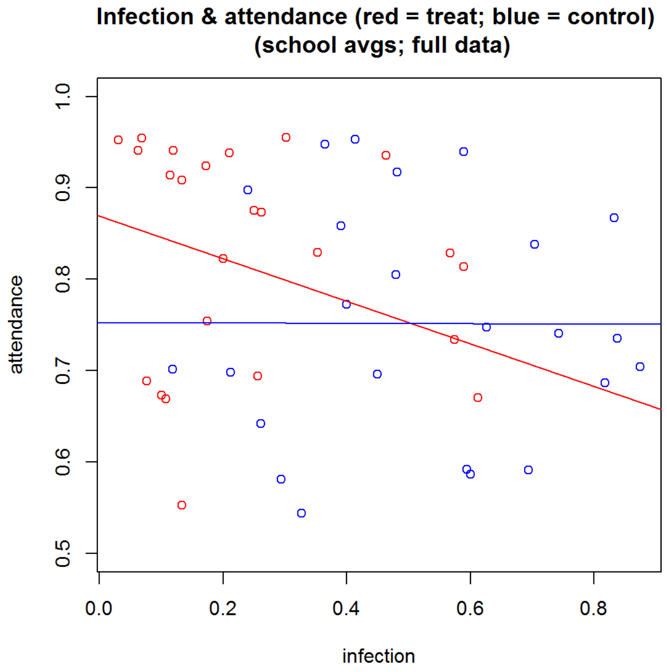
	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)
Year 1 Treatment	0.065	0.000	NA	NA	0.061	0.001
Any ICS 1999	NA	NA	-0.030	0.003	-0.016	0.071
0-3km Externality	0.004	0.891	0.006	0.860	0.003	0.931
3-6km Externality	-0.050	0.024	-0.039	0.077	-0.051	0.020

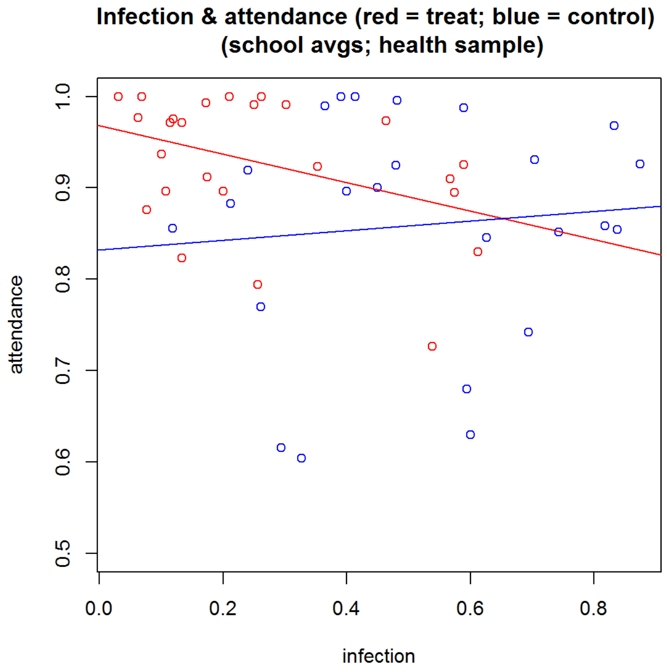
A weakness of this approach is that the attendance data is measured over a period of time while the health data is at a point of time. The analysis implicitly assumes that a person with worms at that point had worms the entire period. In fact they may have been well, attended school, and only recently contracted worms. An alternative approach is to look at the relations at the school level: are schools with the highest level of infections also those with the lowest attendance? Again this is a non-experimental comparison but it more closely reflects the main analyses of the paper. The basic relations are shown below for the unweighted data at the school level. Under a pure mediation hypothesis we might expect to see both treatment and control groups scattered around the same downward sloping line but with treatment groups to the northwest of control groups. Intuitively, treatment will send a point westwards and health benefits will then send it north.

In contrast we see no relation between infection and attendance for the control schools, and a negative relation for the treatment schools driven in large part by the set of very high attendance, low infections schools we noticed before. These are schools that in the epidemiologists' account may have taken their medicine but got better scores in attendance for other reasons. There is no discernible concentration of high infection, low attendance schools in the southeast corner. (Note graphs given here in two versions; the one on the right uses attendance data only for those kids for which there is also worms data; interestingly these are in general more high attendance kids).

```
par(mfrow= c(1,2))
plot(S1$any_ics99.mean, S2$prs.mean, col= ifelse(S1$groups.mean==1, "red","blue"), xlab = "infection", ylab = "attendance", ylim = c(.5,1), main = "Infection & attendance (red = treat; blue = control) \n (school avgs; full data)")
abline(lm(S2$prs.mean[S1$groups==1]~S1$any_ics99.mean[S1$groups==1]), col = "red")
abline(lm(S2$prs.mean[S1$groups==2]~S1$any_ics99.mean[S1$groups==2]), col = "blue")

plot(S1$any_ics99.mean, S1$prs.mean, col= ifelse(S1$groups.mean==1, "red","blue"), xlab = "infection", ylab = "attendance", ylim = c(.5,1), main = "Infection & attendance (red = treat; blue = control) \n (school avgs; health sample)")
abline(lm(prs.mean~any_ics99.mean, data = S1[S1$groups==1, ]), col = "red")
abline(lm(prs.mean~any_ics99.mean, data = S1[S1$groups==2, ]), col = "blue")
```


Infection & attendance (red = treat; blue = control)
(school avgs; full data)


Infection & attendance (red = treat; blue = control)
(school avgs; health sample)

Of course, how serious these concerns around mechanisms and the related concerns around the exclusion restriction are depends in part on what you think the treatment is and what policy you end up advocating for. Is it deworming? or is it a complete package? would delivery be similar to what was achieved under research conditions? and so on. In correspondence MK note that actual interventions come as packages and so the package effect may be of interest. Moreover even in the original article they note mechanisms that do not work uniquely through the worms (notably social mechanisms such as kids missing school to attend to sick siblings). These nuances on what is doing the work have implications for what policy lessons to draw.

3.5 Communication

In most economics papers there are many results but only some of them make it to the abstract and get cited. For this paper the cited number is the 25% (or 7.5 point) reduction in attendance rates. This was a communication decision by the authors. But the weight of that decision stems from the readiness of readers to rely on author's summaries. Was it the right choice? Recall above we saw that there are multiple treatment effects. There is this year-averaged effect including spillovers, which are a strong function of local geography and of partial treatment saturation. There are also other effects implied by the design, partly estimated in Table 9, and shown above. The effect of a one year treatment is strong, and the averaged effect picks part of this up, but it has been used to argue for repeated treatment. Yet the evidence for a second year treatment is much weaker; and indeed groups that had two years of treatment have, if anything, *worse* outcomes than groups with one year of treatment.[17] This might reflect baseline imbalance (see discussion in MK response here (http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000371.pub6/full#CD000371-sec1-0012)) but it seems to me that this would be a thing to highlight in discussions of policy implications.

A quite distinct communication question is around the **communication of errors**. It turns out that a number of these problems were discovered by the original authors by 2007. If this was communicated to the journal it looks like they did not act on it; even now the article page in the journal *Econometrica* does not link to any errata. Others may also have been confused. Even the current JPAL description (http://www.povertyactionlab.org/evaluation/primary-school-deworming-kenya) repeats the claim of spillover effects for people living within 6km of treated schools.[18] (This does not imply anyting nefarious on the part of disseminators of this material: it is natural to refer back to the published work when errata have not been formally posted.) Most surprising perhaps is that it appears that during the heated disputes over this study with the Cochrane team, the authors did not mention these errors; indeed in a published response they refer to the original study and quote the original size effect[19] without indicating that they knew these to be based on mistaken calculations. Perhaps the reasoning of the economist team is that the errors were too minor to warrant a formal erratum or to muddy the discussion, but I imagine if I were an epidemiologist following this now my trust in social science research and transparency standards would feel stretched.

Given the contested nature of the claims of the paper, the policy import, and the considerable communication around it, including many follow up articles and communications with Cochrane, it would have been useful if the knowledge of errors was shared in a more public way.

3.6 Putting it all together

The replication and ensuing discussion has brought up questions about identification, measurement, inference, explanation, and communication. So there are multiple issues. I don't know if this is an unusually large number of issues since few studies are given this kind of scrutiny. Still it's been interesting to see that no economists have voiced particular surprise at all this. The original authors don't seem to give much ground on their original claims either. The conclusions that I take away though are that (a) the magnitude and significance of spillover effects are in doubt because of the measurement issues and the inference issues; (b) the inferences on the main effects are also in doubt because of the problems with identification and explanation. Neither of the main claims is demonstrably incorrect, but there are good grounds to doubt both of them.

4 The case for deworming

A number of commentators have argued that the policy implications are more or less unchanged. This includes organizations that focus specifically on the evidence base for policy (such as CGD and GiveWell).

Perhaps the most important point of confusion is what policy conclusions this discussion *could* affect. Many are defending deworming for non-educational reasons. But the discussion of the MK paper really only matters for the education motivation. And perhaps primarily for the short-term school attendance motivation. Like much other literature in this area it finds only weak evidence for direct health benefits (beyond the strong evidence for the removal of worms). It also does not claim to find evidence on actual performance. Although many groups endorse deworming for health reasons, and rank it as a top priority, this, curiously, goes against the weight of evidence as summarized in the Cochrane reports at least. If the consensus for deworming for health reasons still stands it is not because of this paper.

4.1 The place of MK in the policy discussion

Does the challenge to this paper weaken the case for deworming for educational reasons? I find it hard to see how it cannot. This is the go-to paper to support the argument that deworming affects school attendance. The Copenhagen Consensus for instance, which has given a big boost to deworming (ranking it the fourth most important development priority), draws on a background document that discusses the costs and benefits of deworming. They focus first on the MK study and the 25% reduction number, and argue that using this number “the present value of increased wages associated with increased school achievement gives a benefit:cost ratio of 60:1.” They go on to note “there were external effects on children in non-participating schools” (Hall and Horton, p 18 (http://www.copenhagenconsensus.com/sites/default/files/deworming.pdf)). The 60:1 figure enters into the Executive Summary, along with a discussion of the benefits arising from reducing anemia. The anemia result is one of those that dropped out (note though that MK do not make much of the anemia mechanism in Kenya in the first place).

The World Bank’s Bundy (http://www.londonntd.org/sites/www.londonntd.org/files/content/Worms,%20Wisdom%20and%20Wealth.pdf) and team reviewing evidence, claim that “worm infections are also associated with other significant societal impacts, including school attendance” but MK is the only RCT they cite. GiveWell advocate (http://www.givewell.org/international/technical/programs/deworming) for deworming –two of four top recommended charities are deworming programs — but they also note the evidence base is thin. They “guess that [combination] deworming populations […] does have some small impacts on general health, but do not believe it has a large impact on health in most cases.” For social effects they point to just three studies. One is the MK study, another, also by MK is based on the same experiment as MK, and the third is a historical observational study.

In short, groups arguing for the educational benefits of deworming rely a lot on this study. Indeed one of the study authors reportedly launched Deworm the World (http://www.evidenceaction.org/) on the back of these results (at least as described by the Atlantic Monthly (http://www.theatlantic.com/business/archive/2015/06/what-is-the-greatest-good/395768/)).[20]

There is now increasing reliance on two or three longer term studies (these are as yet mainly unpublished, and include a follow up study to MK (Baird et al (http://www.nber.org/papers/w21428.pdf)); for a discussion of these see Ahuja et al (http://emiguel.econ.berkeley.edu/assets/miguel_research/62/WBER_When_Should_Govts_Subsidize_Health.pdf)), ; even still, if the MK study mattered at all, surely there should be more doubt about these prescriptions now. Duflo and Karlan (http://www.povertyactionlab.org/about-j-pal/news/deworming-informed-debate) are one of the few that have, post-replication, stressed the importance of this study for their own support of deworming, writing "[w]hile we await the results of the re-appraisal of the Kremer and Miguel data, IPA and J-PAL will also continue to support efforts at school based deworming."

4.2 Can we be a bit more Bayesian?

Here's another way to think about the implications of the study and the reanalysis.

Perhaps a sufficient argument for deworming is that it is cheap and worth doing *even if* there is reasonable uncertainty about the benefits. You can think about that argument in two ways: one is that really the policy is driven by principles and not by evidence. Another is that the demands we put on evidence for an intervention should depend on the costs and benefits of the intervention. In other words, we should be a bit more Bayesian about all this.

Here's how a Bayesian might think about this. Say the benefits of an intervention (if it works!) are five times the cost. But say you have doubts about the intervention; in fact you start out thinking that there is only a 10% chance that the intervention works at all.

Then, based on your beliefs about effectiveness, the intervention is a bad bet.

Say now you observe evidence which would arise with only a 5% probability if the intervention was not effective but with an 80% probability if the intervention was effective. What should you infer?

Using Bayes' rule your "posterior probability" that the intervention is effective is given by:

$$\frac{Pr(\text{Data}|\text{Effective})Pr(\text{Effective})}{Pr(\text{Data}|\text{Effective})Pr(\text{Effective})+Pr(\text{Data}|\text{Dud})Pr(\text{Dud})}$$

or

```
(0.80 * .1) / (0.80 * .1 + 0.05 * .9)
```

```
## [1] 0.64
```

So, `64%` . Based on this it clearly makes sense to go ahead. It's a good bet. You can justify your decision using this decision calculus. But you can also draw on classical statistics since, given the data you can reject the null hypotheses of no effect at the 5% level — or informally your finding is "significant" at conventional levels.

Say now that with a reanalysis of the data you find that actually there are new concerns that the patterns in the data might be observed *even if the intervention has no effect.* In particular say that you now think there is a 12% chance you would see data like this even if there were no effect. This means that you can no longer claim that the result is "significant." For many this means that there is now nothing to be learned from the study. For some, this means that the study suggests the intervention is not effective and not worth undertaking.

However from a Bayesian perspective, not all that much has changed. Your posterior is now: `43%` .

So your confidence has gone down a bit but the bottom line stays the same: it’s a good bet. The expected benefit now is a bit over over twice the costs whereas before it was a bit over three times the cost. It is still a great investment.

In practice figuring out the values is harder than this and you have to set up the models differently—in any real model you would have a distribution over effect sizes not simply over whether the intervention works or not. But the key point is that if you use a Bayesian decision calculus then you naturally take account of lots of relevant features that are entirely ignored by the classical statistical tests. First the priors; second the probability of observing the data if in fact the intervention is effective; and third, the actual costs and benefits of the intervention. If you are thinking about it this way then you wouldn’t be too bothered about a *p* value shifting from just below a significance threshold to just above it, after all the difference between significant and not significant is not itself significant (http://www.stat.columbia.edu/~gelman/research/published/signif4.pdf).

Perhaps this is how advocates should be responding to this reanalysis — recognizing that the evidence is weakened and then showing that the bet remains good.

5 The Science

It is not clear that this whole episode has shown decentralized science working well. On the bright side MK providing data and code is obviously good. The broad interest in what the evidence says is also good to see, even if there is not much to suggest that anyone has actually changed their minds very much on either side.

But there are also disappointing aspects. We have learned how slow **researchers** are to listen to others outside their discipline, even experts in the field. And we have seen how slow researchers can be to question leaders in their own discipline. I am not in this field but I still hesitated to write up these comments. It’s no fun to point to weaknesses in each others work in a public debate, even if that is the bread and butter of university seminars. It is especially hard when it involves leaders of the field who have made enormous contributions to how we think about social processes. These are influential researchers that also, rightly, enjoy great respect among their peers. Despite the many weaknesses uncovered, including those recognized as such by the authors, I have not seen economists come out saying that they were surprised by the errors or that this changes their opinions about the paper. The closest thing was the statement by Duflo and Karlan (http://www.povertyactionlab.org/about-j-pal/news/deworming-informed-debate) who called for additional data analysis, focusing I think on the policy, and not on defending the paper.

I expect that much of this is the well established tendency of humans to care most about the opinions of local peer groups. But some of it might suggest deeper disciplinary differences. Along with MK, for example, I simply could not fathom the insistence of authors of the Cochrane reports that the absence of baseline data meant that there was a risk of unknown bias. The claim seems to suggest a very different understanding of bias, or perhaps of the role of randomization in assessing causal effects.

I suspect that the polarization has partly arisen from the difficulties in both disciplines on knowing what inferences to take from uncertain findings. There is a tendency at the level of individual analysis to view results in a very binary way, as either supportive or not supportive of claims. Significance serves as a kind of *filter*. This is a feature of frequentist statistics that my colleague Andy Gelman (http://www.stat.columbia.edu/~gelman/research/unpublished/power4r.pdf) has been complaining about for a long time. But a similar logic seems to be operating at the article level. Among experimentalists, studies are often thought of as identified, and informative, or not-identified, and uninformative. Thus the defense of identification takes the form of an all-or-nothing fight. For deworming, this logic has played out around criteria for inclusion in the Cochrane reviews. There is clearly frustration that additional studies that support MK’s claims have not be judged as meeting these standards. Perhaps the difficulty the original authors seem to have in recognizing that the replication raises important questions about the validity of the finding is because of the binary way we tend to treat study quality.[21] As described above there would be much less at stake around small changes in p values if we used evidence in a more Bayesian way.

Attempts by the **media** to cover the discussion and make sense of it were oftentimes unhelpful. The polarized nature of the discussion brought the focus onto deworming in general when the only new information was about one paper. There was surprisingly little engagement with the paper itself and the replication. This shifted the focus to the question of whether there are benefits to deworming *at all* and not to the policy relevant question which is how large these benefits are and how greatly deworming should be prioritized given extreme pressures on shamefully limited development resources. Nevertheless the media attention did bring *focus* at least: the paper had already been circulated and discussed (notably by Ozler (http://blogs.worldbank.org/impactevaluations/worm-wars-review-reanalysis-miguel-and-kremer-s-deworming-study)) months before but with nothing like the urgency of late July.

I think we have also seen flaws in the feverish digestion of the result through the **blogs**. A lot was written very quickly (the World Bank even has an anthology (http://blogs.worldbank.org/impactevaluations/worm-wars-anthology)). There was clearly a lot of interest, but despite the availability of data and code, the high stakes, and thousands trained in the kinds of methods needed to look at the evidence, almost none of these responses have engaged with either the data or the papers in much depth.

One thing that did partly work, where standard peer review processes failed, was the focus on this evidence brought about by 3ie's somewhat centralized initiative to supplement standard review processes for particularly influential papers. I say only *partly* worked because the innovation has so clearly not been welcomed by so many. Good science gone bad (http://blogs.berkeley.edu/2015/08/03/good-science-gone-wrong/), in the words of one commentator. Perhaps what's needed is a development of the kind of structure being pioneered by 3ie that will support data-informed commentary, updating, and cumulation, and that will enjoy wide legitimacy. There could be a couple of components to this:

1. **Probationary publication**: Prior to publication, journals could put provisionally accepted articles online together with data and code and invite a period of public discussion, informed by data. Errors and inconsistencies could then be identified without embarrassing authors, used by editors to make a final decision on publication (under an assumption of publication unless fundamental errors are found), and provided to authors to allow them to make fixes before the final version is published.

2. **An observatory for post-publication peer review**: For particularly influential pieces (a property revealed often post publication), perhaps what's needed is a *standing panel* of disinterested scholars from different disciplines (at least for cross disciplinary work like this) who independently replicate, probe, and evaluate. The 3ie initiative is a huge step in that direction. But what is also clear is that the research of a single team, even experts in the field, may not be enough to sway researchers or their supporters. Something more like a jury, or an institution with broad legitimacy, independence, and skills may be needed. Such juries would be involved in the qualitative task of weighing up the value of different types of evidence and seeking to assess the policy implications of uncertain evidence; for this I expect a reliance on pre-registered replications only may be unhelpful.[22] Rather any such initiative could seek ways to crowdsource replicable analysis, make data publicly available and invite the broader research community to provide input and code via platforms such as GitHub (https://github.com/) or the Open Science Framework (https://osf.io/) as evidence to help resolve critical points of contention.

3. **Errata and live documents.** Such a process will lead to revisions and these revisions will need to get known. For these we need to handle errata better. At a minimum there should be a strong expectation that errata should be shared with journals and published by journals. Better still, as we move to digital publishing, journals should not only publish the Errata they should archive old version and replace them with live corrected versions (with some form of tracked changes and versioning control for referencing).

4. **Cumulative evidence.** As most agree, in the medium to longer term, confidence in findings will require that we not rely on one or a handful of studies for major policy directions. This is all the more important given that individual studies will have idiosyncratic errors. The aim should be towards more integrated studies of the form that IPA recently completed (http://www.sciencemag.org/content/348/6236/1260799) on a "graduation program" for the ultra poor, and towards reviews similar to those pioneered by the Cochrane Community (http://community.cochrane.org/cochrane-reviews). Although there is broad agreement now on the need for this, it does push against the desire to act quickly on hard earned, if limited, knowledge. Many of the huge deworming rollouts following MK were not conducted in a way that allows for credible impact assessment in large part, perhaps (http://www.newrepublic.com/article/120178/problem-international-development-and-plan-fix-it), because of the confidence placed in this study.

---

[1] See exchange published at the end of the most recent Cochrane (http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000371.pub6/abstract) report.↩

[2] Note this effect is a little larger and significant at the 10% in model 2 of that Table. I prefer a model in which a year term and treatment year interactions are included to take account of a concern on temporal shifts highlighted by the replication team. In that model the second year of treatment effect is slightly larger, closer to 4, though again not significant.↩

[3] I heard a little from both sides in the discussion on this point. The Cochrane authors said that they did not know about these errors until the replication team told them. MK did not claim to have informed the Cochrane team of the errors but did say that later analyses sent to Cochrane used the corrected data.↩

[4] There is a slightly deeper way of thinking about what randomization does: under certain conditions randomization lets you draw random samples from different universes of "potential outcomes"; equipped with these samples, estimates of average *potential outcomes* can be assessed and so estimates of average treatment effects follow from the fact that differences in averages are equal to averages of differences.↩

[5] Interestingly the strategy does not assume there to be balance between treatment and control to justify unbiasedness.↩

[6] I understood it as the sum of the direct effects plus the spillover effects. The related quantity for worms is described as the sum of the direct effect on treated schools plus "average spillover gain" or the "average cross-school externality."↩

[7] On p 196 of the original article there is another calculation that the gains from treating one child is 0.14 school years – not the 0.075 we had at the beginning. For discussion: I cannot figure this out at all since the 0.075 effect is already the ITT on trying to treat one child, conditional on success rates, and the 0.02 number already includes information on the number of neighboring schools. For discussion: What is this calculation?↩

[8] Model 3 of Table VII.↩

[9] Things are a little more complicated for the schistosomiasis infections. On the tables using the corrected data for the schistosomiasis infections there is some evidence for spillovers on the untreated, but only in the 3-6km range; this estimate is now dropped in MK's ``preferred'' tables (p51 (http://www.3ieimpact.org/media/filer_public/2014/10/20/worms-3ie-pure-response_2014-10-10-final_3ie2.pdf).↩

10. In this simulation the mean squared error is taken with respect to the predicted outcome rather than a particular estimate of effects.↩

11. Note that the model below is what MK refer to as the “preferred” or “final” specification (eg Table IX, col 3, p 52 of their response (http://www.3ieimpact.org/media/filer_public/2014/10/20/worms-3ie-pure-response_2014-10-10-final_3ie2.pdf)). Things shift favorably if instead the “any year” treatment var is used, though doing that might suggest the need for further correction.↩

12. This is akin to the logic of the LATE estimator (col 7 of Table 9 in the original article.) Say a is the probability of going to school if you are infected, and b is the probability of going to school if you are not. Then under the homogeneity conditions described above, the numbers we have tell us:

$$\text{share w worms in control} \times a + \text{share without} \times b = \text{share absent in control}$$

$$\text{share w worms in treatment} \times a + \text{share without} \times b = \text{share absent in treatment}$$

This is of course two equations with two unknowns and it has a unique solution.↩

13. Note that for that analysis the presence of worms is not itself randomized and so that difference in attendance rates cannot be considered a causal estimate. In another observational estimate Nokes and Bundy (http://trstmh.oxfordjournals.org.ezproxy.cul.columbia.edu/content/87/2/148.short) find some of the most cited observational evidence for a decline in attendance associated with worm infections (see their Figure 6), estimate an approximate 50% reduction in attendance.↩

14. For an older warning about this approach see Judd and Kenny 1981 (http://davidakenny.net/doc/JuddKenny1981.pdf).↩

15. The key additional ssumption is that the errors on the model on infections are uncorrelated with the errors on a model of the outcome model that includes both the mediator and the treatmeent. See Bullock, Green, and Ha (http://neuron4.psych.ubc.ca/~schaller/528Readings/BullockGreenHa2010.pdf).) When this assumption fails to hold the results may be biased; in particular if the same factors that effect health (ansence of infections) also affect attendance then this strategy will *over*estimate the extent of mediation of the intervention via worms.↩

16. Note that this analysis is restricted to the subjects for whom there is worm data in 1999. For this group interestingly there is no evidence of a local spillover but there is evidence of a more distal spillover.↩

17. One reason consistent with MK’s account for the decline is that stricter consent processes weakened the first stage. A reason consistent with research demand effects is that there could have been less contact with teams in this period, as suggested by temporal patterns in the “obs” weighting variable, which I understand, indicates that in year one up to six visits were made (mean 3.7) and in year 2 up to four visits were made (mean 2.6).↩

18. On the JPAL site it still states “It is estimated that moderate-to-heavy helminth infections among children living up to 6 kilometers away from treatment schools were 23 percentage points lower on average.” Even still it is hard to see some of the implicit errata in recent documents by MK. The most recent “Replication Guide” for example notes the errors in calculation around the key 7.5 figure. But the introduction, rather than stating that data correction nearly cuts their estimate in half, claims that a new reasonable model actually produces bigger numbers than before. The older 2007 guide describes differences such as the loss of significance on treatment in period 2 and the loss in significance for spillovers in the 3-6 km range as slight changes, and that the new tables show “little substantive differences” and “almost identical” results.↩

19. They quote the quarter reduction number but write 7% rather than the usual 7.5, referring to the original study. See paragraph prior to “Costs and Benefits” section, here (http://journals.plos.org/plosntds/article?id=info:doi/10.1371/journal.pntd.0000362)). Note though that in fact I *should* have known about some of these issues long ago because I requested and received the data in 2010. The package I received then included a file with a description of many of these errors.↩

20. June 2015, citing McAskill (http://www.effectivealtruism.com/).↩

21. The Cochrane report does engage in a form of weighting of studies, describing evidence as “low quality”, “very low quality” and so on.↩

22. I am a strong supporter of registration and see the need for registration to avoid a type of “negative fishing” in which researchers can deliberately or inadvertently focus on unrepresentative negative findings in a generally positive study. Registration may help for that. And it could also help make sure that there is a record of replications that do not get published because they find nothing new of note. But registration faces particular challenges also. First unlike with the registration of pre-analysis plans, registration of reanalyses takes place after data collection, often when data is available prior to registration, and almost always after researchers have seen the core results and patterns. Moreover replication has an investigative element that does not lend itself well to registration: the point is that you do not necessarily know what unusual features the replication data will present to you in advance. Finding unusual weights, odd errors structures, unusual patterns of missingness, could all be clues that give insights to why the original work claimed what it did. But they might also be hard to see in advance. The middle ground might be registration with an expectation of transparent deviations from registered plans or perhaps the use of “standard operating procedures (http://www.stat.berkeley.edu/~winston/sop-safety-net.pdf)” for handling unexpected patterns.↩