The GiveWell Blog

GiveWell
REAL CHANGE FOR YOUR DOLLAR

New deworming reanalyses and Cochrane review

JULY 24, 2015

By Alexander

g+1

On Wednesday, the *International Journal of Epidemiology* published two new reanalyses of Miguel and Kremer 2004, the most well-known randomized trial of deworming. Deworming is an intervention conducted by two of our top charities, so we've read the reanalyses and the simultaneously updated Cochrane review closely and are responding publicly. We still have a few remaining questions about the reanalyses, and have not had a chance to update much of the content on the rest of our website regarding these issues, but **our current view is that these new papers do not change our overall assessment of the evidence on deworming, and we continue to recommend the Schistosomiasis Control Initiative and the Deworm the World Initiative.**

Key points:

• We're very much in support of replicating and stress-testing important studies like this one. We did our own reanalysis of the study in question in 2012, and the replication released recently is more thorough and identifies errors that we did not.

• We don't think the two replications bear on the most important parts of the case we see for deworming. Both focus on Miguel and Kremer 2004, which examines impacts of deworming on school attendance; in our view, the more important case for deworming comes from a later study that found impacts on earnings many years later. The school attendance finding provides a possible mechanism through which deworming might have improved later-in-life earnings; this is important, because (as stated below) the mechanism is a serious question.

• However, the replications do not directly challenge the existence of an attendance effect either. One primarily challenges the finding of externalities (effects of treatment on untreated students, possibly via reducing e.g. contaminated soil and water) at a particular distance. The other challenges both the *statistical significance* and the size of the main effect for attendance but we believe is best read as finding significant evidence for a smaller attendance effect. Regardless, the results we see as most important, particularly on income later in life, are not affected.

• The updated Cochrane review seems broadly consistent with the earlier version, which we wrote about in 2012. We agree with its finding that there is little sign of short-term impacts of deworming on health indicators (e.g., weight and anemia) or test scores, and, as we have previously noted, we believe that this does undermine – but does not eliminate – the plausibility of the effect on earnings.

• In our view, the best reasons to be skeptical about the evidence for deworming pertain to external validity, particularly related to the occurrence of El Nino during the period of study, which we have written about elsewhere. These issues are not addressed in the recent releases.

• At the same time, because mass deworming is so cheap, **there is a good case for donating to support deworming even when in substantial doubt about the evidence**. This has consistently been our position since we first recommend the Schistosomiasis Control Initiative in 2011. Our current cost-effectiveness model (which balances the doubts we have about the evidence with the cost of implementing the program) is here.

• While we think that replicating and challenging studies is a good thing, it looks in this case like there was an aggressive media push –

publication of two papers at once coinciding with an update of the Cochrane review and a Buzzfeed piece, all on the same day – that we think has contributed to people exaggerating the significance of the findings.

Details follow. We also recommend the comments on this issue by Chris Blattman (whose post has an interesting comment thread) and Berk Ozler.

<div align="center">**The reanalyses of Miguel and Kremer 2004**</div>

Aiken et al. 2015 and Davey et al. 2015 participated in a replication program hosted by the International Initiative for Impact Evaluation (3ie), in which Miguel and Kremer shared the data from their trials and Aiken, Davey and colleagues reanalysed them. Working paper versions of these reanalyses were published on the 3ie website dated October 2014, and Joan Hamory Hicks, Miguel and Kremer responded to both of them there. The World Bank's Berk Ozler wrote a blog post in January reviewing the reanalyses and Hicks, Miguel, and Kremer's replies.

Aiken et al. 2015 straightforwardly attempts to replicate Miguel and Kremer 2004's results from data and code shared by the authors. They do a much more thorough job than when we attempted something similar in 2012, and find a number of errors.

Amongst a number of smaller issues, Aiken et al. find a coding error in Miguel and Kremer's estimate of the externality impacts of deworming on students in nearby schools, in which Miguel and Kremer only counted the population of the nearest 12 schools. That coding error substantially changes estimates of the impact of deworming on both the prevalence of worm infections in nearby schools and the attendance of students in nearby schools, particularly estimates of the impact of further out schools, between 3 and 6 km away.

Aiken et al. state: "Having corrected these errors, re-analysis found no statistically significant indirect-between-school effect on the worm infection out- come, according to the analysis methods originally used. However, among variables used to construct this effect, a parameter describing the effect of Group 1 living within 0–3 km did remain significant, albeit at a slightly smaller size (original -0.26, SE 0.09, significant at 95% confidence level; updated -0.21, SE 0.10, significant at 95% confidence). The corresponding parameter for the 3–6- km distances became much smaller and statistically insignificant (original -0.14, SE 0.06, significant at 90% confidence; updated -0.05, SE 0.08, not statistically significant)." Aiken et al.'s supplementary material and Hicks, Miguel, and Kremer's response to the 3ie replication working paper clarifies this explanation. In short, fixing the coding error does not much affect estimates of the externality within 3 km of treatment schools, but does significantly change estimated externalities between 3 and 6 km out, and following the original Miguel and Kremer 2004 process for synthesizing those estimates into an overall estimate of the cross-school externality on worm prevalence, the resulting figure is not statistically significant. However, if you simply drop the 3-6 km externality estimate, which is now negative and no longer statistically significant, then you continue to see a statistically significant cross-school externality (see the second to last row of Table 1).

The same coding error also affects estimates of the externality effect on school attendance, in a broadly similar way. Aiken et al. write: "Correction of all coding errors in Table IX thus led to the major discrepancies shown in Table 3. The indirect-between-school effect [on attendance] was substantially reduced (from +2.0% to -1.7%) with an increased standard error (from 1.3% to 3.0%) making the result non-significant. The total effect on school attendance was also substantially reduced (from 7.5% to 3.9% absolute improvement), making it only slightly more than one standard error interval away [from] zero, hence also non-significant." The correction to the coding error significantly increases the standard error of the 3-6km externality estimate, which then increases the standard error of the overall estimate significantly. The increased uncertainty, rather than the change in the point estimate of the externality, is what drives the conclusion that the total effect on school attendance is no longer statistically significant. As in the prevalence externality case, dropping the 3-6km estimate altogether preserves a statistically significant cross-school externality (and total effect).

We are uncertain about what to believe about the externality terms at this point. It seems fairly clear that had Miguel and Kremer caught the coding error prior to publication, their paper would have ignored potential externalities beyond 3km, and the replication done today would have found that the analysis up to 3km was broadly right. The replication penalizes the paper for having initially (incorrectly) found externalities further out. While we continue to be worried about the possibility of specification searching in the externality terms, and we

see a case for treating the initial paper as a form of preregistration, we don't see it as at all obvious that we should penalize the Miguel and Kremer results in the way that Aiken et al. suggest.

The Aiken et al. replication, like the original paper, finds no evidence of an impact on test scores.

Davey et al. 2015 is a more interpretive reanalysis, in which the authors use a more "epidemiological" analytical approach to reanalyze the data. The abstract states:

Results: Quasi-randomization resulted in three similar groups of 25 schools. There was a substantial amount of missing data. In year-stratified cluster-summary analysis, there was no clear evidence for improvement in either school attendance or examination performance. In year-stratified regression models, there was some evidence of improvement in school attendance [adjusted odds ratios (aOR): year 1: 1.48, 95% confidence interval (CI) 0.88–2.52, P = 0.150; year 2: 1.23, 95% CI 1.01–1.51, P = 0.044], but not examination performance (adjusted differences: year 1: −0.135, 95% CI −0.323–0.054, P = 0.161; year 2: −0.017, 95% CI −0.201–0.166, P = 0.854). When both years were combined, there was strong evidence of an effect on attendance (aOR 1.82, 95% CI 1.74–1.91, P < 0.001), but not examination performance (adjusted difference −0.121, 95% CI −0.293–0.052, P = 0.169).
Conclusions: The evidence supporting an improvement in school attendance differed by analysis method. This, and various other important limitations of the data, caution against over-interpretation of the results. We find that the study provides some evidence, but with high risk of bias, that a school-based drug-treatment and health-education intervention improved school attendance and no evidence of effect on examination performance.

Reviewing the key conclusions in order:

- "In year-stratified cluster-summary analysis, there was no clear evidence for improvement in either school attendance or examination performance." The results of the year-stratified cluster-summary analysis are substantively the same as the results of the year-stratified regression models that Davey et al. use (next bullet), with wider confidence intervals resulting from the reduction in sample size of caused by using unweighted school-level data (N=75). Table 2 reports a 5.5 percentage point impact on attendance in 1998 (corresponding to an odds ratio of 1.78) and a 2.2 percentage point impact for 1999 (corresponding to an odds ratio of 1.21). Davey et al.'s regressions find an odds ratio for 1998 of 1.77 (unadjusted, p=0.097) or 1.48 (adjusted, p=0.150) and for 1999 of 1.23 (unadjusted, p=0.047, or adjusted, p=0.044), i.e. the same point estimates with tighter confidence intervals. We don't see it as surprising or problematic that collapsing a large cluster-randomized trials' data to the cluster level results in a loss of statistical significance.

- "In year-stratified regression models, there was some evidence of improvement in school attendance [adjusted odds ratios (aOR): year 1: 1.48, 95% confidence interval (CI) 0.88–2.52, P = 0.150; year 2: 1.23, 95% CI 1.01–1.51, P = 0.044], but not examination performance (adjusted differences: year 1: −0.135, 95% CI −0.323–0.054, P = 0.161; year 2: −0.017, 95% CI −0.201–0.166, P = 0.854)." The lack of a result on exam performance echoes Miguel and Kremer 2004's results. The "some evidence of improvement" result for school attendance is more striking, since the year 2 results are positive and statistically significant while the year 1 results are *more positive but not statistically significant* (due to a wider confidence interval). We read this as the test in year 1 being underpowered; treating years 1 and 2 as two independent randomized control trials, a fixed-effects meta-analysis would find a statistically significant overall effect.

- "When both years were combined, there was strong evidence of an effect on attendance (aOR 1.82, 95% CI 1.74–1.91, P < 0.001), but not examination performance (adjusted difference −0.121, 95% CI −0.293–0.052, P = 0.169)." These results accord with the Miguel and Kremer 2004 results.

- "We find that the study provides some evidence, but with high risk of bias, that a school-based drug-treatment and health-education intervention improved school attendance and no evidence of effect on examination performance." The authors make two main arguments for the high risk of bias. First, they note (in Figure 3) that the correlation across schools between attendance rates and the number of attendance observations appears to differ across the treatment and control groups, with a broad tendency towards positive correlation between observations and attendance rates in the intervention group and a negative correlation in the control group, which would lead to estimates weighted by the number of observations to overestimate the true impact. However, we see three reasons not to regard this evidence as particularly problematic:

  - Hicks, Miguel, and Kremer report conducting a test for the claimed change in the correlation and finding a non-statistically significant result (page 9). As far as we know, Davey et al. have not responded to this point, though we think it is possible that Hicks, Miguel, and Kremer's test is underpowered.

  - As noted above, the unweighted (year-stratified cluster-summary) estimates are not lower than the year-stratified regression models (which Davey et al. report do weight by observation–"we used random-effects regression on school attendance observations, an approach which gives greater weight to clusters with higher numbers of observations"), they just have wider confidence intervals. **In order for the observed correlation to be biasing the weighted results, the weighted estimates would need to be meaningfully different from the unweighted ones, which is not the case here**. Accordingly, we see little reason even in Davey et al.'s framework for preferring the less precise year-stratified cluster-summary results to the year-stratified regressions, which use significantly more information to reach virtually the same point estimates.

  - Hicks, Miguel, and Kremer report results weighted by pupil instead of observation (Table 3), and find results strongly consistent with their attendance-weighted results, without the risk of being biased by attendance observations. However, their results imply treatment effects that are larger than the odds ratios reported in Davey et al.'s year-stratified regression models, which Davey et al. report do weight by observation. We're not sure what to make of this discrepancy, and we haven't see Davey et al. respond on this point.

Second, and relatedly, Davey et al. note that the estimated attendance effect in the combined years analysis is larger than in either of the underlying years, and they suggest that the change is due to the inclusion of a before-after comparison for Group 2 (which switched from control in year one to treatment in year two) in the purportedly experimental analysis. We see this concern as more plausible, and don't have a conclusive view on it at this point, but we think it would affect the magnitude of the observed effect rather than its existence (since we read the year-stratified regressions, which are not subject to this potential bias, as supporting an impact on attendance).

To summarize, we see no reason even based on Davey et al.'s own choices to prefer the year-stratified cluster-summary, which discards a significant amount of information, to the year-stratified regression models, which together point to a statistically significant impact on attendance. Hicks, Miguel, and Kremer make a variety of other arguments against decisions made by Davey et al., and they, along with Blattman and Ozler, argue that many of the changes are jointly necessary to yield non-significant results. We haven't considered this claim fully because we see the Davey et al. results as supporting a statistically significant attendance impact, but if we turn out to be wrong about that, it would be important to more fully weigh the other deviations they make from Miguel and Kremer's approach in reaching a conclusion.

School attendance data has never played a major role in our view about deworming (more on our views below), but we see little reason based on these re-analyses to doubt the Miguel and Kremer 2004 result that deworming significantly improved attendance in their experiment. We see much more reason to be worried about external validity, particularly related to the occurrence of El Nino during the period of study, which we have written about elsewhere.

The new Cochrane Review

The new Cochrane review on deworming reaches largely the same conclusions as the 2012 update, which we have discussed previously.

The new review incorporates the Aiken et al. and Davey et al. replications of Miguel and Kremer 2004 and the results of the large DEVTA trial, but continues to exclude Baird et al. 2011, Croke 2014, and Ozier 2011.

We agree with the general bottom line that there is little evidence for any biological mechanism linking deworming to longer term outcomes, and that that should significantly reduce one's confidence in any claimed long-term effects of deworming. However, the Cochrane authors make some editorial judgments we don't agree with.

They state:

- "The replication highlights important coding errors and this resulted in a number of changes to the results: the previously reported effect on anaemia disappeared; the effect on school attendance was similar to the original analysis, although the effect was seen in both children that received the drug and those that did not; and the indirect effects (externalities) of the intervention on adjacent schools disappeared (Aiken 2015)." As described above, in summarizing the results of Aiken et al. 2015, we would have noted that estimated cross-school externalities remain statistically significant in the 0-3km range.

- "The statistical replication suggested some impact of the complex intervention (deworming and health promotion) on school attendance, but this varied depending on the analysis strategy, and there was a high risk of bias. The replication showed no effect on exam performance (Davey 2015)." We think it is misleading to summarize the results as "[impact on school attendance] varied depending on the analysis strategy, and there was a high risk of bias." Our read is that Davey et al. reported some analyses in which they discarded a significant amount of information and accordingly lost statistical significance, but found attendance impacts that were consistently positive and of the same magnitude (and statistically significant in analyses that preserved information).

- "There have been some recent trials on long-term follow-up, none of which met the quality criteria needed in order to be included in this review (Baird 2011; Croke 2014; Ozier 2011; described in Characteristics of excluded studies). Baird 2011 and Ozier 2011 are follow-up trials of the Miguel 2004 (Cluster) trial. Ozier 2011 studied children in the vicinity of the Miguel 2004 (Cluster) to assess long-term impacts of the externalities (impacts on untreated children). However, in the replication trials (Aiken 2014; Aiken 2015; Davey 2015), these spill-over effects were no longer present, raising questions about the validity of a long-term follow-up." This last sentence seems problematic from multiple perspectives:
  - Davey et al. 2015 does not mention or look for externalities or spill-over effects.
  - Aiken et al. 2015 replicates Miguel and Kremer 2004's finding of a statistically significant externality within 0-3 km, so summarizing it as "these spill-over effects were no longer present" seems to be an over-simplification.
  - The lack of geographic externality is a particularly unpersuasive explanation for excluding Ozier 2011, which focuses on spill-over effects to younger siblings of children who were assigned to deworming, especially given that Aiken et al. confirm Miguel and Kremer's finding of within-school externalities (which seems more similar to the siblings case). More generally, the fact that one study failed to find a result seems like a bad reason to exclude a follow-up study to it that did.

More generally, we agree with many of the conclusions of the Cochrane review, but excluding some of the most important studies on a topic because they eventually treated the control group seems misguided. Doing so structurally excludes virtually all long-term follow-ups, since they are often ethically required to eventually treat their control groups.

<div align="center">

Our case for deworming

</div>

As we wrote in 2012, the last time the Cochrane review on deworming was updated, our review of deworming focuses on three kinds of benefits:

- **General health impacts, especially on haemoglobin**. We currently conclude, partly based on the last edition of the Cochrane review: "Evidence for the impact of deworming on short-term general health is thin, especially for soil-transmitted helminth (STH)-only deworming. Most of the potential effects are relatively small, the evidence is mixed, and different approaches have varied effects. We would guess that deworming populations with schistosomiasis and STH (combination deworming) does have some small impacts on general health, but do not believe it has a large impact on health in most cases. We are uncertain that STH-only deworming affects general health." This last claim continues to be in line with Cochrane's updated finding of no impact of STH-only deworming on haemoglobin and most other short-term outcomes.

- **Prevention of potentially severe effects, such as intestinal obstruction**. These effects are rare and play a relatively small role in our position on deworming.

- **Developmental impacts, particularly on income later in life**. The new Cochrane review continues to exclude the studies we see as key to this question. Bleakley 2004 is outside of the scope of the Cochrane review because it is not an experimental analysis, and Baird et al. 2011 is excluded because its control group eventually received treatment. However, as before, the Cochrane review does discuss Miguel and Kremer 2004, which underlies the Baird et al. 2011 follow-up; in their assessment of the risk of bias in included studies, Miguel and Kremer 2004 continues to be the worst-graded of the included trials.
  We also do not think that the Aiken et al. or Davey et al. papers should substantially affect our assessment of the Baird et al. 2011 results. Aiken et al.'s main finding is about the coding error affecting the 3-6km externality terms. I'm not clear on whether the coding error in the construction of the externality variable extends to Baird et al. 2011, but, regardless, the results we see as most important, particularly on income, do not rely on the externality term. Davey et al.'s key argument is against the combined analysis in which Group 2 is considered control in year one and treatment in year two. I remain uncertain about whether this worry is fundamentally correct, but Baird et al. is not subject to it because their estimates treat Group 2 as consistently part of the treatment group.

Nonetheless, we continue to have serious reservations about these studies and would counsel against taking them at face value.

We think it's a particular mistake to analyze the evidence in this case without respect to the cost of the intervention. Table 4 of Baird et al. 2012 estimates that, not counting externalities, their results imply that deworming generates a net present value of $55.26, against an average cost of $1.07, i.e. that deworming is ~50 times more effective than cash transfers. We do not think it is appropriate to take estimates like these at face value or to expect them to generalize without adjustment, but the strong results leave significant room for cost-effectiveness to regress to the mean and still beat cash. In our cost-effectiveness model, we apply a number of ad-hoc adjustments to penalize for external validity and replicability concerns, and most of us continue to guess that deworming is more cost-effective than cash transfers, though of course these are judgment calls and we could easily be wrong.

The lack of a clear causal mechanism to connect deworming to longer term developmental outcomes is a significant and legitimate source of uncertainty as to whether deworming truly has any effect, and we do not think it would be inappropriate for more risk-averse donors to prefer to support other interventions instead, but we don't agree with the Cochrane review's conclusion that it's the long-term evidence that is obviously mistaken in this case. (We have noted elsewhere that most claims for long-term impact seem to be subject to broadly similar problems.)

The importance of data sharing and replication

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

Posted in Uncategorized   permalink [http://blog.givewell.org/2015/07/24/new-deworming-reanalyses-and-cochrane-review/]