Published on *Impact Evaluations* (http://blogs.worldbank.org/impactevaluations)

Worm Wars: A Review of the Reanalysis of Miguel and Kremer's Deworming Study

Submitted by Berk Ozler On Fri, 07/24/2015

This post was updated on July 24, 2015 in reponse to the increased traffic to the site from Twitter upon the publication of the replication discussed below [1]and the authors' response [2]in the International Journal of Epidemiology. I took a day to re-review the papers in question and not surprisingly, what I said below remains as valid as 6 months ago because all that happened is that the papers got published without much change since they first appeared on the 3ie website. However, I do have a few new thoughts (and one new table from Hamory Hicks, Kremer, and Miguel), which I discuss below. My original post remains unedited. I'd also like to thank Stefane Helleringer for a nice response he wrote about the definion of ITT in public health: see the back and forth here [3].

Despite the differences in various methodological and data handling choices, which I discussed below in my original post, it is clear that the interpretation of whether one believes the results of Miguel and Kremer are robust really rests on whether one splits the data or not. Therefore it is important to solely focus on this point and think about which choice is more justified and whether the issue can be dealt with another way. A good starting point is the explanation of DAHH in their pre-analysis plan as to why they decided to split the data into years and analyze it cross-sectionally rather than the difference-in-difference method in the original MK (2004):

> The data from a stepped wedge trial can be thought of as a one-way cross-over, and treated as such, by comparing before and after in the cross-over schools (group 2) and accounting for the secular trend using the non-crossing schools (groups 1 and 3). However, such an approach requires assumptions about the uniformity of the trend and the ability of the model to capture the secular change, and as such loses the advantage of randomization.

This explanation seems confused to me: the common trend assumptions are something that need to be established in observational studies when we're using Diff-in-Diff as an identification strategy, but in a cluster-randomized trial like this one, we have it by design: Groups 2 & 3 are perfect counterfactuals for each other if the randomization has been done correctlty. Surely, if we look at a bunch of outcomes, we might find random differences in the changes from 1998-1999 between the two groups, but that's no reason to assume that there is something wrong with this approach or that it takes us away from the advantage of randomization. Analyzing everything cross-sectionally by year and not controlling for the lagged value of the outcome variable is costing DAHH some statistical power instead.

However, let's accept for a second DAHH's argument that there's something strange with Group 2 and we're wary of it. Them it seems to me that the solution is simple: why not look at the two clean groups that never change treatment status the whole study period of 1998-1999. In other words, exclude Group 2, pool all the data for 1998 and 1999 and compare the effects between Group 1 and Group 3. Sure, we lose power from throwing out a whole study arm, but if the results stand we're done! Thankfully, Joan Hamory Hicks was able to run this analysis and send me the table below, which is akin to their Table 3 in their original response [4]:

Updated cluster summary results (Groups 1 and 3 only)

	Weight by by Pupil Population		Weight by Num. of Attendance Obs.	
	Difference	P-value	Difference	P-value
Panel A: Treatment indicator and year defined as in Aiken et al. (2015)				
1998+1999[1]	5.80**	[0.016]	5.75**	[0.016]
1998+1999[2]	5.80**	[0.050]	5.75*	[0.052]
Panel B: Treatment indicator and year defined as in Miguel and Kremer (2004)				
1998+1999[1]	5.94**	[0.011]	5.87**	[0.014]
1998+1999[2]	5.94**	[0.036]	5.87**	[0.049]

Note: This analysis is based on the top left panel of Aiken et al. (2015), Table 2. All analysis includes only individuals in Group 1 and Group 3 schools, and eligible, non-transferring pupils. Panel B utilizes the same data as Panel A, but redefines the treatment indicator and year as described in the text. P-values are in square brackets and stars reflect: "***" P-value < 0.01, "**" P-value < 0.05, "*" P-value < 0.10. [1] Includes a year 2 indicator. [2] Includes a year 2 indicator and clusters the standard errors by school.

As you can see, all effect sizes on school participation are about 6 percentage points (pp), which is remarkably close to the effect size of 7 pp in the original study. The p-values went up from <0.01 to <0.05, but that is fully expected having shed a third of the sample. So, even if you think that there is something strange going on with Group 2, for which the visual inspection presented by DAHH in Figure 3 is really not sufficient, you still have similarly-sized and statistically significant effects when making the cleaner comparison of Groups 1 & 3. Problem solved?

I want to conclude by making a bigger picture point about replications. They are really a really expanded version of robustness checks that are conducted for almost any paper. It's just that the incentives are different: authors want robustness and replicators might be tempted to find a hole or two to poke in the evidence and "debunk" the paper (if I had a dime yesterday for every deworming debunked tweet...). But, when that happens, I start worrying about multiple hypothesis testing. We now know and have tools for how to deal with multiple inference corrections, when the worry is Type I errors (false rejections of a correct null). But, what about Type 2 errors? After all this is exactly what a replicator would be after: finding a manner of handling the data/analysis that makes the results go away. But, how do we know whether that is a true "failure to reject" or a Type 2 error? Even in studies with 80% power, there is a 20%chance that each independent test will fail to reject under the null of a positive effect. The more of these you try, the more likely you'll come across one or two estimates that are insignificant. What to do about that?

To be fair to the authors, they were at least aware of this issue, mentioned on page 7 of the PAP:

> We aim to deal with this problem by making a small number of analyses using as much of the original data as possible at each stage and concentrating initially on the direct intervention effects on the major study outcomes.

But, then this is where it would have been really important to have a very clear PAP, describing only a very few, carefully methodologically justified, analyses proposed and sticking very strictly to it. But, every step of the way when the authors decide to weight or not weight the data (cluster summaries), splitting the data by year, adjusted/unadjusted estimates, alternative treatment definitions dropping large numbers of observations, etc. there is a fork and the fork opens up more roads to Type 2 errors. We need replications of studies that are decently powered themselves, where the replicators are careful to hoard all the power that is there and not scatter it along the way.

I hope that this update has brought some clarity to the key issues that are surrounding the debate about the publication of the replication results and the accompanying flurry of articles. I was an unwitting and unwilling participant of the Twitter storm that ensued, only because many of you were responsible for repeatedly pointing out the fact that I had written the blog post below six months ago and linking to it incessantly throughout the day. I remain indebted to our readers who are a smart and thoughtful bunch...

Berk.

This post follows directly from the previous one, which is my response to Brown and Wood [5]'s (B&W) response to "How Scientific Are Scientific Replications? [6]" It will likely be easier for you to digest what follows if you have at least read B&W's post and my response to it. The title of this post refers to this tweet [7] by @brettkeller [8], the responses to which kindly demanded [9] that I follow through with my promise of reviewing this replication when it got published online.

Background: Miguel and Kremer (2004) [10] evaluated a school-based mass deworming treatment on worm loads, school participation, and academic test scores. It showed improvements in the former two outcomes, while finding no effects on achievement. Its importance goes

beyond the finding of direct effects, but extends to displaying the indirect effects of mass deworming on untreated individuals – even those that resided in different clusters (schools) that were sufficiently close to treatment schools. Davey et al. (2014b) [11], DAHH from hereon, reanalyzes the original data making different choices about important aspects of the handling of the data, definition of treatment, and econometric methods and concludes: "We found that the evidence that the intervention improved school attendance differed according to how we analysed the data," continuing "These data provide weak evidence that a school-based drug-treatment and health-education intervention improved school attendance…" Hicks, Kremer, and Miguel (2014b) [4], HKM from hereon, state in response: "We strongly disagree with this conclusion."

[Note to Reader: It is undoubtedly true that the readers will get much more out of reviewing the articles themselves instead of (or in addition to) reading what follows. However, given that many will choose not to, I suggest that you at least examine Table 4 in DAHH and Table 1 in HKM and the surrounding text in each document. The key point of departure from the original analysis and its knock-on effects on other decisions made by DAHH can be seen in the former while a thorough summary of HKM's response in the latter. I provide my views on these key points of departure/disagreement below.]

Key points of disagreement between DAHH and HKM: In their reanalysis of the data from the original study, DAHH make some choices that are significantly different than the ones made by the original study authors. There are many departures but four of them are key: (i) definition of treatment; (ii) ignoring the longitudinal data in favor of cross-sectional analysis of treatment effects by year; (iii) weighting observations differently; and (iv) ignoring spillovers from treatment to control. I address them in order below:

1. Defining the treatment variable: In this step-wedge cluster randomized design of Miguel and Kremer (2004), Group 1 schools started receiving treatment in March 1998 immediately after baseline measurements at all schools, while Groups 2 & 3 awaited treatment. Group 2 joined Group 1 in being treated between March and June 1999, after follow-up data including worm loads and multiple instances of random, unannounced attendance checks were conducted to gauge the one-year effects of the interventions. Miguel and Kremer (2004) call these periods of March 1998-February 1999 and March-December 1999, between which the treatment status of Group 2 switches, 1998 and 1999, respectively. But, it is clear what they mean when they refer to these treatment periods and there is no disagreement between the two sides on this issue.

The disagreement occurs when DAHH decide to redefine the treatment periods to match the calendar years, i.e. January to December 1998 and January to December 1999 (DAHH, page 3). The effect of this is to consider a substantial number of control observations in early 1999 (20% according to HKM, page 1) as treated. Of course, since they were not actually treated, the result is a dampening of the treatment effect on attendance, which is now a weighted average among treatment and control individuals.

DAHH justify this important "redefinition" of the main treatment variable by invoking their preference for intention-to-treat analysis, comparing "…outcomes between clusters (for example, schools) randomly allocated to different treatment conditions irrespective of whether treatment was, in practice, actually implemented or adhered to." They further state "We inferred from the original paper, in the absence of a protocol, that the combined educational and drug-treatment intervention package was intended to be delivered from the start of each year" (page 3).

Yet, this is a very unusual, almost bizarre, interpretation of intention-to-treat. The approach, which is also the method used by Miguel and Kremer (2004), refers to circumstances where outcomes are analyzed according to the treatment group/period assigned to each individual or cluster rather than their actual treatment status. In other words, if I offered Group 1, say, a small amount of cash while Group 2 was offered a large amount of cash, the analysis would be independent of how much money they actually received during the intervention period (due to non-compliance, implementation glitches, contamination between groups, etc.). But, there is no analogy to these circumstances here. Instead, what DAHH seem to be claiming is that the original plans were to treat children at the very beginning of each year and that the treatments in both years slid forward by a couple of months. Even if this were true, and HKM strongly deny that this was ever the case (see more below), it does not justify their redefinition of the treatment groups/periods. Suppose that you wrote a funding proposal and a pre-analysis plan to start a school-based RCT in Liberia in September 2014 before the Ebola crisis shut down all schools indefinitely. Sometime in the future, you were able to resume your research and started treatment in, say, January 2016. Would you include the period September 2014 to December 2015 as a treatment period? The choice made by DAHH seems to suggest that their answer would be "yes." There is no doubt value in informing the readers that this delay took place, but I don't know any development economists who would define this pre-treatment period as treatment.

Furthermore, even that interpretation becomes invalid if you believe HKM's assertion that there was never any intent to treat schools earlier. Their justification makes sense: the study design required collection of data at the schools during the school year – to measure worm loads at baseline and to measure effects on worm loads and attendance right before Group 2 switched treatment status. How could Miguel and Kremer (2004) have intended to treat schools in early 1998 before baseline data were collected or switch Group 2 to treatment before they measured worm loads again to assess intervention effects? Researchers who adopt these "delayed treatment" designs are generally most concerned about making sure that treatment is not offered to the control group in waiting until follow-up data collection has been completed so that the treatment effect estimates are not contaminated (even then we generally have to worry about anticipation effects, adoption of the intervention by the individuals themselves, and the like – all of which would generally lead to conservative estimates of treatment effects.).

One last point here before I move on to the next point: some children in Group 2 were in fact receiving benefits from the program prior to March 1999 and not only between January and March 1999, but for the entire treatment period for Group 1. This is because of the spillover effects of mass deworming on untreated children in the same schools and schools in Group 2 that were sufficiently close to Group 1 schools. As I said above, such effects, which are taken into account in Miguel and Kremer (2004) and DAHH (2014a) [12], but ignored in DAHH (2014b), would imply that the true treatment effects are underestimated because of the violation of the assumption of no interference between schools, aka SUTVA.

2. Unpooling the longitudinal data: The step-wedge study design of Miguel and Kremer (2004) has three groups, only one of which switches treatment status after March 1998: Group 2. Group 1 is treated in both years, but while Group 3 remains untreated in both years. Having groups with unchanging treatment statuses provide time trends, both improving the precision of the estimates and potentially correcting for random differences between the groups at baseline. Miguel and Kremer took full advantage of the panel data, estimating effects by pooling all of the data from 1998 and 1999. They consider this issue so key that this is the only thing they do not vary for their

robustness check in Table 1, where all 32 estimates are statistically significant at the 99% level of statistical confidence -- all using the pooled data.

What seems to have happened, in an apparent departure from their pre-analysis plan, is that DAHH decided to look at cluster level summaries of outcomes in the three groups using unweighted school means and separately for 1998 and 1999 (while using their redefinition of treatment). Splitting the sample into 1998 and 1999, reduces power and also leads to slightly different treatment effect estimates (by no longer adjusting for secular trends) and these differences seem to have concerned/puzzled the reanalysis authors: why is the panel data analysis showing slightly larger effects that are much more precise than the cross-sectional analysis? Rather than considering the obvious power implications of their approach and perhaps trying ANCOVA rather than a difference-in-difference approach, they seem to have gone looking for something that must be biasing the estimates. What they found is summarized in DAHH Figure 3 (and responded to by HKM in section 2.2 and Table 2): I am not going to go into the details here, but their argument is akin to differential changes in measurement of the school attendance data between Group 1 and Group 2 being the explanation for the large treatment effects that are found in the original study rather than a real treatment effect stemming from mass deworming.

Looking at Figure 3 in DAHH, there does seem to be something there – although it is a bit of a stretch. HKM correctly counter that this visual inspection is something that can actually be tested quantitatively: i.e. we can tell whether the correlation between the number of attendance observations and school participation rate changes over time differentially between the treatment and the control group. HKM run this test, present the results in Table 2, and show that the key estimates of interest (the last rows of columns 2 and 3) are not statistically significant. They claim that DAHH should not have embarked on what comes next on the basis of this evidence.

I am not convinced by HKM's argument here. If a study is underpowered to detect treatment effects, robustness tests, tests of baseline differences, etc. will similarly suffer from low power: in other words, in such studies there can be large imbalances in baseline characteristics or their interactions with treatment indicators in relation to an outcome, which don't show up as significant in statistical tests. For example, the triple interaction between treatment (but not treatment and G2), number of attendance observations, and the indicator for 1999 is 0.122 (not small) and has a p-value of 0.14 (not very large). I also would have conducted these tests but, personally, I would not have dismissed this issue completely on the basis of the findings.

However, that issue becomes moot when we consider the next point, which concerns the weights used in the analysis:

3. Weighting of observations in the statistical analysis: The other change that DAHH make to the analysis in Table 4 (Step 1) is to use unweighted school means, whereas the original study by Miguel and Kremer (2004) calculated these means using school population of pupils as weights. DAHH explain that they are doing this because these weights are differentially correlated with attendance between the treatment groups over time – on the basis of what I discussed in point #2 above. Their splitting of the sample by year does also seem related to this point: if we analyze each year separately, we break the differential correlation. But, both of these choices are bad ones: having each school be weighted equally treats a school with 10 students the same as one with 10,000 students, while splitting the data by year severely reduces statistical power. The weighting by population issue has analogues in, say, the inequality debates between countries vs. individuals – i.e. whether we treat China and Vatican as one observation each or whether we give much more weight to China's outcomes – but I don't see that analogy here: if you're trying to estimate deworming impacts in this study population, then the thing to do is to give equal weight to each student in the study sample. This fixes the worry about bias raised by DAHH in point #2 (by eliminating the weighting of the data by the number of attendance observations) and retains the power envisioned in the original study design. In Table 3, HKM show that the statistical significance of the results is gone only when outcomes from each school are weighted equally: furthermore, even this null finding is not robust when treatment is defined as in Miguel and Kremer (2004) as opposed to the redefinition in DAHH (HKM, Table 3, Panel B, final two rows).

4. Ignoring Spillover Effects between Schools: DAHH chose to ignore the issue of interference between schools, stemming from the fact that living near children who have been dewormed benefits children who have not. DAHH explain their decision not to address this issue in pages x-xi of their reanalysis study – I cannot say that I understood it fully: there were mentions of "theory of change," "scientific replication," "pure replication," etc. This decision sticks out for two reasons: first, a quick search for the word externality produces 178 hits in the 59-page original study – a rate of three mentions per page. Second, ignoring the violation of the "no between-cluster interference assumption" produces biased estimates of treatment (see, for example, "Designing Experiments to Measure Spillover Effects [13]," by Baird et al. 2014). HKM object similarly, stating that all of the estimators used in DAHH are "…downward biased."

Bottom line: Based on what I have seen in the reanalysis study by DAHH and the response by HKM, my view of the original study is more or less unchanged. In fact, if anything, I find the findings of the original study more robust than I did before. Tables 1 & 3 in HKM's response demonstrate that a number of unconventional ways of handling the data and conducting the analysis are jointly required to obtain results that are qualitatively different than the original study. Leaving aside whether these decisions made by DAHH are warranted or not, the simple fact that there are a large number of such tests, which produce differences in only a small number of cases, combined with the fact that the reanalysis authors made no attempt to consider the implications of these multiple comparisons (DAHH, page 11), gives little reason to the audience to update their views of the evidence provided by Miguel and Kremer (2004).

Of course, if your issue with the original study was that it caused too much policy change on the basis of a single paper from one region in one country, this is a different matter and a legitimate subject for debate. Same can be said for the motivation to reanalyze the data and present the results found in this influential paper to the public health audience using a language and methods that are more familiar to it. However, none of this, in my view, justifies the conclusions that are drawn by DAHH in the abstract or Table 8 in the concluding section that "the results are sensitive to analytic choices." People who are worried about the strength of the evidence on deworming interventions would be better served by examining the evidence from more recent studies, summarized in Section 3 of HKM.

[Full disclosure: I have no conflicts of interest as they directly relate to Miguel and Kremer (2004), Davey et al. (2014a, 2014b), or Hicks, Kremer, and Miguel (2014a, 2014b). However, Sarah Baird, who is a co-author of Hicks, Kremer, and Miguel on Baird et al. (2014) that "… followed up the Kenya deworming beneficiaries from the Miguel and Kremer (2004) study during 2007-2009 and find large improvements in their labor market outcomes," is my long-time partner as well as my co-author on many papers. She was neither associated with the original study, nor with the response by HKM. I have deliberately avoided talking to her about the replication and my review of it, with the exception of checking with her that she agreed with the wording of this disclosure.]

- Tags:

- school attendance [14]
- Kenya [15]
- deworming [16]
- worms [17]
- replication [18]
- 3ie [19]

Links:
[1] http://ije.oxfordjournals.org/content/early/2015/07/21/ije.dyv128.full.pdf
[2] http://ije.oxfordjournals.org/content/early/2015/07/21/ije.dyv129.full.pdf
[3] http://chrisblattman.com/2015/07/23/dear-journalists-and-policymakers-what-you-need-to-know-about-the-worm-wars/#comment-192756
[4] http://www.3ieimpact.org/media/filer_public/2015/01/07/rps3_worms-3ie-pure-response_2014-12-22-part_2.pdf
[5] http://blogs.worldbank.org/impactevaluations/response-brown-and-woods-how-scientific-are-scientific-replications-response
[6] http://blogs.worldbank.org/impactevaluations/how-scientific-are-scientific-replications
[7] https://twitter.com/search?q=worm wars %40brettkeller&src=typd
[8] https://twitter.com/brettkeller
[9] http://blogs.worldbank.org/impactevaluations/files/impactevaluations/slide1.jpg
[10] http://cega.berkeley.edu/assets/cega_research_projects/1/Identifying-Impacts-on-Education-and-Health-in-the-Presence-of-Treatment-Externalities.pdf
[11] http://www.3ieimpact.org/media/filer_public/2015/01/07/rps_3_part_2_top_copy_reduced_size_1_7_15-top.pdf
[12] http://www.3ieimpact.org/media/filer_public/2015/01/07/3ie_rps3_worms_replication_1.pdf
[13] https://openknowledge.worldbank.org/bitstream/handle/10986/17738/WPS6824.pdf
[14] http://blogs.worldbank.org/impactevaluations/category/tags/school-attendance
[15] http://blogs.worldbank.org/impactevaluations/taxonomy/term/10923
[16] http://blogs.worldbank.org/impactevaluations/taxonomy/term/13018
[17] http://blogs.worldbank.org/impactevaluations/taxonomy/term/13005
[18] http://blogs.worldbank.org/impactevaluations/category/tags/replication
[19] http://blogs.worldbank.org/impactevaluations/category/tags/3ie