Comment on Macartan Humphreys' and Other Recent Discussions of the Miguel and Kremer (2004) Study

Edward Miguel, Michael Kremer, Joan Hamory Hicks

December 21, 2015

The replication of the Miguel and Kremer 2004 *Econometrica* paper, "Worms: Identifying Impacts on Education and Health In the Presence of Treatment Externalities", carried out by Aiken et al. and Davey et al. and our responses to the replication, which can be found here and here, triggered discussions by scholars and media taking an interest in the evidence base for mass deworming policies in regions where intestinal worm infection are endemic.

As we discussed in our *International Journal of Epidemiology* response, the two main statistical findings in the original Miguel and Kremer 2004 paper are (i) positive deworming impacts on school participation, and (ii) positive deworming treatment externalities (or spillovers).

Multiple scholars and organizations have examined the evidence from the Miguel and Kremer 2004 study regarding these two claims, in light of the replications. Most of the recent analyses largely confirm the original conclusions regarding both school participation impacts and deworming externalities, including the discussions by Alex Berger (GiveWell), Chris Blattman (Columbia University), Paul Gertler (University of California, Berkeley), Michael Clemens and Justin Sandefur (Center for Global Development), and Berk Ozler (World Bank). Ozler's post contains particularly detailed discussion of the estimation of school participation impacts, and Clemens and Sandefur carry out a detailed analysis of the deworming externality findings. David Choi (Carnegie Mellon University) applies an alternative statistical methodology to the data, and confirms the existence of the deworming spillover effects. Dr. Aaron Carroll (Healthcare Triage) concludes that the main findings of the original Miguel and Kremer 2004 paper appear robust, and argues that much of the sensationalistic early media coverage of the replication articles was misleading. The statistical analysis of Macartan Humphreys (Columbia University) confirms the existence of deworming impacts on school participation (although he is skeptical of their large magnitude), but he raises concerns about the robustness of the deworming treatment externality results.

Here we focus on the main points raised in Humphreys' post, while also bringing in discussion of the analyses carried out by the other recent commentators. We note two areas in which there is broad agreement between Humphreys' perspective and ours, including on the statistical evidence in favor of positive deworming impacts on school participation. We then discuss the main areas of disagreement, including regarding the strength of evidence for positive deworming treatment externalities.

1. Overview of Areas of Broad Agreement

(a) Adopting a Bayesian Perspective

We are in broad agreement with Humphreys' perspective in his section 4.2 that a Bayesian approach to decision making is warranted.

This approach to incorporating new pieces of research evidence into a policymaker's decision-making problem is highly appropriate in this context, and rightly moves the discussion away from binary interpretations based on whether a P-value is less than 0.05 in one single specification in a particular study. We had earlier called for such a Bayesian approach in our recently published review of deworming research in the *World Bank Economic Review*.

A conclusion of Humphreys' Bayesian analysis is that any small shifts in statistical significance between the original and updated results in the Miguel and Kremer 2004 study (hereafter MK04) are unlikely to have much effect on the attractiveness of deworming as a public policy. The extreme cost-effectiveness of mass deworming through schools (at cents per child each year) means that the expected returns to deworming could be quite high even if there remained some uncertainty around the precise magnitude of its benefits, making it a very good investment.

This is a sensible way to think about the issue, and we think that other scholars could usefully extend this approach, making it more detailed and bringing in results from additional studies, including the growing number of new studies demonstrating large long-run deworming impacts on educational and socio-economic outcomes in the historical U.S. South (here), and in contemporary Uganda and Kenya (in two studies, here and here).

(b) Confirming the Robustness of the Main School Participation Estimates

A second important element of Humphreys' analysis is the affirmation of the robustness of the school participation impact estimates.

Recall that a central finding of MK04 is that deworming leads to large gains in school participation. Humphreys concludes (in section 2.2.4) that "The coefficient on the direct effects [on school participation] is robust to many of the alternative estimation strategies that Davey et al examined", and that the various assumptions Davey et al. make "don't make all that much of a difference".

We discussed the issues around the robustness of the main school participation results and multiple hypothesis testing in some productive email exchanges that preceded the posting of Humphreys' piece. The question of whether the significance levels we report in our updated results are overly optimistic given the possibility of multiple hypothesis testing is important, and we shared results showing that the statistical significance is robust to several methods of multiple testing adjustment, as Humphreys reports in section 3.3.3. Humphreys here concludes that this "confirms the robustness of the direct effect [on school participation]".

As mentioned above, these conclusions of Humphreys' analysis confirming the positive impact of deworming on school participation in the MK04 data are in line with the conclusions reached by other recent analysts who have examined the findings (including Berk Ozler, Chris Blattman, Alexander Berger, and Aaron Carroll). This is a central intellectual issue in the debate, and obviously critical for public policy choices. Elsewhere we lay out why this conclusion on school participation impacts diverges from some of the claims made in the Aiken et al. (2015) and Davey et al. (2015) replication articles, including a detailed scientific discussion of the multiple non-standard analytical choices made in those articles that drive their results.

We graphically portray the robustness of the estimated effect of school participation to a wide range of analytical specification choices (including regression functional form, inclusion of covariates, sample inclusion criteria, and weighting of observations) in Figure 2 of our *IJE* response piece. We reproduce a version of that figure here (below) since it succinctly portrays the robustness of the results. Each vertical gray line denotes a distinct estimate of the effect of deworming on school participation excluding any cross-school externalities. There are results from 32 distinct regression specifications presented in the figure, all of which are statistically significant at 99% confidence.

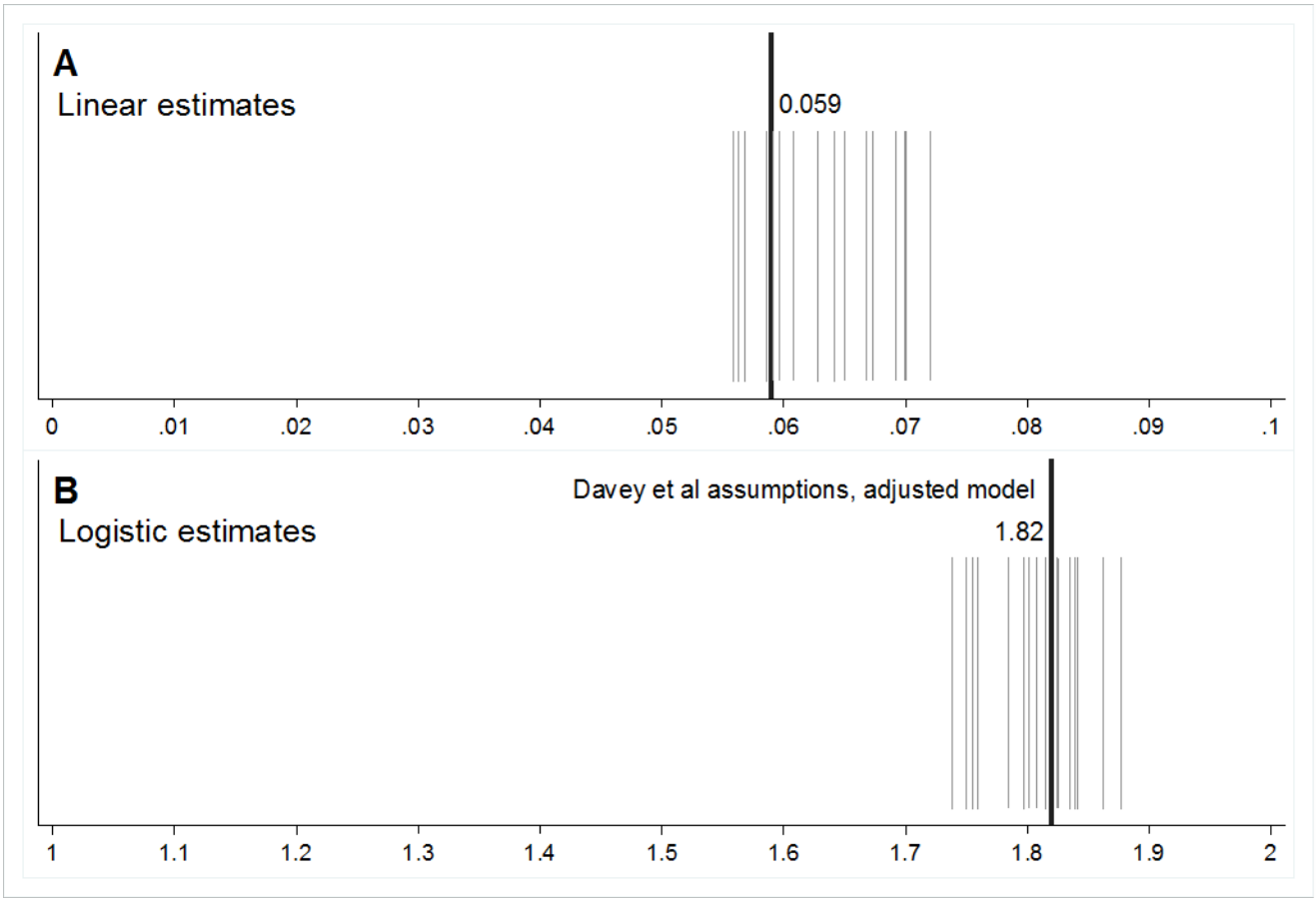


Figure 2: Deworming treatment effect estimates on school participation. Each vertical gray line denotes a coefficient estimate of the effect of deworming on school participation, excluding any cross-school externalities (and thus estimates are lower bounds on true effects). The estimates use both years of data, and differ in: (i) statistical model (the original linear regression model in Panel A, and random effects logistics regression from Davey et al. (2015) in Panel B); (ii) sample (the original full sample, and the sample eligible for treatment in Davey et al.); (iii) regression models adjusted for covariates and unadjusted; (iv) approaches to weighting observations (each attendance observation equally, and each pupil equally); and (v) the dataset that in Davey et al. employ in their analysis, which incorrectly defines treatment and makes additional missing data assumptions (Appendix B), versus data that correctly defines treatment. All 16 coefficient estimates in Panel A are significant at P<0.01; all 16 estimates in Panel B are significant at P<0.001. The bold vertical lines denote the adjusted model estimate using Davey et al.'s (2015) data; the Panel B estimate is from their Table 2, top right panel. This figure is reproduced here from Hicks, Kremer, and Miguel (2015) in the *International Journal of Epidemiology*.

2. Three Main Areas of Disagreement

One area where Humphreys differs from us is in his interpretation of the school participation results. Echoing the replication authors, Humphreys expresses concerns that, while they exist statistically, the school participation effects may be "too large" to be plausible, or that they might be driven by placebo effects or Hawthorne effects (since the original intervention was not blinded to participants). A second main difference lies in Humphreys' interpretation of the deworming treatment externality findings in MK04, which in turn affects his analysis of the magnitude of overall deworming impacts on school participation. We also disagree with elements of Humphreys' discussion of the estimation of externalities in his first worms-related blog post (on 1 August 2015).

(a) Interpreting the school participation impacts

Humphreys raises multiple concerns regarding the interpretation of the school participation impacts of deworming in Kenya. Many are closely related to points made by the replication authors (and we have responded to these already here, here and here). Here is a concise overview of the key issues he raises, and our take on them.

First, Humphreys is concerned that there remains some uncertainty about the precise health channels linking deworming to school participation gains in our data.

This is not to say that there is no evidence of health effects: there are huge reductions in moderate-heavy worm infections in the data, significant improvements in self-reports of whether the student was sick in the last week, and suggestive evidence of small height gains. Yet the reduction in anemia (from 4% to 2%) is not statistically significant (in the updated results) and there is no significant effect on child weight.

As we noted in the original MK04 paper, anemia is rare to begin with in Western Kenya and therefore not likely to be the key driver of school-participation results. The literature suggests other possible channels that might link worm infections to schooling outcomes beyond child weight and anemia. For example, Stephenson et al., 1993, have argued that treating worm infections can improve child appetite and physical fitness, or strengthen the immunological response to other infections, such as malaria (Kirwan et al., 2010). Chronic parasitic infections may generate inflammatory (immune defense)

responses and elevated cortisol levels, and this could lead to impaired intestinal transport of nutrients and other adverse health consequences (Crimmins and Finch, 2005).

The MK04 study focused on educational outcomes, and did not collect data on this full range of health channels, and hence does estimate effects on all of these possible biological channels that could affect school participation. However, the self-reported health measure that was collected suggests a direct mechanism for the school attendance impacts: kids who feel sick stay home from school, and kids who do not feel sick are more likely to show up.

Humphreys' claim that deworming generates minimal health benefits comes in part from the recent Cochrane review of deworming, but there is considerable controversy around that review. The Cochrane review on deworming makes a number of questionable analytical decisions and has important limitations; we discuss some of these below.

Two other recent reviews, Hall et al. (2008) and Smith and Brooker (2010), discuss deworming impacts on health and reach very sharply different conclusions than the Cochrane review. Many leading health bodies that have also carefully surveyed the full body of evidence, including the WHO, conclude that mass drug administration has important public health benefits and is highly cost-effective. So Humphreys' conclusion that it is implausible that deworming has meaningful health and nutritional impacts is far from a consensus view in the health research community.

Second, Humphreys raises the possibility that placebo effects might be behind the school participation gains in treatment schools, as the intervention was not blinded.

One of the core findings of the original MK04 paper, which is confirmed in the Aiken et al. 2015 replication, is the strength of within-school externalities (i.e., gains among untreated pupils in the treatment schools). A deworming treatment placebo effect cannot explain why untreated pupils in a treatment school experienced sharply reduced worm infections. Note that worm infection is an objectively measured outcome, not a self-report (and not even an outcome like school participation, which might be influenced by participant beliefs). Nor can a placebo effect explain why the younger children in Ozier's (2014) analysis, who did not themselves take the deworming drugs, experienced cognitive gains. The same goes for the externality findings of reduced worm infections and higher school participation that we document among schools located within 3 km of treated schools. A placebo effect does not seem a plausible explanation for any of these patterns in the data.

Closely related to the placebo argument is the concern in Humphreys' section 3.1.3 that children might have shown up more in treatment schools because there were "so many services and benefits" provided. However, there were no other services provided to treatment schools in this intervention. Deworming treatment days were pre-announced, and children received a small snack (a biscuit or piece of banana) when administered the deworming pills, so there would be no surprise if more kids showed up on those particular days. But those days are not included in the school participation calculation in MK04. And on other days of the year, there were no additional benefits for the treatment schools. During the attendance checks, there were no services provided. The attendance checks happened in both treatment and control schools, were unannounced and scattered on 4 to 6 days throughout the school year, so the odds of a visit on any given day was small. Note also that school participation effects were found in nearby schools (along with lower worm load).

Given all of the above, it seems much more plausible that the school participation effects are due to the large observed reductions in worm load, rather than to hypothetical placebo effects.

Third, Humphreys argues that it is possible that it is not deworming pills per se that caused the school participation impacts, but rather a combined package of the pills plus the health education messages and lessons that accompanied them.

This is certainly a possibility, as we discuss in MK04, and since real-world deworming programs typically feature health education, understanding the impact of drugs plus messaging is highly relevant for public policy. That said, in the paper we tested and found no evidence for changes on several dimensions of worm prevention behaviors. This makes it far more likely that the main driver of the impact is the deworming drug treatment itself.

Fourth, in sections 3.4.1 and 3.4.2 Humphreys conducts original analysis attempting to pin down the role played by reduced worm infections per se in driving the overall school participation effect, and focuses on mediation analysis.

One of the key pieces of evidence Humphreys offers against the view that worm infections are the key mechanism here is the mediation analysis in section 3.4.2. Here he applies a method that has been sharply criticized by many statisticians and econometricians. Specifically, he regresses school

participation on the treatment indicator plus the post-treatment measure of worm infections (for the subset of individuals with parasitological data), and shows that the estimated treatment effect does not change substantially, interpreting this to mean that worm infections cannot be an important channel.

Leading econometricians have long argued that such an approach, which uses an outcome variable (here, worm infection) as a covariate, will generally lead to biased estimates (Angrist and Krueger, 1999) and should be avoided in most cases. Angrist and Pischke term this the method of "bad control" (Angrist and Pischke, 2009). This approach could potentially provide insights if certain relatively strong assumptions hold, but Humphreys does not provide any evidence that the identifying assumptions are satisfied here. Given that the estimates he provides are likely to be biased, this mediation analysis is highly speculative at best and we do not put much stock in these results. (Humphreys notes that this mediation approach is also controversial in psychology.)

There are good analytical reasons to expect very large differences between experimental estimates and non-experimental cross-sectional estimates (like those Humphreys produces) in this setting. The cross-sectional OLS estimate using baseline data will be badly biased to the extent that there are effects on the enrollment margin, since people are only in the dataset at baseline in January 1998 if they have not dropped out earlier. To see this, consider the extreme thought experiment in which worms act strongly on the dropout margin, but do not strongly affect attendance conditional on not dropping out. At baseline, one will observe only the students who had not dropped out earlier. In this case, there will be little difference in the observed school participation of kids with and without worms, even if in fact there is a very large treatment effect. This is why experimental data like ours that tracks both school dropouts and attendance over time is essential.

Fifth, in analysis closely related to some results in MK04, Humphreys attempts to back out how large the effect of each eliminated infection would need to be to explain the school participation treatment effect (in section 3.4.1). He concludes that the roughly 19% increase in attendance per moderate-heavy infection eliminated is "too large".

One key challenge here analytically is computing the number of infections eliminated over the study period, since infections in the treatment and control groups are only captured at one snapshot in time during the study period (namely, in early 1999). Humphreys' calculation likely understates the gap in infections between treatment and control, due to the timing of the parasitological data collection.

Infection rates were measured roughly 5-6 months after deworming, sufficient time for substantial reinfection to occur, and shortly before the next deworming round took place. This means that the treatment group likely had lower infection rates both before the measurement of infection rates, and again shortly afterwards (after another round of deworming), so the difference between treatment and control is underestimated. This implies a far smaller school participation gain per infection eliminated is needed to account for the observed program effect.

Moreover, the 19% figure is similar to the number estimated in Hoyt Bleakley's historical paper for the U.S. South, implying that the finding is in line with other evidence, rather than "too large". Humphreys' main reason for considering this figure to be too large appears to be the estimated cross-sectional relationship between worm infections and school participation. But such relationships are likely biased for myriad reasons, confounding factors and omitted variables, including those discussed above, and this is exactly the reason why experimental estimates are preferred. Humphreys has argued in favor of experimental methods and research designs in much of his own political science research; the same methodological arguments apply in this setting.

Are there are other factors to consider to help explain the large school participation impacts of deworming of mass deworming? There might be. We discussed a possible factor in our original 2004 paper on p. 197:

> [T]he exclusion restriction—that the program only affects pupils' school attendance by changing their health—may not hold, due to complementarities in school participation. For example, if the pre-schoolers, first-graders, and second-graders for whom we estimate the largest school participation effects stay home sick with worms in the comparison schools, their older sisters may also stay home to take care of them, and this may partly explain the relatively large treatment effects we find for older girls. More generally, there may be complementarity in school attendance if children are more inclined to go to school if their classmates are also in school, so school participation gains in treatment schools may partially reflect increased school participation among children who were not infected with worms. Such effects would influence the impact of a large-scale deworming program on school participation and are captured in a prospective evaluation (like ours) in which treatment is randomized at the school level, but they would not be picked up in an individual-level regression of school participation on worm levels, or in a prospective study in which treatment is randomized at the individual level.

In other words, from a policy perspective, these sorts of complementarities or social effects are highly relevant, and are caused by a mass deworming treatment program, but they work through a channel other than just the number of worm infections eliminated. This sort of social effect falls outside the standard channels considered by health researchers accustomed to carrying out individual level RCTs, but it is highly relevant for social science and important for policy. (Bobonis and Finan (2009) found a

related pattern in PROGRESA cash transfer program data from Mexico, in which the school participation of program non-beneficiaries in the treatment villages increased significantly, and they have interpreted this similarly in terms of complementarities and changing social norms.) This is a strength of the cluster randomized design in Miguel and Kremer (2004).

(b) Robustness of deworming treatment externalities

Humphreys questions the robustness of the deworming treatment externality estimates in MK04. His discussion is incomplete, ignoring many of the pieces of evidence that we present in the original paper, and we feel that his conclusions are misleading and based on a misunderstanding of our original procedure.

It is worth first recapping how we tested for the existence of treatment externalities in the original MK04 article. Within each of a set of specific populations, we first tested for epidemiological externalities (focusing on the "any moderate to heavy infection" variable). We then tested for externalities on school participation. We first tested for externalities within treatment schools (among those who did not receive treatment). We then tested for externalities among schools located within 0-3 km of treatment schools, and then among schools 3-6 km away, distances at which we judged transmission to be possible based on the underlying epidemiology and on the mobility of school children in the area.

Later, as part of an exercise that could be used in cost effectiveness analysis, we summed up the direct and the externality impacts of deworming over the distances within which we felt we could reliably measure such effects. We emphasize that this weighted sum of effects at various distances was *not* our main test for externalities. Among other things, as we noted in the original MK04 paper, focusing on any pre-specified distance misses externalities beyond that distance.

However, since school children are less likely to spend time at greater distances from their home, one would expect epidemiological externalities to drop off gradually with distance, so that at a sufficient distance from a treated school, it would likely be impossible to detect externality effects. In the original paper, we noted that health externalities for soil transmitted helminths are likely to be fairly localized, while schistosomiasis externalities could plausibly take place over longer distances, since children travel to the lake to swim or fish, and the snails within which schistosomiasis lives part of its life cycle can themselves move.

We did not specify a particular distance at which to look in a pre-analysis plan, since such plans were not established in the social sciences at the time (and in fact would not be widely used until over a decade later), and since there was no clear a priori reason to focus on a particular distance. Instead, we first examined externality relationships within schools, then out to 3 kilometers, and finally out to 6 kilometers, looking separately at worm infection levels and school participation, for a total of six different and independent externality tests. The appropriate way to assess the existence and extent of externalities is to examine each of these externality measures individually, and then to consider the resulting body of evidence holistically.

The empirical approach to detecting within-school externalities in MK04 is to compare those students in treatment schools who did not receive deworming treatment when offered, i.e. "non-compliers", to those students in control schools who we know did not receive treatment when they were offered it in later years. This approach deals with the issue of selection into treatment, since comparable "non-compliers" are compared across the treatment and control schools.

We show in the original paper that there are large reductions in worm infection rates, presumably due to less reinfection, among untreated students in the treatment schools. These reductions are smaller than the gains experienced by children who took the deworming drugs themselves, but still over half the gains experienced by those who took the drugs. Providing further assurance, these untreated children also show large and statistically significant school participation gains. All of these results have been confirmed by the replication team. This provides evidence that there are large positive deworming treatment externalities within treatment communities.

Perhaps because there has been no controversy around these within-school externality estimates, they do not feature much in the discussion in Humphreys' post. However, the (unchallenged) existence of externalities within schools is sufficient to drive the key methodological and economic conclusions of the Miguel and Kremer 2004 paper.

Methodologically, the existence of large spillovers means that existing studies that randomize treatment across individuals within the same community underestimate the impact of mass treatment programs. This is not an idle concern: the majority of studies in the recent Cochrane review on deworming are individual-level RCT's with designs of exactly this kind, leading some observers to conclude that many of these studies are grossly understating the true health, nutritional and educational impacts of deworming (for instance, see here). Cluster randomized designs, like the one used in the MK04 study,

help to overcome this central issue. Moreover, from a public finance point of view, positive externalities provide a rationale for public subsidies.

After we had found evidence of within school externalities in MK04, we next turned to estimating if there were externalities across schools, and to do so we developed an empirical approach that relies on the study's experimental design. Since we developed this approach in our 2004 *Econometrica* paper, this method to estimating externalities has since been widely used in other empirical studies across fields, and is generally seen as one of the main intellectual contributions of the paper.

When estimating the spillovers from treated schools to other nearby schools, the differences between schools in high-density treatment areas and low-density treatment areas is captured by controlling for the total number of pupils located nearby. It is important to condition on the total number of pupils located nearby to make sure that the treatment density measure is not simply picking up differences between areas that have higher concentrations of population (i.e., market centers versus more remote rural areas). Conditional on this total local pupil population variable, the numbers of treatment pupils located nearby is determined by the experimental design, and thus should not be systematically correlated with other school characteristics. See the discussion in section 4.1 of the original paper for more on this approach.

MK04 found epidemiological spillovers on any moderate to heavy infection outcomes at 0-3 km, and found school participation externalities in that range as well. At the time, we believed that there was also evidence for epidemiological externalities on the infection outcome at 3-6 km, but we did not find evidence for school participation externalities at this distance, with a point estimate that was negative and close to zero. The absence of school participation externalities in a population with epidemiological externalities could be considered a slight anomaly in the original set of results.

The main change in the results from correcting the programming error (discussed in our replication materials), is that the infection externality estimates using our main "any moderate to heavy infection" variable in the 3-6 km range are no longer statistically significant.[1]

[1] It is worth noting that there is some evidence of significant effects on one type of infection, schistosomiasis. Schistosomiasis externalities at 3-6 km are significant at the 1% level, consisted with differences in the underlying transmission mechanism between schistosomiasis and soil-transmitted helminths. However, schistosomiasis levels were only high enough to justify mass treatment near Lake Victoria, with only approximately one-third of the sample receiving treatment.

Why this change? The original coding error led us to include only the 12 closest schools in the calculation of school densities; correcting it does not affect the 0-3 km calculations, since there were never more than 12 schools in that range, but many more schools are now included in the calculation of the 3-6 km density terms. With the revised results, the anomaly of finding positive externalities on the “any moderate to heavy infection” variable, but not on school participation, at 3-6 km, disappears, as there is no evidence for externality effects along either dimension at this distance.

Nonetheless, the estimated spillover effects on "any moderate to heavy worm infection” and on the school participation outcomes remain significant within schools, and in the 0-3 km range, as in the original paper, confirming the existence of deworming treatment externalities.

Calculation of Total Deworming Effects

Having found evidence of externalities, we then performed a calculation of the overall impact of deworming on school participation, including our best estimate of the impact on schooling. This was not our statistical “test” for externalities, but primarily an input into the cost effectiveness analysis.

Following the logic of the original MK04 article, since there are not statistically significant externalities on either the “any moderate to heavy infection” or school participation variables at 3-6 km, and given the sensible perspective that externality effects will disappear at sufficiently large distances, it does not make methodological sense to include them in the overall estimate of the impact of deworming on school participation. The replication authors and Humphreys argue that it is appropriate to include a heavily weighted negative, but not statistically significant, estimated effect on school participation. This is not justified based on the underlying scientific logic, but rather through an almost legalistic argument that this quantity had been generated in MK04. In fact, MK04 did not state that there was any a priori reason to focus on school participation externalities out to 3-6 km, as opposed to examining school participation effects where analysis first revealed impacts on worm infections.

To justify the inclusion of the 3-6 km estimate in the overall school participation effect estimate, one would need to believe that that deworming reduced worm loads and this increased school participation among the treated, among untreated people in the treatment schools, and among people 0- 3 km away, but that deworming somehow reduced school participation (without affecting worm loads) from 3-6 km away. (The negative point estimate on school participation effects at 3-6 km is not statistically significant, in any case.) This entire line of argument is based on copying the form, but not the

methodological substance, of the original MK04 analysis, combined with an incorrect underlying presumption that conducting the analysis out to 6 km constituted our original analytical plan.

Why would including a not statistically significant externality estimate at 3-6 km affect the overall estimate? The main reason is because it is assigned a very high weight due to the large number of schools located at greater distances (i.e., on a uniform plane, the number of schools increases at rate distance^2). In our IJE response, we use the term "average externality" over some distance (i.e., 0-3 km) to refer to the per-pupil externality effect multiplied by the average number of treatment pupils located at that distance. An alternative term for this effect, the "weighted externality effect" at that distance, is arguably more descriptive and clearer, and we also use that term in what follows here.

We reproduce a version of Figure 1 from our IJE response here (below) since it succinctly represents the changes in effect estimates between the original paper and the updated results. It is immediate that these differences are minor, with the exception of the larger confidence interval on the 3-6 km weighted externality term.

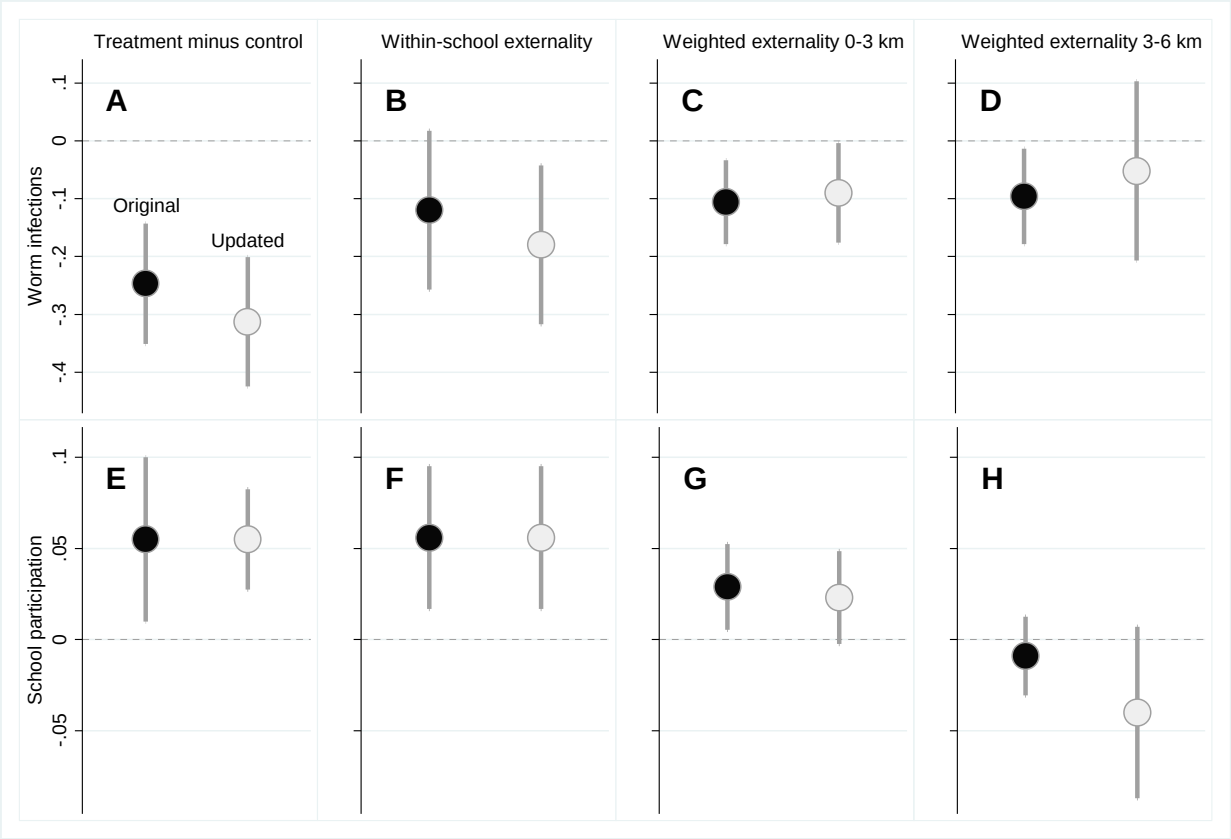


Figure 1: Deworming treatment effect estimates. Estimates from the original Miguel and Kremer (2004) article (black circles) and updated estimates from the Aiken et al (2015) re-analysis (light gray circles), with bars denoting associated 95% confidence intervals. Moderate-heavy intestinal worm infection is the dependent variables in Panels A-D, and the school participation rate is the dependent variable in Panels E-H. The estimated effect is: the difference between treatment schools and control schools in Panels A and E; the within-school externality effect for untreated pupils in the treatment schools in Panels B and F; cross-school weighted (average) externality effect for schools within 3 kilometers of treatment schools in Panels C and G, and the cross-school weighted (average) externality effect between 3 to 6 kilometers of treatment schools in Panels D and H. This figure is reproduced here from Hicks, Kremer, and Miguel (2015) in the *International Journal of Epidemiology,* albeit with the updated terminology.

A central argument in Humphreys' piece, echoing the replication authors, is that no change should be made in the procedure used to estimate the overall deworming treatment effect or overall cross-school infection externality effects with the updated data. This begs the question of what the procedure is. Aiken et al. and Humphreys write as if the procedure was pre-specified to go to 6 km independent of the data rather than acknowledging that this was a data-dependent decision. (Of course, the fact that this was a data dependent decision should legitimately be taken into account in forming posterior beliefs about the impact of deworming, but in this way our approach is no different than most empirical research conducted without a pre-analysis plan.) In the absence of a pre-analysis plan, it makes more sense to look seriously at all the data and come to an informed judgment. However, even if one were trying to infer an "implicit" pre-analysis plan from the article, the plan was certainly not to include 0-3 km and 3-6 km externality estimates weighted by the number of schools within those radii as the test for the existence of deworming treatment externalities.

The distance over which we considered school participation externalities was driven by the extent of worm infection externalities. The procedure that made sense at the time, and that still makes sense in our view, is to estimate school participation externalities at distances where we find evidence for externalities on worm infections, since this is the leading channel for impact. As noted above, with the updated data, these externality effects on "any moderate to heavy infection" are no longer significant in the 3-6 km range, and thus there is little rationale for examining school participation externality effects in this range.

In the supplementary appendix to our IJE response piece, we show that an estimator of overall impact including weighted externality effects in the 3-6 km range is not likely to be the most informative estimate, using a standard mean squared error (MSE) logic to guide our thinking on the appropriate estimator to choose in this setting. The goal of the MSE approach is to examine whether, under some reasonable assumptions, the MSE criterion might indicate that the estimator of "total deworming impact" on school participation that excludes the weighted 3-6 km externality term is preferable. This is done by comparing the likely reduction in bias achieved when including this term versus the increase in the variance of the estimator.

In particular, we show that the increase in MSE due to additional noise from including the weighted externality effect at 3-6 km is likely to be more than six times greater than any decrease in the MSE due to reducing bias. In other words, adding the weighted the 3-6 km effect likely adds a lot of noise to the estimate without much reduction in bias. This is not a proof, but we find it highly suggestive that the

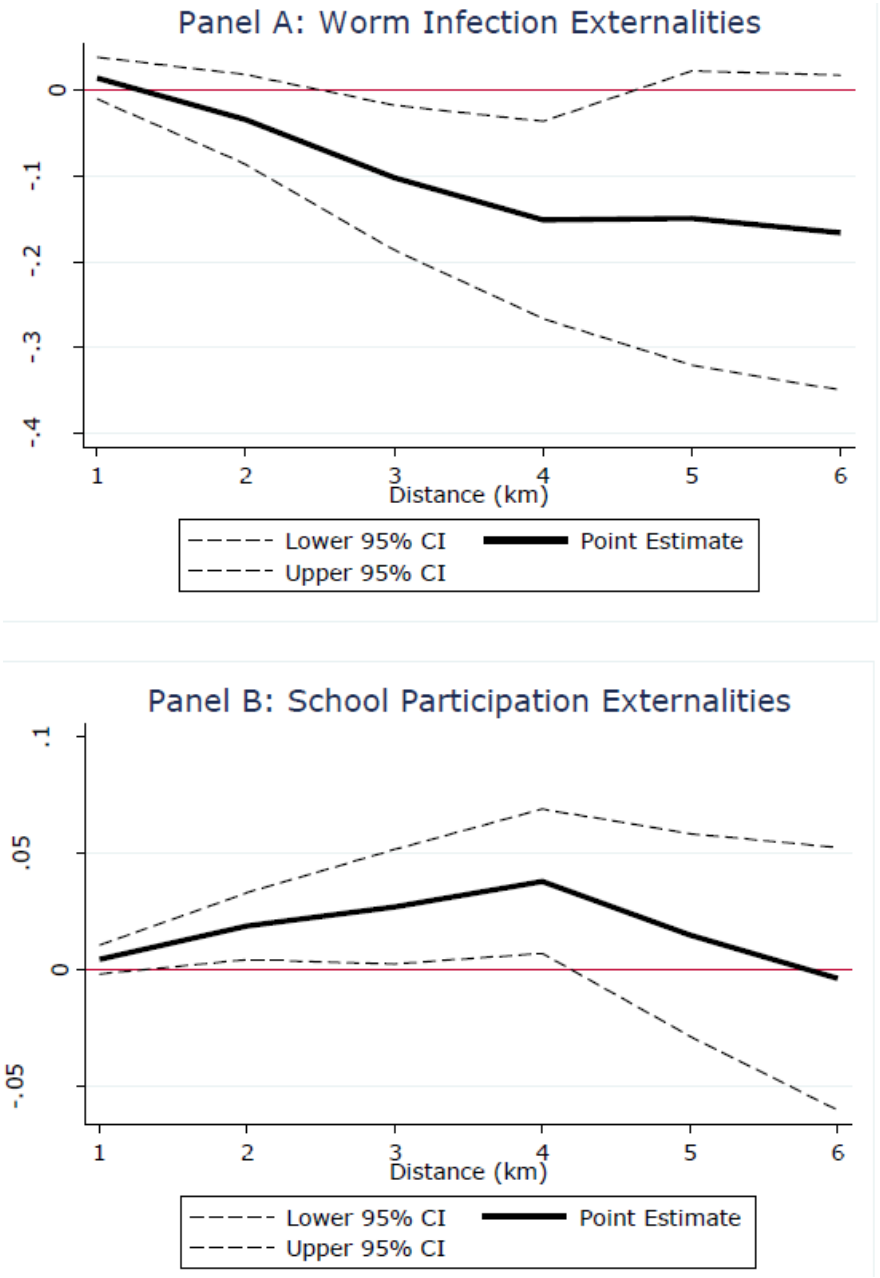
estimator that excludes the weighted 3-6 km externality effect on school participation is likely to be preferable. This has direct implications for what the "headline" number (as Humphreys puts it) should be for both worm infection reductions and school participation gains from the program.

It is also possible to show that going out to larger distances *d* does not in general improve the estimator in the MSE sense. Recall that the number of schools at distance *d* is proportional to d^2. The standard error of the estimate of externalities at a given distance is proportional to *1/d*. However, the estimate of the weighted externality effect is proportional to the estimated externality effect (proportional to *1/d*) times the number of schools (proportional to d^2), and rather than converging as *d* increases, this clearly diverges. Noise in the estimate becomes arbitrarily large as *d* increases, and under the plausible epidemiological assumption that the magnitude of the externality effects diminish with distance, will asymptote to zero, any benefits from reduced bias at larger *d* will eventually be exceeded by losses from increased imprecision.

The estimator that considers the direct effect plus weighted externality effects out to a distance of 3 km from each treatment school yields large estimated total deworming impacts on school participation. In fact, using the corrected data, the total deworming effect is somewhat larger than previously thought. Instead of reducing absenteeism by one-quarter, absenteeism is reduced by one-third (or 8.5 percentage points). Deworming is thus more cost-effective, costing just $2.92 per additional year of schooling rather than $3.50. As discussed above, this result is obtained using the same basic methodological approach employed in the original paper.

The best way of viewing the estimation of cross-school externalities in the MK04 paper is that the original regression specification choice was driven by our exploration of the data, and not by an ex ante plan to examine externality effects out to exactly 6 km for some reason. It is clear that if one averages the externality impact over a sufficiently large enough distance the impact will always approach zero and become imprecisely estimated in expectation. Thus it seems appropriate to step back and look at the data in its entirety, as we did with the updated data in Figure S1 of our supplementary appendix. We reproduce Figure S1 below since it succinctly summarizes the updated externality results, and illustrates both the increasingly wide confidence intervals, and declining weighted externality impacts (especially for school participation) at greater distances from treatment schools.

Figure S1: Weighted (average) externality impacts at various distances



Notes: Panel A plots the weighted (average) externality effect estimates presented in Table S3 (for worm infections) in the supplementary appendix to our IJE response**.** Panel B plots the weighted (average) externality estimates from Table S4 (for school participation). See notes to these tables for details on the regressions.

As Humphreys' Columbia colleague Andrew Gelman argues in his work on p-hacking here, "When pre-registered replication is difficult or impossible (as in much research in social science and public health), we believe the best strategy is to move toward an analysis of all the data rather than a focus on a single comparison or small set of comparisons." When we do that in Figure S1, it shows clear evidence of a positive externality over a significant range (between roughly 2 to 4 km for school participation), and then declines, consistent with externalities becoming smaller at greater distances, as expected.

Michael Clemens and Justin Sandefur have also carefully re-examined the cross-school externality estimates using the updated data, and they concur with this interpretation, concluding that there is indeed considerable evidence for these externalities, albeit most likely out to 4 km rather than the 6 km emphasized in the original MK04 paper (see here).

The central finding of deworming externality effects also holds under alternative statistical approaches. David Choi (here) examined the data and rejects the hypothesis of no cross-school deworming externalities at high levels of confidence ($P<0.05$), using an entirely different statistical approach than MK04.

Interactions between treatment and spillovers

One last issue merits a brief discussion here. Humphreys emphasizes the difference in cross-school externality impacts experienced by other treatment schools, versus by control schools. This is an issue that we discuss extensively in the original article and analyze through an interaction term in Table VII columns 3, 6 and 9. The original paper explicitly notes theoretical reasons to expect that externalities may apply both to children who were and were not themselves treated. The updated version suggests that much of the cross-school externality gains are concentrated among treatment schools, but the coefficient estimate on the key interaction term ("Group 1 * Group 1 pupils within 3 km") is relatively imprecisely estimated (and is not always statistically significant), so this pattern should not be over-interpreted.

The abstract of the original paper states that spillovers affected "untreated children in both treatment schools and neighboring schools." To the extent that we interpret the results of the updated Table VII to mean that cross-school gains were concentrated in other treatment schools, then this sentence should perhaps have eliminated the word "untreated" to the within-school externality. This change does not diminish the main methodological and policy-relevant findings of the MK04 paper. If anything, to the

extent that this finding of a positive interaction between own treatment status and the treatment of nearby schools is robust, it could imply that the benefits of a mass treatment program are even larger than we currently estimate, due to this complementarity between own and others' treatment.

What is the bottom line on externalities in this data, then? There is a range of evidence that there are in fact significant positive deworming treatment externalities: within schools and in a 0-3 km range to neighboring schools (and probably out to 4 km). We got the exact distances wrong in the original paper due to a coding error. But that does not mean there is no evidence for deworming treatment externalities.

To our mind, the most significant new evidence on the existence of deworming treatment externalities is from an "out of sample" test on a different population. Ozier (2014) examined younger children in the treatment and control communities in Kenya, who were themselves too young to be treated. Many were younger siblings of participants in the MK04 analysis. Ozier finds significant positive effects on the cognitive outcomes of children in the treatment communities roughly a decade later.

(c) Concerns about the discussion of the estimation of externalities in Humphreys' first worms-related blog post (on 1 August 2015)

Humphreys lays out an example in his Annex B that, he claims, calls into question the linear regression modeling approach to the estimation of externalities that we develop in MK04. It is certainly true that our estimation approach will not correctly estimate a non-linear relationship, but we never claim that our linear regression model does. Humphreys constructs an example in which deworming additional children could make some other children less likely to be infected, and others more likely to be infected. It is not completely clear whether the externalities he posits at different distances are meant to represent a particular realization of a stochastic process, in which case it is not clear what the example is trying to prove (or could prove) unless the underlying process is specified. Or he might be laying out an example in which the expected externalities take different signs at different distances away from a treatment school. If the latter, the example he constructs does not represent a false rejection of the null hypothesis of no externalities.

Note also that it is epidemiologically implausible that the incremental effect of additional exposure to local treatment schools is sometimes harmful, i.e., more local exposure to treated people raises infection risk, rather than lowering it. It seems to us that there is no scientific mechanism under which

deworming children in one school might increase infection levels in another school. Rather standard epidemiological models would predict that infection externalities are at least weakly beneficial, and likely to fall in magnitude at greater distances from treatment schools. Essentially Humphreys' example in his Annex B is a case with non-monotonic externality effects (i.e., the incremental effect is sometimes positive and sometimes negative), with heterogeneity across units. It is immediate in such a case that the estimation of externalities using a linear model is inappropriate. However, in our view there is simply no biological basis for the form that externalities take in his "rigged" example.

(Incidentally, it is straightforward to show formally that under the assumption that externalities are weakly of the same "sign" as direct effects, that the treatment minus control difference is a lower bound on the overall effect of deworming including both direct effects and externalities; see our discussion of this result here.)

We also tested for various types of nonlinearity in our data, and did not reject the linear specification in Miguel and Kremer 2004. Others are of course welcome to estimate more complicated nonlinear models, or alternative specifications, on the data. As noted above, Choi (2015) rejects the hypothesis of no cross-school externality effects at $P<0.05$ using an entirely different statistical approach, so the result is not an artifact of the linear regression specification in MK04.

3. Additional Specific Comments

(A) There is extensive discussion of Year 1 versus Year 2 effects in Section 2.1, and again in 2.2.3. The basic pattern in the data is that Year 1 treatment effects are somewhat larger (at around 6 percentage points) than Year 2 effects (at around 4 percentage points). The difference between these estimates is not statistically significant, as Humphreys himself notes, meaning this difference could be the result of chance or sampling variation. We thus do not view this as a very meaningful difference. Nor is there a strong logic for why Year 2 impacts should be larger than Year 1 impacts, which appears to be Humphreys' strong prior.

A natural possibility is that treatment has a constant effect on current school participation regardless of how long treatment has been in place. Treatment kills the worms currently in the body, but reinfection can be rapid. A person who was dewormed once, two months ago and a person who was dewormed twice, fourteen months ago and again two months ago, are likely to have similar worm infection levels in

their bodies. If current school participation depends on current worm infection, then one would not expect larger effects on school participation in Year 2 of treatment than in Year 1.

Note that Humphreys uses some confusing language in his discussion here, especially in 2.2.3, where he writes that "the implied effect of a second year of treatment is actually negative". What we think he means here is that the *additional* effect of a second year of treatment is negative, although as mentioned above, not statistically significant. So it is important to be aware that his "second year of treatment effect" is not the same as the Year 2 effect.

If one is interested in longer term effects of deworming, there is growing evidence from this Kenyan program as well as other contexts. For instance, our own follow-up work (Baird et al., 2015) finds substantial impacts of this childhood deworming program on adult labor market and educational outcomes. Other papers have also looked at long run effects in other contexts – Bleakley, 2007 in the historical U.S. South; Croke, 2014 in Uganda; and Ozier, 2014 in this same Kenyan program but examining a different population – and also find impacts of deworming on earnings, academic test scores, and cognitive test scores, respectively, nearly a decade or more after initial treatment.

(B) Communication (Section 3.5): As we have discussed elsewhere, we put together a detailed replication guide in early 2007 and in the process we uncovered some mistakes. We created a new version of all of the tables in the paper and we characterized our view that updating these led to only minor changes in the results. Figure 1 of our IJE response (reproduced above) makes it clear just how similar the original and updated results are. Since then we have distributed these materials (data, data manual, and replication manual with fully updated tables) to numerous scholars across multiple disciplines and institutions, many of whom apparently used the data for replication exercises in their graduate classes. None of these scholars, or the students carrying out replications in their classes, contacted us to discuss the differences between the original and updated results, nor did any apparently believe that they merited publication. We provided this same set of materials to Aiken, Davey and coauthors in 2013. When the Cochrane review authors requested a number of statistics, tables and results from us in 2008 for use in their review, we sent them results using the updated data. Nonetheless, in retrospect, we wish we had done more to disseminate these results, for instance, by distributing the replication guide with updated tables through a working paper series or publishing them in a journal.

(C) In section 3.2, Humphreys claims that, in our updated results (here), we choose estimators to get statistically significant results on the cross-school externalities. This is not true. As discussed at length in our section A.2.3.2, there is a strong statistical rationale for the choice of estimators given the updated data. The standard errors on the weighted 3-6 km externality effect, taking into account the larger number of schools in this range, become much larger, nearly doubling in the worm infection case and more than doubling in the estimation of the average effect of school attendance externalities. Including all schools, instead of only the nearest twelve, is what adds "noise" to the estimated weighted 3-6 km externality effects. With such large standard errors on the overall effect, the degree of noise in the estimates of overall externalities becomes very large, and the estimates are relatively uninformative about the underlying signal in the data. (We present a more detailed discussion of these issues in Section 2b above.)

(D) In section 3.3.1, Humphreys calls the "overall" program estimate that we present a "curious quantity". We disagree: the average effect of the program we are studying is arguably the most natural quantity to estimate and focus on. He goes on to claim that this approach "puts pressure on the model", but our approach provides a local externality effect near the actual treatment density values found in the data, while his proposed quantity (which is equivalent to a mass treatment program for all local schools) arguably relies much more heavily on functional form assumptions. Extrapolating the externality effect estimates out to 100% coverage is where any non-linearities or regression specification choices would likely become more important.

(E) In section 1, there is brief mention of an India study with one million children (here) that "found little evidence of impact on health outcomes". One million children sounds impressive, but this is a somewhat misleading claim with little relevance for the current discussion. The million child analysis was for a study of mortality; the related study of morbidity used a sample a tiny fraction of this size. The study area had low levels of infection prevalence, and most infections were light, making it a far healthier environment in relation to worms than the region we study in Kenya (or that studied in Uganda by Croke, 2014). The India study also did not collect data on educational outcomes, so does not even speak directly to one of the core issues in MK04, namely, the impact of deworming on school participation.

(F) Humphreys makes several different observations about the recent Cochrane Review of deworming. Regarding one of these points in particular, we share Humphreys' concerns with the approach taken in the recent Cochrane Review of deworming regarding the risk of "bias" in particular studies, which also sometimes affects study inclusion in the review. He writes in his Section 5: "Along with MK [Miguel and Kremer], 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."

Beyond the specific issue Humphreys raises here, the Cochrane review excludes several studies in which the control group eventually received deworming treatment (typically several years later) even though this would tend to dampen treatment effects, resulting in more conservative estimates. Several of these studies are long-run analyses based on cluster randomized designs that document significant positive impacts of deworming on educational and economic outcomes, as we discuss in our review (Ahuja et al., 2015). The Cochrane review also makes some seemingly odd choices regarding the analytical weight to place on estimates from the studies they do include. For example, they appear to weight a large cluster randomized study (Alderman et al. 2006) with 27,995 observations similarly to a very small individually randomized trial with N=70 (Dossa et al. 2001), and this is especially problematic as individually randomized trials will yield systematically downward biased estimates in the presence of treatment externalities, as we discussed above.

(G) Several commentators have noted that much of the early media coverage of the Aiken et al and Davey et al replication studies was sensationalistic and misleading, including Aaron Carroll (Healthcare Triage), Clemens and Sandefur (Center for Global Development), and Chris Blattman (Columbia University). The initial media coverage, most notably pieces in the Guardian (here) and Buzzfeed (here) write dramatically about the "debunking" of the findings of Miguel and Kremer 2004. This goes beyond the written claims in the replication pieces themselves (as even a glance at the abstracts of those articles indicates), but this initial round of articles set the tone for much of the subsequent debate; first impressions matter. In our view, it would be impossible for any objective observer of Figures 1 and 2 (reproduced from our IJE response above) to conclude that the findings of Miguel and Kremer 2004 had been debunked.

Yet it does not appear that our IJE response, or our earlier 2014 responses to the replication teams 3ie papers, figured at all in the influential initial Guardian or Buzzfeed pieces. Neither of the journalists who wrote those pieces contacted us, and neither of them referenced our IJE response (even though it can be found right next to the Aiken et al and Davey et al articles on the IJE website) or our 3ie responses. After the publication of her article, we emailed and spoke by phone with Sarah Boseley, the author of the Guardian piece, and she acknowledged that she was unaware that we had even written a response, and that our response piece had never come up during her interviews with the replication authors. Ben Goldacre, the author of the Buzzfeed piece, also appears to have been unaware of our response when he wrote his article. This is highly unfortunate, since even a cursory Google search would have revealed our responses, as well as the earlier blog discussion on the replication articles, including Ozler's detailed blog post from January 2015, which concluded that the main results in Miguel and Kremer 2004 continue to hold.

It is telling, in our view, that the journalists and commentators who were aware of our responses or interviewed us directly, wrote about the research in a much less sensationalistic fashion, and reached conclusions much closer to the consensus view among the majority of scholars who have evaluated the evidence. These media include pieces in the Financial Times (here) and the Economist (here), as well as many others (some of which are listed here).

(H) In Section 2.2.2 and the start of section 2.3, Humphreys makes a claim about "imbalance in the handling of treatment and control groups which could lead to biased estimates". The central assertion along these lines by the replication team (in Davey et al., 2015) is that we collected different numbers of attendance observations from treatment and control groups in a way that was correlated with school attendance levels, and that this pattern changed over time for Group 2. Humphreys briefly mentions that we respond to this claim, but does not provide the substantive details of our response or evaluate our points. This may leave the reader with the impression that we do not have a credible response worth mentioning.

In fact, we show that there are no statistically significant differences in the number of attendance observations collected between treatment groups as a function of attendance levels, nor do these patterns change over time (see Table S6 in our supplementary appendix).

One additional note here: even if there would have been an unbalanced pattern of data collection in terms of attendance observations (which we show there is not), it could be addressed analytically by using a population-weighted attendance measure, rather than weighting by the number of attendance observations. Doing so would maintain the analysis as the average impact in the sample *population*, a meaningful quantity.

Instead the replication authors choose to focus on specifications that weight each school equally regardless of pupil population. The impact weighting each school equally is not standard in the health economics or public health literature, nor is it appropriate in a setting in which some schools only have 100 pupils and others have over 700 pupils. Davey et al. (2015) do not provide any rationale for why they would arbitrarily over-weight pupils in the smaller schools up to seven times more than comparable pupils in larger schools, nor do we feel that there is a plausible rationale for such a decision. Ozler's comment is similarly critical of the replication authors' choice regarding the weighting of observations in this analysis.

It is also worth noting that the approach of weighting each school equally was not mentioned in the replication authors' pre-analysis plan (Aiken et al., 2013), where they emphasize individual level analysis.

Beyond this choice, in our IJE response we raise multiple concerns regarding the analytical choices made by the replication authors. Many of their choices deviate from (or were not described in) their pre-analysis plan and we believe many have important methodological shortcomings that lead to erroneous conclusions.