Supplementary Appendix for

"Deworming externalities and schooling impacts in Kenya: A Comment on Aiken *et al.* (2015) and Davey *et al.* (2015)"

Joan Hamory Hicks, Michael Kremer and Edward Miguel*

*Corresponding author. E-mail: emiguel@berkeley.edu

May 2015

This PDF file includes:

Contents

A. Detailed response to Aiken et al. (2015) pure replication

A.1 Summary of Main Points

Aiken et al. (2015) undertake a replication of Miguel and Kremer (2004), which evaluates a Kenyan project in which mass treatment with deworming drugs was randomly phased into schools, rather than to individuals, allowing estimation of overall effects even in the presence of epidemiological effects due to reduced transmission of disease. We thank the authors for undertaking this work and are pleased to be part of a continuing conversation regarding the health and development impacts of school-based deworming. We are supportive of the process of replication as a normal part of scientific research, and have been active supporters of growing efforts to promote greater transparency and reproducibility in the social sciences (Miguel et al., 2014b).

This appendix (Appendix A) comments on the replication analysis presented in Aiken et al. (2015). The tables in Aiken et al. (2015) confirm the main empirical findings of the Miguel and Kremer (2004) paper, namely (1) that deworming creates positive epidemiological externalities, which implies that individually randomized studies will underestimate the impact of deworming; and (2) that deworming increases school participation.

In particular, Aiken et al. (2015, Appendix Table VII-Updated) find substantial epidemiological externalities on worm infections among untreated classmates (P-value < 0.05), and externalities on worm infections among schools within 0-3 km (P-value < 0.05). With regard to school participation, Aiken et al. (2015, Appendix Table IX-Updated) find externalities on school participation among classmates (P-value < 0.01), and externalities on school participation in neighboring schools within 3 km (P-value < 0.10). Aiken et al. (2015, Table 5) also find that deworming increases school participation by 5.7 percentage points in treatment schools relative to control schools (P-value < 0.01); the comparable deworming impact on school participation in Miguel and Kremer (2004) was 5.1 percentage points (P-value < 0.01). The strong evidence of within-school externalities in Aiken et al. (2015) implies that one of the key conclusions of the Miguel and Kremer (2004) paper – that individually randomized studies of the impact of deworming will underestimate the true impact – remains valid. Aiken et al. (2015, Appendix Table IX-Updated) also find that worm infections impact school participation using an instrumental variables approach (P-value < 0.05).

Aiken et al. (2015) helpfully correct a number of issues in the Miguel and Kremer (2004) paper, including: (1) a number of rounding errors in reported coefficients, some of which led to associated errors in reported P-values, and (2) some tables reporting regressions run on intermediate, rather than final, versions of data sets. These inconsistencies were introduced during the editing process when the paper was being prepared for publication, and neither of these lead to substantial changes in coefficient estimates. Aiken et al. (2015) also discuss cases of inaccurately labeled statistical significance. The effect on anemia was originally reported as significant with P-value < 0.05 but is found in re-analysis to have a P-value of 0.19. The coefficient estimate and standard error in Miguel and Kremer (2004) were reported correctly, but we believe the significance level was misreported due to a calculation of the t-statistic using rounded coefficients.

The replication also corrects errors in the original code used to estimate the externalities associated with deworming. As a result of these errors, Miguel and Kremer (2004) measure externalities among only a subset of schools within 3-6 km, rather than all schools in this radius.

The externality effect on moderate-to-heavy worm infections from treated pupils attending schools 3-6 km away was statistically significant in the original Miguel and Kremer (2004) analysis, but is not significant in the updated analysis in Aiken et al. (2015). The point estimate on the 3-6 km externality term in the school participation analysis was negative but not statistically significant in the original Miguel and Kremer (2004) analysis, and remains so in the updated analysis. The fact that there are no infection externalities in the 3-6 km range (with the updated data) means there is little reason to expect school participation externalities at this distance.

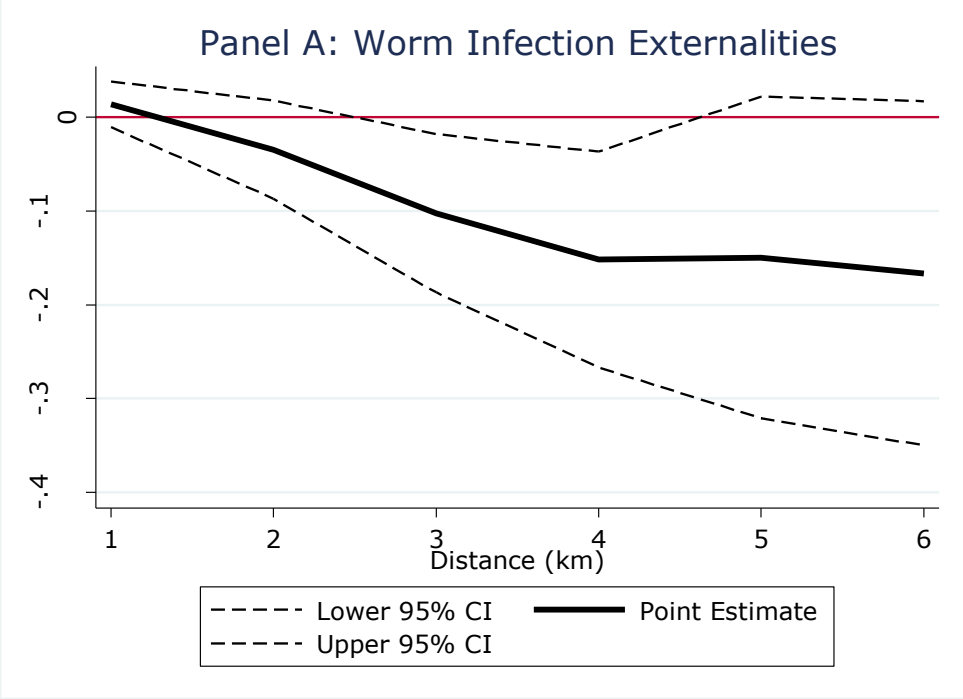
When the 3-6 km externality terms are omitted, externality effects are strong both within schools, and across schools up to 3 km away, both for worm load and for school participation. Estimated overall externality effects that go out to 3 km or to 4 km are all also strong. However, an estimator for overall externalities that goes out beyond this distance, and that puts extensive weight (due to the large numbers of schools at that distance) on the not statistically significant 3-6 km externality estimate adds large amounts of "noise" to the overall externality estimate.

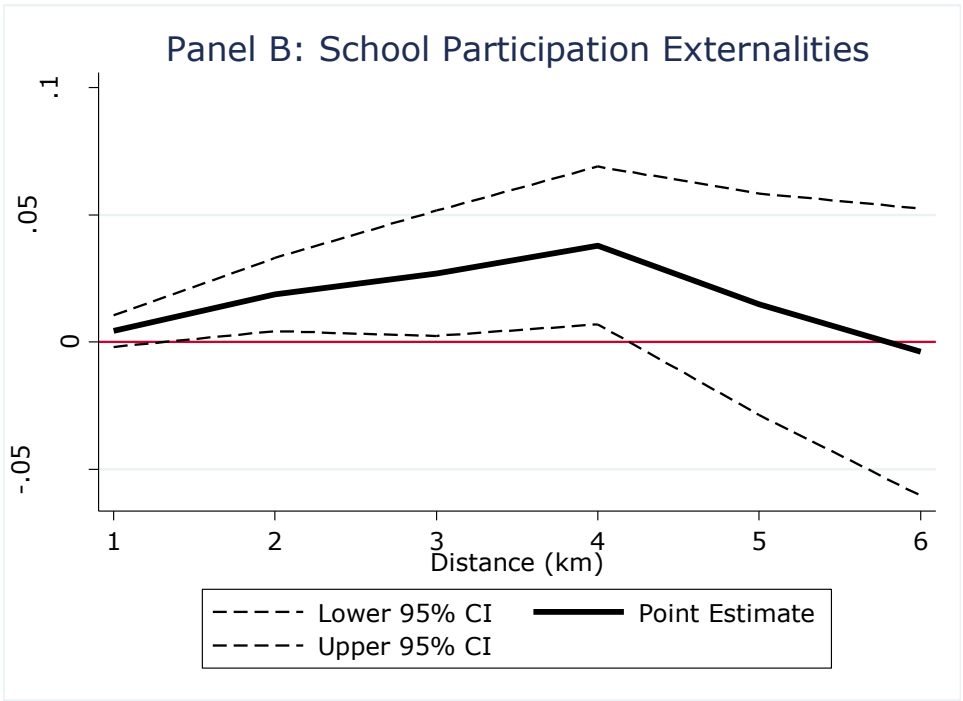
Aiken et al. (2015, Methods of original study) note that they "*reproduced the analytic steps to re-determine the results as originally calculated,*" but they did not re-evaluate these steps in light of their re-analysis findings. In Section A.2 (below), we demonstrate that, under reasonable assumptions, the estimator that excludes the 3-6 km externalities is preferred under the standard statistical criterion of minimizing mean squared error. We thus differ with Aiken et al. (2015) over the appropriate way to calculate overall deworming externalities on school participation and the overall impact of deworming on school participation in the updated data.

Figure S1, Panel B demonstrates how standard errors on school participation externality estimates become large when one considers schools beyond 4 km using the updated data. The average cross-school externality impact on school participation is positive and statistically significant at 95% confidence at distances of 0-2, 0-3 and 0-4 km. This is evidence of deworming externalities for schools within up to 3 to 4 km of treatment schools, but not for more distant schools.

The "cost-effectiveness" of deworming in terms of boosting school participation is nearly unchanged relative to the original paper, using the updated data and considering the direct effects and the externalities up to 3 km, with 34.3 additional years of school participation per $100 of spending on deworming with the updated data (and 29.1 additional years per $100 in the original analysis). Focusing on the most conservative treatment effect estimate, the "naïve" T – C difference, also implies that deworming is a highly cost-effective approach to reducing school absenteeism in this setting, with 17.8 additional years of school participation per $100 of deworming spending, placing it among the most cost-effective interventions yet evaluated in education studies (see Figure S2).

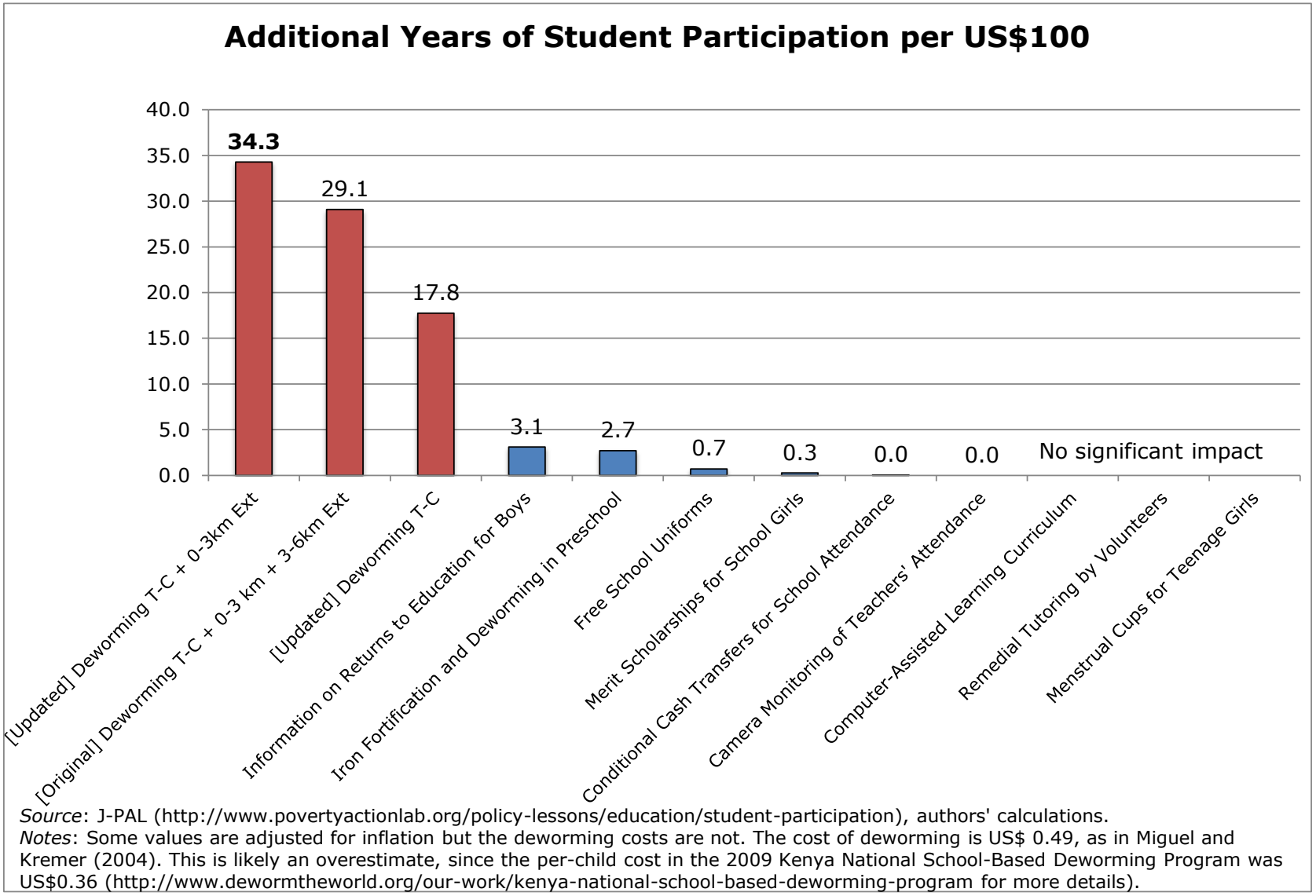
Figure S1. Average externality impacts at various distances





Note: Panel A plots the "average externality effect" estimates presented in Table S3 (for worm infections) and Panel B plots the "average externality effect" estimates from Table S4 (for school participation). See the notes to these tables for details on the regressions.

Figure S2: Cost-effectiveness of school participation interventions



Source: J-PAL (http://www.povertyactionlab.org/policy-lessons/education/student-participation), authors' calculations.
Notes: Some values are adjusted for inflation but the deworming costs are not. The cost of deworming is US$ 0.49, as in Miguel and Kremer (2004). This is likely an overestimate, since the per-child cost in the 2009 Kenya National School-Based Deworming Program was US$0.36 (http://www.dewormtheworld.org/our-work/kenya-national-school-based-deworming-program for more details).

New evidence is accumulating on the educational and socio-economic impacts of child deworming. A key lesson of Miguel and Kremer (2004) is that traditional individual-level randomized designs will miss any spillover benefits of deworming treatment, and this could contaminate estimated treatment effects. Thus cluster randomized designs provide better evidence. Three new working papers with such cluster randomized designs estimate long-run impacts of child deworming up to 10 years after treatment; these effects on long-run life outcomes are arguably of greatest interest to public policymakers.

Croke (2014) finds positive long-run educational effects of a program that dewormed a large sample of 1 to 7 year olds in Uganda, with statistically significant average test score gains of 0.2 to 0.4 standard deviation units on literacy and numeracy 7 to 8 years later. The Ugandan program is one of the few studies to employ a cluster randomized design, and earlier evaluations of the program had found large short-run impacts on child weight (Alderman et al., 2006; Alderman, 2007). Croke (2014, p. 16) also surveys the emerging deworming literature and concludes that "*the majority of clustered trials show positive effects*".[1]

Two other new working papers explore the long-run impacts of the Kenya program we study. While the primary school children in the Miguel and Kremer (2004) sample were probably too old for deworming to have major impacts on brain development, and there was no evidence of such impacts, Ozier (2014) estimates cognitive gains 10 years later among children who were 0 to 2 years old when the deworming program was launched and who lived in the catchment area of treatment schools. These children were not directly treated themselves but could have benefited from the positive within-community externalities generated by mass school-based deworming. Ozier (2014) estimates average test score gains of 0.3 standard deviation units, which is equivalent to roughly half a year of schooling and similar to the effect magnitudes estimated by Croke (2014). This provides further strong evidence for the existence of large, positive, and statistically significant deworming externality benefits within the communities that received mass treatment.

Finally, Baird et al. (2014) 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. Ten years after the start of the deworming program, men who were eligible to participate as boys work 3.5 more hours each week, spend more time in entrepreneurship, are more likely to hold manufacturing jobs with higher wage earnings, and have higher living standards. Women who were eligible as girls have better educational outcomes (including higher rates of passing the primary school completion exam and enrolling in secondary school), are more likely to grow cash crops, and reallocate labor time from agriculture to entrepreneurship. The impacts of deworming subsidies on labor hours are sufficiently large that the social internal rate of return is very high, with an annualized rate of at least 32.2%.

Taken together, and building on Miguel and Kremer (2004), Alderman et al. (2006), and Alderman (2007), this new wave of studies promises to bring considerable new evidence to bear on the long-run impacts of childhood deworming on important life outcomes in areas with high worm infection rates.

[1] One exception is Awasthi *et al.* (2013), who use a clustered randomized design and find positive, but not statistically significant, effects of deworming on infant mortality and weight in a lightly infected preschool population in India. This study does not track later educational or labor market outcomes.

We focus on the most important technical issues of Aiken et al.'s (2015) replication analysis in Section A.2, and address additional points raised in their report in Section A.3. In Section A.5, we present all of the tables from the original Miguel and Kremer (2004) paper, updated using the final data and correcting any coding errors discussed in Aiken et al. (2015), and in Section A.6 we present our preferred final tables using the updated data. The tables we present in this appendix should be considered the fully "updated" version of the analysis in the 2004 paper, and these may be of interest to scholars, non-profit organizations, and policymakers. The full replication dataset, code, and documentation are available from the authors and the Harvard Dataverse Network, and we welcome further analysis by other interested researchers.

A.2 Technical response to Aiken et al. (2015)

In this section, we first provide an overview of the Miguel and Kremer (2004) study, and then go on to discuss the cross-school externality findings and other issues raised in Aiken et al. (2015).

A.2.1 Background on Miguel and Kremer (2004)

It is useful to briefly summarize Miguel and Kremer (2004)'s approach and findings up front. The abstract to the paper summarizes its main goals, results and contributions, and we reproduce it here:

> *"Intestinal helminths—including hookworm, roundworm, whipworm, and schistosomiasis—infect more than one-quarter of the world's population. Studies in which medical treatment is randomized at the individual level potentially doubly underestimate the benefits of treatment, missing externality benefits to the comparison group from reduced disease transmission, and therefore also underestimating benefits for the treatment group. We evaluate a Kenyan project in which school-based mass treatment with deworming drugs was randomly phased into schools, rather than to individuals, allowing estimation of overall program effects. The program reduced school absenteeism in treatment schools by one-quarter, and was far cheaper than alternative ways of boosting school participation. Deworming substantially improved health and school participation among untreated children in both treatment schools and neighboring schools, and these externalities are large enough to justify fully subsidizing treatment. Yet we do not find evidence that deworming improved academic test scores."*

Miguel and Kremer (2004) evaluate a deworming program conducted by the non-governmental organization ICS in 75 Kenyan primary schools. Schools were divided into three groups of 25 schools each, and these groups were phased into deworming treatment over time, thus allowing the data to be analyzed using stepped-wedge methods. Deworming treatment began in March 1998 among the 25 Group 1 schools, and took place between March and June 1999 for both Group 1 and Group 2 schools; Group 3 schools did not receive deworming treatment in either of these two years.

It is worth reviewing the nature of disease transmission since these bear on the potential for epidemiological externalities. Geohelminths are deposited in stool, and while adults in the area typically use latrines, children are more likely to defecate in the open. This can lead to transmission of geohelminths when children defecate near their school or home. Schistosomiasis involves transmission

through fresh water (via intermediate hosts) and in the study area can be transmitted when children travel to Lake Victoria to bathe or fish. It is thus likely to be transmissible over somewhat larger distances than geohelminths, particularly as part of the life cycle of the parasite occurs in snails and the snails themselves are mobile. Treatment for geohelminths was provided in all treatment schools, while treatment for schistosomiasis was only provided in those schools with sufficient prevalence of the disease, typically in schools that were located near Lake Victoria.

It was only after evidence of externalities among untreated children in the treatment schools, both in terms of worm infections and school attendance, was detected, that the decision was made to investigate the existence of externalities across neighboring schools. This analysis initially focused on the schools closest to the treatment schools. Finding evidence for positive deworming treatment effects on both worm infections and school participation at those distances, impacts were then estimated at even greater distances from each school. Externality results were presented up to 6 km away from each school, and no farther, not because there were *a priori* reasons to expect effects at 6 km *ex ante*, but rather because having found effects at 3 km – and knowing that effects could be biased downward if spillover effects were not included – we thought it worth checking for effects further out, as long as they could be estimated with sufficient precision. Note that the key test in Miguel and Kremer (2004) for the existence of externality effects lies in the statistical significance of externalities at various distances, rather than being based on a weighted sum of these externalities.

A.2.2 Results common to Aiken et al. (2015) and Miguel and Kremer (2004)

Miguel and Kremer (2004) conclude that deworming reduced worm infections and improved school participation in Kenyan primary schools, when deworming treatment schools are compared to control schools that did not receive deworming drugs. The paper also finds evidence of large externality (spillover) benefits in these two dimensions among untreated children (those who did not receive deworming drugs) in treatment schools. It presents evidence for large externality benefits on worm infections for those attending other schools located near treatment schools (within 0 to 3 km) and for those located farther away from treatment schools (3 to 6 km away). It presents evidence for large externality benefits on school participation within 3 km of treatment schools, but finds no statistically significant externality effect from 3-6 km away.

The Aiken et al. (2015) replication affirms most of these findings in the Miguel and Kremer (2004) paper. Epidemiological externalities on worm infections within schools, and across schools located up to 3 km away remain strong. Direct effects of deworming on school participation and externality effects within schools remain strong. As in Miguel and Kremer (2004), there are no statistically significant externality effects on school participation beyond 3 km. As in Miguel and Kremer (2004), there is no statistically significant effect on test scores within the time period examined.

However, the replication was useful in highlighting some discrepancies, and we thank the replication team for enabling us to jointly update the scientific record.[2] A key difference is one of

[2] We note that we produced a full set of data and documentation in 2007 that replicated and updated the tables from Miguel and Kremer (2004), including a "replication manual" that detailed typos, rounding errors, and coding errors in the tables. We have distributed these materials (data, data manual, and replication manual with fully updated tables) since that time to numerous scholars across multiple disciplines and institutions. We provided this same set of materials to Aiken and coauthors in 2013 *prior* to the registration of their pre-analysis plan. In fact,

interpretation of the cross-school externalities on school participation. We interpret the results as indicating statistically significant externalities at 0-3 km and no statistically significant effects at 3-6 km. Aiken et al. (2015) note that the confidence interval on a weighted sum of the two coefficients (with weights given by the average number of schoolchildren at each distance) includes zero, and therefore conclude that there are no cross-school externalities on school participation.

A.2.3 Errors and discrepancies addressed in Aiken et al. (2015)

Aiken et al. (2015) helpfully re-analyze the data in Miguel and Kremer (2004), and discuss a number of errors. We review these below, starting with rounding errors and minor changes to the data set (which accounted for the majority of the discrepancies), and then considering the coding errors that led to measurement of externalities in the 3-6 km range only among a subset of schools near the reference school.

A.2.3.1 Rounding errors and data updates

A leading reason for these errors had to do with the rounding of some figures after reducing the number of significant figures from three to two (for aesthetic reasons) during the journal revision process. For instance, a figure of 0.7745 was initially presented as 0.775 in tables, but then incorrectly rounded up to 0.78 (rather than down to 0.77) when we moved to presenting only two digits in the published version of the paper. By definition, rounding errors are small in magnitude, and they lead to only small changes in the results.

Aiken et al. (2015) additionally discuss several cases of inaccurately labeled statistical significance. We believe that some of these were also the result of rounding in coefficient estimates and standard errors, which led to inaccuracies in t-statistics. Some of these led to results reported as at traditional levels of confidence becoming insignificant. The most important among these is that presented in Miguel and Kremer (2004), Table V – "Proportion anemic", which was originally reported as statistically significant with 95% confidence, but is found in reanalysis to have a p-value of 0.19. Note that the coefficient estimate and standard error in the original Miguel and Kremer (2004) paper were reported correctly, so the magnitude of the effect is unchanged at -2 percentage points, but the statistical significance level was misreported. While it was important to include an examination of anemia from a medical perspective, Miguel and Kremer (2004) note that anemia is not likely to be a main channel of impact in the setting examined because only 4% of the population was anemic. Correspondingly, this is not one of the major findings of the original paper.

A second reason for these errors is that intermediate versions of several datasets were used in production of the paper, and not all of the tables were fully updated with final versions of the data during the journal revision process. This accounts for the largest number of discrepancies with the original paper. However, the extent of final data cleaning was only moderate over that time, so that using different versions of the data leads to very similar results.[3] We support the growing trend among

---

Aiken et al. (2015) reproduce portions of our own replication manual in their report. The replication manual and data user's guide can be found in Miguel and Kremer (2014) and Miguel et al. (2014a).

[3] Data cleaning, in both Kenya and the United States, was an ongoing process on these large, original data sets during 1998-2002, and this led to the existence of various "intermediate" versions of data, versions that were

journals to require authors to prepare online replication data materials prior to publication, since we believe that this will make it less likely that these sorts of errors will happen going forward.

Aiken et al. (2015) note at several points that the changes in results due to these rounding errors and data updates are generally small (in the range of 0.01 for many estimates), and do not substantively change the results in Miguel and Kremer (2004).

A.2.3.2 Externality effects 3 to 6 km away from treatment schools

The biggest issue is coding errors in the construction of the local population density terms at a radius of 3-6 km from each school. The main error meant that whereas Miguel and Kremer (2004) reported externalities between 3-6 km away from a school, it actually measured externalities only for those schools within 3–6 km that were among the 12 closest schools.[4] Some deworming treatment effects are marginally larger in magnitude and somewhat more precisely estimated when all schools within 3-6 km are included, and some are smaller or less precisely estimated. There are a few noteworthy changes and we focus on those here.

This issue did not affect the construction of the 0-3 km externality terms, but in a number of cases it did affect the construction of the 3-6 km externality terms. In no case did a school have more than 12 schools within a 4 km radius, so externality terms up to that radius were correct. Three quarters of schools had twelve or fewer schools within 5 km. However, at distances greater than 5 km many schools are affected.

Once all schools within a 3-6 km radius are included, Aiken et al. (2015) find direct effects (namely, the Treatment vs. Control difference) and within-school externality impacts for worm infections that are marginally larger in magnitude than the original study. Furthermore, the replication confirms Miguel and Kremer (2004)'s findings of cross-school epidemiological externality impacts within 3 km, as well as the direct effects and within-school externality impacts for school attendance. It is mainly the cross-school externality estimates beyond 3 km that are affected. With regard to worm infections, Miguel and Kremer (2004) find reductions within 3-6 km, but this finding is not statistically significant upon re-analysis.[5] There is no evidence of externality effects on school participation among the full set of schools within 3–6 km.

The standard errors on the "overall" 3-6 km externality effect 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 overall 3-6 km externality effects. With such large standard errors, the degree

---

progressively cleaner over time. Cleaning typically took the form of eliminating duplicate observations, correcting data entry errors through hard copy checks, and better matching across files. Economics journals ask authors for specific revisions, and in revising the paper we also discovered and corrected minor errors in our dataset. However, we did not systematically update all of the other tables, so different tables in Miguel and Kremer (2004) were based on slightly different versions of the dataset.

[4] There was a second, and much more minor, error in the construction of the externality measures, which affected only two schools. We explain this error in detail in Section A.3.

[5] However, there is evidence that these longer-range 3-6 km externalities exist for schistosomiasis infection, as shown in Appendix Table VII-Updated of Aiken et al. (2015) and Table VII of Section A.5 below, but schistosomiasis drugs were given in only a minority of schools (where the disease was common).

of noise in the estimates of overall externalities becomes very large, and the estimates are relatively uninformative about the underlying signal in the data.

Note that the 3-6 km externality effect for school participation was not statistically significant in the original Miguel and Kremer (2004) paper. At a distance over which overall externalities can be precisely estimated (up to 3 km), the main finding remains that there are large and highly significant cross-school externalities for both worm infections and school attendance. Using the updated data, the estimated average cross-school externality effect of deworming on worm infections is a reduction of 10.2 percentage points (s.e. 4.3, P-value < 0.05), shown in column 2 of Table S1. The estimated average cross-school externality effect of deworming on school participation is a gain of 2.7 percentage points (s.e. 1.3, P-value < 0.05), shown in column 2 of Table S2.

Aiken et al. (2015) follow the original paper in focusing on externalities out to 6 km, and calculate the "overall effect" of deworming on school attendance by taking the weighted sum of the two coefficients (on 0-3 km and 3-6 km, with weights given by the average number of schoolchildren at each distance). The weight given to the 3-6 km externality term increases substantially once all schools in the 3-6 km range are included. The authors go on to conclude that "*the 'total effect' on school attendance resulting from the intervention … was more modest and less precisely estimated than previously reported and was also not statistically significant*" (Aiken et al., 2015, Discussion).

We disagree with this claim, and believe it is a misinterpretation of the statistical evidence presented in their tables. Given the updated data, a regression specification different from that in the original paper is necessary to precisely estimate the overall externality effect of deworming. While it is natural to first replicate the exact specification used in the original paper, the changes to the data mean that this estimator is no longer appropriate. More reliable conclusions can be reached by excluding the 3-6 km externality effect from the calculation of overall effects, since it is adding a tremendous amount of "noise" to the estimate.

Miguel and Kremer (2004) demonstrate that the "naïve" mean difference between Treatment and Control units, what we call the T-C difference, underestimates the total impact of treatment in the presence of epidemiological externalities and propose a simple and tractable methodology for estimating cross-unit externalities. The idea behind the estimation strategy in Miguel and Kremer (2004) is that the "naïve" T-C difference – and in fact any estimator that only considers externalities up to a certain distance away from each school – would serve as a lower bound on the true overall impact of deworming due to the presence of positive spillovers.

The original paper presented externality results up to 6 km away from each school, and no farther, not because we had conceived of this exact test *ex ante*, but because we could not precisely estimate overall externality effects at greater distances. Page 186 of the original paper explains why we chose to focus on externality impacts out to 6 km from each school – but not beyond – at that time:

> *"Due to the relatively small size of the study area, we are unable to precisely estimate the impact of additional treatment school pupils farther than six kilometers away from a school, and thus cannot rule out the possibility that there were externalities at distances beyond six kilometers and possibly for the study area as a whole, in which case the estimates presented in Table VII (and discussed below) would be lower bounds on actual externality benefits."*

However, the effect of the variable construction issue was that instead of measuring externalities between 3-6 km, we were in fact measuring externalities over a narrower range (typically a subset of schools within the 3-6 km range). The key issue that arises when we expand the measure of externalities to all schools within 3-6 km is that the precision of the overall externality estimate goes down dramatically. There is a natural statistical interpretation for this reduction in precision using the updated data. The externality coefficient estimates are multiplied by the average number of treatment pupils in the appropriate range (either 0-3 km or 3-6 km), and this number increases dramatically in the updated data that includes all schools in the 3-6 km range. Since the updated 3-6 km externality terms are not statistically significant for worm infections (Table S1, column 3) or school participation (Table S2, column 3), this means that a lot of "weight" in the calculation of the overall externality effect is placed on distant schools with an imprecisely estimated "zero" externality effect.[6]

As shown in Table S1, the standard error on the average overall 3-6 km externality effect nearly doubles in the estimation of infection externalities; you can see this in Table S1 by comparing the standard error of 0.042 in column 6 (results from the original paper, with the coding errors) to the standard error of 0.079 in column 3 (results using the updated and corrected data). Similarly, it more than doubles in the estimation of school participation effects (comparing the standard error of 0.011 in column 6 to the standard error of 0.024 in column 3 of Table S2). This marked reduction in statistical precision is also clear visually in Figure S3, where the 95% confidence intervals increase substantially once the updated 3-6 km externality effects are included, for both infection outcomes and school participation outcomes. These large confidence intervals are relatively uninformative, and also lead the estimate of total deworming impacts to be much less precisely estimated.

If we impose a sensible decision rule and exclude externality estimates that are simply too imprecisely estimated to be informative (as we did in the original analysis), then including the 3-6 km effect is inappropriate with the updated data. The best way to think about it is that including these 3-6 km externalities is like adding a very "noisy zero" estimate to what is otherwise quite a precise estimate. It is appropriate to focus on the estimator that includes the "naïve" treatment minus control difference plus the 0-3 km externalities, since these are both precisely estimated, and these together constitute a lower bound on the overall effect of deworming under the reasonable assumption that deworming externality effects are non-negative. Even focusing on the precisely estimated "naïve" estimator – the simple T minus C difference – which is downward biased since it excludes all cross-school externality effects, would be preferable to employing the estimator that incorporates externalities from 3-6 km, since the naïve estimator is precisely estimated and provides a lower bound on the magnitude of the true effect.

It is useful to think about including additional externality estimates in terms of the usual goal of choosing an estimator that minimizes "mean squared error". Recall that mean squared error is the sum of the variance of an estimator plus the square of its bias. Including further externality terms in the analysis helps reduce bias in the estimation of the overall effect (by capturing more of the externalities)

[6] A second issue is that, while more data is utilized by bringing in all schools between 3-6 km, precision falls because there is relatively less idiosyncratic variation in the number of treatment school pupils (relative to total pupils) in larger geographic areas.

but the analyst faces a trade-off if their inclusion increases the variance of the resulting estimator. In cases where standard errors increase dramatically with the inclusion of additional terms, mean squared error is reduced by focusing on precisely estimated effects that constitute a lower bound on the true overall effect. Aiken et al.'s (2015) conclusion that there is no significant evidence of a deworming effect on school participation is driven by their decision to take a precisely estimated effect that is a lower bound on the true impact – and indicates large school participation gains – and add lots of "noise" to it, by including the 3-6 km externality effects. In our view, this is a statistically inappropriate approach given the updated data.

The patterns in the tables illustrate this point. Using the original data, including the 3-6 km externality effect in the overall deworming effect does not appreciably increase the standard error on the overall effect: in Table S1, the standard error remains unchanged at 0.055 when the 3-6 km term is included in the worm infection analysis (shown in the bottom row of columns 5 and 6), and similarly the standard error on the overall effect remains nearly unchanged in the school participation analysis (comparing columns 5 and 6 of Table S2). With the original data, there does not appear to be much of a trade-off between bias and statistical precision. Moreover, with the original data the 3-6 km externality effect is statistically significant on its own (Table S1, column 6), so it is natural to include it in the calculation of overall effects. While the 3-6 km externality effect is not significant for school participation using the original data (Table S2, column 6), it is reasonable to consider the possibility that there might be schooling externalities at that distance, given the worm infection externality gains at 3-6 km.

(Note that in the original working paper version of the paper (Miguel and Kremer, 2001), we did not consider the 3-6 km externality effects in our calculation of overall deworming impacts on school participation since they were not statistically significant, and in fact we did not even present them in the analysis (in Table 11 of that paper). During the paper revision process at the journal *Econometrica*, we later incorporated the 3-6 km externality effects into the school participation regressions to maintain analytical consistency with the infection externality regressions, and given the existence of statistically significant 3-6 km worm infection effects using that data.)

In contrast, the pattern of results using the updated data indicates that it is not appropriate to include the 3-6 km externality effects in the calculation of overall deworming impacts. First, the 3-6 km externality effect is not statistically significant for either worm infections (with a coefficient estimate of -0.050 and standard error of 0.077 implying a P-value of 0.52, in Table S1 column 3), or for school participation (Table S2, column 3). Second, there is a tremendous loss of statistical precision in the overall effect estimate when 3-6 km externality effects are included in the calculation. For worm infections, the standard error on the overall effect estimate increases by 50% (from 0.061 to 0.091, Table S1, columns 2 and 3) when the 3-6 km externality effect is included. For school participation, the standard error on the overall effect estimate nearly doubles, from 0.017 to 0.032 (Table S2, columns 2 and 3). This doubling of the standard error in the school participation analysis is equivalent to increasing the variance of the estimator roughly four-fold, so the reduction in bias from including the 3-6 km externality effect would have to be very large to justify its inclusion under the criterion of minimizing the mean squared error (MSE). Yet it is unlikely that the 3-6 km externality effect on school participation is substantial given the lack of worm infection externality impacts at 3-6 km.

Some straightforward calculations suggest that the estimator that excludes the 3-6 km externality terms from the calculation of overall deworming impacts on school participation is preferable under the criterion of minimizing MSE. In particular, we show that the increase in MSE due to additional noise from including the 3-6 km term is likely to be more than six times greater than any decrease in the MSE due to reducing bias.To see this, define the estimator that includes the Treatment minus Control effect plus the 0-3 km externality effect as β_1 (this is the estimate presented in the bottom row of Table S2, column 2), and the estimator that also includes the 3-6 km externality effect as β_2 (column 3). An estimate of the variance of β_1 is the square of its standard error, or 0.017^2, and similarly for the variance of β_2 (0.032^2). For simplicity, we conservatively assume that Bias(β_2) = 0, in other words, all deworming externality effects are captured within 6 km. The estimator that excludes the 3-6 km externality terms is preferred under the mean squared criterion – in other words, MSE(β_1) < MSE(β_2) – as long as Bias(β_1)^2 < ($0.032^2 - 0.017^2$) = 0.000735, or equivalently, if Bias(β_1) – Bias(β_2) < $(0.000735)^{1/2}$ = 0.027.

Recall that the direct effect of being in a treatment school is to reduce moderate-heavy worm infections by -0.31 (Aiken et al. 2015's Appendix Table VII-Updated). Note also that they estimate that the direct effect of being in a treatment school is to increase school participation by 0.057 (Aiken et al., 2015, Appendix Table IX-Updated). This suggests that every percentage point reduction in moderate-heavy infection increases school participation by roughly (0.057)/(0.31) = 0.184 percentage points. Aiken et al. (2015)'s Appendix Table VII-Updated also shows that the point estimate of the reduction in worm infections within 3-6 km is only 0.050, which is not statistically significant. This implies that the predicted gain in school participation per 1000 treated pupils within 3-6 km is approximately 0.050 * 0.184 = 0.0092, and the average externality effect (given the average number of treated pupils within 3-6 km) is 0.011. This is the predicted change in bias from including the 3-6 km externalities in the estimation of overall deworming impacts, Bias(β_1) – Bias(β_2). It is immediate that Bias(β_1) – Bias(β_2) = 0.011 < 0.027, and thus that the estimator excluding the 3-6 km externality term is preferred under the MSE criterion.

This means that the predicted decrease in MSE due to reduced bias is 0.011^2 = 0.000115. Recall from above that the increase in MSE from the additional “noise” contributed by including the 3-6 km externality effect is $(0.027)^2$ or 0.000735. Hence the increase in MSE due to the extra “noise” from including the 3-6 km externality term (0.000735) is likely to be approximately 6.4 times larger than the predicted decrease in MSE due to reduced bias (0.000115).

Table S1: Worm infection results from Miguel and Kremer (2004), updated and original

	UPDATED			ORIGINAL		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Indicator	-0.347***	-0.333***	-0.313***	-0.347***	-0.311***	-0.247***
	(0.052)	(0.052)	(0.057)	(0.052)	(0.052)	(0.053)
Treatment pupils w/in 3 km		-0.234**	-0.212**		-0.249***	-0.256***
(per 1000 pupils)		(0.097)	(0.104)		(0.085)	(0.087)
Treatment pupils w/in 3 - 6 km			-0.050			-0.140**
(per 1000 pupils)			(0.077)			(0.060)
Total PSDP 'eligible' students w/in 3 km		0.069*	0.046		0.074**	0.109***
(per 1000 pupils)		(0.037)	(0.036)		(0.033)	(0.040)
Total PSDP 'eligible' students w/in 3-6 km			-0.022			0.133**
(per 1000 pupils)			(0.039)			(0.056)
School average of mock score, 1996	-0.208***	-0.216***	-0.188***	-0.208***	-0.220***	-0.093
	(0.055)	(0.052)	(0.073)	(0.055)	(0.048)	(0.068)
Calculated Effects						
Average 0-3 km externality effect		-0.102**	-0.090**		-0.111***	-0.106***
		(0.043)	(0.044)		(0.038)	(0.037)
Average 3-6 km externality effect			-0.052			-0.096**
			(0.079)			(0.042)
Average overall cross-school externality effect		-0.102**	-0.146		-0.111***	-0.212***
		(0.043)	(0.110)		(0.038)	(0.065)
Overall deworming effect	-0.347***	-0.435***	-0.459***	-0.347***	-0.421***	-0.460***
	(0.057)	(0.061)	(0.091)	(0.057)	(0.055)	(0.055)

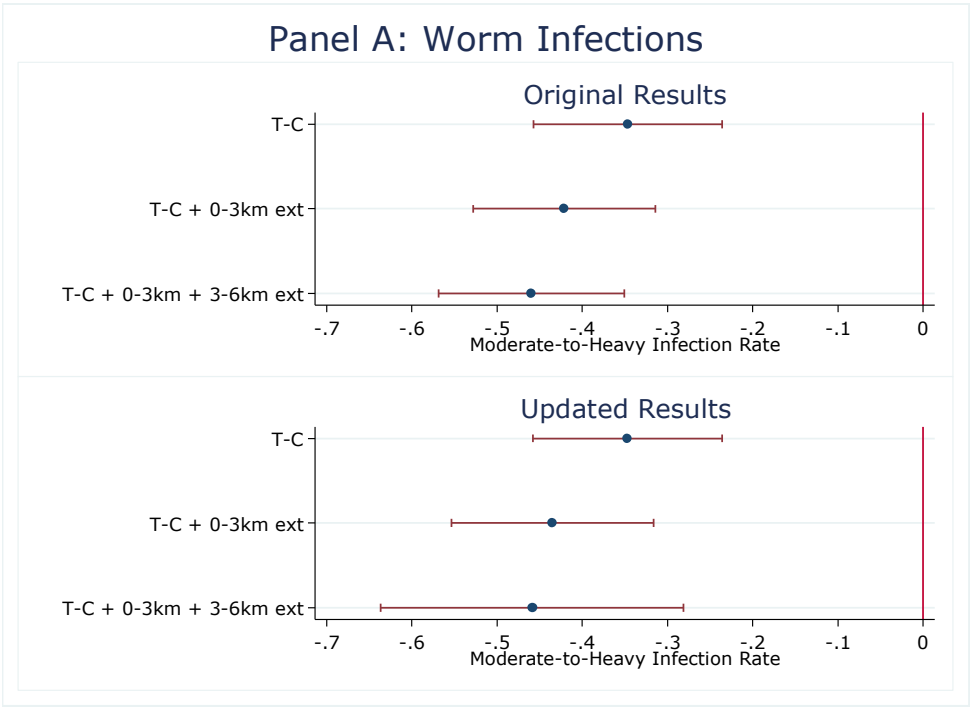
Note: The sample size in columns (1)-(3) is 2,330, and in (4)-(6) is 2,328. The sample includes pupils in grades 3–8, in 1999 Group 1 and Group 2 schools. Results are from probit estimation, where observations are weighted by total school population. The dependent variable is an indicator for moderate-to-heavy infection. Eligible pupils include girls less than 13 years old and all boys. Additional explanatory variables include indicators for 1998 grade and school SAP participation. Robust standard errors are in parentheses, and disturbance terms are clustered within schools. Stars denote statistical significance at 99 (***), 95 (**), and 90 (*) percent confidence.

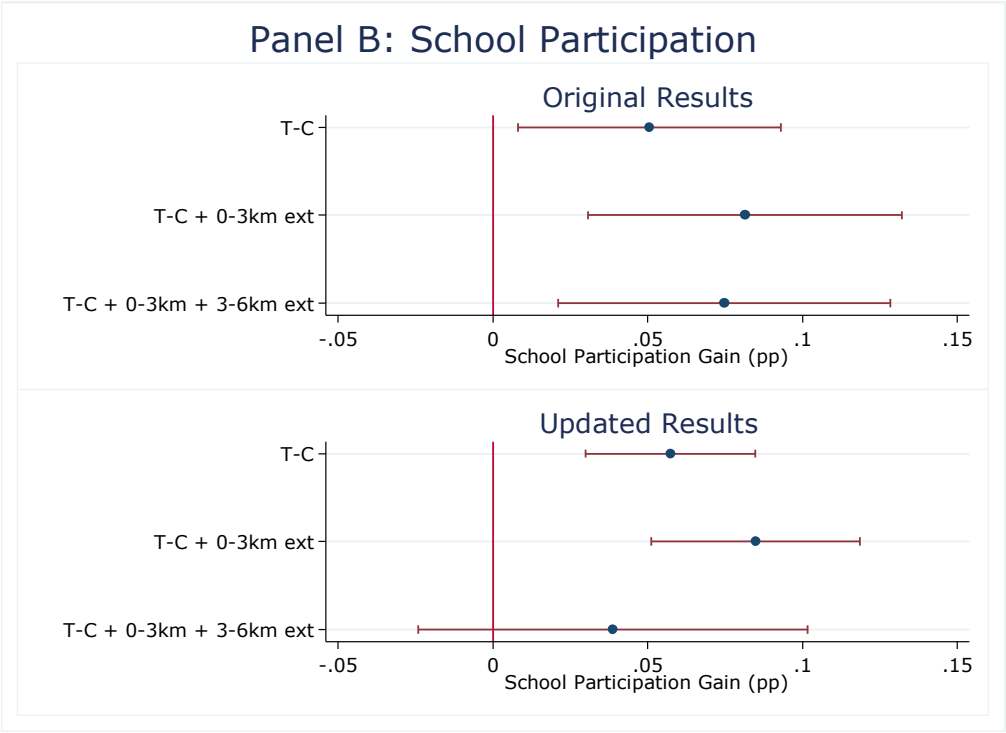
Table S2: School participation results from Miguel and Kremer (2004), updated and original

	UPDATED			ORIGINAL		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Indicator	0.057***	0.058***	0.055***	0.051**	0.054**	0.055**
	(0.014)	(0.014)	(0.014)	(0.022)	(0.023)	(0.023)
Treatment pupils w/in 3 km		0.045**	0.038*		0.046**	0.048**
(per 1000 pupils)		(0.021)	(0.021)		(0.018)	(0.019)
Treatment pupils w/in 3 - 6 km			-0.024			-0.013
(per 1000 pupils)			(0.015)			(0.015)
Total PSDP 'eligible' students w/in 3 km		-0.030**	-0.030**		-0.031***	-0.037***
(per 1000 pupils)		(0.013)	(0.012)		(0.012)	(0.012)
Total PSDP 'eligible' students w/in 3-6 km			0.012			-0.014
(per 1000 pupils)			(0.009)			(0.012)
School average of mock score, 1996	0.071***	0.071***	0.078***	0.063***	0.064***	0.055***
	(0.021)	(0.022)	(0.022)	(0.021)	(0.021)	(0.021)
Calculated Effects						
Average 0-3 km externality effect		0.027**	0.023*		0.028**	0.029**
		(0.013)	(0.013)		(0.011)	(0.012)
Average 3-6 km externality effect			-0.040			-0.009
			(0.024)			(0.011)
Average overall cross-school externality effect		0.027**	-0.017		0.028**	0.020
		(0.013)	(0.030)		(0.011)	(0.013)
Overall deworming effect	0.057***	0.085***	0.039	0.051**	0.081***	0.075***
	(0.014)	(0.017)	(0.032)	(0.022)	(0.026)	(0.027)

Note: The sample size in columns (1)-(3) is 56,496, and in (4)-(6) is 56,487. The dependent variable is average school participation in each year (Year 1: May 1998 - March 1999; Year 2: May 1999 - November 1999). Participation is computed among all pupils enrolled at the start of the 1998 school year; pupils present during an unannounced NGO school visit are considered participants. Additional controls include an indicator for girls < 13 years and all boys; the rate of moderate-heavy infections in geographic zone, by grade (zonal infection rates among grade 3 and 4 pupils are used for pupils initially recorded as drop-outs; rates among grade 5 and 6 pupils are used for grades 5 and 6, and similarly for grades 7 and 8); 1996 school average test score; indicators for participation in the SAP, alone and interacted with an indicator for 1998; indicators for 1998 grade of pupil; and indicators for semester of observation. Robust standard errors are in parentheses, and disturbances are clustered within schools. Stars denote statistical significance at 99 (***), 95 (**), and 90 (*) percent confidence.

Figure S3: Original vs. updated "overall effect", with 95% confidence intervals





Note: Panel A displays the "overall effect" of deworming, as calculated in the bottom panel of Table S1 (for worm infections) and Panel B displays the "overall effect" of deworming from Table S2 (for school participation). See the notes under these tables for details on the regressions.

Even if one makes the far weaker assumption that the overall externality effect on school participation at 3-6 km is simply equal to or smaller than that from 0-3 km, one reaches the same conclusion that MSE decreases when the 3-6 km externality term is excluded. Recall from Table S2, column 2 that the overall 0-3 km externality effect on school participation is also (coincidentally) 0.027. Thus the estimator that excludes the 3-6 km externality effects (β_1) has a smaller mean squared error if the overall externality effect at 3-6 km is smaller than the 0-3 km effect. This is a natural "monotonicity" assumption given the nature of worm transmission and reinfection, which tend to be locally concentrated and should fall at greater distances from a treatment school.

The comparison of columns 2 and 3 in Table S2 further illustrates this point. The total estimated effect incorporating the "naïve" treatment minus control difference plus the 0-3 km effect is 0.085 (s.e. 0.017), significant at 99% confidence. The total estimated effect incorporating externalities out to 6 km has a standard error of 0.032, nearly twice as large as the standard error only considering externalities out to 3 km. Regarding the negative 3-6 km point estimates, there is no obvious epidemiological reason to our knowledge why the 3-6 km effects on school participation would be negative, especially given the large, positive and significant externality effects we estimate both within-schools and within 3 km of treatment schools. We instead believe the negative and very far from statistically significant point estimates on the 3-6 km school density are most likely to be "noisy zeros", as mentioned above. It is worth mentioning again that even in the original Miguel and Kremer (2004) paper the 3-6 km externality effect on school participation was not statistically significant, but this "zero" effect becomes considerably noisier with the updated data.

In fact, once the 3-6 km variable construction is corrected, the "naïve plus 0-3 km" effect is nearly unchanged for worm infections (comparing the column 2 and column 5 results at the bottom of Table S2), and the school participation effect is slightly larger in the updated case with a somewhat smaller standard error than in the original estimation. Both the infection and school participation effects are large in magnitude and statistically significant at over 99% confidence considering externalities out to 3 km (see column 2 of Table S1 and column 2 of Table S2). Thus there remains considerable evidence that deworming led to reductions in worm infections and large improvements in school participation. But the effects beyond 3 km are simply too imprecisely estimated to be usefully employed in the analysis.

As noted above, the externality analysis was not pre-specified in advance of analyzing the data. Readers might be concerned about the possibility of data mining and selective presentation of analytical results, and wonder just how robust the externality results truly are. It is straightforward to show that the positive deworming externality results across nearby schools are robust to using different distances and specifications; it is not the case that the 3 km distance was "cherry-picked" from among the set of possible distances over which to estimate externality effects. For worm infections, the externality effects are statistically significant at 95% confidence at distances of both 0-3 and 0-4 km (Table S3, columns 3 and 4) and significant at 90% confidence at distances of 0-5 and 0-6 km (columns 5 and 6). Note that as one gets further away, one would expect the spillovers from any given school to be smaller, but the "overall" effect from multiplying the average spillover times the number of schools to stay constant or grow. The magnitude of the "overall" cross-school externality benefits become larger at increasing distances, although they are estimated with considerably less precision, especially beyond 4

km (Figure S1, Panel A). (Externality estimates are also imprecisely estimated for schools within 1 km from the reference school, since very few schools are located this close together.)

The same pattern is evident for school participation externalities. The impact of cross-school externalities is positive and statistically significant at 95% confidence at distances of 0-2, 0-3 and 0-4 km (Table S4, columns 2-4), and the magnitude is largest for the 4 km radius. Once again externality effects increase at larger distances, in this case up to 4 km, after which confidence intervals become considerably wider. In all of these regression specifications, the naïve effect on treatment schools is nearly unchanged, ranging between gains of 0.057 and 0.063 and is significant at over 99% confidence.

There is a simple bottom line on deworming externalities. As Aiken et al. (2015) show in their appendix tables, there are large, positive and significant deworming externalities for worm infections and school participation within schools (i.e., for untreated pupils in the treatment schools). Externalities are large, positive and significant for both worm infections and school participation for schools up to 3 km and 4 km of treatment schools. This is all that is needed to show that cross-school externalities "exist": they need not hold at *all* distances in order to exist, they simply need to hold at *some* distances. Given the epidemiology of worm infections, it is reasonable that they would be more pronounced closer to treatment schools, as comes through in the updated data. In fact, the 3-6 km worm infection externalities which were statistically significant in the original paper when analysis was restricted to a subset of nearby schools no longer come through once the variable construction errors are corrected. The externality impacts on school participation at a distance of 3-6 km from treatment schools were not statistically significant in the original Miguel and Kremer (2004) analysis, and they remain not significant in the updated analysis.

There is an important alternative approach to estimating the impact of worms on school participation presented in Miguel and Kremer (2004), namely an instrumental variables (IV) approach. The IV method is attractive because it simultaneously exploits multiple sources of experimental variation, including both school treatment status and proximity to treatment schools (in both the 0-3 km and 3-6 km ranges), to identify a single impact of worm infections on school participation. This approach is not discussed in the Aiken et al. (2015) report, although the updated results are presented in their Appendix Table IX-Updated (column 7), where they show that the estimated impact of a moderate-heavy worm infection on school participation is large and statistically significant (coefficient estimate -0.195, s.e. 0.095, P-value < 0.05). (Note that the sign here is negative since it is parameterized in terms of the effect of a moderate-heavy worm infection on school participation.) This result is robust to other specifications: we show in Appendix A.6 Table IX (column 7) that the effect is even larger in magnitude at -0.282 (s.e. 0.111, P-value < 0.05), when only the school treatment status and externality effects within 0-3 km are used as instrumental variables. Note that these effects are statistically significant despite the use of a much smaller sample size, since this analysis relies on the year 1 individual worm infection data. This result provides a further piece of evidence that worm infections have a large and statistically significant impact on school participation in our setting.

Given these findings, we disagree with the claim in Aiken et al. (2015) that there is little evidence for cross-school externality impacts on school participation, and no overall effect of deworming on school participation.

For those interested in policy implications, the estimated overall average effect of deworming on worm infections using the finalized data is a reduction of 43.5 percentage points (s.e. 6.1, P-value <

0.01), shown in column 2 of Table S1. The estimated overall average effect of deworming on school participation is a gain of 8.5 percentage points (s.e. 1.7, P-value < 0.01), shown in column 2 of Table S2. We show in Figure S2 that the "cost-effectiveness" of deworming in terms of boosting school participation is nearly unchanged, relative to the original paper, using the updated data and considering the direct effects and the externalities up to 3 km, with 34.3 additional years of school participation per $100 of spending on deworming with the updated data (versus 29.1 additional years per $100 in the original analysis). Focusing on the most conservative treatment effect estimate, the "naïve" T – C difference, also implies that deworming is a highly cost-effective approach to reducing school absenteeism in this setting, with 17.8 additional years of school participation per $100 of deworming spending, placing it among the most cost-effective interventions yet evaluated in education studies, as shown in the figure.

A.2.4 Presentation of Non-Worm Infection Health Results

The key health outcome measure emphasized in Miguel and Kremer (2004) is the helminth infection rate. This is the most natural health outcome to focus on given the intervention. As shown above in Table S1, Miguel and Kremer (2004) find large and highly statistically significant decreases in worm infections due to deworming, and this result is unchanged upon re-analysis with the finalized data. As shown in Table A.5-V (which reproduces a table from Miguel and Kremer (2004) using the updated data), there are substantial decreases in worm infection rates for "any moderate-heavy infection", for hookworm infection, and for roundworm infection after one year of treatment (in fact, no concerns were noted in reporting of these results). Although point estimates suggest a substantial decline in schistosomiasis infection, the treatment effect is no longer statistically significant at traditional confidence levels; recall that treatment for schistosomiasis was only provided in the subset of schools with sufficient prevalence of the disease, typically in schools that were close to Lake Victoria, and thus overall infection levels and treatment effects are likely to be more evident in those schools. Table S1 and Table S3 above also present evidence that there are large, positive epidemiological externalities associated with deworming across schools.

Beyond worm infection, Miguel and Kremer (2004) present six other health outcomes, including self-reported health in the past week, self-reports of being "sick often", height- and weight-for-age Z-scores, hemoglobin concentration, and proportion anemic.

The height result was reported as a modest improvement in the original paper, and this result is unchanged in re-analysis. The weight-for-age Z-score and hemoglobin concentration outcomes were not found to be statistically significant either in the original study or in re-analysis.

Table S3. Worm infection regressions, with externalities at various radii

	(1)	(2)	(3)	(4)	(5)	(6)
	w/in 1 km	w/in 2 km	w/in 3 km	w/in 4 km	w/in 5 km	w/in 6 km
Treatment indicator	-0.325***	-0.354***	-0.333***	-0.296***	-0.283***	-0.306***
	(0.047)	(0.051)	(0.052)	(0.057)	(0.064)	(0.056)
Treatment pupils within XX km	0.581	-0.236	-0.234**	-0.201***	-0.124*	-0.112*
(per 1000 pupils)	(0.535)	(0.180)	(0.097)	(0.077)	(0.072)	(0.063)
Total pupils within XX km	-0.248	0.110	0.069*	0.044	-0.011	-0.001
(per 1000 pupils)	(0.357)	(0.085)	(0.037)	(0.036)	(0.030)	(0.032)
Calculated Effects						
Average XX km externality effect	0.013	-0.035	-0.102**	-0.152***	-0.150*	-0.166*
	(0.012)	(0.027)	(0.043)	(0.059)	(0.087)	(0.094)
Overall effect	-0.311***	-0.389***	-0.435***	-0.448***	-0.432***	-0.472***
	(0.052)	(0.062)	(0.061)	(0.062)	(0.068)	(0.085)

Note: This table uses the fully corrected, updated data from Miguel and Kremer (2004). Regressions are as specified in Table S1, with the exception that we allow the radius at which externalities are considered to vary across the columns as indicated.

Table S4. School participation regressions, with externalities at various radii

	(1)	(2)	(3)	(4)	(5)	(6)
	w/in 1 km	w/in 2 km	w/in 3 km	w/in 4 km	w/in 5 km	w/in 6 km
Treatment indicator	0.061***	0.063***	0.058***	0.059***	0.058***	0.057***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Treatment pupils within XX km	0.179	0.093**	0.045**	0.034**	0.009	-0.002
(per 1000 pupils)	(0.131)	(0.037)	(0.021)	(0.014)	(0.013)	(0.013)
Total pupils within XX km	-0.117	-0.064***	-0.030**	-0.022**	-0.009	-0.002
(per 1000 pupils)	(0.109)	(0.025)	(0.013)	(0.009)	(0.009)	(0.008)
Calculated Effects						
Average XX km externality effect	0.004	0.019**	0.027**	0.038**	0.015	-0.004
	(0.003)	(0.007)	(0.013)	(0.016)	(0.022)	(0.029)
Overall effect	0.065***	0.081***	0.085***	0.097***	0.073***	0.053
	(0.014)	(0.015)	(0.017)	(0.020)	(0.024)	(0.033)

Note: This table uses the fully corrected, updated data from Miguel and Kremer (2004). Regressions are as specified in Table S2, with the exception that we allow the radius at which externalities are considered to vary across the columns as indicated.

Aiken *et al.* (2015) note an error in the reporting of the "proportion anemic" outcome, which is not statistically significant in the updated analysis. We thank the authors for updating the scientific record on this point. Note that the coefficient estimate on anemia in the original Miguel and Kremer (2004) paper was reported correctly (a reduction of 2 percentage points in anemia), so the magnitude of the effect and the standard error are unchanged, but the statistical significance level was misreported. We believe that this was due to a calculation of the t-statistic using the rounded coefficients. While anemia is interesting to study from a medical perspective, Miguel and Kremer (2004) paper note that anemia is not likely to be a main channel of impact in the setting examined because only 4% of the population was anemic. Correspondingly, this is not one of the major findings of the original paper. As we write on p. 174 (Miguel and Kremer, 2004):

> *"Severe anemia is relatively rare in Busia: fewer than 4 percent of pupils in Group 2 schools (comparison schools in 1998) fell below the Kenya Ministry of Health anemia threshold of 100 g/L in early 1999 before deworming treatment. This is low relative to many other areas in Africa, of which many have substantial helminth problems: a recent survey of studies of anemia among school children in less developed countries (Hall and Partnership for Child Development (2000)) indicates that there is considerably less anemia in Busia than in samples from Ghana, Malawi, Mali, Mozambique, and Tanzania"*

Aiken *et al.* (2015) downplay the importance of measures of self-reported health in their re-analysis. However, self-reported health measures are widely used in studies set in less developed countries, and other research has found that self-reported health often predicts later morbidity and mortality even when other known health risk factors are accounted for (Idler and Benyamini, 1997; Haddock *et al.*, 2006; Brook *et al.*, 1984). In Miguel and Kremer (2004), statistically significant reductions of meaningful magnitude are estimated in both of the self-reported measures ("sick in past week" and "sick often") and both continue to hold in the Aiken et al. (2015) re-analysis.

A.3. Additional points raised in Aiken et al. (2015)

This section provides detailed responses to other points raised in Aiken et al. (2015). For legibility, we have included the original text from Aiken et al. (2015) in ***bold italics***, followed by our response. Square brackets denote text added to the quotes for clarity.

Abstract, Results: "For school attendance, re-analysis showed benefits similar to those originally found in intervention schools for both children who did and did not receive deworming drugs. However, after correction of coding errors, there was little evidence of an indirect effect on school attendance amongst children in schools close to intervention schools. Combining these effects gave a total increase in attendance of 3.9% amongst treated children, which was no longer statistically significant."

We address this claim extensively in Section A.2 above. The results after correction of variable construction make clear that it is impossible to precisely estimate overall deworming externalities on school participation out to a distance of 6 km, as this results in very wide and largely uninformative

confidence intervals, although it is worth noting that the point estimate on the 0-2 km, 0-3 km, and 0-4 km externality terms remains negative, large and statistically significant at 95% confidence (in Table S4, columns 2-4). When we instead explore overall externality effects only up to distances which are precisely estimated, we find large, positive and statistically significant between-school externality impacts (see Tables S2 and S4 above, as well as Figure S2, Panel B). The estimated overall effect of deworming on school participation, considering externality impacts out to a distance of 3 km, is 8.5% percentage points (s.e. 1.7, P-value < 0.01), shown in column 2 of Table S2.

Abstract, Conclusions: "Re-applying analytic approaches originally used but correcting various errors, we found little evidence for some previously-reported indirect effects of a deworming intervention."
We believe the authors' own tables dispute this finding. Aiken et al. (2015) Appendix Table VII-Updated clearly shows worm infection externality benefits to untreated individuals in treatment schools (column 2, p-value<0.05) and externality benefits to individuals living within 3 km of treatment schools (column 1, p-value<0.05). Aiken et al. (2015) Appendix Table IX-Updated clearly shows school participation externality benefits to untreated individuals in treatment schools (column 5, p-value<0.01). Furthermore, as we detail extensively in Section A.2 above, there are substantial and statistically significant (p-value<0.05) school participation externality benefits to individuals living within 4 km of treatment schools (Table S4 above).

Methods of original study, Intervention: "The original paper also describes other school-based interventions occurring concurrently in 27 of 75 schools."

Local schools participating in the intensive CSP/SHP program were dropped from the sample of eligible schools, while 27 primary schools with less intensive NGO programs were retained in the sample. These 27 schools were receiving assistance in the form of either free classroom textbooks, grants for school committees, or teacher training and bonuses. It is worth emphasizing that randomized evaluations of these various interventions did not find statistically significant average project impacts across a wide range of educational outcomes (Glewwe, Kremer, and Moulin, 2009). The schools that benefited from these previous programs were found in all eight geographic zones in the study area. The results in Miguel and Kremer (2004) are robust to including controls for inclusion in these other NGO programs.

Methods of original study, Categories of effect: "As worm infections are transmitted by excretion of worm eggs in faeces, and as faecal contamination of the environment was known to be common, it was hypothesized that there would be a local reduction of transmission of worm infection around the intervention schools. Effects were calculated based on composites of results at 0-3 km and 3-6 km. It is not clear to us from the original paper how these intervals were decided upon."

As we describe in detail in the Section A.2 above (including a direct quote from the 2004 paper which describes why we estimated effects out to 6 km in that paper), we did not have *a priori* assumptions regarding the precise radius over which there would be epidemiological externalities related to worm infections, nor was there any existing quantitative research to our knowledge that would guide this choice. In fact, we did not even plan to study cross-school externality effects at the start of the study. Thus the focus on a particular radius around each school was guided by analysis of the

data. The data suggest that epidemiological and school participation externalities extend out to at least about 4 km (see Tables S3 and S4 above).

Methods of original study, Categories of effect: "In this pure replication, we did not evaluate the appropriateness of separating effects into the different categories described above. Instead, we reproduced the analytic steps to re-determine the results as originally calculated."

As we detail in Section A.2 above, we disagree with Aiken et al. (2015)'s calculations of the average externality effects and overall effects of deworming.

Results, Table VII: "Having corrected these errors, re-analysis found no statistically significant indirect-between-school effect on the worm infection outcome, according to the analysis methods originally used. However, amongst variables used to construct this effect, a parameter describing the effect of Group 1 living within 0-3km 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 3-6km 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)."

As Aiken et al. (2015) point out in the above quote, worm infection externalities out to a radius of 3 km are substantial and statistically significant at 95% confidence. This is all that is needed to show that cross-school externalities "exist": they need not hold at *all* distances in order to exist, they simply need to hold at *some* distances. As Aiken et al note, there is also strong evidence (using the updated data) for within-school externalities on both worm infections and school participation.

Results, Table VIII: "Both Table VIII and IX included weighted regression analyses: these were inaccurately described as being weighted by number of pupils, whereas these were actually weighed by numbers of pupil observations."

The large, positive, and statistically significant impacts of deworming on school participation hold whether weighted by the number of pupil-observations or the number of pupils. In fact, both of these approaches are standard in the related research literature and have their merits. Pupil weighting is attractive since it generates the population average, while pupil-observation weights increase power and precision. The school participation results are robust to either approach (as we discuss in more detail in Appendix B).

Results, Table IX: "The indirect-between-school effect 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 zero, hence also non-significant."

We address this claim extensively in Section A.2 above. The results after correction of variable construction make clear that it is impossible to precisely estimate overall deworming externalities on school participation out to a distance of 6 km, as this results in very wide and largely uninformative confidence intervals, although it is worth noting that the point estimate on the 0-2 km, 0-3 km, and 0-4 km externality terms remains negative, large and statistically significant at 95% confidence (Table S4,

columns 2-4, above). When we instead explore overall externality effects only up to distances which are precisely estimated, we find large, positive and statistically significant between-school externality impacts (see Tables S2 and S4 above, as well as Figure S2, Panel B). Using the updated data and exploring externalities out to 3 km, we calculate (in Table S2) a cross-school deworming impact of 2.7 percentage points (s.e. 1.3, p-value<0.05) and an overall deworming impact of 8.5 percentage points (s.e. 1.7, p-value<0.01).

Results, Presentation of missing data: ***"Throughout the original paper, there is limited description of the extent of missing data, especially for baseline parameters presented in Table I. In this table, there is, in fact, a large amount of missing data for year of birth – this information is missing for 17% of children in Group 1, 19% of children in Group 2 and 31% of children in Group 3. The extent of these missing data is not described in the table or the accompanying text."***

The year of birth data was much more likely to be missing for children in grades 0, 1, and 2. It is not unexpected for many children in this young age group in the study setting of rural western Kenya to be unaware of their exact birth year or birthdate.

Discussion: "Our most important finding was that after correction of coding errors in the original authors' analysis files, there was little evidence for previously described 'positive externalities' (or indirect effects) from the deworming intervention on school attendance in untreated schools... We found beneficial effects on school attendance similar to or greater than those originally reported for the direct, indirect-within-school and "naïve" effects…In corrected re-analysis, the indirect-between-school effect on school attendance had shifted in direction and was less precisely estimated – there was now little evidence for an effect of this kind in the format of analysis originally employed. We have not re-examined for evidence of indirect-between-school effect at a distance other than that used in original paper (up to 6 km from schools) as this would deviate from our stated pre-analytic plan. We do note that some parameters suggest effects may be present at distances of up to 3 km. It remains unclear how the distance intervals used for these spatial effects in the original paper were decided upon."

We discuss this in detail in Section A.2, and show large and statistically significant (p-value <0.05) positive externalities on school attendance among schools within 3 km of treatment schools. In cases like this one, where an estimator can be shown to be less attractive under the standard criterion of minimizing mean squared error, a deviation from the pre-analysis plan would be well-justified. In the second part of this replication endeavor – the "statistical replication" – Davey et al. (2015) choose to make multiple deviations from their pre-analysis plan. We further show in Table S4 that these externality impacts are statistically significant (p-value<0.05) at distances of 0-2 km and 0-4 km.

As we describe in detail in the Section A.2 above (including a direct quote from the 2004 paper which describes why we estimated effects out to 6 km in that paper), we did not have *a priori* assumptions regarding the precise radius over which there would be epidemiological externalities related to worm infections, nor was there any existing quantitative research to our knowledge that would guide this choice. In fact, we did not even plan to study cross-school externality effects at the start of the study. Thus the focus on a particular radius around each school was guided by analysis of the

data. The data suggest that epidemiological and school participation externalities extend out to at least about 4 km (see Tables S3 and S4 above).

Discussion: "In contrast to the original study, we found limited evidence of non-worm-related health benefits as the prevalence of anaemia was not significantly affected by the intervention."

This conclusion is as much due to a focus by the replication authors on particular health measures as to a change in the results of Miguel and Kremer (2004). The only variable with significance changes was the "proportion anemic", where the coefficient estimate is unchanged but where the P-value was reported as being less than 0.05 when it is actually 0.19. The estimated effects on the level of Hb, the self-reported health outcomes, and the HAZ (height) result are unchanged in the replication results presented in Aiken et al. (2015).

Discussion: "The "total effect" on school attendance resulting from the intervention described by the original authors, a combination of the naive and indirect-between-school effects, was more modest and less precisely estimated than previously reported and was also not statistically significant. This counter-intuitive finding – strong evidence for a naïve effect but no evidence of a total effect – derives from the additive logic used by the original authors to calculate the total effect result and the reversal in direction of the indirect between- school effect."

We disagree with Aiken et al. (2015)'s characterization of this finding (i.e., strong evidence for a naïve effect but no evidence for a total effect) as counter-intuitive, given an understanding of how the total effect was calculated by the replication authors. Including the 3-6 km externality terms leads standard errors to double in size, making the resulting total effect estimator largely uninformative with the updated data. As we show in Table S2, the total effect is large, meaningful, and highly statistically significant (p-value<0.01) once unnecessary noise is removed from this calculation.

Appendix, Page 4, Table I-Updated results:

We disagree with Aiken et al. (2015) regarding the means reported for "Grade progression" in the columns for Group 1, Group 2, and Group 3 in this table. We have provided updated values for these figures in Appendix A.5 of this document. We find that no substantive results are changed.

Appendix, Page 5: "It is not clear how this random selection [for stool sample testing] was performed."

The random sampling for stool sample testing was conducted nearly 18 years ago, and unfortunately we are unable to locate the statistical code that produced it. We do know from our recollection of the field work that this was meant to be a representative sample of the students in each grade in the surveyed schools.

Appendix, Page 5: "Thresholds used for moderate/heavy infection with hookworm, whipworm and schistosomiasis are different from those suggested by the World Health Organization (1). The supporting reference provided makes a case for and uses locally-defined thresholds for heavy infection (2), but does not mention moderate/heavy infection thresholds. Therefore, a more appropriate description of how these thresholds for moderate/heavy infection were selected would be "personal

communication with Dr Simon Brooker and Professor Donald Bundy", according to the authors' own report of how this was actually done."

We thank the replication authors for clarifying this point. The Brooker *et al.* (2000) article uses alternative thresholds (that do not correspond exactly to the WHO standard) for defining heavy infections in the study area; namely, infection levels at the 90th percentile level. These thresholds are: Hookworms 1,250+ epg; A.lumbricoides 20,000+ epg; T.trichiura 1,000+ epg; S.mansoni 500+ epg, somewhat lower than the WHO standard for heavy infections. Both the heavy infection prevalence used in the Brooker *et al.* (2000) paper and the moderate-to-heavy infection levels used in Miguel and Kremer (2004) were developed in personal communication with Simon Brooker for the specific context of Busia District, Kenya during 1998 and 1999. As these were designed in close consultation with Dr. Brooker and Dr. Donald Bundy, both global experts on intestinal worms, we felt that deviating from the WHO standard was an appropriate adjustment for the setting of the study.

Appendix, Page 11: "A second coding error was present that miscalculated local density figures for three of the schools – these were School numbers 108 (in Group 3), 109 (in Group 2), and 115 (in Group 1)… This code was problematic as it miscalculated the local population densities for these three schools by omitting some other schools from the calculations."

We would like to take this opportunity to clarify the nature of this coding error. For school 108 (a Group 3 school), the coding error resulted in ignoring all Group 1 schools in calculation of the local density terms – however, we note that there were no Group 1 schools located within 6 km of school 108, so the coding error literally had no effect on the data in this case. For school 109 (a Group 2 school), all Group 2 schools were ignored in calculation of the local density terms. There was only 1 Group 2 school located with 6 km of school 109 (and no Group 2 schools located within 3 km of school 109), so this error affected the 3-6 km density term only (by missing one school), not the 0-3 km term for this school. Finally, for school 115 (a Group 1 school), all Group 3 schools were ignored in calculation of the local density terms, and there were seven such schools within 6 km of school 115. Hence, only two schools were affected by this coding error.

Appendix Page 13, Table VII-Updated results:

We disagree with Aiken et al. (2015) regarding the results reported in column (6) of this table. In particular, the result on "Group 1 pupils within 3 km (per 1000 pupils)" should read -0.08 (s.e. 0.07). This is not statistically significant at traditional levels of confidence. We have provided an updated version of this table in Appendix A5.

Appendix, Page 21, Table X-Updated results:

We disagree with Aiken et al. (2015) regarding the results reported in column (2) of this table of their report. In particular, the standard error on "Second year as treatment school (T2)" is 0.079. We have provided an updated version of this table in Appendix A5. No substantive results are changed.

A.4 Additional References

Aiken AM, Davey C, Hayes RJ, Hargreaves J. Deworming schoolchildren in Kenya - Replication plan. International Institute Impact Evaluation (3ie) website: 2013.

Aiken, A, Davey, C, Hargreaves, J, and Hayes, R. (2015). "Re-analysis of health and educational impacts of a school-based deworming program in western Kenya: a pure replication", *International Journal of Epidemiology*.

Aiken, A, Davey, C, Hayes, R and Hargreaves, J. (2014). "Re-analysis of health and educational impacts of a school-based deworming program in western Kenya: a pure replication", 3ie Replication Paper 3, part 1. Washington, DC: International Initiative for Impact Evaluation (3ie).

Alderman, H., J. Konde-Lule, I. Sebuliba, D. Bundy, A. Hall. (2006). "Increased weight gain in preschool children due to mass albendazole treatment given during 'Child Health Days' in Uganda: A cluster randomized controlled trial", *British Medical Journal*, 333, 122-6.

Alderman, Harold. (2007). "Improving nutrition through community growth promotion: Longitudinal study of nutrition and early child development program in Uganda", *World Development*, 35(8), 1376-1389.

Awasthi, Shally, et al. (2013). "Population deworming every 6 months with albendazole in 1 million pre-school children in north India: DEVTA, a cluster-randomized trial", *Lancet*, 381(9876): 1478-1486.

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel. (2014). "Worms at Work: Long-run impacts of a child health investment", unpublished working paper, University of California, Berkeley.

Brook, R.H., et al. (1984). *The Effect of Coinsurance on the Health of Adults: Results from the RAND Health Insurance Experiment*. RAND: Santa Monica, CA.

Brooker, S, Miguel, EA, Moulin, S, Luoba, AI, Bundy DA & Kremer, M, 2000. Epidemiology of single and multiple species of helminth infections among school children in Busia District, Kenya. *East African Medical Journal*, 77, 157-61.

Croke, Kevin. (2014). "The long run effects of early childhood deworming on literacy and numeracy: Evidence from Uganda", unpublished working paper, Harvard University.

Haddock, C.K., et al. (2006). "The validity of self-rated health as a measure of health status among young military personnel: evidence from a cross-sectional survey", *Health and Quality of Life Outcomes*, 4(57).

Hicks, JH, Kremer, M and Miguel, E (2014). "Estimating deworming school participation impacts and externalities in Kenya: A Comment on Aiken et al. (2014)". Original author response to 3ie Replication Paper 3, part 1. Washington, DC: International Initiative for Impact Evaluation (3ie).

Idler, Ellen L., and Yael Benyamini. (1997). "Self-rated health and mortality: A review of twenty-seven community studies", *Journal of Health and Social Behavior*, 38(1).

Miguel, Edward and Michel Kremer (2001). "Worms: Education and Health Externalities in Kenya", National Bureau of Economic Research Working Paper #8481.

Miguel, Edward and Michael Kremer (2004). "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica*, 72(1), 159-217.

Miguel, Edward and Michael Kremer (2014). “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, Guide to Replication of Miguel and Kremer (2004).” CEGA Working Paper #39.

Miguel, Edward, Michael Kremer, Joan Hamory Hicks and Carolyne Nekesa (2014a). “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, Data User’s Guide.” CEGA Working Paper #40.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, M. Van der Laan. (2014b). “Promoting Transparency in Social Science Research”, *Science*, 10.1126/science.1245317.

Ozier, Owen. (2014). “Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming”, World Bank Policy Research Working Paper #7052.

Section A.5: Updated tables for Miguel and Kremer (2004)

This appendix includes all tables in Miguel and Kremer (2004), updated to use the "final" versions of all datasets and corrected of all rounding, typographical and coding errors.

Table A.5-I: 1998 Average pupil and school characteristics, pre-treatment[†]

	Group 1 (25 schs)	Group 2 (25 schs)	Group 3 (25 schs)	G1-G3	G2-G3
Panel A: Pre-school to Grade 8					
Male	0.53	0.51	0.52	0.01	-0.01
				(0.02)	(0.02)
Proportion girls < 13 years, and all boys	0.89	0.89	0.88	0.00	0.01
				(0.01)	(0.01)
Grade progression (= Grade – (Age – 6))	-2.0	-1.8	-2.0	-0.0	0.1
				(0.1)	(0.1)
Year of birth	1986.2	1986.5	1985.8	0.4**	0.8***
				(0.2)	(0.2)
Panel B: Grades 3 to 8					
Attendance recorded in school registers (during the 4 weeks prior to the pupil survey)	0.973	0.963	0.969	0.003	-0.006
				(0.004)	(0.004)
Access to latrine at home	0.82	0.81	0.82	0.00	-0.01
				(0.03)	(0.03)
Have livestock (cows, goats, pigs, sheep) at home	0.66	0.67	0.66	-0.00	0.01
				(0.03)	(0.03)
Weight-for-age Z-score (low scores denote undernutrition)	-1.39	-1.40	-1.44	0.05	0.04
				(0.05)	(0.05)
Blood in stool (self-reported)	0.26	0.22	0.19	0.07**	0.03
				(0.03)	(0.03)
Sick often (self-reported)	0.10	0.10	0.08	0.02	0.02*
				(0.01)	(0.01)
Malaria/fever in past week (self-reported)	0.37	0.38	0.40	-0.03	-0.02
				(0.03)	(0.03)
Clean (observed by field workers)	0.60	0.66	0.67	-0.07**	-0.01
				(0.03)	(0.03)
Panel C: School characteristics					
District exam score 1996, grades 5-8[‡]	-0.10	0.09	0.01	-0.11	0.08
				(0.12)	(0.12)
Distance to Lake Victoria	10.0	9.9	9.5	0.6	0.5
				(1.9)	(1.9)
Pupil population	392.7	403.8	375.9	16.8	27.9
				(57.6)	(57.6)
School latrines per pupil	0.007	0.006	0.007	0.001	-0.000
				(0.001)	(0.001)
Proportion moderate-heavy infections in zone	0.37	0.37	0.36	0.01	0.01
				(0.03)	(0.03)
Group 1 pupils within 3 km[††]	430.4	433.2	344.5	85.9	88.7
				(116.2)	(116.2)
Group 1 pupils within 3-6 km	1157.6	1043.0	1297.3	-139.7	-254.4
				(199.3)	(199.3)
Total primary school pupils within 3 km	1272.7	1369.1	1151.9	120.8	217.2
				(208.1)	(208.1)
Total primary school pupils within 3-6 km	3431.3	3259.8	3502.1	-70.8	-242.3
				(366.0)	(366.0)

[†]School averages weighted by pupil population. Standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. Data from the 1998 ICS Pupil Namelist, 1998 Pupil Questionnaire and 1998 School Questionnaire. [‡]1996 District exam scores have been normalized to be in units of individual level standard deviations, and so are comparable in units to the 1998 and 1999 ICS test scores (under the assumption

that the decomposition of test score variance within and between schools was the same in 1996, 1998, and 1999).
[††] This includes girls less than 13 years old, and all boys (those eligible for deworming in treatment schools).

Table A.5-II: January 1998 helminth infections, pre-treatment, Group 1 schools[†]

	Prevalence of infection	Prevalence of moderate-heavy infection	Avg worm load, in eggs/gram (s.e.)
Hookworm	0.77	0.15	426 (1055)
Roundworm	0.42	0.16	2337 (5156)
Schistosomiasis, all schools	0.22	0.07	91 (413)
Schistosomiasis, schools < 5km from Lake Victoria	0.80	0.39	487 (879)
Whipworm	0.55	0.10	161 (470)
At least one infection	0.92	0.37	-
Born since 1985	0.93	0.40	-
Born before 1985	0.91	0.34	-
Female	0.91	0.34	-
Male	0.93	0.38	-
At least two infections	0.65	0.10	-
At least three infections	0.34	0.01	-

The data were collected in January to March 1998 by the Kenya Ministry of Health, Division of Vector Borne Diseases (DVBD). The moderate infection thresholds for the various intestinal helminths are: 250 epg for *S. mansoni*, and 5,000 epg for Roundworm, both the WHO standard, and 750 epg for Hookworm and 400 epg for Whipworm, both somewhat lower than the WHO standard. Refer to Brooker, et al. (2000) for a discussion of this parasitological survey and the infection cut-offs. All cases of schistosomiasis are *S. mansoni*. [†]These are averages of individual-level data, as presented in Brooker, et al. (2000); correcting for the oversampling of the (numerically smaller) upper grades does not substantially change the results. Standard errors in parentheses. Sample size: 1894 pupils. Fifteen pupils per standard in grades 3 to 8 for Group 1 schools were randomly sampled. The bottom two rows of the column "Prevalence of moderate-heavy infection" should be interpreted as the proportion with at least two or at least three moderate-to-heavy helminth infections, respectively.

Table A.5-III: Proportion of pupils receiving deworming treatment in PSDP[†]

	Group 1		Group 2		Group 3	
	Girls < 13 years, and all boys	Girls ≥ 13 years	Girls < 13 years, and all boys	Girls ≥ 13 years	Girls < 13 years, and all boys	Girls ≥ 13 years
	Treatment		*Comparison*		*Comparison*	
Any medical treatment in 1998	0.77	0.20	0	0	0	0
(For grades 1-8 in early 1998)						
Round 1 (March-April 1998), Albendazole	0.68	0.11	0	0	0	0
Round 1 (March-April 1998), Praziquantel[‡]	0.64	0.34	0	0	0	0
Round 2 (Oct.-Nov. 1998), Albendazole	0.56	0.07	0	0	0	0
	Treatment		*Treatment*		*Comparison*	
Any medical treatment in 1999	0.58	0.07	0.54	0.09	0.01	0
(For grades 1-7 in early 1998)						
Round 1 (March-June 1999), Albendazole	0.44	0.06	0.35	0.05	0.01	0
Round 1 (March-June 1999), Praziquantel[‡]	0.47	0.06	0.38	0.06	0.00	0
Round 2 (Oct.-Nov. 1999), Albendazole	0.52	0.06	0.50	0.07	0.01	0
Any medical treatment in 1999	0.73	0.10	0.71	0.14	0.02	0
(For grades 1-7 in early 1998),						
among pupils enrolled in 1999						
Round 1 (March-June 1999), Albendazole	0.55	0.08	0.46	0.08	0.01	0
Round 1 (March-June 1999), Praziquantel[‡]	0.54	0.08	0.46	0.07	0.00	0
Round 2 (Oct.-Nov. 1999), Albendazole	0.65	0.09	0.66	0.11	0.01	0

[†]Data for grades 1-8. Since month of birth information is missing for most pupils, precise assignment of treatment eligibility status for girls born during the "threshold" year is often impossible; all girls who turn 13 during a given year are counted as 12 year olds (eligible for deworming treatment) throughout for consistency. [‡]Praziquantel figures in Table 3 refer only to children in schools meeting the schistosomiasis treament threshold (30 percent prevalence) in that year.

Table A.5-IV - Proportion of pupil transfers across schools

	1998 transfer to a			1999 transfer to a		
School in early 1998 (pre-treatment)	Group 1 School	Group 2 School	Group 3 school	Group 1 school	Group 2 school	Group 3 school
Group 1	0.005	0.007	0.007	0.032	0.026	0.027
Group 2	0.006	0.007	0.008	0.026	0.033	0.027
Group 3	0.010	0.010	0.006	0.022	0.036	0.022
Total transfers	0.020	0.024	0.020	0.080	0.095	0.076

Table A.5-V: January to March 1999, Health and Health Behavior Differences Between Group 1 (1998 Treatment) and Group 2 (1998 Comparison) Schools [†]

	Group 1	Group 2°	Group 1 – Group 2°
Panel A: Helminth Infection Rates			
Any moderate-heavy infection, January – March 1998	0.38	-	-
Any moderate-heavy infection, 1999	0.27	0.52	-0.25***
			(0.06)
Hookworm moderate-heavy infection, 1999	0.06	0.22	-0.16***
			(0.03)
Roundworm moderate-heavy infection, 1999	0.09	0.24	-0.15***
			(0.04)
Schistosomiasis moderate-heavy infection, 1999	0.08	0.18	-0.10
			(0.06)
Whipworm moderate-heavy infection, 1999	0.13	0.17	-0.04
			(0.05)
Panel B: Other Nutritional and Health Outcomes			
Sick in past week (self-reported), 1999	0.40	0.45	-0.05**
			(0.02)
Sick often (self-reported), 1999	0.12	0.15	-0.03**
			(0.01)
Height-for-age Z-score, 1999	-1.13	-1.22	0.08*
(low scores denote undernutrition)			(0.05)
Weight-for-age Z-score, 1999	-1.25	-1.25	-0.00
(low scores denote undernutrition)			(0.04)
Hemoglobin concentration (g/L), 1999	124.9	123.3	1.6
			(1.4)
Proportion anemic (Hb < 100g/L), 1999	0.02	0.04	-0.02
			(0.01)
Panel C: Worm Prevention Behaviors			
Clean (observed by field worker), 1999	0.59	0.60	-0.01
			(0.02)
Wears shoes (observed by field worker), 1999	0.24	0.26	-0.02
			(0.03)
Days contact with fresh water in past week (self-reported), 1999	2.4	2.2	0.2
			(0.3)

[†]These are averages of individual-level data for grade 3-8 pupils; disturbance terms are clustered within schools. Robust standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. Obs. for parasitological results: 2328 (862 Group 1,466 Group 2). Obs. for hemoglobin results: 769 (290 Group 1, 479 Group 2). Obs. for 1999 Pupil Questionnaire health outcomes: 9,039 (3545 Group 1, 5497 Group 2 and Group 3). Following Brooker et al. (2000) and personal communications with the authors, moderate-to-heavy infection thresholds for the various intestinal helminths are: 250 epg for *S. mansoni*, and 5,000 epg for Roundworm, both the WHO standard, and 750 epg for Hookworm and 400 epg for Whipworm, both somewhat lower than the WHO standard. Kenya Ministry of Health officials collected the parasitological data from January to March 1998 in Group 1 schools, and from January to March 1999 in Group 1 and Group 2 schools. A random subset of the original 1998 Group 1 parasitological sample was re-surveyed in 1999. Hb data were collected by Kenya Ministry of Health officials and ICS field officers using the portable Hemocue machine. The self-reported health outcomes were collected for all three groups of schools as part of Pupil Questionnaire administration. °Note that for the outcomes collected in the 1999 Pupil Questionnaire, statistics in these columns also include Group 3 individuals.

Table A.5-VI: Deworming health externalities within schools, January to March 1999 [†]

	Group 1, Treated in 1998	Group 1, Untreated in 1998	Group 2, Treated in 1999	Group 2, Untreated in 1999	(G1 T 1998) – (G2, T 1999)	(G1, UT 1998) – (G2, UT 1999)
Panel A: Selection into Treatment						
Any moderate-heavy infection, 1998	0.39	0.44	-	-	-	-
Proportion of 1998 parasitological sample tracked to 1999 sample [‡]	0.36	0.35	-	-	-	-
Access to latrine at home, 1998	0.85	0.80	0.81	0.86	0.03	-0.06
					(0.04)	(0.05)
Grade progression (=Grade – (Age – 6)), 1998	-2.0	-1.8	-1.8	-1.8	-0.2	-0.0
					(0.1)	(0.2)
Weight-for-age (Z-score), 1998 (low scores denote undernutrition)	-1.58	-1.52	-1.57	-1.46	-0.01	-0.06
					(0.06)	(0.11)
Malaria/fever in past week (self-reported), 1998	0.37	0.41	0.40	0.39	-0.03	0.02
					(0.04)	(0.06)
Clean (observed by field worker), 1998	0.53	0.59	0.60	0.66	-0.07	-0.07
					(0.05)	(0.10)
Panel B: Health Outcomes						
Girls < 13 years, and all boys						
Any moderate-heavy infection, 1999	0.24	0.34	0.51	0.55	-0.27***	-0.21**
					(0.06)	(0.10)
Hookworm moderate-heavy infection, 1999	0.04	0.11	0.22	0.20	-0.19***	-0.10*
					(0.03)	(0.05)
Roundworm moderate-heavy infection, 1999	0.08	0.12	0.22	0.30	-0.14***	-0.18**
					(0.04)	(0.07)
Schistosomiasis moderate-heavy infection, 1999	0.09	0.08	0.20	0.13	-0.11*	-0.05
					(0.06)	(0.06)
Whipworm moderate-heavy infection, 1999	0.12	0.16	0.16	0.20	-0.04	-0.05
					(0.05)	(0.09)
Girls ≥ 13 years						
Any moderate-heavy infection, 1998	0.31	0.30	-	-	-	-
Any moderate-heavy infection, 1999	0.27	0.44	0.32	0.54	-0.05	-0.09
					(0.17)	(0.09)
Panel C: School Participation						
School participation rate, May 1998 to March 1999 [††]	0.872	0.774	0.808	0.690	0.064*	0.084**
					(0.033)	(0.037)

[†]These are averages of individual-level data for grade 3-8 pupils in the parasitological survey subsample; disturbance terms are clustered within schools. Robust standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. The data are described in the footnote to Table 5. Obs. for the 1999 parasitological survey: 669 Group 1 treated 1998, 76 Group 1 untreated 1998, 874 Group 2 treated 1999, 349 Group 2 untreated 1999. [‡]We attempted to track a random sample of half of the original 1998 parasitological sample. Because some pupils were absent, had dropped out, or had graduated, we were only able to re-survey 72 percent of this subsample. [††]School averages weighted by pupil population. The participation rate is computed among pupils enrolled in the school at the start of 1998. Pupils present in school during an unannounced NGO visit are considered participants. Pupils had 3.8 participation observations per year on average. Participation rates are for grades 1 to 7; grade 8 pupils are excluded since many graduated after the 1998 school year, in which case their 1999 treatment status is irrelevant. Preschool pupils are excluded since they typically have missing compliance data. All 1998 pupil characteristics in Panel A are for grades 3 to 7, since younger pupils were not administered the Pupil Questionnaire.

Table A.5- VII: Deworming health externalities within and across schools, January to March 1999[†]

	Any moderate-heavy helminth infection, 1999			Moderate-heavy schistosomiasis infection, 1999			Moderate-heavy geohelminth infection, 1999		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Indicator for Group 1 (1998 Treatment) School	-0.31***	-0.18**	-0.21*	-0.09***	-0.06	-0.03	-0.30***	-0.19***	-0.26***
	(0.06)	(0.07)	(0.11)	(0.04)	(0.05)	(0.06)	(0.05)	(0.06)	(0.09)
Group 1 pupils within 3 km (per 1000 pupils)	-0.21**	-0.22**	-0.10	-0.12***	-0.12***	-0.08	-0.12	-0.13	-0.06
	(0.10)	(0.11)	(0.14)	(0.05)	(0.05)	(0.07)	(0.09)	(0.10)	(0.12)
Group 1 pupils within 3-6 km (per 1000 pupils)	-0.05	-0.04	-0.08	-0.15***	-0.15***	-0.13**	0.06	0.08	0.03
	(0.08)	(0.08)	(0.11)	(0.04)	(0.04)	(0.05)	(0.06)	(0.06)	(0.09)
Total pupils within 3 km (per 1000 pupils)	0.05	0.05	0.05	0.08***	0.08***	0.08***	-0.01	-0.01	-0.01
	(0.04)	(0.04)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Total pupils within 3-6 km (per 1000 pupils)	-0.02	-0.03	-0.02	0.04*	0.04*	0.04*	-0.04	-0.05	-0.04
	(0.04)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)		-0.06*			0.04**			-0.10***	
		(0.03)			(0.02)			(0.03)	
(Group 1 Indicator) * Received treatment, when offered		-0.15**			-0.04			-0.11**	
		(0.06)			(0.04)			(0.05)	
(Group 1 Indicator) * Group 1 pupils within 3 km (per 1000 pupils)			-0.27**			-0.07			-0.16
			(0.14)			(0.08)			(0.11)
(Group 1 Indicator) * Group 1 pupils within 3-6 km (per 1000 pupils)			0.01			-0.03			0.03
			(0.09)			(0.06)			(0.07)
Grade indicators, school assistance controls, district exam score control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	2330	2329	2330	2330	2329	2330	2330	2329	2330
Mean of dependent variable	0.41	0.41	0.41	0.16	0.16	0.16	0.32	0.32	0.32

[†]Grade 3-8 pupils. Probit estimation, robust standard errors in parentheses. Disturbance terms are clustered within schools. Observations are weighted by total school population. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. The 1999 parasitological survey data are for Group 1 and Group 2 schools. The pupil population data is from the 1998 School Questionnaire. The geohelminths are hookworm, roundworm, and whipworm. We use the number of girls less than 13 years old and all boys (the pupils eligible for deworming in the treatment schools) as the school population for all schools.

Table A.5- VIII: School participation, school-level data[†]

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)		
Panel A: First year post-treatment (May 1998 to March 1999)	*1st Year Treatment*	*Comparison*	*Comparison*	*Group 1 – (Groups 2 & 3)*	*Group 2 – Group 3*
Girls < 13 years, and all boys	0.841	0.731	0.766	0.093***	-0.035
				(0.030)	(0.035)
Girls ≥ 13 years	0.868	0.804	0.820	0.056*	-0.016
				(0.031)	(0.036)
Preschool, Grade 1, Grade 2 in early 1998	0.797	0.689	0.707	0.100***	-0.019
				(0.037)	(0.043)
Grade 3, Grade 4, Grade 5 in early 1998	0.877	0.788	0.827	0.071***	-0.039
				(0.024)	(0.029)
Grade 6, Grade 7, Grade 8 in early 1998	0.934	0.859	0.891	0.058***	-0.032
				(0.021)	(0.025)
Recorded as "dropped out" in early 1998	0.066	0.051	0.030	0.024	0.022
				(0.018)	(0.017)
Females[‡]	0.855	0.771	0.789	0.076***	-0.018
				(0.027)	(0.032)
Males	0.844	0.736	0.780	0.088***	-0.044
				(0.031)	(0.037)
Panel B: Second year post-treatment (March to November 1999)	*2nd Year Treatment*	*1st Year Treatment*	*Comparison*	*Group 1 – Group 3*	*Group 2 – Group 3*
Girls < 13 years, and all boys	0.716	0.718	0.664	0.051*	0.054*
				(0.027)	(0.027)
Girls ≥ 14 years[††]	0.627	0.649	0.588	0.039	0.061*
				(0.035)	(0.035)
Preschool, Grade 1, Grade 2 in early 1998	0.692	0.725	0.641	0.051	0.084**
				(0.034)	(0.034)
Grade 3, Grade 4, Grade 5 in early 1998	0.749	0.766	0.720	0.029	0.046**
				(0.022)	(0.023)
Grade 6, Grade 7, Grade 8 in early 1998	0.781	0.790	0.754	0.027	0.036
				(0.025)	(0.026)
Recorded as "dropped out" in early 1998	0.188	0.130	0.062	0.126*	0.068
				(0.066)	(0.056)
Females[‡]	0.716	0.746	0.649	0.067**	0.097***
				(0.027)	(0.027)
Males	0.698	0.695	0.655	0.043	0.040
				(0.028)	(0.029)

[†]The results are school averages weighted by number of pupil observations. Standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. The participation rate is computed among all pupils enrolled in the school at the start of 1998. Pupils who are present in school on the day of an unannounced NGO visit are considered participants. Pupils had 3.8 participation observations per year on average. The figures for the "Preschool-Grade 2"; "Grade 3-5"; "Grade 6-8"; and "Dropout" rows are for girls < 13 years, and all boys. [‡]Some pupils in the sample are missing information on gender. For this reason, the average of the female and male participation rates does not equal the overall average. [††]Examining girls ≥14 years old eliminates the cohort of girls in Group 1 schools (12 year olds in 1998) who were supposed to receive deworming treatment in 1998.

Table A.5-IX: School participation, direct effects and externalities[†]
Dependent variable: Average individual school participation, by year

	OLS (1)	OLS (2)	OLS (3)	OLS (4) May 98-Mar 99	OLS (5) May 98-Mar 99	OLS (6) May 98-Mar 99	IV-2SLS (7) May 98-Mar 99
Moderate-heavy infection, early 1999						-0.025**	-0.195**
						(0.010)	(0.096)
Treatment school (T)	0.057***						
	(0.014)						
First year as treatment school (T1)		0.063***	0.062***	0.062***	0.056***		
		(0.015)	(0.014)	(0.022)	(0.020)		
Second year as treatment school (T2)		0.039*	0.033				
		(0.021)	(0.021)				
Treatment school pupils within 3 km (per 1000 pupils)			0.040*		0.022		
			(0.022)		(0.032)		
Treatment school pupils within 3-6 km (per 1000 pupils)			-0.024		-0.067***		
			(0.015)		(0.020)		
Total pupils within 3 km (per 1000 pupils)			-0.031**		-0.040**	0.014	-0.029*
			(0.012)		(0.016)	(0.014)	(0.016)
Total pupils within 3-6 km (per 1000 pupils)			0.012		0.035***	0.016*	0.008
			(0.009)		(0.011)	(0.009)	(0.009)
Indicator received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)					0.104***		
					(0.014)		
(First year as treatment school Indicator)* (Received treatment, when offered)					-0.013		
					(0.020)		
1996 district exam score, school average	0.071***	0.070***	0.077***	0.058*	0.106***	0.020	-0.000
	(0.021)	(0.021)	(0.022)	(0.032)	(0.034)	(0.024)	(0.022)
Grade indicators, school assistance controls, and time controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.22	0.22	0.22	0.33	0.37	0.29	-
Root MSE	0.279	0.279	0.278	0.223	0.217	0.150	0.069
Number of observations	56496	56496	56496	18215	18215	2327	49 (schools)
Mean of dependent variable	0.747	0.747	0.747	0.793	0.793	0.884	0.884

[†] The dependent variable is average individual school participation in each year of the program (Year 1 is to March 1999, and Year 2 is May 1999 to November 1999); disturbance terms are clustered within schools. Observations are weighted by the number of times the pupil was observed in that year. Robust standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. Additional explanatory variables include an indicator variable for girls < 13 years and all boys, and the rate of moderate-heavy infections in geographic zone, by grade (zonal infection rates among grade 3 and 4 pupils are used for pupils in grades 4 and below and for pupils initially recorded as drop-outs as there is no parasitological data for pupils below grade 3; zonal infection rates among grade 5 and 6 pupils are used for pupils in grades 5 and 6, and similarly for grades 7 and 8). Participation is computed among all pupils enrolled at the start of the 1998 school year. Pupils present during an unannounced NGO school visit are considered participants. Pupils had approximately 3.8 attendance observations per year. Regressions 6 and 7 include pupils with parasitological information from early 1999, restricting the sample to a random subset of Group 1 and Group 2 pupils. The number of treatment school pupils from May 1998 to March 1999 is the number of Group 1 pupils, and the number of treatment school pupils after March 1999 is the number of Group 1 and Group 2 pupils. The instrumental variables in regression 7 are the Group 1 (treatment) indicator variable, Treatment school pupils within 3 km, Treatment school pupils within 3-6 km, and

the remaining explanatory variables. We use the number of girls less than 13 years old and all boys (the pupils eligible for deworming in the treatment schools) as the school population for all schools.

Table A.5-X: Academic examinations, individual-level data[†]

	Dependent variable: ICS Exam Score (normalized by standard)		
	(1)	(2)	(3) Among those who filled in the 1998 pupil survey
Average school participation (during the year of the exam)	0.63***		
	(0.07)		
First year as treatment school (T1)		-0.035	-0.036
		(0.047)	(0.049)
Second year as treatment school (T2)		-0.015	-0.013
		(0.079)	(0.088)
1996 District exam score, school average	0.74***	0.72***	0.75***
	(0.07)	(0.07)	(0.07)
Grade indicators, school assistance controls, and local pupil density controls	Yes	Yes	Yes
R^2	0.14	0.13	0.15
Root MSE	0.919	0.923	0.916
Number of observations	24979	24979	19072
Mean of dependent variable	0.019	0.019	0.039

The ICS tests for 1998 and 1999 were similar in content, but differed in two important respects. First, the 1998 exam featured multiple-choice questions while the 1999 test featured short answers. Second, while each grade in 1998 was administered a different exam, in 1999 the same exam – featuring questions across a range of difficulty levels – was administered to all pupils in grades 3 to 8. Government district exams in English, Maths, Science-Agriculture, Kiswahili, Geography-History, Home Science, and Arts-Crafts were also administered in both years. Treatment effect estimates are similar for both sets of exams (results not shown). [†]Each data point is the individual-level exam result in a given year of the program (either 1998, or 1999); disturbance terms are clustered within schools. Linear regression, robust standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. Regression 3 includes only pupils who completed the 1998 Pupil Questionnaire. Additional explanatory variables include an indicator variable for girls < 13 years and all boys, and the rate of moderate-to-heavy infections in geographic zone, by grade (zonal infection rates among grade 3 and 4 pupils are used for pupils in grades 4 and below and for pupils initially recorded as dropouts as there is no parasitological data for pupils below grade 3; zonal infection rates among grade 5 and 6 pupils are used for pupils in grades 5 and 6, and similarly for grades 7 and 8). The local pupil density terms include treatment school pupils within 3 km (per 1000 pupils), total pupils within 3 km (per 1000 pupils), treatment school pupils within 3-6 km (per 1000 pupils), and total pupils within 3-6 km (per 1000 pupils). We use the number of girls less than 13 years old and all boys (the pupils eligible for deworming in the treatment schools) as the school population for all schools.

Table A.5-AII: Local densities of other primary schools and deworming compliance rates[†]

	Dependent variable:	
	1998 Compliance rate (any medical treatment) OLS (1)	1999 Compliance rate (any medical treatment) OLS (2)
Treatment school pupils within 3 km (per 1000 pupils)	-0.03	-0.07
	(0.06)	(0.08)
Treatment school pupils within 3-6 km (per 1000 pupils)	0.10*	-0.01
	(0.05)	(0.04)
Total pupils within 3 km (per 1000 pupils)	0.09**	0.05
	(0.03)	(0.06)
Total pupils within 3-6 km (per 1000 pupils)	-0.04	0.00
	(0.03)	(0.02)
Grade indicators, school assistance controls, district exam score control	Yes	Yes
R^2	0.69	0.68
Root MSE	0.070	0.108
Number of observations	25	49
Mean of dependent variable	0.76	0.51

[†]Robust standard errors in parentheses. Observations are weighted by total school population. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. The 1998 compliance data is for Group 1 schools, and the 1999 compliance data is for Group 1 and Group 2 schools. The pupil population data is from the 1998 School Questionnaire. We use the number of girls less than 13 years old and all boys (the pupils eligible for deworming in the treatment schools) as the school population for all schools. The number of treatment school pupils in 1998 is the number of Group 1 pupils, and the number of treatment school pupils in March 1999 is the number of Group 1 and Group 2 pupils.

Table A.5- AIII: Deworming health externalities– Robustness Checks [†]

	Any moderate-heavy helminth infection, 1999				Moderate-heavy schistomiasis infection, 1999			
	Probit	OLS, spatial s.e.	Probit	Probit (G1 only)	Probit	OLS, spatial s.e.	Probit	Probit (G 1 only)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Indicator for Group 1 (1998 Treatment) School	-0.31***	-0.28***	-0.32***		-0.09***	-0.13**	-0.08**	
	(0.06)	(0.06)	(0.06)		(0.04)	(0.06)	(0.04)	
Group 1 pupils within 3 km (per 1000 pupils)	-0.21**	-0.20**		-0.28***	-0.12***	-0.17***		-0.06**
	(0.10)	(0.09)		(0.08)	(0.05)	(0.04)		(0.03)
Group 1 pupils within 3-6 km (per 1000 pupils)	-0.05	-0.11		-0.02	-0.15***	-0.14*		-0.06***
	(0.08)	(0.07)		(0.06)	(0.04)	(0.07)		(0.02)
Total pupils within 3 km (per 1000 pupils)	0.05	0.05	0.00	0.02	0.08***	0.12***	0.06***	0.02**
	(0.04)	(0.06)	(0.04)	(0.02)	(0.02)	(0.04)	(0.02)	(0.01)
Total pupils within 3-6 km (per 1000 pupils)	-0.02	0.02	-0.05*	-0.02	0.04*	0.04	-0.01	0.01
	(0.04)	(0.05)	(0.03)	(0.02)	(0.02)	(0.04)	(0.02)	(0.01)
(Group 1 pupils within 3 km) / (Total pupils within 3 km)			-0.21*				-0.10	
			(0.12)				(0.09)	
(Group 1 pupils within 3-6 km) / (Total pupils within 3-6 km)			-0.10				-0.46***	
			(0.23)				(0.12)	
Any moderate-heavy helminth infection, 1998				0.25***				
				(0.03)				
Moderate-heavy schistosomiasis infection, 1998								0.25***
								(0.10)
Grade indicators, school assistance controls, district exam score control	Yes	No	Yes	Yes	Yes	No	Yes	Yes
R^2	-	0.46	-	-	-	0.48	-	-
Root MSE	-	0.200	-	-	-	0.169	-	-
Number of observations	2330 (pupils)	49 (schools)	2330 (pupils)	603 (pupils)	2330 (pupils)	49 (schools)	2330 (pupils)	512 (pupils)
Mean of dependent variable	0.41	0.41	0.41	0.25	0.16	0.16	0.16	0.09

[†] Grade 3-8 pupils. Robust standard errors in parentheses. Disturbance terms are clustered within schools for regressions 1, 3, 4, 5, 7, and 8. Disturbance terms are allowed to be correlated across spaces using the method in Conley (1999) in regressions 2 and 6. Observations are weighted by total school population. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. The 1999 parasitological survey data are for Group 1 and Group 2 schools. The pupil population data is from the 1998 School Questionnaire. We use the number of girls less than 13 years old and all boys (the pupils eligible for deworming in the treatment schools) as the school population for all schools.

Table A.5-AIV: IV estimates of health and school participation externalities[†]

	Any moderate-heavy helminth infection, January - March 99		Average individual school participation, May 98-March 99	
	Probit	IV-2SLS	OLS	IV-2SLS
	(1)	(2)	(3)	(4)
Indicator for Group 1 (1998 Treatment) School	-0.18**	-0.07	0.056***	0.024
	(0.07)	(0.10)	(0.020)	(0.027)
Group 1 pupils within 3 km (per 1000 pupils)	-0.22**	-0.19**	0.022	0.019
	(0.11)	(0.09)	(0.032)	(0.032)
Group 1 pupils within 3-6 km (per 1000 pupils)	-0.04	-0.03	-0.067***	-0.065***
	(0.08)	(0.07)	(0.020)	(0.020)
Total pupils within 3 km (per 1000 pupils)	0.05	0.05	-0.040**	-0.037**
	(0.04)	(0.03)	0.016)	(0.017)
Total pupils within 3-6 km (per 1000 pupils)	-0.03	-0.02	0.035***	0.034***
	(0.04)	(0.04)	(0.011)	(0.011)
Indicator received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)	-0.06*	-0.06	0.104***	0.022
	(0.03)	(0.05)	(0.014)	(0.031)
(First year as treatment school Indicator)* (Received treatment, when offered)	-0.15**	-0.26**	-0.016	0.056
	(0.06)	(0.12)	(0.020)	(0.045)
Grade indicators, school assistance controls, district exam score control	Yes	Yes	Yes	Yes
Time controls	No	No	Yes	Yes
R^2	-	-	0.37	-
Root MSE	-	0.450	0.217	0.218
Number of observations	2329	2329	18215	18215
Mean of dependent variable	0.41	0.41	0.793	0.793

[†] Disturbance terms are clustered within schools. Robust standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. The two instrumental variables are an indicator for girls under age 13 and all boys (ELG), and (ELG)*(Group 1 indicator). The coefficient on the Group 1 school indicator variable serves as an estimate of the within-school externality effect in 1998. This IV approach could overestimate the treatment effect if the treatment effect is heterogeneous, with sicker pupils benefiting most from treatment, and if among the girls over 13, the sickest girls are most likely to be treated in treatment schools. However, among the sub-sample of older girls, the compliance rate was not significantly related to infection status in 1998 (Table 6), and in 1999 under ten percent of older girls were treated (Table 3). We find similar effects even when we exclude the schools near the lake where older girls were likely to be treated (results not shown). Note that the IV estimates of within-school participation externalities should be interpreted as local average treatment effects for the older girls. Since school participation treatment effects are largest for younger pupils, it is not surprising that the IV externality estimates among the older girls are smaller than the OLS estimates, which are for the entire population. We use the number of girls less than 13 years old and all boys (the pupils eligible for deworming in the treatment schools) as the school population for all schools.

A.6: Preferred Updated Miguel and Kremer (2004) tables

This section includes the relevant tables from Miguel and Kremer (2004), updated to use the final versions of all datasets, which contain our "preferred" analysis. As we argue in section A.2, it is not possible to precisely estimate externalities out to 6 km in this study. Thus, this set of tables includes externalities only out to a distance of 3 km. This change affects Tables I, VII, IX, and X from Miguel and Kremer (2004).

Table A.6-I: 1998 Average pupil and school characteristics, pre-treatment

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)	G1–G3	G2–G3
Panel A: Pre-school to Grade 8					
Male	0.53	0.51	0.52	0.01	-0.01
				(0.02)	(0.02)
Proportion girls < 13 years, and all boys	0.89	0.89	0.88	0.00	0.01
				(0.01)	(0.01)
Grade progression (= Grade – (Age – 6))	-2.0	-1.8	-2.0	-0.0	0.1
				(0.1)	(0.1)
Year of birth	1986.2	1986.5	1985.8	0.4**	0.8***
				(0.2)	(0.2)
Panel B: Grades 3 to 8					
Attendance recorded in school registers (during 4 weeks prior to pupil survey)	0.973	0.963	0.969	0.003	-0.006
				(0.004)	(0.004)
Access to latrine at home	0.82	0.81	0.82	0.00	-0.01
				(0.03)	(0.03)
Have livestock (cows, goats, pigs, sheep) at home	0.66	0.67	0.66	-0.00	0.01
				(0.03)	(0.03)
Weight-for-age Z-score (low scores denote undernutrition)	-1.39	-1.40	-1.44	0.05	0.04
				(0.05)	(0.05)
Blood in stool (self-reported)	0.26	0.22	0.19	0.07**	0.03
				(0.03)	(0.03)
Sick often (self-reported)	0.10	0.10	0.08	0.02	0.02*
				(0.01)	(0.01)
Malaria/fever in past week (self-reported)	0.37	0.38	0.40	-0.03	-0.02
				(0.03)	(0.03)
Clean (observed by field workers)	0.60	0.66	0.67	-0.07**	-0.01
				(0.03)	(0.03)
Panel C: School characteristics					
District exam score 1996, grades 5-8‡	-0.10	0.09	0.01	-0.11	0.08
				(0.12)	(0.12)
Distance to Lake Victoria	10.0	9.9	9.5	0.6	0.5
				(1.9)	(1.9)
Pupil population	392.7	403.8	375.9	16.8	27.9
				(57.6)	(57.6)
School latrines per pupil	0.007	0.006	0.007	0.001	-0.000
				(0.001)	(0.001)
Proportion moderate-heavy infections in zone	0.37	0.37	0.36	0.01	0.01
				(0.03)	(0.03)
Group 1 pupils within 3 km††	430.4	433.2	344.5	85.9	88.7
				(116.2)	(116.2)
Total primary school pupils within 3 km	1272.7	1369.1	1151.9	120.8	217.2
				(208.1)	(208.1)

Note: School averages weighted by pupil population. Standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. Data from the 1998 ICS Pupil Namelist, 1998 Pupil Questionnaire and 1998 School Questionnaire.

‡1996 District exam scores have been normalized to be in units of individual level standard deviations, and so are comparable in units to the 1998 and 1999 ICS test scores (under the assumption that the decomposition of test score variance within and between schools was the same in 1996, 1998, and 1999).

†† This includes girls less than 13 years old, and all boys (those eligible for deworming in treatment schools).

Table A.6-VII: Deworming health externalities within and across schools, January to March 1999

	Any moderate-heavy helminth infection, 1999			Moderate-heavy schistosomiasis infection, 1999			Moderate-heavy geohelminth infection, 1999		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Indicator for Group 1 (1998 Treatment)	-0.33***	-0.20***	-0.24***	-0.12***	-0.08	-0.10*	-0.29***	-0.18***	-0.22***
School	(0.05)	(0.07)	(0.06)	(0.04)	(0.05)	(0.06)	(0.04)	(0.06)	(0.05)
Group 1 pupils within 3 km (per 1000 pupils)	-0.23**	-0.25**	-0.14	-0.13**	-0.13**	-0.10	-0.14	-0.15	-0.07
	(0.10)	(0.10)	(0.12)	(0.05)	(0.05)	(0.08)	(0.09)	(0.10)	(0.12)
Total pupils within 3 km (per 1000 pupils)	0.07*	0.08**	0.07**	0.10***	0.10***	0.10***	-0.01	-0.00	-0.01
	(0.04)	(0.04)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)		-0.06**			0.04*			-0.10***	
		(0.03)			(0.02)			(0.03)	
(Group 1 Indicator) * Received treatment, when offered		-0.14**			-0.05			-0.11**	
		(0.07)			(0.04)			(0.05)	
(Group 1 Indicator) * Group 1 pupils within 3 km (per 1000 pupils)			-0.23*			-0.06			-0.18
			(0.13)			(0.08)			(0.12)
Grade indicators, school assistance controls, district exam score control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	2330	2329	2330	2330	2329	2330	2330	2329	2330
Mean of dependent variable	0.41	0.41	0.41	0.16	0.16	0.16	0.32	0.32	0.32

Note: Grade 3-8 pupils. Probit estimation, robust standard errors in parentheses. Disturbance terms are clustered within schools. Observations are weighted by total school population. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. The 1999 parasitological survey data are for Group 1 and Group 2 schools. The pupil population data is from the 1998 School Questionnaire. The geohelminths are hookworm, roundworm, and whipworm. We use the number of girls less than 13 years old and all boys (the pupils eligible for deworming in the treatment schools) as the school population for all schools.

Table A.6-IX: School participation, direct effects and externalities
Dependent variable: Average individual school participation, by year

	OLS (1)	OLS (2)	OLS (3)	OLS (4) May 98-March 99	OLS (5) May 98-March 99	OLS (6) May 98-March 99	IV-2SLS (7) May 98-March 99
Moderate-heavy infection, early 1999						-0.028***	-0.282**
						(0.009)	(0.111)
Treatment school (T)	0.057***						
	(0.014)						
First year as treatment school (T1)		0.063***	0.065***	0.062***	0.044*		
		(0.015)	(0.014)	(0.022)	(0.024)		
Second year as treatment school (T2)		0.039*	0.036*				
		(0.021)	(0.021)				
Treatment school pupils within 3 km (per 1000 pupils)			0.046**		0.027		
			(0.022)		(0.040)		
Total pupils within 3 km (per 1000 pupils)			-0.031**		-0.034*	0.016	-0.032*
			(0.013)		(0.019)	(0.015)	(0.017)
Indicator received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)					0.104***		
					(0.014)		
(First year as treatment school Indicator)* (Received treatment, when offered)					-0.013		
					(0.020)		
1996 district exam score, school average	0.071***	0.070***	0.070***	0.058*	0.060*	0.016	-0.004
	(0.021)	(0.021)	(0.022)	(0.032)	(0.031)	(0.024)	(0.021)
Grade indicators, school assistance controls, and time controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.22	0.22	0.22	0.33	0.36	0.28	-
Root MSE	0.279	0.279	0.278	0.223	0.218	0.150	0.071
Number of observations	56496	56496	56496	18215	18215	2327	49 (schools)
Mean of dependent variable	0.747	0.747	0.747	0.793	0.793	0.884	0.884

Note: The dependent variable is average individual school participation in each year of the program (Year 1 is to March 1999, and Year 2 is May 1999 to November 1999); disturbance terms are clustered within schools. Robust standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. Additional explanatory variables include an indicator variable for girls < 13 years and all boys, and the rate of moderate-heavy infections in geographic zone, by grade (zonal infection rates among grade 3 and 4 pupils are used for pupils in grades 4 and below and for pupils initially recorded as drop-outs as there is no parasitological data for pupils below grade 3; zonal infection rates among grade 5 and 6 pupils are used for pupils in grades 5 and 6, and similarly for grades 7 and 8). Participation is computed among all pupils enrolled at the start of the 1998 school year. Pupils present during an unannounced NGO school visit are considered participants. Pupils had approximately 3.8 attendance observations per year. Regressions 6 and 7 include pupils with parasitological information from early 1999, restricting the sample to a random subset of Group 1 and Group 2 pupils. The number of treatment school pupils from May 1998 to March 1999 is the number of Group 1 pupils, and the number of treatment school pupils after March 1999 is the number of Group 1 and Group 2 pupils. The instrumental variables in regression 7 are the Group 1 (treatment) indicator variable, Treatment school pupils within 3 km, and the remaining explanatory variables. We use the number of girls less than 13 years old and all boys (the pupils eligible for deworming in the treatment schools) as the school population for all schools.

Table A.6-X: Academic examinations, individual-level data

	Dependent variable: ICS Exam Score (normalized by standard)		
	(1)	(2)	(3) Among those who filled in the 1998 pupil survey
Average school participation (during the year of the exam)	0.63***		
	(0.07)		
First year as treatment school (T1)		-0.042	-0.043
		(0.048)	(0.051)
Second year as treatment school (T2)		-0.014	-0.011
		(0.075)	(0.085)
1996 District exam score, school average	0.74***	0.75***	0.78***
	(0.07)	(0.06)	(0.07)
Grade indicators, school assistance controls, and local pupil density controls	Yes	Yes	Yes
R^2	0.14	0.13	0.14
Root MSE	0.919	0.924	0.918
Number of observations	24979	24979	19072
Mean of dependent variable	0.019	0.019	0.039

Note: Each data point is the individual-level exam result in a given year of the program (either 1998, or 1999); disturbance terms are clustered within schools. Linear regression, robust standard errors in parentheses. Significantly different than zero at 99 (***), 95 (**), and 90 (*) percent confidence. Regression 3 includes only pupils who completed the 1998 Pupil Questionnaire. Additional explanatory variables include an indicator variable for girls < 13 years and all boys, and the rate of moderate-to-heavy infections in geographic zone, by grade (zonal infection rates among grade 3 and 4 pupils are used for pupils in grades 4 and below and for pupils initially recorded as dropouts as there is no parasitological data for pupils below grade 3; zonal infection rates among grade 5 and 6 pupils are used for pupils in grades 5 and 6, and similarly for grades 7 and 8). The local pupil density terms include treatment school pupils within 3 km (per 1000 pupils), and total pupils within 3 km (per 1000 pupils). We use the number of girls less than 13 years old and all boys (the pupils eligible for deworming in the treatment schools) as the school population for all schools. The ICS tests for 1998 and 1999 were similar in content, but differed in two important respects. First, the 1998 exam featured multiple-choice questions while the 1999 test featured short answers. Second, while each grade in 1998 was administered a different exam, in 1999 the same exam – featuring questions across a range of difficulty levels – was administered to all pupils in grades 3 to 8. Government district exams in English, Maths, Science-Agriculture, Kiswahili, Geography-History, Home Science, and Arts-Crafts were also administered in both years. Treatment effect estimates are similar for both sets of exams (results not shown).

B. Detailed response to Davey et al. (2015) statistical replication

B.1 Summary of Main Points

Miguel and Kremer (2004) evaluate a deworming program in 75 Kenyan primary schools using a stepped-wedge research design, in which groups of schools were phased into treatment over time. Methodologically, the paper shows that deworming treatment lowered worm counts not only among treated pupils, but also among untreated pupils within the same school, and among pupils in nearby schools – consistent with the hypothesis that deworming interrupts the chain of disease transmission, what economists would term an "epidemiological externality" or "spillover". The paper shows that in these circumstances, "naïve" estimators of the impact of the program based on examining the simple difference between treatment and comparison schools will be biased downwards, and the paper introduces an estimator of program impact that takes into account effects on neighboring schools. Miguel and Kremer (2004) also show that the Kenya deworming program increased school participation and did so very cost effectively relative to other known approaches. No effect was detected on academic test scores during the time period examined.

Aiken et al. (2015) and Davey et al. (2015) re-analyze the data from Miguel and Kremer (2004). In a separate paper that composes the first part of their replication exercise, Aiken et al. (2015) utilize the statistical methods of the original paper. In that study, the re-analysis authors obtain results consistent with the key claims of Miguel and Kremer (2004); they report substantial, positive impacts of deworming on treated pupils, untreated pupils in treatment schools, and pupils in schools near treatment schools (within 3 km) for both worm infection and for school attendance outcomes. We discuss this "pure replication" re-analysis in detail in Appendix A of this document.

In the present appendix (Appendix B), we comment on Davey et al. (2015), which re-analyzes the original data after changing the definition of treatment, splitting the data into subsamples, re-weighting, and various other adjustments. The re-analysis authors argue that evidence for school participation impacts is not robust, but this conclusion is based on a series of errors in their analysis.

One error is recoding of the treatment measure to include pre-treatment "control" periods in both years of the study (1998 and 1999). To illustrate, Group 2 schools began receiving deworming in March 1999. The correct coding of "treatment" for Group 2 begins after March 1999, and this is the coding discussed and employed in Miguel and Kremer (2004) as well as in the re-analysis presented in Aiken et al. (2015); however, Davey et al. (2015) incorrectly consider the Group 2 school attendance observations from the pre-treatment period in the first months of 1999 as "treatment" observations, leading to the incorrect classification of a sizeable 20% of control observations in Year 2 of the study.

Beyond the miscoding of the treatment variable, the analytic approach taken by Davey et al. (2015) has a number of other important flaws. First, since all of their estimators are based on the "naïve" estimation approach of comparing treatment and control schools in a context where the stable unit treatment value assumptions (SUTVA) are violated by positive disease transmission externalities, their estimates are all downward biased (in the statistical sense of the term).

Second, many of the estimators in Davey et al. (2015) ignore the study's stepped wedge design, in which some schools change treatment status during the course of the study. They instead focus on cross-sectional estimates, completely neglecting the time series aspect of the data, and moreover, split

the data into year subsets and report results separately for the subsets, sacrificing statistical precision and unnecessarily introducing additional noise. The re-analysis authors' own power calculations imply that such approaches are very underpowered (Aiken et al., 2013, p. 7; Davey et al., 2014, Appendix 1). Confirming a key result in Miguel and Kremer (2004), they find a large, statistically significant effect of deworming on school attendance when they pool the data (Davey et al., 2015, Table 2).

Davey et al. (2015) raise a number of concerns about the data and analysis (which we discuss below) and adopt other changes in statistical procedures to address them, most importantly, re-weighting the data. One central finding of our present Commentary is that this central empirical finding of Miguel and Kremer (2004), namely, that deworming increased school participation rates, is robust across a range of statistical estimators once the treatment term is correctly coded and the research design is appropriately utilized.[7] In particular, we show that when treatment is correctly defined to include only periods when deworming treatment had actually occurred, there is a statistically significant impact of deworming on school attendance even in the statistical models which Davey et al. (2015) argue contain the "weakest" evidence. Moreover, in the two pieces of analysis that employ both years of data and use the original study's stepped-wedge research design – the specification which represents the culmination of their pre-specified analysis (Aiken et al., 2013) – the re-analysis authors estimate the same finding as the original paper, namely, a large, positive and statistically significant impact of deworming on school participation. They write: *"When both years were combined, there was strong evidence of an effect on attendance."* (Davey et al., 2015, Abstract).

Nonetheless, it is worth noting and responding to some of the concerns raised by Davey et al. (2015). In particular, they raise concerns about the cross-sectional correlation between the number of attendance observations per school and average school participation rates, in the treatment versus control schools, which they apparently observe by "eyeballing" a plot of the relationship; we present statistical evidence that this correlation does not bias treatment effect estimates. Davey et al. (2015) also base part of their conclusion on a cluster-level analysis making use of a non-standard approach to "weighting" observations, which is contrary to the approach described in their pre-analysis plan (Aiken et al., 2013). We show that deworming has a robust, positive effect on school participation even when considering each year separately (1998 and 1999) using this cluster summary approach once a standard weighting approach (i.e., either weighting each individual equally or each attendance observation equally) is applied.

The bottom line assessment reached by Davey et al. (2015) is that the results in Miguel and Kremer (2004) are not robust to different analytical approaches; they write: *"The evidence supporting an improvement in school attendance differed by analysis method."* (Davey et al., 2015, Abstract). We disagree with this conclusion.

In order to assess the purported "sensitivity" of the school participation results to different analytical assumptions, in Table S5 (below) we present the results in 32 different ways that are common in both the economics and medical literatures (and all of which relate to analytical choices mentioned in the re-analysis authors' pre-analysis plan, Aiken et al., 2013). The key takeaway is that in all 32

[7] The findings in Miguel and Kremer (2004) that receive by far the most attention in Davey et al. (2015) are the impacts of deworming on school participation. This comment focuses almost entirely on this issue, although we also discuss the evidence regarding other deworming impacts at several points.

specifications the coefficient estimate on the deworming treatment indicator variable is large, positive, and statistically significant at 99% confidence. The specifications: (i) use different statistical models (the linear regression model preferred by Miguel and Kremer (2004) and the random effects logistics regression preferred by Davey et al. (2015)); (ii) different samples of pupils (the full sample preferred by Miguel and Kremer (2004) and the sample eligible for deworming treatment as preferred by Davey et al. (2015)); (iii) regression models unadjusted for covariates and adjusted for covariates (the latter of which is preferred by Davey et al. (2015)); (iv) use two different approaches to weighting observations (weighting each attendance observation equally, as in Davey et al. (2015) and in Miguel and Kremer (2004), as well as weighting each pupil equally to obtain the population average); and finally, (v) use the final dataset that Davey et al. (2015) employ in their analysis, even though it incorrectly defines treatment (as described above) and despite the fact that we disagree with some of the assumptions made regarding missing observations (as we detail in Section B.3 below), versus using the correct definition of treatment and our version of the data. The one thing we keep fixed across all of the results in Table S5 is that we use both years of data (1998 and 1999) throughout, as envisioned in the project's original prospective stepped wedge research design, emphasized as the culmination of analysis in the replication authors' own pre-analysis plan (Aiken et al., 2013), and which is the appropriate way to analyze these data.

In all, Table S5 contains 32 different coefficient estimates allowing the five factors mentioned above to vary across the cases. This produces a striking set of results that demonstrate just how remarkably robust the positive impact of deworming on school participation is in this data. In all 32 specifications, the point estimate is positive and large in magnitude, with point estimates in the linear regressions ranging from 5.6 to 7.2 percentage point gains. Furthermore, in all 32 specifications the point estimate is statistically significant at 99% confidence (P-value < 0.01). Note that the coefficient estimates are generally somewhat smaller in specifications using the Davey et al. (2015) version of the data that miscodes treatment, as expected given the measurement error that this induces. A coefficient of particular interest is the culmination of the proposed primary analysis in Aiken et al.'s (2013) pre-analysis plan, which is highlighted with a dark "box" (in column 1 of Panel A). This coefficient estimate is large, positive, and statistically significant with P-value < 0.001. These results presented in Table S5 run counter to the unfounded claim in Davey et al. (2015) that the relationship shows "sensitivity" depending on how the data is analyzed.

Section B.2 of this appendix explores these key points in detail, and addresses the main claims raised in Davey et al. (2015). Section B.2.6 summarizes, and discusses the current state of evidence on the educational and economic impact of deworming. A point-by-point treatment of Davey et al. (2015) is contained in Section B.3, and references for this Appendix B are included in Section B.4.

Table S5: Deworming impacts on school participation (1998-1999)

Analytical approach: Data and variable construction:	Random-effects logistic regression Davey et al. (2015) (1)	Random-effects logistic regression Original (2)	Linear regression Davey et al. (2015) (3)	Linear regression Original (4)
Panel A: Eligible pupils, adjusted				
- weight by attendance observations	1.82*** [p<0.001]	1.88*** [p<0.001]	0.059*** [p=0.002]	0.060*** [p=0.001]
- weight all pupils equally	1.84*** [p<0.001]	1.86*** [p<0.001]	0.059*** [p=0.003]	0.064*** [p<0.001]
Panel B: Eligible pupils, unadjusted				
- weight by attendance observations	1.78*** [p<0.001]	1.84*** [p<0.001]	0.065*** [p=0.005]	0.070*** [p=0.003]
- weight all pupils equally	1.80*** [p<0.001]	1.82*** [p<0.001]	0.069*** [p=0.008]	0.072*** [p=0.003]
Panel C: All pupils, adjusted				
- weight by attendance observations	1.81*** [p<0.001]	1.81*** [p<0.001]	0.056*** [p=0.001]	0.056*** [p=0.001]
- weight all pupils equally	1.83*** [p<0.001]	1.80*** [p<0.001]	0.057*** [p=0.002]	0.061*** [p<0.001]
Panel D: All pupils, unadjusted				
- weight by attendance observations	1.74*** [p<0.001]	1.76*** [p<0.001]	0.063*** [p=0.005]	0.067*** [p=0.004]
- weight all pupils equally	1.76*** [p<0.001]	1.75*** [p<0.001]	0.067*** [p=0.008]	0.070*** [p=0.005]

Notes: These analyses all use both 1998 and 1999 data, finalized and updated, reflecting our own replication documentation (Miguel and Kremer, 2014) as well as comments in Aiken et al. (2015). The Davey et al. (2015) data contains several additional modifications regarding the inclusion of transfer students, and assumptions on missing data, which are described in Davey et al. (2015), as well as erroneously defining treatment to include pre-treatment "control" periods in each year of the deworming program. The original version of the data is as employed by Miguel and Kremer (2004), with the exception that missing age data is imputed using average age within 1998 grade, as detailed in Davey et al. (2015); this is done in order to maintain the same sample while controlling for age in the "adjusted" estimates. All analyses contain covariates for school pupil population size and geographic zone. "Adjusted" estimates follow Davey et al. (2015) in also including covariates for pupil age and SAP program. "Eligible pupils" are those potentially eligible for deworming treatment, as described in Miguel and Kremer (2004). Logistic analyses in columns 1 and 2 present odds ratios and employ school random effects, following Davey et al. (2015); in the linear regression analyses in columns 3 and 4, disturbance terms are clustered by school, following Miguel and Kremer (2004). P-values are in square brackets and stars reflect: "***" P-value < 0.01, "**" P-value < 0.05, "*" P-value < 0.10.

B.2 Technical response to Davey et al. (2015)

Aiken et al. (2013)'s pre-analysis plan culminates in the analysis of the combined 1998 and 1999 data using individual-level random effect logistic regression, either with or without adjustment (i.e., additional covariates), and these results are presented in the top right panel of Table 4 of Davey et al. (2015). The two main results are the finding of an odds ratio of 1.78 (P-value<0.001) and an adjusted odds ratio of 1.82 (P-value<0.001), and we reproduce these in our Table S5 (column 1) above. Both are positive and statistically significant, and they are also very large in magnitude.

It is worth noting up front that Davey et al. (2015) focus entirely on the simple difference between treatment and control schools, and ignore the important issue of deworming externalities. We disagree with this approach. In the presence of positive deworming treatment externalities such as those estimated in Miguel and Kremer (2004) and Aiken et al. (2015), all of the estimators used in Davey et al. (2015) are downward biased, yielding lower bounds on true deworming treatment effects.

In this section, we explore key aspects of the analysis presented in Davey et al. (2015) in detail, and address the main concerns raised by the re-analysis authors.

B.2.1 Miscoding of the treatment measure in Davey et al. (2015)

Nearly all the analysis presented in Aiken et al. (2015) and Davey et al. (2015) was produced through an organized replication program run by the International Initiative for Impact Evaluation (3ie), and as part of that process the replication authors contacted us (as authors of the original paper) with questions and draft reports for comments. We were also provided with the final versions of the reports prior to their publication on 3ie's website in order to prepare a comment that would be posted alongside them.

In the process of studying the school participation analysis presented in Davey et al. (2015) after it had been submitted to 3ie for publication, we discovered what we assumed to be a coding error in the definition of the treatment indicator. Specifically, the replication authors define Group 1 individuals to be "treated" for the entire calendar year for both 1998 and 1999, even though the first attendance visit in 1998 was conducted prior to any Group 1 school receiving deworming treatment or health education (treatment took place between March and April 1998); and they define Group 2 individuals to be "treated" for the entire 1999 calendar year, even though the first two attendance visits in 1999 were conducted prior to any Group 2 schools receiving deworming or health education (treatment took place between March and June 1999). We thought this to be a coding error, as the re-analysis authors had made no mention of this important recoding of the treatment variable in their report (the version of Davey et al., 2014 that was originally submitted to 3ie for publication) or in their pre-analysis plan (Aiken et al., 2013), nor did they object to the original coding in Miguel and Kremer (2004) as it was employed in their "pure replication" report (Aiken et al., 2014, or Aiken et al., 2015), nor had they directly raised the issue in our multiple emails and conference calls during 2013 and 2014. However, subsequent to our bringing this important issue to the attention of the replication authors, they added text to their report justifying this coding choice, and added a new table of results (now their Table 4) with associated discussion.

Davey et al. (2015) purport to justify this choice using an "intention-to-treat" statistical framework. Such a framework is typically utilized in situations where a population was assigned to treatment, but in practice only some individuals within that population actually received treatment

(compliers) while others did not (non-compliers). Davey et al. (2015) incorrectly apply this framework to a different situation – one in which no individuals were actually treated (i.e., Group 2 prior to March 1999) and none were supposed to be treated, but it is claimed (by the re-analysis authors themselves) that individuals could have or should have been treated. This entire argument rests on the assumption that there was some intention to provide deworming treatment at the exact start of the calendar year to each group of schools assigned to treatment later that year. However, as we detail in Section B.3 below, there was never any such intention, and in fact the study's core research design necessitated treatment *not* starting immediately at the start of each calendar year. Davey et al. (2015)'s decision to impose their own notion of what the "planned" timeline of data collection and deworming treatment should have been is inappropriate.

In fact, if we follow Davey et al. (2015)'s assumption on what constitutes a treatment observation to its logical conclusion, then any analysis on the worm infection and health outcomes needs to be discarded, since according to them, Group 2 schools are all already "treatment schools" in early 1999, and thus the comparison between Group 1 and Group 2 using data collected in early 1999 is meaningless. Yet this does not make sense since no Group 2 schools were treated, nor was there ever any intention of treating them, in the early months of 1999. Rather, extensive data collection was carried out in all schools in the early months of 1999 precisely *because* Group 2 had not yet been phased into treatment, allowing for analysis of health impacts after one year of treatment.

We show that when treatment is correctly defined to include only periods when deworming treatment had actually started, there is a statistically significant impact of deworming on school attendance even in the statistical models which Davey et al. (2015) argue contain the "weakest" evidence. In particular, we show this for both the cluster summary and individual-level analyses. The individual-level results presented in Table S5, columns (2) and (4), already correct this miscoding of the treatment term, as we mention above. Table S7, Panel B (below) explores the implications of the miscoding of the treatment term in the cluster summary analyses. The results in this panel utilize exactly the same data and weighting methodology as in Panel A (which we go through in more detail in Section B.2.3), but we have redefined treatment appropriately. Specifically, Group 1 individuals are considered "treated" starting at attendance check visit #2 in 1998 (attendance check visit #1 is dropped from the analysis for simplicity, although it could also be included without changing the results), and for the rest of 1998 and 1999; Group 2 individuals are considered "treated" starting at attendance check visit #3 in 1999, and for the rest of 1999. Making only this change, the cluster summary results weighted by either pupil population or number of attendance observations remain large and highly statistically significant (P-value < 0.05) in all cases, as before. But interestingly, even in the Davey et al. (2015) analysis that weights each school equally, which we argue below is inappropriate, the impact of deworming on school participation in 1998 alone result is marginally significant (P-value=0.056) and the pooled 1998 and 1999 year results are highly significant (P-value < 0.05).

Davey et al. (2015) make an unfounded decision to recode the treatment variable in their analysis. Once this blatant error is corrected, the estimated impact of deworming on school participation using the correctly coded treatment variable but otherwise using their analytical methods is large, positive and statistically significant, as detailed below.

B.2.2 Davey et al. (2015) concern #1: possible relationship between number of observations and attendance

The main concerns raised by Davey et al. (2015) appear to revolve around data collection and data quality. One claim is that there are some unusual correlations between the number of school attendance observations per school and the average school participation rate. However, the existence of a simple correlation of this kind is not sufficient to introduce bias into the study. In our data the key driver of the total number of school participation observations is the school population, i.e., large schools have many more pupil-level observations than small schools, as expected. School participation rates could correlate with school population (or with any of a number of other demographic and social characteristics) for many different reasons, and the existence of such a correlation alone is not a source of bias. For instance, larger schools could be located in more densely populated areas, have a different disease environment, or be located closer to (or farther from) Lake Victoria; better schools may attract more pupils and also have lower attendance rates; and in denser areas, schools may be larger but closer together, affecting the average walking distance to school, etc.

So the argument in Davey et al. (2015) is more subtle. For there to be bias in the analysis, the correlation between school participation observations and the average school participation rate would have to differ systematically between treatment and control schools. The authors are particularly focused on the case of the Group 2 schools, which start out in the control group in 1998 and "phase in" to deworming treatment in 1999. In the case of the Group 2 schools, their concern is that there is a time-varying difference (between 1998 and 1999) in how the correlation between the number of school participation observations and the average school participation rate differs between treatment and control schools. In their own words: *"We are particularly concerned about the reliability of this before-after comparison because, as Figure 3 shows, in Group 2 the association between the number of pupil observations and mean school attendance changed between years. This would potentially lead to over-estimation of the effect on attendance in a weighted analysis."* (Davey et al., 2015, Discussion).

This is the central critique of the Miguel and Kremer (2004) data and analysis in Davey et al. (2015), as we read it. This potential for "bias" in the estimation of deworming treatment effects would be due to "excessive" data collection in "high" school participation treatment schools relative to "low" school participation treatment schools.

We are puzzled by this assertion since no statistical test was provided in Davey et al. (2015) about whether there actually is "excessive" data collection in certain types of schools than in others. Rather, the assertion is apparently based on "eye-balling" their Figure 3, and the visual evidence does not look compelling to us: all three groups of schools have a downward sloping (negative) relationship in 1998, and the relationships in 1999 appear flatter, with some upward sloping. Yet the test that Davey et al. (2015) allude to is straightforward to run with the data in hand: one can test (using data at the school-year level) if there is a significant difference in the correlation between school participation and the number of school participation observations between treatment and control schools, and moreover, if this correlation changes over time (which is critical to the replication authors' claim that they cannot reliably exploit the study's stepped wedge research design, which includes the incorporation of the Group 2 schools into the treatment group in 1999).

We run this test in Table S6 (below). We first note that we find no statistically significant correlation between school participation and the number of school-year attendance observations

overall pooling both years of data (column 1). The point estimate is very close to zero, at -0.024, with a P-value of 0.14. The test alluded to by Davey et al. (2015) is presented in column 2, and further includes indicators for year 2 (1999) and treatment schools (= Group 1 in 1998 and Groups 1 and 2 in 1999), as well as interactions between these two terms and the measure of attendance observations. In the table, we bold the two key interaction terms that they allude to, namely, the interaction between the treatment indicator and the number of observations, and then the triple interaction of these terms with the 1999 indicator. We find that there are no significant interaction effects of treatment with the number of observations, and once again the point estimate is very close to zero with a large P-value (P-value = 0.71), nor does this correlation change over time, in the triple interaction term (P-value = 0.14). We then investigate whether this relationship differs between the Group 1 and Group 2 schools in column 3, but once again find no statistically significant interaction effects between these deworming group indicators and the number of attendance observations, nor do these effects differ across years (once again P-value > 0.10 in all cases). The coefficient estimate that Davey et al. (2015) specifically focus on is the triple interaction of the Group 2 indicator with the Number of observations and the 1999 indicator (to capture whether the nature of data collection these schools that "switched" treatment status due to the stepped wedge design changed over time) and this estimate is very close to zero (0.045) with a large P-value of 0.56.

The bottom line is that there is no statistically significant – or even suggestive – evidence that there is a differential correlation between the number of observations and school participation rates across treatment and control schools, nor that this relationship changes over time. We are not surprised by this pattern, since we were involved in the original data collection and know that approximately equal numbers of visits were made to schools in treatment and control schools throughout. In the absence of this evidence, Davey et al. (2015)'s assertion that it is inappropriate to pool data from 1998 and 1999 and utilize the project's research design is unfounded.

B.2.3 Davey et al. (2015) concern #2: Appropriate weighting

Even if one were to accept their assertions about potential bias (based on the broad visual patterns the re-analysis authors claim to discern in Figure 3), the suggested remedy proposed by Davey et al. (2015) – namely, using an approach that weights each school equally in their (not pre-specified) cluster-level analysis – is inappropriate in our view. The correct way to address this issue would be to weight each pupil equally. Doing so would maintain the analysis as the average impact in the sample *population*, a meaningful quantity. The school average impact 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 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 rationale for such a decision. 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.

Table S6: Relationship between school attendance and observations

Dep var: School attendance	(1)	(2)	(3)
	-0.024	-0.061***	-0.115***
Number of attendance observations (by school-year)	[0.144]	[0.003]	[0.007]
		-0.132**	-0.204**
Indicator for 1999		[0.050]	[0.014]
		0.067	-0.005
Indicator for treatment school (col 3 = G1 only)		[0.398]	[0.955]
			-0.106
Indicator for Group 2 school			[0.142]
		-0.023	0.030
Treatment indicator * Number attendance obs		**[0.713]**	[0.680]
			0.071
G2 indicator * Number attendance obs			[0.137]
		-0.163	-0.065
Treatment indicator * 1999 indicator		[0.123]	[0.597]
			-0.013
G2 indicator * 1999 indicator			[0.899]
		0.027	0.080
Number attendance obs * 1999 indicator		[0.581]	[0.189]
		0.122	0.051
Treatment * Number attendance obs * 1999 indicator		**[0.138]**	[0.592]
			0.045
G2 indicator * Number of attendance obs * 1999 indicator			[0.557]

Note: The dependent variable is average school attendance in a school-year. Controls are as shown. Number of attendance observations is presented in thousands. P-values are in square brackets and stars reflect: "***" P-value < 0.01, "**" P-value < 0.05, "*" P-value < 0.10.

Given the importance that Davey et al. (2015) attach to their cluster summaries analysis (which was not pre-specified) in driving their claim that the results in Miguel and Kremer (2004) are sensitive to analytical choices, we explored the analysis they present in the top left panel of their Table 2. Following that analysis, we focus on simple school average outcomes year-by-year but simply re-weight each of these observations by the school population at baseline in 1998. This solves the potential problem they point to about "excessive" school participation observations in some schools relative to others, but maintains the analysis in terms of population averages, which is attractive and standard.

As we show in Panel A of Table S7, below, the cluster summaries analysis with this standard weighting approach generates large and statistically significant deworming treatment effects in 1998 alone (P-value < 0.05), in 1999 alone (P-value < 0.05), and in 1998 and 1999 combined (P-value < 0.01). We also present these results weighting by the number of attendance observations (which we feel is also appropriate given the lack of evidence above on the purported "excessive" observations in high participation treatment schools), and weighting each school equally, as in Davey et al. (2015).

To take a step back and summarize the argument in Davey et al. (2015), they claim that there was excessive data being collected in "high" school participation treatment schools relative to lower

attendance schools, and that this may have led to bias in the school participation estimates. They use this purported pattern to justify both: (i) weighting each school observation equally (rather than using population averages or weighting by the number of attendance observations), and (ii) to not pool data across 1998 and 1999 (a decision which greatly reduces the statistical power of the original study design).

However, we showed in Table S6 that there is in fact no statistically significant difference between the correlation of school participation rates and school participation observations in treatment versus control schools, nor does this relationship change over time. So there is no statistical evidence for the purported data problem that forms the centerpiece of the argument in Davey et al. (2015). Moreover, even if one were to accept their argument based on more informal evidence, such as broad visual inspection of their Figure 3, the solution they propose is inappropriate, since it is preferable on all dimensions to weight each pupil equally and obtain the population average rather than weight each school equally, and arbitrarily weight some students seven times more than others. When one does so, the cluster summary results in Table S7 indicate that deworming led to large, positive and statistically significant impacts on school participation in 1998 alone, in 1999 alone, and in 1998 and 1999 together.

Taking Tables S5 and S7 together and considering all deworming treatment effect estimates that (i) pool both years of data (since we have shown there is no justification not to do so), and (ii) correct Davey et al. (2015)'s incorrect recoding of the treatment indicator, to us it appears hard to avoid the conclusion that school-based deworming in this Kenyan sample has positive, large, and highly statistically significant impacts that are robust to a wide range of sensitivity analyses, including regression models (random effects, linear regression), weighting schemes (at the school-level, pupil-level, and attendance observation-level), covariates (adjusted and unadjusted), samples (all pupils and only those eligible for the deworming drug), and assumptions on the data (including the treatment of missing data as preferred by Davey et al., 2015, or by Miguel and Kremer, 2004). All 32 of these estimates are presented graphically in Figure 2 of the main Commentary text.

B.2.4 Davey et al. (2015) concern #3: blinding and data quality

Another concern for Davey et al. (2015), which they mention in multiple places in their report, is the fact that the original study was not blinded. The authors suggest that: *"Allocation to the intervention arm could therefore plausibly have affected school attendance through behavioural pathways, such as the placebo or Hawthorne effects"* (Davey et al., 2015, Discussion).

First, we would point out that the deworming program may indeed have led to behavioral changes (i.e., changes in family or school practices) that in turn affected schooling outcomes. This is not a problem for the study: one would certainly want to capture and understand these behavioral changes caused by the program. Improved health can affect life outcomes and choices in many ways, and the school participation effect is the combined effect across potentially multiple behavioral channels.

Table S7: Cluster summary results, with different weighting schemes

	Weight by Pupil Population		Weight by Num. of Attendance Obs.		Weight each School equally	
	Difference	P-value	Difference	P-value	Difference	P-value
Panel A: Treatment indicator and year defined as in Davey et al. (2015)						
1998	8.57**	[0.011]	7.86**	[0.019]	5.48	[0.121]
1999	5.15**	[0.028]	5.84**	[0.011]	2.16	[0.483]
1998+1999[1]	6.87***	[0.001]	6.84***	[0.001]	3.81	[0.102]
1998+1999[2]	6.87***	[0.004]	6.84***	[0.004]	3.81	[0.105]
Panel B: Treatment indicator and year defined as in Miguel and Kremer (2004)						
1998	9.25**	[0.033]	8.86***	[0.004]	7.38*	[0.056]
1999	4.99**	[0.046]	5.35**	[0.037]	3.57	[0.150]
1998+1999[1]	7.15***	[<0.001]	7.46***	[<0.001]	5.48**	[0.017]
1998+1999[2]	7.15***	[0.002]	7.46***	[0.002]	5.48**	[0.017]

Note: This analysis is based on the top left panel of Davey et al. (2015), Table 4, and in fact the first two rows of "unweighted" results (for 1998 and 1999 in Panel A) replicate those results. All analysis includes only 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.

Rather Davey et al. (2015) appear concerned that receiving drug treatment changes behavior due mainly to placebo effects. The re-analysis authors advance this claim without providing statistical evidence that these effects are in fact meaningful. However, there are several ways to explore these issues in the data. First, there are sizeable numbers of students in treatment schools who did not receive deworming treatment either due to absence on the day of deworming or because they were adolescent girls (who were meant to be excluded from treatment due to potential drug side effects). If the effect were mainly driven by placebo effects, rather than real deworming impacts, then there would be no meaningful effects on those students who did not themselves take the deworming pills. This would be true both for the untreated within treatment schools, and for control school students located within 3 km of a treatment school. Yet these populations, who we show benefit from reduced worm infection burden (due to epidemiological externalities) also show gains in school participation even though a placebo effect is not plausible for them (Miguel and Kremer, 2004; Aiken et al., 2015).

There is also a related possibility, namely that it was health education rather than the deworming drugs themselves that drove impacts. However, we show in both Miguel and Kremer (2004) and Kremer and Miguel (2007) that there are no significant differences in a range of worm prevention behaviors between the treatment and control schools, including wearing shoes, contact with fresh water, or observed cleanliness.

There is a final point regarding "blinding" in the context of a deworming study that is important to consider, namely the fact that it may be impossible to carry out such a study using a cluster randomized design. (Recall that a key point of Miguel and Kremer (2004) is that individually randomized studies will underestimate the impact of treatment in the presence of epidemiological externalities.) One of the immediate consequences of taking deworming drugs for those with worm infections is that worms are expelled from the body, usually in stool (although more rarely also through vomiting). This is a highly visible outcome and one that is much commented upon in communities receiving mass deworming. While individual participants in a study that randomized treatment at the individual level to a subset of children in a school, say, may not know if they received deworming drugs or placebo (since many but not all those who are infected and treated will see worms expelled), participants in a study that randomizes treatment at the cluster level, as in Miguel and Kremer (2004), will immediately know if they are a "treatment" or "placebo" school: in treatment schools, a sizeable group of students (approximately 12% in our data) will immediately experience gastrointestinal discomfort, worms will be expelled in stool and some will vomit; in placebo schools, there will be no such outcomes. Similarly, it would be impossible for enumerators to avoid finding out the school's treatment status, since enumerators interview and speak with hundreds of pupils, teachers and parents during a school visit, and side effects are a common topic of conversation. Thus a direct, but quite unattractive, implication of Davey et al.'s concern with blinding would be that it is impossible to carry out a "high quality" deworming cluster randomized study. Since Miguel and Kremer (2004) demonstrate that the violations of SUTVA in an individually randomized study in the context we examine are real, while the concern that lack of blinding affected reporting or data collection remains hypothetical, we believe that the use of cluster randomization remains appropriate.

It is also important to draw attention to the sharply different norms in social science and medical research on the appropriate way to report results, even conditional on the exact same research design. Indeed, Eble et al. (2013) review all randomized experiments in economics published since 2000 against the biomedical CONSORT trial reporting standards and conclude that nearly all economics studies would be considered "low quality" and at "high risk of bias" under these reporting guidelines. The emphasis on blinding leads to the almost immediate conclusion that data from most real-world social science experiments provide "low quality" evidence that is at "high risk of bias", since participants in real programs are typically aware of their treatment status – and in fact social scientists are often very interested in the endogenous behavioral change that results from that knowledge. The existing criteria on both blinding and reporting have the unfortunate implication that Cochrane Reviews appear to systematically down-weight new evidence from the social science disciplines, where the most rigorous evidence on the socio-economic impacts of health interventions arguably lies. It also means that replication efforts (like the present one) that rely on medical researchers such as Davey et al. to carry out the replication of social science studies are very likely to lead to conclusions that the evidence is "weak" or of "low quality" for similar reasons, in large part due to disciplinary differences.

Davey et al. (2015) also refer to the original study's lack of pre-documented data collection plans. Beyond disciplinary differences, there are also issues of timing. Back in 1997 when this study was being set up, the state of pre-registration in public health and health fields was far less developed than today. To illustrate, the CONSORT guidelines were only conceived of in 1996 and did not become "standard" until a number of years later. The NIH "clinicaltrials.gov" website was only launched in 2000,

after the data collection for the Miguel and Kremer (2004) study was completed. It was only after that point that pre-registration of trials was widely required in the medical literature.

A related point has to do with the 18 year (1997 to 2015) time lag between the setup of the Kenya deworming project and the Aiken et al. (2015) and Davey et al. (2015) papers. That is clearly a long time, and despite our best efforts, not all documentation has been easily accessible. We did not have access to Dropbox or scanners in 1997 when project planning for Miguel and Kremer (2004) was taking place; in fact, making an international phone call and getting basic email access was a challenge in the field. The replication authors have benefitted from extensive personal access to all of us; we have also shared numerous original documents, surveys, etc. with them when we have had them available (and they do refer to some of these in their reports). We believe readers should keep these issues in mind when the replication authors discuss data quality, as many of their concerns have to do with their inability to access detailed *ex ante* data collection plans, protocols and field notes, rather than any evidence of bias within the data itself (or in the field plans as we recall them). The lack of this documentation 18 years later does not constitute evidence of bias. In fact, a range of measures, tests, and statistical patterns discussed above demonstrate that the data in Miguel and Kremer (2004) was collected in an even-handed way. These patterns, our own experience designing the field data collection procedure in Kenya, and the lack of any statistical evidence for biased data, together imply that the data quality assertions in Davey et al. (2015) are largely without basis.

B.2.5 Concerns regarding missingness noted in Davey et al. (2015)

Davey et al. (2015) provide a discussion of missing data. A particular concern relates to missing data school attendance, as the primary outcome variable of interest. However, an earlier version of their report (Davey et al., 2014) concluded that: *"As the extent of missingness in attendance data was similar in each of the groups, we believe that this risk [of bias in the primary analysis] is low."* (Section 4.6, para 1, bracketed text added for clarity). On average, roughly 20% of attendance observations are missing, with nearly equal rates across the three treatment groups, and this level of attrition is reasonably low for longitudinal data collection in a rural low-income setting. The re-analysis authors also reproduce the Miguel and Kremer (2004) finding that the three treatment groups are largely balanced on baseline observable characteristics (Aiken et al., 2015), providing further confidence in the validity of the experimental design and the data collection procedures.

B.2.6 Discussion

To summarize, in Section B.1 we discuss the results in Davey et al. (2015), and argue that their statistical evidence is consistent with the conclusions in Miguel and Kremer (2004). In particular, their statistical evidence provides support for the conclusion that mass school-based deworming leads to higher school participation. This is true across a range of specifications, samples, adjustment, weighting and data choices (as shown in our Table S5 and Table S7, as well as Figure 2 in the main Commentary text), when the full dataset is used (and a miscoding of the treatment term in the replication analysis is corrected), including the key specifications emphasized as the primary analysis in Aiken et al. (2013)'s pre-analysis plan. In Section B.2, we respond concerns about the data and approach, and argue that these do not change the main conclusions of the Miguel and Kremer (2004) study, or its implications in terms of the cost-effectiveness of school-based deworming in the study setting.

In an overview of their results related to school participation, the re-analysis authors write: *"When both years were combined, there was strong evidence of an effect on attendance…"* (Davey et al., 2015, Abstract). Our Table S5 (above) shows that this full model can be specified any number of ways and the effect is still strong. It is only when the re-analysis authors slice the data into underpowered subsamples, mis-define the treatment measure, and perform incorrectly weighted analysis that they obtain results that do not suggest a strong impact of deworming on school participation.

The central issue raised in Davey et al. (2015) in our view is the possibility that there is bias in the estimation of school participation treatment effects because of potentially "excessive" data collection (i.e., more observations collected) in high participation treatment schools relative to low participation treatment schools, and especially that this relationship changed over time. This purported relationship is the re-analysis authors' justification for not pooling both years of data in the analysis, and for using an alternative, non-standard and, we argue, inappropriate approach to weighting observations.

We first show that there is actually no statistical evidence for the purportedly "biased" data collection patterns in the data (Table S6, above). Second, even if one were to accept this assertion, the appropriate solution would be to weight each pupil equally (rather than each school equally), and the school participation results in Miguel and Kremer (2004) are robust to doing so (Tables S5 and S7).

Aiken et al. (2015) and Davey et al. (2015) comment on the broader deworming literature and policy debate, and we briefly do so as well here. New evidence is rapidly accumulating on the educational and socio-economic impacts of child deworming. A key lesson of Miguel and Kremer (2004) is that traditional individual-level randomized designs will miss any spillover benefits of deworming treatment, and this could contaminate estimated treatment effects. Thus cluster randomized designs provide better evidence. Three new working papers with such cluster randomized designs estimate long-run impacts of child deworming up to 10 years after treatment; these effects on long-run life outcomes are arguably of greatest interest to public policymakers, as discussed in Ahuja et al (2015).

Croke (2014) finds positive long-run educational effects of a program that dewormed a large sample of 1 to 7 year olds in Uganda, with statistically significant average test score gains of 0.2 to 0.4 standard deviation units on literacy and numeracy 7 to 8 years later. The Ugandan program is one of the few studies to employ a cluster randomized design, and earlier evaluations of the program had found large short-run impacts on child weight (Alderman et al., 2006; Alderman, 2007). Croke (2014, p. 16) also surveys the emerging deworming literature and concludes that "*the majority of clustered trials show positive effects*".

Two other new working papers explore the long-run impacts of the Kenya program we study. While the primary school children in the Miguel and Kremer (2004) sample were probably too old for deworming to have major impacts on brain development, and there was no evidence of such impacts, Ozier (2014) estimates cognitive gains 10 years later among children who were 0 to 2 years old when the deworming program was launched and who lived in the catchment area of a treatment school. These children were not directly treated themselves but could have benefited from the positive within-community externalities generated by mass school-based deworming. Ozier (2014) estimates average test score gains of 0.3 standard deviation units, which is equivalent to roughly half a year of schooling and similar to the effect magnitudes estimated by Croke (2014). This provides further evidence for the existence of large, positive, and statistically significant deworming externality benefits within the communities that received mass treatment.

Finally, Baird et al. (2014) 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. Ten years after the start of the deworming program, men who were eligible to participate as boys work 3.5 more hours each week, spend more time in entrepreneurship, are more likely to hold manufacturing jobs with higher wage earnings, and have higher living standards. Women who were eligible as girls have better educational outcomes (including higher rates of passing the primary school completion exam and enrolling in secondary school), are more likely to grow cash crops, and reallocate labor time from agriculture to entrepreneurship. The impacts of deworming subsidies on labor hours are sufficiently large that the social internal rate of return is very high, with an annualized rate of at least 32.2%.

Taken together, and building on Miguel and Kremer (2004), Alderman et al. (2006), and Alderman (2007), as well as economic history work that estimates similar long-run deworming impacts on socio-economic outcomes (Bleakley 2007), this new wave of studies promises to bring considerable new evidence to bear on the long-run impacts of childhood deworming on important life outcomes in areas with high worm infection rates.

B.3. Additional points raised in Davey et al. (2015)

This section provides detailed, point-by-point responses to points raised in Davey et al. (2015). For legibility, we have included the original text from that report in ***bold italics***, followed by our response. Square brackets denote text added to the quotes for clarity.

Abstract: "We estimated intention-to-treat effects using year-stratified cluster-summary analysis and observation-level random-effects regression, and combined both years with a random-effects model accounting for year."

Neither the "intention-to-treat" recoding of the treatment term nor the year-stratified cluster summary analysis were detailed in the authors' pre-analysis plan (Aiken et al., 2013). We further note that the authors received our data and documentation (including our "replication manual" already detailing the coding errors and the vast majority of rounding errors and other typos reported in Aiken et al., 2015) prior to the registration of that plan.

Abstract: "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: year1: 1.48, 95% confidence interval 0.88–2.52, p=0.150; year 2: 1.23, 95%CI 1.01-1.51, p = 0.044), but not for 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 (difference -0.121, 95%CI -0.293-0.052, p=0.169)."

We found it unusual that Davey et al. emphasized subset results (the year-by-year analysis) and analysis that was not pre-specified (the cluster summary analysis) up front as "main" findings. The study took place over two years, and the original study – as well as the pre-analysis plan for the reanalysis (Aiken et al., 2013) – both emphasize the importance of combining the estimates across both years. In particular, the pooled estimation is the culmination of the pre-specified analysis, and the power

calculations on page 7 of that plan showing only moderate power in the stepped wedge design also appear to be based on the two years of data together, indicating that an analysis year-by-year would be severely underpowered. It is thus not surprising that the results are less statistically significant once only subsets of the data are analyzed – the most immediate reason for higher P-values when you split the data into year 1 and year 2 separately is that there are smaller samples (roughly half) that are being analyzed. The stepped wedge design may also contribute, since it contains a valuable change in treatment status for the Group 2 schools, which can increase statistical power.

When one focuses on the pooled results, which efficiently utilize all of the data, there is evidence that deworming led to large, positive and statistically significant impacts on school participation across dozens of regression specifications, as shown in Table S5 and Table S7 (above), as well as Figure 2 in our main Commentary text.

Abstract: "The evidence supporting an improvement in school attendance differed by analysis method."

As we detail in Sections B.1 and B.2 above, we disagree with this interpretation of the results. Under the specifications laid out in the authors' own pre-analysis plan (Aiken et al., 2013), the combined estimates across the two years is always statistically significant at high levels of confidence (P-value <0.01), and this is true with different covariates (Davey et al., 2015, Table 2), with slightly different samples (i.e., all children, or just those targeted with deworming; Davey et al., 2014, Appendix 4), and even diluting the treatment effect by mis-defining the treatment measure (Davey et al., 2015, Table 4). So in fact the evidence presented in Davey et al. (2015) itself overwhelming demonstrates a robust large increase in school attendance. Ignoring the original study's prospective research design by focusing only on cross-sectional variation, and then splitting the sample into halves leads to under-powered statistical analysis that is less convincing than the approach used in Miguel and Kremer (2004) and in the pre-analysis plan of Aiken et al. (2013).

Abstract: "This, and various other important limitations of the data, cautions against over interpretation of the results."

We discuss this issue in section B.2 above. In particular, we make the point that the re-analysis authors provide no evidence for biased measurement procedures, and in fact there is some evidence that data collection across the three program treatment groups was carried out in an even-handed manner.

Introduction: "We focus on the 'naïve' results of the original study (as described in the pure replication) specifying school attendance and examination performance as the co primary outcomes, as these were the major focus of the original study."

We expected to see externalities included in the analysis, as the re-analysis authors' own pre-analysis plan (Aiken et al., 2013) was clear in its intent to include a study of indirect effects of deworming in the statistical replication arm (which followed the pure replication arm) of the analysis:

> *"In addition, and depending on the results of the primary analyses, we will conduct further analyses that look at … the indirect effects of the intervention on all three*

> *outcomes domains (school attendance, exam performance, health indicators). We aim to replicate the spatial method used in the original study to estimate the indirect effects of the intervention, using the same distances (up to 6 km from schools) employed in the original study, as these are plausible distances for the scale of such an effect. However, our plan for analysis of these indirect effects is dependent on first demonstrating a direct effect – following the standard reporting practice for clinical trials, if our analysis does not demonstrate direct effects, we will not pursue analyses looking for indirect effects."*
> (Aiken et al., 2013, page 5)

Aiken et al. (2015) find externalities on worm prevalence within schools and up to 3 km away, and large direct deworming treatment effects on school participation. Yet Davey et al. (2015) focus on the simple difference between treatment and control schools alone, and ignore the important issue of cross-school deworming externalities. We disagree with this approach. In the presence of positive deworming treatment externalities such as those estimated in Miguel and Kremer (2004) and Aiken et al. (2015), all of the estimators used in Davey et al. (2015) are downward biased, yielding lower bounds on true deworming treatment effects.

Methods: "Concurrently, ICS were also evaluating five other interventions under their 'School Assistance Programme' in 27/75 study schools (SAP schools)."

The research paper that estimated impacts of this other program (the School Assistance Program, or SAP) on educational outcomes (including school attendance) finds no significant average educational impacts (Glewwe, Kremer, and Moulin, 2009).

Methods: "We inferred, in the absence of a protocol, that the complex intervention was intended to be delivered from the start of each calendar/academic year."

This justification for the re-analysis authors' decision to recode the treatment term is unfounded. We do not view the intervention as complex, simply two rounds of deworming drug administration per year plus health education talks. Description of the timing of treatment in each year is provided in Miguel and Kremer (2004); all references to the timing of treatment (pages 170, 192, and 210) note that treatment took place in March-April in 1998 and March-June in 1999. Moreover, the construction of the (post-treatment) school participation measure for each year was clearly defined in the original authors' STATA analysis do files, which were provided to the replication authors at the time they embarked upon this project. Davey et al. (2015) carefully studied this code and did not raise any objections to that definition, or confusion, in their pure replication report (Aiken et al., 2015). Nor did the re-analysis authors make any explicit mention of any redefinition of the treatment measure in their pre-analysis plan (Aiken et al., 2013) – which was registered after receipt of the data and do files from the original paper – or in the original version of the present report (Davey et al., 2014) that was initially submitted for publication by 3ie. It was only after we were provided the analysis files underlying that report, and discovered what we assumed to be a major coding error, that the re-analysis authors eventually added any text justifying the recoding of treatment.

We find Davey et al.'s use of an intention-to-treat justification unusual and non-standard here. Such a framework is typically employed to study treatment impacts for a group in which, among those

assigned to treatment, some received the treatment and others did not. In our case, not a single school began receiving treatment prior to March of either program year, so there is no situation in the early months of 1998 or 1999 when, among schools that were supposed to receive treatment, some had already done so and others had not.

Moreover, treatment in the first several weeks of each year would have been impossible due to the project's research design. As the timeline described in Miguel and Kremer (2004), Appendix Table A1 makes clear, it is central to the design of the original study that administration of deworming drugs not begin immediately at the start of each year. In both 1998 and 1999, the early part of the calendar year was devoted to conducting meetings introducing the program to each community, and to collecting pupil questionnaire and parasitological data. In Year 1, this pupil questionnaire and parasitological data serves as a baseline. In Year 2, this data collection was critical to the study of health impacts: the pupil questionnaire and parasitological data collection in the first 3 months of 1999 provide the only opportunity to study the impacts of deworming on worm loads, height, weight, and hemoglobin concentrations, comparing outcomes in Group 1 (which had already been treated in 1998) to Group 2 (which had not yet been treated, but was about to be phased into treatment). Hence, the timing of treatment following the collection of this data was central to the research design of Miguel and Kremer (2004), and much of the analysis in the original paper would not be possible without it.

The collection of parasitological data at the start of each year before treatment was also necessary to determine which drugs would be administered in each school, i.e., albendazole and/or praziquantel (based on the prevalence of geohelminths and schistosomiasis, respectively).

In fact, if we follow Davey et al.'s assumption on what constitutes a treatment observation to its logical conclusion, then all of the worm infection and health outcomes program estimates need to be discarded since, according to them, Group 2 schools are all already treatment schools by January 1st 1999, and thus the comparison between Group 1 and Group 2 is meaningless. Yet this does not make sense since no Group 2 schools were treated, nor was there ever any intention of treating them, in the early months of 1999. Rather, extensive data collection was carried out in all schools in the early months of 1999 precisely *because* Group 2 had not yet been phased into treatment, allowing for analysis of health impacts after one year of treatment.

There is no basis for the assertion in Davey et al. (2015) that schools were supposed to be phased into treatment at the start of each calendar year.

Methods: "School intervention status was not concealed from fieldworkers collecting outcome data."

As we explain in Section B.2.4 above, this would have been impractical if not impossible due to the cluster randomization design of the study. Furthermore, most data collection focused on objective, rather than subjective, measures, making subtle field worker data collection biases less of a concern.

Methods: "In accordance with our interpretation of the intention to treat of this complex intervention, school attendance observations of pupils in 1998 were assigned to the treatment condition in Group 1 and the control condition in Groups 2 and 3, and in 1999 observations were assigned to the treatment condition in Groups 1 and 2, and control condition in Group 3."

We note that this text was not present in the original version of this report as it was initially submitted for publication as part of the 3ie Replication Paper Series, nor was there any mention of this

recoding of the treatment variable in the replication authors' pre-analysis plan (Aiken et al., 2013). As we discuss above, the justification for recoding the treatment measure in this way is unfounded. Tables S5 and S7 of this Appendix present the primary results of Davey et al. (2015) – both the individual-level and cluster summary results – correctly defining treatment, and show substantial, highly significant, and robust impacts of deworming on school participation.

Methods: "Initial analyses identified an unexpected cluster level association between the level of school attendance and the total number of pupil observations performed, which was influenced by whether or not schools were involved in the SAP programme. To describe and investigate this association further we plotted the proportion of pupils observed as present in each school against the number of observations made in a school, stratified by year and by allocation Group, and fitted ordinary least squares regression lines."

The observed patterns of correlation between attendance and cluster size are not necessarily unexpected: there might be a correlation between the number of observations per schools (which is driven mainly by pupil population) and average attendance rates. School population might correlate with many different things, including school quality, local socioeconomic status, etc. The fact that such a correlation exists in no way affects the validity of the research design, as we describe in detail in Section B.2.2 above.

Davey et al. (2015) never clearly explain why either of these issues creates a problem for the analysis. For instance, school attendance may be correlated with school population when we look across schools. Larger schools may be richer (or poorer), or more or less isolated, etc. Finding this correlation is interesting but orthogonal to our understanding of treatment effects, and it does not undermine the research design.

Methods: "The primary outcome analyses were conducted in three steps that increased progressively in complexity to reflect the cluster allocated stepped wedge design of the trial. First, in each year the means of the school level summary outcomes for each Group were calculated, and also for each intervention arm. The latter were compared within years using the unpaired t test."

The cluster-level analysis presented in the left panel of Davey et al. (2015), Table 2 is not mentioned anywhere in the authors' pre-analysis plan. In fact, the authors did not even pre-specify that they would present intervention versus control statistics in the cluster summary table; they write: "*Summarize and display the outcomes clearly for each intervention arm in each year. For example, the proportion of children absent in the 25 schools in each group in 1998, and in 1999.*" (Aiken et al., 2013, p. 10, point 1). Instead, all of the analysis was to be carried out using "*individual-level analysis … using regression models with random effects*" (Aiken et al., 2013, p. 10, point 2). They call this their "primary analysis of school attendance". This pre-specified individual-level analysis corresponds to the results reported on the right hand panel of Davey et al. (2015), Table 2. This is also made clear in Aiken et al. (2013, p. 10) where they say:

> "*For the primary analysis of school attendance we will compare observations of attendance or non-attendance across treatment arms, within years. Each child, in each school, will have a number of observations that are either 'present' or 'absent' and*

coded as 1 and 0, respectively. Therefore, this analysis will use logistic regression to model the effect of treatment condition on the outcome at each observation. We will include a 'treatment' variable in the model that will take the value '1' if the child under observation was enrolled at a school receiving treatment in that year and '0' if the child was in a school not receiving treatment in that year. The primary result will be an odds ratio that a child is present between treatment and non-treatment arms."

Given that the cluster-level analysis was not pre-specified, we did not expect so much importance to be placed on these results. In particular, these results are featured in the "primary outcomes" table (Davey et al., 2015, Table 2), alongside individual-level pre-specified analysis, and are used by the re-analysis authors to make claims about the supposed robustness of the school attendance results. We believe that two decisions in particular related to this non-pre-specified cluster-level analysis are inappropriate, and we show that a standard approach to a cluster-level analysis generates a substantial and significant relationship between deworming treatment and school attendance.

First, we believe Davey et al.'s decision to present an unweighted cluster-level analysis (that implicitly weights each school equally, rather than each individual or each attendance observation) is non-standard. As we discuss in detail in Section B.2.3 above, cluster-level analysis weighted by either pupil observations or pupil population have meaningful interpretations, and these are standard analytical approaches. We show in Table S7 that either of these standard weighting methods suggests a substantial and highly statistically significant (P-value < 0.05) impact of deworming on school participation in the year-by-year analysis.

Second, there is no clear justification given for why Davey et al. (2015) chose not to present pooled estimates (accounting for a secular trend over time) in the cluster summaries, mirroring what they did in the individual-level analysis. As we describe in detail in Section B.2 above, pooling the years makes maximum use of the data available, providing analysis that is adequately powered to detect impacts. As we show in Table S57, the pooled results are statistically significant (P-value < 0.01) when either standard weighting approach is used, and even when the cluster-level analysis is unweighted but the treatment measure is correctly defined. It is only when the replication authors simultaneously make multiple analytical errors – in weighting observations, defining the treatment variable, and failing to pool both years of data – that they find results that are not statistically significant.

Methods: "In analyses that were not pre planned, we investigated the sensitivity of our school attendance results to the assumption of the intention to treat applying from the start of each year. This was based on information in the study timeline (Appendix 4) that indicated the drug component of the intervention was not delivered at the start of each year."

This so-called sensitivity analysis is essential. As we describe in Section B.2.1 above, in their main analysis Davey et al. (2015) recode the key treatment measure, assigning over 10,000 observations to a treatment condition when they were in fact not yet treated. The re-analysis authors did not raise any issues regarding the correct coding of the treatment measure in their re-analysis of the Miguel and Kremer (2004) do files (as presented in Aiken et al., 2015), nor did they explicitly mention this recoding in their pre-analysis plan (Aiken et al., 2013) or the original version of the present report that was submitted to 3ie for publication (Davey et al., 2014). Moreover, the justification added on to the present

version of the report misuses the "intention to treat" terminology, as we discuss in the main text to this Commentary.

There was never any intention to treat children at the very start of each calendar year in the Miguel and Kremer (2004) study, as we have mentioned. That would not have been possible given our research design, which required the collection of parasitological and anthropometric data prior to deworming treatment in each calendar year. Indeed, the analysis of health outcomes is only possible using the data that was collected from Group 1 and Group 2 individuals during the first three months of 1999.

This mis-classification of individuals has important implications for the analysis, as a comparison of Davey et al. (2015) Table 2 (using the miscoded treatment term) and Table 4 (the "Alternative Scenario 1" panel, which correctly codes the treatment term and makes maximum use of the data by not dropping the early visits in 1999 unnecessarily) shows. Specifically, in their incorrectly coded "primary" school participation analysis presented in the top right panel of their Table 2, the impact of deworming on school participation is not statistically significant in 1998 (P-value = 0.150), while the 1999 impact and the pooled impacts are both statistically significant (P-value = 0.044 and < 0.001, respectively). In contrast, the correctly coded so-called "sensitivity" school participation analysis presented in the center panel of Table 4 indicates that there are significant results for both years separately and pooled together (P-values = 0.036, =0.088, and <0.001, respectively).

Finally, Davey et al. (2015) incorrectly claim that there is no information on the timing of deworming treatment visits, but this data is available and has been shared with the replication authors, and confirms the timeline of data collection and deworming treatment described in Miguel and Kremer (2004).

Results: "The mean school size was similar in the three Groups, but the range was much larger for Group 2 (min 37, max 1,392)."

This outlier in school population among Group 2 schools is because two study sample schools (both Group 2 schools) were flooded in late 1997, and were not open during the 1998 school year. Most of these students ended up enrolling in another nearby PSDP school, which was a Group 2 school, temporarily swelling its enrolment. Most of the pupils returned to their original schools in 1999. Note that we followed a standard intention to treat (ITT) approach and continue to assign each pupil to her/his original school throughout the study, as noted in Miguel and Kremer (2004).

Results: "[School attendance] Data were available for 74% of pupils during conducted visits in year 1 and 86% in year 2, and within years the proportions were broadly similar between Groups."

Missing data in a multi-year longitudinal study on the order of 14 to 26% per data collection round is quite typical in field studies, especially in low income settings. There are many reasons for missing data, from lost paper copies (as the data collection was recorded on print-outs), information lost when the sheets were transferred to the data entry team, data entry errors, and so on. There were also many cases (that we can personally recall from fieldwork) where the field team only had time to collect namelist information for a subset of grades in a particular school, because they simply ran out of time or something else came up that forced them to leave the school. For that reason, too, there will be "missing" observations for some pupils even on days when other students in a school had their

attendance observed. As long as these errors are occurring at approximately the same rate in treatment and control schools, they should not induce systematic bias. Indeed, the proportions missing are quite similar and not statistically significantly different across the three intervention groups, which is reassuring.

Results: "In year 1, but not year 2, there were several schools that had more than 95% attendance."

This pattern makes a lot of sense. The sample in 1998 was selected on those who were enrolled in school, so we would expect quite high attendance in 1998. By 1999, many students in all groups dropped out. Dropout rates in primary school are currently quite high in Kenya, and were even higher in 1998 and 1999. So the fact that no school had over 95% attendance in 1999 is not surprising either.

Results: "The means of the cluster-summaries of school attendance for intervention schools in year 1 and year 2 were both higher than the corresponding control school means, but there was no statistical evidence for the differences (year 1 difference +5.48%, 95%CI -1.48–12.44, t-test p=0.12; year 2 difference +2.16%, 95%CI -3.39–8.27, t-test p=0.48) (see Table 2). These cluster-level risk differences were equivalent to Odds Ratios (OR) of 1.78 (year 1) and 1.21 (year 2)."

As noted above, the cluster-level analysis was not pre-specified – the replication authors did not suggest that they would present treatment versus control group statistics at the cluster summary level (Aiken et al., 2013). Furthermore, and more importantly, this analysis is presented in an unusual way, weighting each school equally rather than weighting either by number of observations or by pupil population, and not pooling the data to make use of the research design and maximize statistical power.

Creating a "school-weighted impact estimate" is not of general interest; the "individual-weighted impact estimate" is of general interest, both intellectually and in terms of public policy, when we care about health or education outcomes in a *population*, and weighting by attendance population has attractive properties in terms of improving statistical precision. Davey et al. (2015) provide no rationale for presenting estimates which weight all schools equally, and we find this approach unattractive in a setting with such large differences across schools in pupil population, with seven-fold differences in populations across schools in some cases. If we do consider the cluster summary analysis, but weight the clusters with any standard weighting approach (either by number of observations, or by population), we find large, positive and significant impacts of deworming on school attendance, for each year separately or pooled for both years (see Table S7, above).

We are also unsure of why the pooled 1998 and 1999 results are not shown here. Table S7 shows that there are large effects with much greater statistical precision in that case, too. I.e., no matter how you do the analysis, if you pool data across both years there is always a large, positive and statistically significant impact of deworming on school attendance in this data. Davey et al. (2015) do show here that looking at 1998 and 1999 separately, and using non-standard weighting, using a specification that was not pre-specified, does sometimes lead to only marginally significant results.

If we focus on the individual-level results presented in the right-hand panel of Davey et al. (2015), Table 2, we observe statistically significant improvements in school attendance due to deworming in 5 out of the 6 estimates presented. So far, we see that in both 1998 and 1999, there are large positive point estimates in this analysis, which are sometime statistically significant on their own. But of course each of these only uses a piece of the data for the study as a whole. In the limit, we could

analyze data separately month by month (or week by week) and none of the individual treatment effect estimates would be statistically significant. But that would not imply that there is no impact of the study using all of the data at hand. When the authors present the results cut up year by year, they owe it to the reader to mention that each of these is a subset of the data, and thus is underpowered relative to the overall data set and research design. I.e., a not significant effect within a subset of the data does not constitute meaningful evidence for a "non-effect".

When the prospective research design is utilized, there is a large, positive, and statistically significant estimated impact of deworming on school attendance. This holds with and without controls (age and SAP), and holds for either the full sample or the eligible population sample, so is quite robust, as shown in Figure 2 of our Commentary. It is not surprising that when you look at each year separately (i.e., using only half the data, and not exploiting the full research design, with Group 2 changing treatment status) that statistical precision falls somewhat – although in the pre-specified analysis on the eligible subsample each year (1998, 1999) on its own is significant at either 95 or 90% confidence. Given this, we disagree with Davey et al.'s interpretation of the evidence on deworming impacts on school attendance.

Results: "The key results of a sensitivity analysis exploring effects of the handling of the treatment condition on school attendance results are shown in Table 4 (full results in Appendix 5). In scenario one, 11,588 attendance observations performed at the start of 1998 were excluded, and 31,404 observations occurring during the first two visit-periods in year 1999 were handled as 'year 1' observations. In this scenario, the cluster summary mean differences were slightly larger in both 'years', and had smaller p-values than in our pre-specified analysis. In adjusted regression models, the OR for 'year 1' was slightly closer to the null, whilst the result for 'year 2' was virtually unchanged. The adjusted combined-year logistic regression OR was larger with similarly strong evidence. In scenario two, 11,588 observations at the start of 1998 were excluded, as well as the 31,404 observations during the first two visits in 1999. In comparison with our prespecified primary analysis, the year-specific results were largely unchanged, with the cluster-summary mean different in year 2 being slightly larger. For the combined-year logistic regression analysis, the adjusted OR was larger than in the pre-specified analysis."

In scenario two, which drops the observations corresponding to the miscoded periods of treatment, actually dropping data led to impacts that are generally larger in magnitude (in 6 out of 8 specifications, also considering the unadjusted odds ratio results). Furthermore, the cluster summary results are closer to significance (from P-values of 0.121 and 0.483 to P-values of 0.109 and 0.150, for 1998 and 1999 respectively).

In scenario one, which makes full use of the data collected and correctly classifies treated individual, the results are much stronger. Impacts of deworming on school participation in 1998 come across in both the cluster summary analysis (P-value = 0.056) and the individual-level analysis (P-value = 0.036), where these effects had been non-significant in the miscoded analysis. Overall, the results in the center panel of Davey et al. (2015) Table 4 suggest positive and statistically significant impacts of deworming on school participation in 4 out of 5 models, with the odd case being an unweighted cluster mean for 1999; our Table S7, above, shows that even that result is statistically significant (P-value < 0.5) when an appropriate weighting approach is applied.

Results: "There was an unexpected association between the number of school-attendance observations in a school and the school's mean attendance which depended on the intervention arm. As indicated by the slope of the lines in Figure 3, in 2/3 intervention Group-years school attendance was higher in schools where more observations were undertaken. Conversely the opposite relationship was seen in all three of the control Group years."

As discussed in detail in section B.2.2, there are no statistical significant differences in this relationship across treatment groups over time.

Discussion: "The stepped wedge design appeared to exacerbate the influence of the unexpected patterns in the data. The combined years model estimated an effect that was higher than either of the two year specific effects. "

This pattern is a natural possibility in the analysis of panel data using a stepped wedge design. For instance, different intervention groups of schools are likely to start out with slightly different school attendance levels at baseline simply due to sampling variation. Stepped wedge analytical designs are able to account for these minor baseline differences, and the additional statistical power they provide is a major strength of the analysis in Miguel and Kremer (2004). This analytical approach may in general lead to pooled estimates that differ from each individual cross-sectional estimate.

Discussion: "Regarding generalizability of the intervention effect, worm burden would need to be high for schools to be eligible for the same level of treatment. Burden may also affect the magnitude of effects: low burden may explain why a large trial in India evaluating the effect of de-worming and vitamin A supplementation on pre-school mortality found no effect (11). Without clear articulation of a causal pathway it is unclear what other factors would need to be similar in other settings to generalize the results of this study."

The Awasthi et al. (2013) study referred to here is not relevant since it does not estimate impacts on educational outcomes, and thus does not speak to the debate at hand. As noted in Section B.2.6 above, there is growing evidence from multiple cluster randomized studies in areas with widespread worm infections that deworming treatment leads to substantial gains in both educational and labor market outcomes in the medium to long-run (Baird et al., 2014; Croke, 2014; Ozier, 2014).

Appendix 1: "We handled missing-ness in the outcome data on pupil attendance by applying the following steps sequentially. First, we removed from the dataset any data that had been collected during a visit that was not scheduled according to the visit plan."

We note that many of the seemingly "stray" observations for particular schools in the original data were for students who transferred across schools, and hence were picked up in other schools that had a different data collection schedule than their original school. Right now this data is portrayed as "bad" data that is related to missingness, data quality problems, etc. In reality, this is a major strength of the data collection: few education datasets directly observe pupil attendance in school at all (instead depending on school registers of unknown reliability), and fewer successfully track most pupils across schools over multiple years. That is why we include these observations in our analysis. Dropping them

does not make a major difference to the results (as shown by the re-analysis authors), but we still believe Davey et al's (2015) characterization of this data is inappropriate.

B.4 Additional References

Ahuja, Amrita, Sarah Baird, Joan Hamory Hicks, Michael Kremer, Edward Miguel, and Shawn Powers. (2015). "When should governments subsidize health? The case of mass deworming", National Bureau of Economic Research Working Paper #21148.

Aiken A, Davey C, Hayes R, Hargreaves J. (2013). "Deworming schoolchildren in Kenya - Replication plan", International Institute Impact Evaluation (3ie) website.

Aiken, A, Davey, C, Hayes, R and Hargreaves, J. (2014). "Re-analysis of health and educational impacts of a school-based deworming program in western Kenya: a pure replication", 3ie Replication Paper 3, part 1. Washington, DC: International Initiative for Impact Evaluation (3ie).

Aiken, A, Davey, C, Hargreaves, J, and Hayes, R. (2015). "Re-analysis of health and educational impacts of a school-based deworming program in western Kenya: a pure replication", *International Journal of Epidemiology,* forthcoming.

Alderman, H., J. Konde-Lule, I. Sebuliba, D. Bundy, and A. Hall. (2006). "Increased weight gain in preschool children due to mass albendazole treatment given during 'Child Health Days' in Uganda: A cluster randomized controlled trial", *British Medical Journal*, 333: 122-6.

Alderman, Harold. (2007). "Improving nutrition through community growth promotion: Longitudinal study of nutrition and early child development program in Uganda", *World Development*, 35(8): 1376-1389.

Awasthi, Shally, et al. (2013). "Population deworming every 6 months with albendazole in 1 million pre-school children in north India: DEVTA, a cluster-randomized trial", *Lancet*, 381(9876): 1478-1486.

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel. (2014). "Worms at Work: Long-run impacts of a child health investment", unpublished working paper, University of California, Berkeley.

Bleakley, Hoyt. (2007). "Disease and Development: Evidence from Hookworm Eradication in the American South." *Quarterly Journal of Economics*, 122(1):73-117.

Croke, Kevin. (2014). "The long run effects of early childhood deworming on literacy and numeracy: Evidence from Uganda", unpublished working paper, Harvard University.

Davey, C, Aiken, A, Hargreaves, J, and Hayes, R. (2015). "Re-analysis of health and educational impacts of a school-based deworming program in western Kenya: a pure replication", *International Journal of Epidemiology,* forthcoming.

Davey, C, Aiken, A, Hayes, R and Hargreaves, J. (2014). "Re-analysis of health and educational impacts of a school-based deworming program in western Kenya: a statistical and scientific replication", 3ie Replication Paper 3, part 2. Washington, DC: International Initiative for Impact Evaluation (3ie).

Eble, Alex, Peter Boone, and Diana Elbourne. (2013). "Risk and evidence of bias in randomized controlled trials in economics", *CEP Discussion paper #1240*.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. (2009). “Many Children Left Behind? Textbooks and Test Scores in Kenya”, *American Economic Journal: Applied Economics*, 1(1): 112-35.

Hicks, Joan Hamory, Michael Kremer, and Edward Miguel. (2014a). “Estimating deworming school participation impacts and externalities in Kenya: A Comment on Aiken et al. (2014)”. Original author response to 3ie Replication Paper 3, part 1. Washington, DC: International Initiative for Impact Evaluation (3ie).

Hicks, Joan Hamory, Michael Kremer, and Edward Miguel. (2014b). “Estimating deworming school participation impacts in Kenya: A Comment on Aiken et al. (2014b)”. Original author response to 3ie Replication Paper 3, part 2. Washington, DC: International Initiative for Impact Evaluation (3ie).

Kremer, Michael, and Edward Miguel. (2007). “The Illusion of Sustainability”, *Quarterly Journal of Economics*, 112(3), 1007-1065.

Miguel, Edward and Michael Kremer (2004). “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities.” *Econometrica*, 72(1), 159-217.

Miguel, Edward and Michael Kremer (2014). “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, Guide to Replication of Miguel and Kremer (2004)”, *CEGA Working Paper #39*.

Ozier, Owen. (2014). “Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming”, *World Bank Policy Research Working Paper WPS7052.*