Estimating deworming school participation impacts in Kenya: A Comment on Aiken et al. (2014b)[*]

Joan Hamory Hicks, University of California, Berkeley CEGA
Michael Kremer, Harvard University and NBER
Edward Miguel[†], University of California, Berkeley and NBER

December 2014

Abstract: Aiken et al. (2014b) draw the conclusion that the evidence for a relationship between deworming and school attendance is "weak" based on two fundamental errors in their data analysis. First, the authors redefine treatment to include pre-treatment control periods. Second, while the original research design was based on a stepped-wedge analysis that was adequately powered, the re-analysis authors undertake a clearly under-powered alternative analysis which ignores the time series element of the data, and then splits the cross-sectional analysis into two separate components, each of which has inadequate power. Examining the fully powered analysis, they report: *"In a fully-adjusted logistic regression model making maximum use of the data available, there appeared to be strong evidence of an improvement in school attendance"*. If either error is corrected, deworming significantly increases school attendance under the full range of statistical analyses considered by Aiken et al. Their analysis also underestimates the impact of deworming on school attendance by neglecting violations of the SUTVA assumption generated by transmission of worm infection to nearby schools (Miguel and Kremer 2004). We also respond to concerns raised by Aiken et al. regarding data collection processes and blinding.

[*] Acknowledgements: We thank Kevin Audi, Evan DeFilippis, Felipe Gonzalez, Leah Luben, and especially Michael Walker for excellent research assistance. All errors remain our own.
[†] Corresponding author.

1 Executive Summary

Miguel and Kremer (2004) evaluates 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, it showed 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 shows 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. (2014a, b) re-analyze the data from Miguel and Kremer (2004) through the new replication process established by the International Initiative for Impact Evaluation (henceforth, 3ie). In a separate paper that composes the first part of their replication exercise, Aiken et al. (2014a) 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 a previous note (Hicks, Kremer, and Miguel, 2014) and an earlier paper (Miguel and Kremer, 2014).

This note is a response to Aiken et al. (2014b), which re-analyzes the original data after changing the definition of treatments splitting the data into subsamples, re-weighting, and various other adjustments. The re-analysis authors argue that evidence for school participation impacts is weak, but this conclusion is based on a series of errors in their analysis of the data. 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. (2014a); however, Aiken et al. (2014b) 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 Aiken et al. (2014b) 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 Aiken et al. (2014b) 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 extremely underpowered (Aiken et al., 2013, p. 7; Aiken et al., 2014b, 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 (Aiken et al., 2014b, Table 4).

Aiken et al. also raise a number of concerns about the data and analysis (which we argue below are spurious) and adopt other changes in statistical procedures to address them, most importantly, re-weighting the data. One central finding of the current note 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.[1] 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 Aiken et al. (2014b) 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: *"In a fully adjusted logistic regression model making maximum use of the data available, there appeared to be strong evidence of an improvement in school attendance."* (Aiken et al., 2014b, p. 26).

Nonetheless, it is worth noting and responding to some of the concerns raised by Aiken et al. (2014b). 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. Aiken et al. 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 Aiken et al. (2014b) is that the results in Miguel and Kremer (2004) are not robust to different analytical approaches; they write: *"We found that the evidence that the intervention improved school attendance differed according to how we analysed the data."* (Aiken et al., 2014b, p. iv). We strongly disagree with this conclusion.

[1] The findings in Miguel and Kremer (2004) that receive by far the most attention in Aiken et al. (2014b) are the impacts of deworming on school participation. This note focuses almost entirely on this issue, although we also discuss the evidence regarding other deworming impacts at several points.

In order to assess the purported "sensitivity" of the school participation results to different analytical assumptions, in Table 1 (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 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 Aiken et al. (2014b)); (ii) different samples of pupils (the full sample preferred by Miguel and Kremer (2004) and the sample eligible for deworming treatment as preferred by Aiken et al. (2014b)); (iii) regression models unadjusted for covariates and adjusted for covariates (the latter of which is preferred by Aiken et al. (2014b)); (iv) use two different approaches to weighting observations (weighting each attendance observation equally, as in Aiken et al. (2014b) 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 Aiken et al. (2014a, b) 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 4 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 1 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 1 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 Aiken et al. (2014b) 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 1 run strongly counter to the unfounded claim in Aiken et al. (2014b) that the relationship shows "sensitivity" depending on how the data is analyzed.

Section 2 of this note explores these key points in detail, and addresses the main claims raised in Aiken et al. (2014b). Section 3 summarizes, and discusses the current state of evidence on the educational and economic impact of deworming. A point-by-point treatment of Aiken et al. (2014b) is contained in the final section.

The 3ie replication process differs in important ways from the standard research community-led peer-review process in academic journals. We have been explicitly instructed by 3ie staff not to discuss our experiences with the replication process at any length in this note, including our views on the weaknesses of their current system and the review

standards they employ. We do have a number of observations based on our experience, as well as suggestions for how the process could be improved, and we look forward to sharing these insights with 3ie staff and with the broader research community in the future.

Table 1: Deworming impacts on school participation (1998-1999)

Analytical approach: Data and variable construction:	Random-effects logistic regression Aiken et al. (2014b) (1)	Random-effects logistic regression Original (2)	Linear regression Aiken et al. (2014b) (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. (2014a). The Aiken et al. (2014b) data contains several additional modifications regarding the inclusion of transfer students, and assumptions on missing data, which are described in Aiken et al. (2014b), 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 Aiken et al. (2014b); 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 Aiken et al. (2014b) 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 Aiken et al. (2014b); 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.

2 Technical Response to Aiken et al. (2014b)

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 their report. 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 1 (column 1) above. Both are positive and statistically significant, and they are also very large in magnitude.

It is worth noting up front that Aiken et al. (2014b) 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. (2014a), all of the estimators used in Aiken et al. (2014b) are downward biased, yielding lower bounds on true deworming treatment effects.

In this section, we explore key aspects of the analysis presented in Aiken et al. (2014b) in detail, and address the main concerns raised by the re-analysis authors.

2.1 Miscoding of the treatment measure in Aiken et al. (2014b)

In the process of studying the school participation analysis presented in Aiken et al. (2014b) after it had been initially 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 whatsoever of this important recoding of the treatment variable in their report (the version of Aiken et al., 20014b that was originally submitted to 3ie for publication) or in their pre-analysis plan (Aiken et al., 2013), and they did not object to the original coding in Miguel and Kremer (2004) as it was employed in their "pure replication" report (Aiken et al., 2014a). However, subsequent to our bringing this important issue to the attention of the authors, they added text to their report justifying this coding choice (the last two paragraphs of Section 2.3 of their report, and the second paragraph of Section 2.5), and added a new table of results (their Appendix 7) with associated discussion.

The re-analysis authors 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). Aiken et al. (2014b) 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 4 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. The replication authors' decision to impose their own notion of what the "planned" timeline of data collection and deworming treatment should have been, which runs counter to reality as described in the published paper (Miguel and Kremer, 2004), is extremely puzzling to us.

In fact, if we follow the re-analysis authors' 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 is nonsensical 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.

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 Aiken et al. (2014b) argue contain the "weakest" evidence. In particular, we show this for both the cluster summary and individual-level analyses. The individual-level results presented above in Table 1, columns (2) and (4), already correct this miscoding of the treatment term, as we mention above. Table 3, 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 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 Aiken et al. (2014b) analysis that weights each school equally, 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).

The bottom line is that the re-analysis authors make an unfounded and incorrect decision to recode the treatment variable in their analysis, despite the lack of any evidence to suggest that deworming treatment was supposed to have been introduced at the start of each calendar year. 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.

2.2 Aiken et al. (2014b) concern #1: possible relationship between number of observations and attendance

The main concerns raised by the re-analysis authors in Aiken et al. (2014b) appear to revolve around data collection and data quality, and these are summarized in their sections 4.2 and 4.3 and their Figure 3. Their basic claim is that they believe 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. It is easy to show in our data that 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, as the re-analysis authors recognize. 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 re-analysis authors' argument 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. They 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:

> *"The Group 2 comparison across years also reduces the level of between-cluster variation and may therefore have greater statistical power. The increase in power ordinarily represents an advantage of the stepped-wedge design. We are concerned about the reliability of this combined estimate of effect across the two study years, because it depends strongly on the 'horizontal' comparison of outcomes between year 1 (1998) and year 2 (1999) in Group 2. Figure 3 shows that there was probably a bias towards more pupil observations in schools with low attendance in year 1 (1998) (control condition), while we saw the opposite bias in year 2 (1999) (intervention condition). This would potentially lead to overestimation of the effect of the intervention on attendance, particularly in an analysis weighted, in part, by the number of observations."* (p. 27)

This is the central critique of the Miguel and Kremer (2004) data and analysis in the Aiken et al. (2014b) report, 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 were surprised by this assertion for two reasons. First, we were involved in the data collection and know that approximately equal numbers of visits were made to all types of schools, with the data collection protocol explicitly "balanced" across the treatment and control groups at all times. There was absolutely no explicit or implicit "bias" towards visiting schools that would help support an ex ante research hypothesis.

Second, we were surprised that this assertion was made when no statistical test was provided by the replication authors 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" 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 Aiken et al. allude to is straightforward to run with the data in hand: one simply needs to 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 2 (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 Aiken et al. (2014b) 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 Aiken et al. (2014b) 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 any differential correlation between the number of observations and school participation rates across treatment and control schools, nor that this relationship changes significantly 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, the re-analysis authors' assertion that it is inappropriate to pool data from 1998 and 1999 and utilize the project's research design appears entirely unfounded.

Table 2: 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.

2.3 Aiken et al. (2014b) 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, but which are not apparent to us), the suggested remedy proposed by Aiken et al. (2014b) – 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. Aiken et al. (2014b) 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 justifiable rationale for such a decision. It is worth noting that the approach of weighting each school equally was never mentioned in the replication authors' pre-analysis plan (Aiken et al., 2013), where they consistently emphasize individual level analysis.

Given the importance that Aiken et al. (2014b) attach to their cluster summaries analysis (which was not pre-specified) in driving the implication that the results in Miguel and Kremer (2004) are sensitive to analytical choices, we decided to explore the analysis they present in the top left panel of their Table 4. 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 3, below, the cluster summaries analysis with this standard weighting approach generates results very similar to the replication team's own individual-level random effects logistic analysis, namely 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 show these results weighting by the number of attendance observations (which we feel is appropriate given the lack of evidence above on the purported "excessive" observations in high participation treatment schools), and weighting each school equally, as in Aiken et al. (2014b).

To take a step back and summarize the argument in Aiken *et al.* (2014b), 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 2 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 evidence for the purported data "problem" that forms the centerpiece of the argument in Aiken et al. (2014b). 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 seems 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 3 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 1 and 3 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 the replication authors' 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 Aiken et al., 2014b or by Miguel and Kremer, 2004).

2.4 Aiken et al. (2014b) concern #3: data collection quality and blinding

A leading concern for Aiken et al. (2014b) is the fact that the study was not "double-blinded", raising the possibility of both "performance bias" and "detection bias" in the terminology of biomedical trials. The re-analysis authors raise these concerns in their section 4.9 (p. 30), with a particular focus on wanting to verify that "*fieldworker data-collection practices were the same in all schools*." They go on to write that "*In practice, this is hard to verify retrospectively, so it is a possibility that there were, consciously or unconsciously, variations in data collection between groups.*"

We were surprised to read this for two reasons. First, as we noted in the original Miguel and Kremer (2004) paper and discussed directly with the re-analysis authors during this replication process, there was emphasis on balanced data collection procedures and timing in all three groups of schools, and the professional field staff were extensively trained in appropriate and consistent data collection procedures. In fact, neither of the two key outcome measures – worm infection rates and school attendance – are subjectively measured questions that ask the respondent or the enumerator to make a judgment call. There is simply not much room for unconscious bias to enter into the collection of these variables. Rather there would have to be malfeasance in the data collection process, i.e., either manipulation of the parasitology lab results that generated the worm counts data, or enumerators who made a decision to mark a student "present" who was absent (or vice versa) in some systematic way across treatment groups.

Second, we were again surprised that Aiken et al. (2014b) provide no statistical evidence to corroborate this assertion about possible problems in the data collection. If there were systematic discrepancies in the nature of data collection across treatment groups, the re-analysis authors could illuminate these patterns. In fact there are multiple pieces of evidence suggesting that data collection was in fact carried out in an even-handed and balanced way across the treatment and control groups. The finding shown above in Table 2 that the timing of visits across schools was the same across treatment groups is consistent with balanced data collection procedures. The fact that pupil school transfer rates, attrition rates, and baseline characteristics are all "balanced" across the three program groups (which the re-analysis authors confirm in Aiken et al., 2014a) is further evidence that data collection was carried out in an even-handed way for the entire sample.

The finding of externality effects both within schools and across schools (as confirmed in Aiken et al., 2014a) is also incompatible with the re-analysis authors' claims about potentially biased data collection by enumerators. The study of externalities was not central to the original research design of the Miguel and Kremer (2004) study, nor was it an issue that was ever discussed with the field data collection team during 1998 and 1999. It is simply inconceivable that biased data collection could have generated the results that there are positive externalities within 0-3 km of treatment schools, and the same holds for the measured health and school participation effects among untreated children within treatment schools. Thus the extensive evidence for positive deworming treatment externalities taken together provides further evidence against biased data collection procedures.

Table 3: 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 Aiken et al. (2014b)						
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 Aiken et al. (2014b), 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.

A major reason that the school participation evidence from the Miguel and Kremer (2004) study is viewed as "low quality" by the re-analysis authors is because the deworming intervention was publicly revealed (i.e., non-blinded). The unsubstantiated (but impossible to disprove) assertion in Aiken et al. (2014b) – that data from non-blinded studies is inherently at risk of "bias" – has dramatic implications not just for this study but for the rapidly expanding body of new experimental evidence based on real-world programs, in which few if any studies are "blinded". In fact, in their conclusion (p. 38), Aiken et al. (2014b) make it explicit that the lack of "blinding" is a major reason why they have doubts about the study, when they write: *"Even if the result showing the strongest effect of the intervention on school attendance were accepted... [o]ne possible explanation is that behavioural changes unrelated to drug treatment occurred in this unblinded study that led to the observed changes in school attendance."*

The deworming program may indeed have led to behavioral changes (i.e., changes in family or school practices) that in turn affected schooling outcomes. One would 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.

Rather Aiken et al. (2014b) appear concerned that receiving drug treatment changes behavior due mainly to placebo effects. The re-analysis authors advance this claim once again without providing any 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. This discussion does not feature in Aiken et al.'s analysis.

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. Aiken et al (2014b) choose not to refer to these results.

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 Aiken 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 not only among treated pupils but also among untreated classmates and students in nearby schools remains hypothetical, we believe that the use of cluster randomization is appropriate.

There are also 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 Aiken 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, essentially due to disciplinary differences.

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. Aiken et al. (2014b) are holding the Miguel and Kremer (2004) study to 2014 standards in public health, rather than 1997 standards in public health, let alone to 1997 standards in economics. (As this paper was one of the first published field experiments in development economics, there was no "standard practice" around these issues within economics at the time.)

A related point has to do with the 17 year (1997 to 2014) time lag between the setup of the Kenya deworming project and the Aiken et al. (2014a, b) replication reports. 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 Aiken et al. team have benefitted from extensive 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 many of these in their reports). We believe readers should keep these issues in mind when the Aiken et al. team 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 17 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 for all treatment groups and over time. These patterns, our experience designing a valid and balanced field data collection procedure in Kenya, and the lack of any statistical evidence for biased data, together imply that the assertions in Aiken et al (2014b) are unfounded.

2.5 Additional concerns noted in Aiken et al. (2014b)

One concern the replication authors briefly mention is missing data, but they conclude (in section 4.6, p. 29) that: *"As the extent of missingness in attendance data was similar in each of the groups, we believe that this risk [of bias] is low."* 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 multi-year panel (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, providing further confidence in the validity of the experimental design and the data collection procedures.

3 Discussion and conclusion

To summarize, we discuss the results in Aiken et al. (2014b), and argue that their statistical evidence is overwhelmingly consistent with the conclusions in Miguel and Kremer (2004). In particular, their statistical evidence provides strong evidence 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 1 and Table 3), 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.'s 2013 pre-analysis plan. In Sections 2 and 4, we also respond in detail to a range of other 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 own results related to school participation, the re-analysis authors write:

> *"In a fully adjusted logistic regression model making maximum use of the data available, there appeared to be strong evidence of an improvement in school attendance. However, the size of the point estimates and the strength of the evidence were not consistent in the analytic steps progressively building up to this fully adjusted model. That is, we found no evidence of effect with cluster summaries, some evidence with individual analysis stratified by year and a larger point estimate of effect when both years were combined than we found in either individual year. This inconsistency, as well as other concerns related to the quality of data and an unexpected pattern of correlations in the observations, raises uncertainty about the reliability of the fully adjusted result."* (Aiken et al., 2014b, p. 26)

Thus, in their own words, the full individual-level model provides strong evidence of an effect. Our Table 1 (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 Aiken et al. (2014b) 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. 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 completely robust to doing so.

The other main concerns raised in Aiken et al. (2014b) relate to the fact that the study was "non-blinded", and that the study does not conform to current 2014 reporting standards in public health research. We discuss a range of patterns in the data that suggest data collection was not affected by enumerator biases or placebo effects, including the strong externality treatment effects among children who never received treatment. We also argue that cluster randomization is necessary to pick up epidemiological externalities and that it may actually be logistically impossible to carry out a truly double-blinded cluster randomized deworming study, given how the salient side effects of treatment (namely, the visible expulsion of worms from the body). It is also worth emphasizing that the critique that unblinded data is inherently "low quality" applies equally to nearly all recent economics and social science field experiments, not just to the Miguel and Kremer (2004) study, and we believe that researchers in these fields would join us in rejecting this simplistic characterization.

The re-analysis authors also use their replication reports (Aiken et al., 2014a, b) as an opportunity to 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.

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 subsidies on labor hours are sufficiently large that the net present value of government revenue generated by deworming subsidies exceeds the cost of the subsidies, creating an "expenditure Laffer effect". In the preferred estimate, each additional $1 in child deworming subsidies increases the net present value of government revenue by $13.

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.

4 Point-by-point treatment of Aiken et al. (2014b)

This section provides detailed, point-by-point responses to points raised in Aiken et al. (2014b). 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.

Page iv, Results: "We found unexpected patterns in the school-attendance data, including a correlation between the amount of attendance data from a school and the level of attendance."

As we discuss at length in section 2.2 above, there is no statistical evidence that this pattern exists in the data.

Page iv, Results: "In cluster-summary analysis, neither school attendance nor examination performance differed between arms in either study year. (School-attendance risk differences: 1998 5.48, 95 per cent CI −1.48–12.44, p = 0.121; 1999 2.16, 95 per cent CI −3.39–8.27, p = 0.483. Examination-performance mean difference: 1998 −0.109, 95 per cent CI −0.332–0.115, p = 0.336; 1999 −0.028, 95 per cent CI −0.228–0.171, p = 0.777.) We found some evidence of improvement in age-adjusted regression models for each year (adjusted OR 1998 1.48, 95 per cent CI 0.88–2.52, p = 0.15; aOR 1999 1.23, 95 per cent CI 1.01-1.51, p = 0.04) but not for examination performance. When we combined data from both study years in an observation-level model, the effect on school attendance was stronger than in either year (aOR 1998+1999 1.82, 95 per cent CI 1.74–1.91, p<0.001), but it had no effect on examination performance."

It is unusual to emphasize 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 at all 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 abundant evidence that deworming led to large, positive and statistically significant impacts on school participation across literally dozens of regression specifications, as shown in Table 1 and Table 3 (above).

Page iv, Results: "We found evidence of reduction in hookworm and roundworm infections but not in schistosomiasis or whipworm."

These results confirm what was found in Miguel and Kremer (2004), and are not surprising. 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. Also, the baseline rate of "moderate-to-heavy" whipworm infections is somewhat lower than for the other geohelminths.

Page iv, Discussion: "We found that the evidence that the intervention improved school attendance differed according to how we analysed the data."

As we detail extensively in Sections 1 and 2 above, we are puzzled by and 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 (Aiken et al., 2014b, Table 4), with slightly different samples (i.e., all children, or just those targeted with deworming; their Appendix 4), and even diluting the treatment effect by mis-defining the treatment measure (their Table 4 versus Appendix 7). So in fact the evidence presented in Aiken et al. (2014b) 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 inferior to the approach used in Miguel and Kremer (2004) and in the pre-analysis plan of Aiken et al.

Page iv, Discussion: "Our inability to review the sampling strategy guiding data collection and the potential for bias in measurement procedures necessitate caution in interpreting these results."

We discuss this issue at length in section 2.4 above. In particular, we make the point that the re-analysis authors provide no evidence for biased measurement procedures, and

in fact there is evidence that data collection across the three program treatment groups was carried out in an even-handed and professional manner.

Page iv, Conclusion: "These data provide weak evidence that a school-based drug-treatment and health-education intervention improved school attendance and no evidence of an effect on examination performance."

As we detailed previously, we are puzzled by and strongly disagree with this assessment of the school participation results, which appears to be driven by the combination of underpowered design and misclassification of treatment status. All analyses that make appropriate use of the prospective experimental design, and that appropriately define the treatment variable, imply that there are substantial, and highly statistically significant (P-value<0.01) impacts of deworming on school participation.

Page vii, Glossary definition for 'indirect effect': "The difference between the outcome in an individual not receiving the intervention in a population with an intervention programme and what the outcome would have been in that individual in a comparable population with no intervention programme."

This is incorrect, as externalities can also accrue to the treated.

Page x: "The results from the pure replication are summarised as follows [in Table 1], with results derived from both study years unless otherwise annotated. Effects that we found to be beneficial and significant in the pure replication are shaded."

We do not agree with several of the results reported in Table 1 of Aiken et al. (2014b). Hicks, Kremer, and Miguel (2014) details how the "indirect effect: between school" and "overall effect" are miscalculated by the re-analysis authors for all outcomes. Corrected calculations result in estimated between-school indirect effect and overall effect sizes that are highly statistically significant for moderate-to-heavy worm infections and school participation (P-value < 0.05 and < 0.01, respectively), and overall effect sizes that are similar to or larger than those reported in Miguel and Kremer (2004).

Page x: "We have not examined the indirect between-school effect (or the overall effect that includes this) in this statistical and scientific report for the following reason: in our pre-analysis plan, we stated that we would investigate the between-school indirect effects using 'the same analytic approach as described in the original paper' (Aiken et al. 2013). In our pure replication report (Aiken et al. 2014), we reanalysed the between-school indirect effects according to precisely the methods used in the original study, and we recorded our results there, reproduced in the column headed 'indirect effect: between school'. Therefore, we have already fulfilled our stated intentions with regards to these types of effects and have not pursued them further."

We were surprised to see no inclusion of the externalities in analysis. The re-analysis authors' own pre-analysis plan 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. (2014a) find externalities on worm prevalence within schools and up to 3 km away. Yet Aiken et al. (2014b) focus 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. (2014a), all of the estimators used in Aiken et al. (2014b) are downward biased, yielding lower bounds on true deworming treatment effects.

Page xi: "Since the original authors did not make their theory of change explicit, we cannot know whether our theory of change differs from theirs."

This claim is a mischaracterization of Miguel and Kremer (2004), which includes a lengthy discussion of the likely mechanisms (i.e., "theory of change") underlying the education results.

Page 3: "We aimed to analyse the trial using the principle of 'intention to treat'. An intention-to-treat approach compares outcomes between clusters (for example, schools) randomly allocated to different treatment conditions irrespective of whether treatment was, in practice, actually implemented or adhered to. Commonly, the intended treatment is described in a protocol, while actual treatment received by both treatment and control groups may be described post hoc in the results. Often, some form of 'per protocol analysis' focused on comparing those that did and did not receive the intended treatment is also conducted, although the intention to treat is typically considered the primary analysis as it is both unbiased by selection into treatment condition and the comparison that is most likely to reflect expected outcomes under real-life implementation.

We inferred from the original paper, in the absence of a protocol, that the combined educational and drug-treatment intervention package was intended to be delivered from the start of each year. This inference was based on the statement that, 'Due to ICS's administrative and financial constraints, the health intervention was phased in over several years. Group 1 schools received free deworming treatment in both 1998 and 1999, Group 2 schools in 1999, while Group 3 schools began receiving treatment in 2001. Thus in 1998, Group 1 schools were treatment schools, while Group 2 and Group 3 schools were comparison

schools, and in 1999, Group 1 and Group 2 schools were treatment schools and Group 3 schools were comparison schools' (p.165). The paper also states, 'In what follows, "treatment" schools refer to all twenty-five Group 1 schools in 1998, and all fifty Group 1 and Group 2 schools in 1999' (p.170). We note that when reporting the results of the analysis, '1998' was operationalised as May 1998–March 1999, while '1999' was operationalised as March 1999–November 1999 (p.191), April 1999–November 1999 (p.193) or May 1999–November 1999 (p.195)."

This justification for the re-analysis authors' decision to recode the treatment term is completely unfounded, and frankly quite strange. Description of the timing of treatment in each year is provided clearly in Miguel and Kremer (2004); all references to the timing of treatment (pages 170, 192, and 210) correctly 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 (which the re-analysis authors describe in the latter part of the above quote) was clearly and consistently defined in the original authors' STATA analysis do files, which were provided to Aiken et al. at the time they embarked upon this project. The re-analysis authors carefully studied this code and did not raise any objections to that definition, or confusion, in their pure replication report (Aiken et al., 2014a). 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 (Aiken et al., 2014b) 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 added any text describing the recoding of treatment.

We find the re-analysis authors' 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 basic research design of the project. 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 the replication authors' 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 "thrown out", 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 is nonsensical 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.

Simply put, as far as we can tell there is no basis for the assertion in Aiken et al. (2014b) that schools were "supposed" to be phased into treatment at the start of each calendar year.

Page 4: "The quasi-randomisation procedure for the deworming trial did not ensure that there was equal balance in the number of SAP schools in each group."

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

Page 4: "In accordance with our interpretation of the intention to treat, school-attendance observations of pupils in year 1 (1998) were interpreted as corresponding to the treatment condition in Group 1 and the control condition in Groups 2 and 3, and in year 2 (1999) observations were interpreted as corresponding to the treatment condition in Groups 1 and 2 and the 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 describe in detail above, the justification for recoding the treatment measure in this way is entirely unfounded. Tables 1 and 3 of this note present the primary results of Aiken et al. (2014b) – both the individual-level and cluster summary results – correctly defining treatment, and show substantial, highly significant, and robust impacts of deworming on school participation.

Page 5: "At the start of each year of the study, worm-infection rates were assessed among subsamples of pupils from intervention schools for that year. Thus in year 1 (1998), a sample was drawn from pupils across all grades in Group 1 schools prior to the drug treatment. In year 2 (1999), pupils from both Group 1 (after one year of intervention) and Group 2 (1999) (pre-intervention) schools were selected."

This is in direct contradiction to the so-called "intention-to-treat" assumption the replication authors just made a few lines earlier, in which they assume that Group 2 schools are treatment schools for the entirety of 1999. Yet here they conduct the worm infection analysis assuming that "Group 2 (1999) (pre-intervention) schools" were the control group. This is an example of a fundamental lack of coherence in the statistical analysis in Aiken et al (2014b).

Note that only pupils in grades 3-8 were eligible to be selected for stool samples in either year.

Page 5: "In later analysis, egg counts from the two readers were averaged and converted into eggs per gram of stool values."

During the helminth egg counts, each reader examined 50 mg of stool from a sampled child. The two separate egg counts were then added together for egg counts per 100 mg, which were then converted to eggs per gram (multiplying by ten).

Page 6: "We performed a sample-size calculation before commencement of this replication (Aiken et al. 2013), which is reproduced in Appendix 2. On this basis, we judged that these data would have adequate power to detect an approximate 5 per cent improvement in school attendance, as per the naïve result in the pure replication."

We note that these power calculations appear to have been calculated based on the pooled data following the study's original research design. This suggests that splitting the data to perform a year-by-year analysis would leave the resulting analysis severely underpowered.

Page 7: "We handled missingness 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 did this to try to increase the likelihood that the data used were prespecified."

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. Very few datasets directly observe pupil attendance in school at all (instead depending on school registers of unknown reliability) and fewer still attempt to track 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 Aiken et al's (2014b) approach is inappropriate.

Page 9, Statistical analysis Step 1: "We summarised the outcomes by calculating the mean of the school-level summary measures for each group and for each intervention arm in each year. We calculated the school-level summary attendance figures from the observations without first summarising pupil-level attendances. We compared the summary measures for each intervention arm within years using

a t-test. This approach is in accordance with the vertical conceptualisation of the stepped-wedge design referred to in the PAP, although the PAP did not prespecify the use of a statistical test. The cluster-summary approach accounts for the correlation between repeat observations and within schools but does not weight according to the precision of the cluster-summary estimates."

The cluster-level analysis presented in the left panel of Aiken et al. (2014b), Table 4 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). This pre-specified individual-level analysis corresponds to the results reported on the right hand panel of Aiken et al. (2014b), Table 4. This is also made clear in Aiken et al., 2013, p. 10, where it says:

> "*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 were surprised to see so much importance being placed on these results. In particular, these results are featured in the "primary outcomes" table (Aiken et al., 2014b, Table 4), alongside individual-level pre-specified analysis, and are used by the re-analysis authors to make claims about the supposed (non-)robustness of the school attendance results. We believe that two decisions in particular related to this non-pre-specified cluster-level analysis are rather unusual, and we show that a standard approach to a cluster-level analysis yields results that suggest a substantial, highly significant relationship between deworming treatment and school attendance.

First, we find the decision of the re-analysis authors to present an unweighted cluster-level analysis (that implicitly weights each school equally, rather than each individual or each attendance observation) to be unusual and non-standard. As we discuss in detail in Section 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 3 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 justification given for why the re-analysis authors 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 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 3, the pooled results are highly 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 at traditional confidence levels.

Page 10: "Our primary analyses identified an unexpected finding: the combined-year multilevel model for school attendance produced an effect estimate that was larger than either of the year-specific effects. On further investigation of the data, we found patterns of correlation between attendance and cluster size that we felt might explain this. Consequently, we plotted the proportion of pupils observed in attendance in each school against the number of observations made in a school stratified by year and by allocation group."

We were surprised by the claim that it is "unexpected" that the pooled-year estimate could be larger than either of the year-specific effects. This 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). It may lead to pooled estimated that differ from each individual cross-sectional estimate; this is standard statistics and nothing "unexpected".

Furthermore, the patterns of correlation between attendance and cluster size are not "unexpected" either - 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 2.2 above.

The re-analysis authors never explain why either of these issues create 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.

Page 11: "We investigated the sensitivity of our school-attendance results to the decision about which school-attendance observations corresponded with pupils being in treatment condition and which corresponded with the control condition. This analysis was not preplanned but was undertaken following a final correspondence with the original authors in October 2014. We investigated two scenarios, based on the suggestion from the original authors that the first school visits occurred before the drug treatment was delivered in year 1 (1998) and that the drug treatment was delivered in Group 2 only after the second visit period in

year 2 (1999). As described above, we do not have information from a protocol about when the drug treatment was intended to be delivered, about when the original authors made the decision to consider these first visits in each year as under control conditions or about when the educational component of the intervention was intended to be or was actually delivered. We did not have sufficient data to perform this analysis according to calendar dates. We also did not have information to explain how the timing of visits and deworming were linked."

This so-called sensitivity analysis is absolutely essential. As we describe in Section 2.1 above, in their primary analysis the re-analysis authors 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., 2014a), nor did they explicitly mention this recoding in their pre-analysis plan (Aiken et al., 2013) or the original version of the present report (Aiken et al., 2014b) that was submitted to 3ie for publication. Moreover, the justification added on to the present version of the report misuses the "intention to treat" terminology.

Simply put, there was never any intention to treat children at the very start of each calendar year in the Miguel and Kremer (2004) study. 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. An implication of the replication authors' claims about the underlying "intention" to treat Group 2 schools starting on January 1st, 1999 is that the original study somehow never intended to estimate impacts on health outcome measures; this is clearly false and simply makes no sense.

This misspecification of individuals has important implications for the analysis, as a comparison of Aiken et al. (2014b) Table 4 (using the miscoded treatment term) and Aiken et al., (2014b) Appendix Table 7 (the bottom panel of 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 4, 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 lower right-hand panel of Appendix 7 indicates that there are statistically significant results for both years separately and pooled together (P-values = 0.036, =0.088, and <0.001, respectively).

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

Page 11: "In scenario two, we excluded observations of attendance in the first visit period in year 1 (1998) and added observations in the first and second visit periods in year 2 (1999) to the analysis for the first year, analysing observations

in Group 2 during these two visit periods as corresponding to the control condition. Therefore, year 1 comprised observations in the second to the eighth visit periods in 1998, plus observations in the first and second visit periods in 1999. Year 2 comprised observations in the third to the eighth visit periods in 1999, which is the same as in scenario one. This data handling most closely approximates that used by the original authors but differs most from our original conception of the design of the stepped-wedge trial shown in Table 2 and published in our preanalysis plan. In effect, this handling of the data can be thought of as changing the time of the crossover from control to intervention from the beginning of 1999 (as in Table 2) to a time point later in 1999."

The pre-analysis plan in Aiken et al. (2013) did not indicate any recoding of the treatment term, so we do not consider that analysis to be pre-specified at all. There is simply no text in the pre-analysis document indicating that the replication authors' "original conception" of the study design bears any resemblance to the coding of the treatment variable that they employ in this replication study.

All of the evidence from project documents, analysis data, and the published paper (Miguel and Kremer 2004) – not to mention our own personal experience working on the project – indicates that the original plan was to introduce deworming treatment starting in March of each year. The replication authors' decision to use different timing in their definition of treatment is simply an error, and one that unnecessarily introduces measurement error into a key variable in the analysis.

Page 14: "There was substantial missingness for WAZ data in all three groups."

This missingness was due to weight not being collected from children who were not in school on the day of the pupil questionnaire data collection. (Note that these proportions missing are similar to daily absenteeism.)

Page 18: "Examining the visits that were successfully conducted, data were available for approximately 74 per cent of the pupils in these visits in year 1 (1998) and approximately 86 per cent in year 2 (1999). Within each year, there were broadly similar proportions of missing data across the three groups for attendance observations in visits that were successfully conducted."

Missing data in a multi-year longitudinal study on the order of 15 to 29% 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 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. Once again, as long as these errors are occurring at approximately the same rate in treatment and control schools, which appears to be the case, then they should not induce systematic bias. This is a point emphasized by Aiken et al. (2013). And indeed, the proportions missing are quite similar across the three intervention groups, which is reassuring.

Page 18: "In year 2 (1999), there was a higher proportion of missing examination data …"

This is presumably in part due to rising drop-out rates over time (an issue that also affects average school attendance in year 2 relative to year 1).

Page 18: "A total of 544 (1.7 per cent) pupils had moved schools by the end of year 1 (1998), and 2,376 (7.6 per cent) pupils had moved by the end of year 2 (1999)…"

We have this data because we systematically tracked students as they moved across schools over time, updating the school namelist data collection to reflect this (i.e., collecting information on students in their new school). This ability to track across schools over time is a strength of the data collection, in our view.

Page 19, Step 1 results: "In year 1 (1998), intervention schools had a mean attendance 5.48 per cent (95 per cent CI −1.48–12.44) higher than control schools, although this was not statistically significant (t-test p-value 0.12). In year 2 (1999), the intervention schools had a 2.16 per cent (95 per cent CI −3.39–8.27) greater mean attendance than control schools, but there was no statistical evidence of a difference (t-test p-value 0.48). These risk differences correspond to odds ratios of 1.78 and 1.21, respectively. In year 1 (1998) and year 2 (1999), there was no evidence of an association between intervention and examination performance in the cluster-means analysis."

As noted above, the cluster-level analysis was not pre-specified – the replication authors did not even suggest that they would present treatment versus control group statistics at the cluster summary level (Aiken et al., 2013). Moreover, 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. The re-analysis authors provide no rationale for presenting estimates which weight all schools equally, and we find this strange 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 statistically significant impacts of deworming on school attendance, for each year separately or pooled for both years (see Table 3, above).

We are also puzzled as to why the pooled 1998 and 1999 results are not shown here. Table 3 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. Aiken et al. (2014b) do show here that looking at 1998 and 1999 separately, and using non-standard weighting, and using a specification that was not pre-specified, does sometimes lead to only marginally significant results. In our view, it is only

when multiple non-standard deviations from the pre-analysis plan are made simultaneously that the result loses statistical significance at traditional confidence levels.

If we focus on the individual-level results presented in the right-hand panel of Aiken et al. (2014b), Table 4, we see statistically significant improvements in school attendance due to deworming in 5 out of the 6 estimates presented. The random-effect logistic regression is step 2 in the pre-analysis plan. 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 statistically significant effect within a subset of the data does not constitute meaningful evidence for a "non-effect".

It is clear here that when the full research design and both years of data are used, there is a large, positive, and statistically significant 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. 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, the conclusion here that there is no meaningful evidence of an impact of deworming on school attendance is puzzling to us.

Page 20: "In year 1 (1998), there were several schools in all of the groups that had more than 95 per cent attendance; in year 2 (1999), no schools had such high levels of attendance."

This 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.

Page 20: "The results of the analysis exploring the sensitivity of the school-attendance results to the handling of the treatment condition are shown in Appendix 7. In scenario one, 11,588 observations at the start of year 1 (1998) were excluded, as well as 31,404 observations during the first two visit periods in year 2 (1999). In comparison with our prespecified analysis, the year-specific results were approximately unchanged…"

In scenario one, 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). Furthermore, the cluster summary results move much closer to statistical 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 two, 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 bottom panel of Aiken et al. (2014b) Appendix 7 suggest positive and statistically significant impacts of deworming on school participation in 5 out of 6 models (with the odd case being an unweighted cluster mean for 1999 – our Table 3, above, shows that even that result is statistically significant (P-value < 0.5) when an appropriate weighting approach is applied).

Page 24: "At the start of year 2 (1999), substantially more pupils were tested in Group 2, the control group, than in Group 1, the intervention group."

The oversampling of Group 2 pupils in 1999 was done deliberately.

Pages 24-25: Results presented in Tables 5 and 6

Again, the re-analysis authors' own pre-analysis plan (Aiken et al., 2013) specifies that the analysis of secondary outcomes (worm infection, HAZ and WAZ) will be performed in an analogous fashion as that for the individual-level regressions presented in Table 2 for school attendance and exam scores, but using OLS instead of logistic regressions (page 10). It is unclear why the authors deviated from their plan, and only presented worm infection outcomes in an unweighted school-level analysis (Aiken et al., 2014b, Tables 5 and 6).

Page 26: "The absence of a clearly specified protocol for collection of these data initially compromised our confidence in the results relating to school attendance. We therefore approached the data cautiously by starting with simpler but arguably more robust and transparent analyses and progressively building up to more complex forms of analysis."

It is unclear to us why a cluster summary analysis weighting each school equally analysis is any "more robust and transparent" than a cluster summary analysis that weights each pupils (or attendance observation) equally. They are equally simple and transparent. It does seem obvious to us that a "simple" analysis that ignores the study's prospective stepped wedge research design and mis-weights observations is inferior to a more "complex" approach that utilizes all of the data, exploits all of the variation in the data, and weights the data appropriately.

Page 27: "In further analysis, we observed that there was a relationship between the number of attendance observations performed in a school and the overall rate of attendance in that school (Figure 2). The association between the number of pupil observations and the overall attendance in schools was noticeably different by intervention status; these were directly related in two out of three intervention-group years but were inversely related in all of the control-group years. In Group 2, which changed from control to intervention status between study years, the direction of this association switched between years."

As discussed in detail in section 2.2, there is no statistical evidence for differences in this relationship across treatment groups or over time.

Page 30: "We investigated the sensitivity of the school-attendance results to decisions about which observations corresponded to the intervention condition and which corresponded to the control condition. In particular, we incorporated information highlighted to us by the original authors concerning the timing of the deworming treatment in schools and, related to this, their opinions about whether some school-attendance observations should be considered as corresponding to control rather than intervention conditions. We explored two scenarios. In neither of the two scenarios were the results substantially different from the pattern of the main results of the prespecified analyses."

We disagree with the implication in this passage that there is an "opinion" on the definition of treatment status. As described clearly and consistently in Miguel and Kremer (2004) and the associated documentation and analysis code (STATA do files), all of which the re-analysis authors had access to since the very start of their work on this endeavor in early 2013, treatment in Group 1 schools did not begin until March 1998 (after the first round of school data collection visits), and treatment in Group 2 schools did not begin until March 1999 (after the second round of school data collection visits). The first few months of each year were utilized for community sensitization meetings, and collection of parasitological, anthropometric, and health behavior data. In 1998, this data forms the baseline, and in 1999 it is used to study the difference between Group 1 (treated) and Group 2 (untreated) schools. Without this data, any analysis of health outcomes (worm loads, hemoglobin, anthropometrics) would have been impossible – so we do not see how the re-analysis authors can claim that there was any "intention" that Group 2 schools should be treated at the start of 1999. An implication of their "view" on the timing of treatment is that we simply never intended to estimate health impacts of the study, since Group 2 was "really" supposed to be a treatment school starting on January 1st, 1999; but this claim about the "intended" timing of treatment is simply false. We also note that the collection of parasitological data at the start of each year from the schools that were being phased into treatment later that year was necessary to determine which drugs would be administered.

We also disagree with the re-analysis authors' summary that the results are not substantively different when treatment is defined correctly. In particular, the estimated impact of deworming on school participation is substantially weaker in the analysis in which they mis-define the treatment variable, as expected since doing so introduced unnecessary measurement error into a key variable in the analysis (which can be seen by comparing their Table 4 and Appendix 7).

Page 31: "Without more information about how the sampling was performed, and the degree of success in relocating the subsample in Group 1, the substantial difference in the number of pupils sampled in Group 1 and Group 2 at the start of year 2 (1999) raises concerns that the samples may not be comparable."

We note that it was a deliberate decision to choose a smaller sample for Group 1 in 1999 compared to Group 2. This was due to budget limitations, in part. Furthermore, it is unclear why the re-analysis authors chose to raise a concern about potential lack of comparability across the Group 1 and Group 2 samples, when they could use the data to simply test for these differences.

Finally, we again note a fundamental intellectual incoherence in the analysis in Aiken et al (2014b), namely, that they consider Group 2 schools to be "control" schools in early

1999 in the worm infection analysis, but "treatment" school in early 1999 in the school participation analysis. In other words, they have elected to mis-code the treatment variable in some pieces of their analysis but not in others, a decision that is puzzling to us.

Page 33-34: "Furthermore, the pure replication found improvements in school attendance to be similar whether or not children had received drug treatment, and in this analysis, all results were similar whether applied to drug-eligible pupils (main analysis) or all school pupils (see Appendix 4). This further undermines confidence in a causal relationship between drug administration and changes in school attendance."

As discussed in Miguel and Kremer (2004), there is strong evidence of deworming externalities on both worm infections and school participation both within the treatment schools and to neighboring schools within 3 km. There are large school participation gains for both the treated and untreated pupils in treatment schools, and we cannot reject that the impacts are equal for these two groups (in Table IX of Miguel and Kremer 2004, and in Aiken et al. 2014a), in part because this is a relatively statistically underpowered test. Deworming breaks the cycle of transmission for worm infections, and we show in Miguel and Kremer (2004) that it reduces reinfection for individuals within and in the vicinity of treatment schools. Intestinal worms have quite short average lifespans, on the order of one to two years, so sharp reductions in reinfection could quickly translate into a lower worm disease burden among both the untreated and the treated. Other work also suggests substantial epidemiological externalities among the untreated in treatment communities (Bundy et al. 1990, Ozier 2014).

Page 34: "A number of plausible pathways to increase school attendance exist that operate through behaviour change in children that are unrelated to the actual removal of worm infections. Causes of pupil behaviour change might include the educational component of the intervention, the placebo effect associated with receiving drug treatment, being in an intervention school (Hawthorne effect), or a desire to please parents or teachers who were aware of the study aims. Behaviour changes could subsequently cause changes in new worm infections or change how children perceive their health. All of these could lead to changes in school attendance without changing health status. It is also plausible that the removal of worm infections could lead to alteration in behaviour patterns mediated through some other biological mechanism that was not examined in this study, such as the alteration of immune-system activity, which has been described as an effect of helminth infections. There are also a number of plausible causal pathways that act outside of the child, such as at the level of the family or school."

We strongly disagree with the concerns raised here regarding placebo effects and other behavioral changes, as laid out in detail in section 2.4 above.

Page 35-36: "For any trial of a public-health intervention, the generalisability of the findings is an important question: would the same intervention lead to the same results if applied in a similar setting outside the context of a formal trial?

What would constitute a similar setting? This study was conducted in rural western Kenya in 1998–1999, and the researchers found that all schools tested

had > 50 per cent baseline prevalence of worm infection. This suggests that Busia District was a 'high worm burden' setting at that time. For the results of this trial to be applied to other settings, there would have to be a similarly high burden. As the nature of the causal pathway operating here is uncertain, it is unclear what other aspects of the setting would need to be similar for the intervention to work in the same way. For example, poverty and gender bias are two other factors that almost certainly impact school attendance, but these effects operate in complex ways that vary substantially from place to place, which might alter the effects of this intervention. A recent high-profile publication reported from a large trial looking at the effect of deworming and vitamin A supplementation on preschool mortality in north India found that deworming had no effect in this lightly infected area (Awasthi et al. 2013)."

The Awasthi et al. (2013) study 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 3 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 (Croke, 2014; Ozier, 2014; Baird et al., 2014).

References

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

Aiken, A, Davey, C, Hayes, R and Hargreaves, J. (2014a). "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, Hayes, R and Hargreaves, J. (2014b). "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).

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: Public finance implications of a child health investment", unpublished working paper, University of California, Berkeley.

Bundy, D.A.P., M.S. Wong, L.L. Lewis, and J. Horton. (1990). "Control of Geohelminths by Delivery of Targeted Chemotherapy through Schools", *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 84: 115–120.

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

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. (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).

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.*