Alexander M Aiken
Calum Davey
James R Hargreaves
Richard J Hayes

Reanalysis of health and educational impacts of a school-based deworming program in western Kenya

Part 1: pure replication

October 2014

Replication
Paper 3
Part 1

Health and education

3ie
International Initiative for Impact Evaluation

About 3ie

3ie is an international grant-making NGO promoting evidence-informed development policies and programmes. We are the global leader in funding and producing high-quality evidence of what works, how, why and at what cost. We believe that better and policy-relevant evidence will make development more effective and improve people's lives.

3ie Replication Paper Series

The 3ie Replication Paper Series is designed to be a publication and dissemination outlet for internal replication studies of development impact evaluations. Internal replication studies are those that reanalyse the data from an original paper in order to explore original evaluation questions. The series seeks to publish replication studies with findings that reinforce an original paper, as well as those that challenge the results of an original paper. To be eligible for submission, a replication study needs to be of a paper in 3ie's online Impact Evaluation Repository and needs to include a pure replication. 3ie invites formal replies from the original authors. These are published on the 3ie website together with the replication study.

The 3ie Replication Programme also includes grant-making windows to fund replication studies of papers identified on the candidate studies list. Requests for proposals are issued one to two times a year. The aim of the 3ie Replication Programme is to improve the quality of evidence from development impact evaluations for use in policymaking and programme design.

About this report

This paper was funded through 3ie's Replication Window with generous funding from an anonymous donor. All content is the sole responsibility of the authors and does not represent the opinions of 3ie, its donors or the 3ie Board of Commissioners. Any errors and omissions are the sole responsibility of the authors. Comments and queries should be directed to the corresponding author, Dr Alexander Aiken, at alexander.aiken@lshtm.ac.uk.

Suggested citation: Aiken, AM, Davey, C, Hargreaves, JR and Hayes, RJ 2014. *Reanalysis of health and educational impacts of a school-based deworming program in western Kenya: Part 1, pure replication*, 3ie Replication Paper 3, part 1. Washington, DC: International Initiative for Impact Evaluation (3ie)

3ie Replication Paper Series executive editor: Annette N Brown
Managing editor: Benjamin DK Wood
Assistant managing editor: Jennifer Ludwig
Copy editor: Pamela Tatz
Cover design: John F McGill
Proof reader: Yvette Charboneau
Layout assistant: Hisham Esper
Printer: Mimeo.com

Reanalysis of health and educational impacts of a school-based deworming program in western Kenya Part 1: pure replication

Alexander M Aiken
Centre for Evaluation, Department of Social and Environmental Health Research, London School of Hygiene and Tropical Medicine, UK

Calum Davey
Centre for Evaluation, Department of Social and Environmental Health Research, London School of Hygiene and Tropical Medicine, UK

James R Hargreaves
Centre for Evaluation, Department of Social and Environmental Health Research, London School of Hygiene and Tropical Medicine, UK

Richard J Hayes
Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, UK

3ie Replication Paper 3
Part 1
October 2014

International Initiative for Impact Evaluation

Acknowledgements

Funding

This replication has been funded and facilitated by the International Initiative for Impact Evaluation (3ie) as part of their replication programme. The broad aim of this programme is to improve the quality of evidence for development policy by reappraising a wide range of influential studies in the development field, seeking to verify and examine the robustness of the original findings in these studies. The funders had no role in writing the analysis plan, the draft or final reports.

Reproduction of images

This report contains reproductions of text published as part of the original paper. This remains subject to the original copyright restrictions. These are shown as part of this report by the kind permission of the journal *Econometrica*. The copyright statement accompanying these images is as follows:

> The copyright to this article is held by the Econometric Society, http://www.econometricsociety.org/. It may be downloaded, printed and reproduced only for personal or classroom use. Absolutely no downloading or copying may be done for, or on behalf of, any for-profit commercial firm or for other commercial purpose without the explicit permission of the Econometric Society. For this purpose, contact the Editorial Office of the Econometric Society at econometrica@econometricsociety.org.

Original authors

This replication would not have been possible without the cooperation and disclosure of the original authors. We would like to thank Professor Edward Miguel, Dr Joan Hamory Hicks and Michael Walker for making available the original data for the study, for providing their analysis files, for providing detailed supporting documentation and for taking part in conference calls. We recognise that they have committed significant hours of work to comply with our requests, as well as reading and commenting on draft reports.

Abstract

Introduction: Helminth infections cause morbidity amongst poor communities worldwide. It is unknown whether removing these infections results in improvements in school attendance and educational achievement.

Methods: We reanalysed data earlier researchers originally collected in a trial of a drug and education intervention that Internationaal Christelijk Steunfonds conducted in 75 schools in western Kenya in 1998–1999. Outcome measures included worm infections, other health-related outcomes (haemoglobin and nutritional parameters), school attendance and exam performance. The study describes different types of effects (direct, indirect-within-school, naïve, indirect-between-school and overall), with each applying to particular groups of school children.

Results: We noted various discrepancies between the published results and those from this reanalysis. These ranged in importance from minor (for example, rounding errors) to moderate (for example, inaccurately labelled significance) to major (for example, coding errors). For the worm-infection outcome, the results were broadly similar to those originally reported. For non-worm-related health outcomes, in contrast to the original study, this reanalysis only found limited evidence of benefit — there was weak evidence of a small benefit on height-for-age z-score only. For school attendance, this reanalysis found beneficial effects similar to or stronger than those originally reported for the direct, indirect-within-school and naïve effects. However, after correcting coding errors in the original analysis, there was no evidence that the intervention had a between-school indirect effect or an overall effect on school attendance. As in the original study, this reanalysis found that the intervention had no effect on examination scores. Other discrepancies that the original results detected did not appear systematic in any way and did not affect the major findings of the study as originally conducted.

Discussion: According to the methods of analysis that the original study used, the intervention did appear to have some beneficial effects in relation to worm infection and school attendance. It had no between-school indirect effect on school attendance. There was only weak evidence of non-worm-related health benefits. These results from reanalysis included important differences from those published in the original analysis.

Contents

List of figures and tables

Figures

Tables

Glossary

Term used in this report	Explanation, for purposes of this report
Missingness	The extent to which data intended to be collected in the study are unavailable for the purposes of analysis.
Bias	Collection or calculation of data or variables that are systematically different from values in the population of interest.
Pure replication	A reanalysis of a study, without collection of new data.
Indirect effect (= externality)	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.

Abbreviations and acronyms

3ie	International Initiative for Impact Evaluation
HAZ	Height-for-age z-score
Hb	Haemoglobin
ICS	Internationaal Christelijk Steunfonds
sd	Standard deviation
se	Standard error
WAZ	Weight-for-age z-score
WHO	World Health Organization

1. Introduction

This is a report on the replication (reanalysis) of work originally describing the impact of a school-based deworming programme in Kenya on the health, school attendance and academic performance of school pupils (Miguel and Kremer 2004). This report forms the first section of this reanalysis — a 'pure replication', which requires one to 'duplicate, repeat, as in a statistical experiment' or 'to make or do something again in exactly the same way' (Hamermesh 2007). Pure replication should always precede any other forms of replication, as it allows verification that applying the methods that the original authors used to their raw data does lead to the outcomes as reported. The second part of the replication, 'Statistical and scientific replication' (Davey *et al*. 2014), applies alternative statistical methods and conceptual frameworks to the same raw data to see how the conclusions reached compare with those of the original study — this follows after the 'pure replication' described here.

2. Description of original study

2.1 Overview

The original study analysed data collected as part of a school-based deworming programme that Internationaal Christelijk Steunfonds (ICS), a Dutch charitable organisation, conducted in 75 schools in Busia District in western Kenya in 1998–1999. The researchers stratified the schools by administrative area and involvement in other ICS programmes and then quasi-randomised them into three groups as follows: first sorted geographic zones alphabetically within each administrative division and then sorted schools within each zone by population size and assigned them to Groups 1, 2 and 3, consecutively. ICS introduced the deworming intervention to the three groups (25 schools per group, averaging 400 pupils per school) in stages over two years, as shown in Figure 1. This phased introduction across randomised groups is known in medical literature as a 'stepped wedge' cluster randomised trial. In this study, the researchers analysed just two 'steps', though Group 3 schools also received the intervention at a later time. Schools are the unit of clustering.

	1998	1999
Group 1	Intervention	Intervention
Group 2	Control	Intervention
Group 3	Control	Control

Figure 1: Stepped wedge design of study

2.2 Types of worm infection

The original analysis looked at an intervention package to prevent and treat four different types of worm infection: hookworm, roundworm, whipworm (all geohelminths — literally 'ground worms') and schistosomiasis. Appendix 1 gives key biological features of these infections.

2.3 Intervention

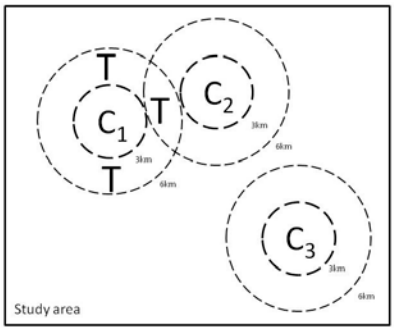
The intervention was composed of two elements: First, ICS administered anti-helminthic (deworming) drug treatments in appropriate doses at spaced intervals. Second, they administered a package of educational interventions. Girls over the age of 12 were not intended to receive the drug intervention due to concerns of potential teratogenicity, although in practice some did. All schools in the intervention group received treatment against soil-mediated helminth infections (geohelminths), but only a subset of schools (6 out of 25 in 1998 [Group 1 only], 16 out of 50 in 1999 [Groups 1 and 2]) additionally received treatment against schistosomiasis, a freshwater-mediated infection. This additional treatment was allocated on the basis of the local prevalence of schistosomiasis

infections, as found in parasite surveys — schools where schistosomiasis rates were previously found to be low were not eligible for treatment for this infection. Educational measures consisted of worm-prevention education, including stressing the importance of hand washing, wearing shoes and not swimming in fresh water. The original study also mentions that there were other ICS-led, school-based interventions occurring concurrently in 27 of 75 schools (p.165 in the original report). Previous analyses by the original study authors (Kremer *et al*. 2002) have stated that these interventions had no substantial effects. We describe these other interventions and their potential effects on this study in detail in the 'Statistical and scientific replication'.

2.4 Outcome measures

ICS measured the impact of the intervention in three different domains: health (principally worm infection but also nutritional and haematological parameters), school attendance and exam performance. All of these outcomes were measured at the level of the individual. ICS fieldworkers measured school attendance by performing multiple unannounced visits to all schools during the course of the year. ICS administered exams in a number of subjects at the end of the academic year and measured exam performance. As for the health impacts, worm infection was only measured in treatment schools immediately before deworming, as it was felt unethical to test for worm infection without offering treatment. Therefore, they did not measure the worm-infection rate of Group 2 in 1998 or of Group 3 in either year. Haemoglobin and nutritional status were only measured in randomly selected small subsets of children. In the original analysis, they obtained an estimate of the direct benefit of the intervention in each of the three domains by comparing outcomes in schools that received the intervention to those that did not receive it in that year. This format of analysis makes the assumption that there was no cumulative effect of the intervention, such that the second year of the intervention in Group 1 had the same effect as the first year of the intervention in Groups 1 and 2.

Figure 2: Schematic representation of study area



The original analysis also assessed the indirect benefits (positive externalities or spillover effects) of the intervention — that is, whether the intervention prevented the transmission of worm infections in nearby schools. The analysts determined this using a spatial approach, illustrated schematically in Figure 2. The control schools (C_n) were located at different physical proximities to treatment schools (T). The excretion of worm eggs in faeces transmits all of these worm infections, and as faecal contamination of the environment was known to be common, the authors assumed that there would be a local reduction of transmission of worm infection in an approximately circular area up to six kilometres around the intervention schools, where they assumed that the children attending the schools lived. The authors hypothesised that the schools would receive greater benefit from indirect reduction in worm infection if they were close to many intervention schools: hence (in the schematic figure), school C_1 would receive a greater indirect benefit than school C_2, which was in turn greater than school C_3 and so on. The analysts also used an additional independent term in their modelling process to account for variation in local population density: schools C_1 and C_2 are in areas of greater local population density than school C_3 (Miguel and Kremer 2004 p.176). They could then use the variation in indirect benefit across a gradient of

exposure created by the variation in spatial proximities to estimate the overall scale of the indirect benefit. The original authors also looked for evidence of an indirect benefit within schools (untreated pupils in treatment schools).

2.5 Original main findings

The main effects of the original study can be considered to fall into several different categories, each describing distinct effects on particular groups of individuals, as follows:

- **Direct effect**: the effect of drug treatment on the treated pupils.
- **Within-school indirect effect**: the indirect effect on all children in treatment schools arising from the drug treatment of children within those schools. This applies to both treated and untreated children in treatment schools. ICS did not treat children in treatment schools who declined or were not eligible for drug treatment.
- **Naïve effect**: the effect found by comparing all children in treatment schools to all children in control schools, irrespective of whether or not the children were themselves treated. This is a combination of the direct effect and the within-school indirect effect, though not a simple addition of effects. This is the type of effect that would typically be analysed in medical literature in an intention-to-treat evaluation of a cluster-randomised trial.
- **Between-school indirect effect**: the average effect of having children treated in schools nearby, across all children in the study. This is the only effect that applies to children in control schools; children in treatment schools accrue this in addition to the other effects.
- **Overall effect:** the combination of the naïve effect (which is applied to treatment schools only) and the between-school indirect effect (which applies to all schools).

Halloran and Struchiner (1991) give a wider discussion of the terminology of dependent happenings in interventional studies. In this pure replication exercise, we did not evaluate the appropriateness of separating effects into the different categories as described above. Instead, we merely reproduced the analytic steps to redetermine the results as originally calculated.

We mainly presented the results as average risk difference at the level of the individual, though we adjusted the calculation of the standard error for the clustered nature of the data. These figures represent average effects across all applicable children. The results are across both years of the study combined, unless otherwise stated. As the number of individuals to whom the effects apply varies, the overall effect is not a simple addition of different components effects. Unless otherwise stated, all results are based on data from all three groups and both years of the study combined.

Table 1: Summary of results from original paper

Measure	Direct effect	Indirect effect: within school	Naïve effect	Indirect effect: between school	Overall effect
Health Worm infection (any mod/hvy inf) pp.184–188	−14% (se 7%)	−12% (se 7%)	−25% (se 5%)	−23% (se 7%)	−35% (se 9%)
Health Anaemia (Hb<100g/L) pp.173–174	Not reported	Not reported	−2% absolute prop'n (se 1%)	Not reported	Not reported
Health Nutritional status pp.173–174	Not reported	Not reported	WAZ: −0.00 (se 0.04) HAZ: 0.09 (se 0.05)	Not reported	Not reported
School attendance (% increase) pp.195–196	+6.2%† (se 2.2%)	+5.6%† (se 2.0%)	+5.1% (se 2.2%)	+2.0% (se 1.3%)	+7.5% (se 2.7%)
Exam performance (average difference) p.201	Not reported	Non-significant result, data not shown	Not reported	−0.049 (se 0.052)	Yr 1: −0.032 (se .046) Yr 2: 0.001 (se 0.073)

Note: abbreviations: sd = standard deviation; se = standard error; Hb = haemoglobin; WAZ = weight-for-age z-score; HAZ = height-for-age z-score. Effects that the original authors felt were beneficial and statistically significant are shaded. Page numbers refer to the original study. † = year 1 data only.

The original analysis concluded that there were both substantial direct and indirect benefits to health and school attendance arising from the deworming programme. As a key finding, the original authors noted that over the course of the programme, overall school absenteeism fell from 28 to 21 per cent, an improvement in attendance of 7.5 per cent — or, alternatively, a reduction in absenteeism of approximately 25 per cent. The authors saw lesser degrees of improvement in attendance for different component effects that made up this overall effect. The original report also described that the authors saw no benefit (either direct or indirect) from deworming on academic test scores.

2.6 Influences of original study

The original study has been enormously influential in development economics. It provided a major piece of the evidence that led to the Copenhagen Consensus Center ranking the deworming of children as fourth among the sixteen most cost-effective investments to overcome the world's biggest challenges in 2012 (CCC 2012). The Center quotes Nobel laureate economist Robert Mundell as saying, 'Deworming is an overlooked intervention deserving of greater attention and resources. This simple, cheap investment can mean a child is healthier and spends more time in school.' The methodological approach of this paper is regarded as seminal in recognising the importance of indirect effects (positive externalities) in assessment of the overall impact of health interventions, and the paper won the 2005 Kenneth Arrow Award for Health Economics, presented by the International Health Economics Association.

However, this paper has received much less acclaim in the medical literature and was notably omitted from a Cochrane Review (the gold standard of evidence-based medicine) on this subject in 2008. More recently, an updated Cochrane Review (in 2012) now included this paper (Taylor-Robinson *et al*. 2012) but described limitations present in the work that could have led to bias in data collection. In particular, the reviewers were concerned about risk of bias from baseline imbalance, incomplete outcome data

and sequence generation — and, overall, graded the paper as having a 'high risk of bias'. Despite these limitations, the review frequently mentioned findings arising from the analysis of these data, although the reviewers ultimately concluded that 'it is probably misleading to justify contemporary deworming programmes based on evidence of consistent benefit on nutrition, haemoglobin, school attendance or school performance as there is simply insufficient reliable information to know whether this is so.' A discussion piece in the *British Medical Journal* entitled 'Deworming debunked' (Hawkes 2012) has sharpened the debate.

The original study has been highly influential for health policy in Kenya. In 2009, the nation launched a school-deworming programme, which is currently ongoing. In that year, over 3.6 million children across 8,200 schools were dewormed. At a wider level, a London Declaration on Neglected Tropical Diseases was signed in 2012 — endorsed by agencies including the World Bank, USAID and several major pharmaceutical firms — agreeing, amongst other goals, to 'sustain, expand and extend drug access programmes to ensure the necessary supply of drugs and other interventions to help control by 2020 schistosomiasis [and] soil-transmitted helminthes' (2012).

On this basis, we felt that it was highly important to determine whether the original results held up to the scrutiny of an independent reanalysis. Hence, we undertook this replication.

3 Methods employed for pure replication

3.1 Overview of replication

This pure replication report intends to provide a comprehensive description of discrepancies identified between the original paper and this reanalysis, coupled with commentary on methodological issues in the authors' original report. Many of the discrepancies we identify are relatively unimportant, such as rounding errors or inaccurately labelled denominators that are unlikely to influence the interpretation of the overall study, though we have listed these in full for completeness. However, in Tables I, V, VII and XI and in calculations described in the text following these tables, there were important discrepancies of various types that appear to have had substantial influence on major study findings; these are described in greater detail.

In keeping with the pure replication approach, we neither perform new calculations nor introduce new concepts and we reproduce the authors' calculation steps after making appropriate corrections when necessary. In the later commentary (Section 4), we discuss how various decisions in the format of study design and presentation could affect the interpretation of analyses.

3.2 Data sharing and set-up

The original study authors kindly supplied us with copies of their original raw data, files for processing and output of data (do-files and log files, as used in their own unpublished internal replication exercise in 2007) and explanatory notes. They also supplied us with a document describing various issues in the original manuscript of which they were already aware, which was helpful for cross-checking against discrepancies that we identified. The data files they provided were configured to run as if in STATA version 9, since the authors used this version in their analysis. We note that the authors report sharing the data and materials relating to their own 2007 replication exercise with various other interested parties (including graduate and doctoral students, professors, policy analysts

and nonprofit organisation analysts) up to the point that they now consider these to be 'effectively public documents'.

3.3 Performance of pure replication

We ran the original analysis files using the original raw data for each of the major output tables of the original paper. We have checked the individual values for all the outputs in each of tables and the associated calculations and provide a brief commentary.

4 Results

4.1 Presentation of pure replication

Each of the 10 tables from the original paper has been reproduced in this report, with shading to show the location of any discrepancies detected between the original report and results from reanalysis. We show the latter result outside the original table. The types of discrepancies that this report describes are as follows, roughly arranged in the order of least to most likely to influence the findings of the study:

- Unclear labelling. For example, in Table II, the final column is labelled as 'Average infection intensity in eggs/gram', when a clearer description of what the report actually calculated would be 'Average worm burden in whole population tested, in eggs/gram'.
- Rounding errors. Some of these may have occurred when progressively shortening the number of digits displayed (for example, 0.7745 →0.775 → 0.78), whilst others may represent transcription errors.
- Inaccurately reported denominators (for example, where the report states that it based a result on n = 24,958 observations when it was actually based on n = 24,979 observations).
- Mislabelled levels of significance (for example, where the report annotates a result with a p-value of 0.06 as significant at the 95 per cent confidence level (**), when it is actually significant at the 90 per cent confidence level (*).
- Coding errors in STATA analysis (do) files. Where possible, we have reproduced the code that led to these errors. One coding error affects part of the output for Table I and all the outputs for Tables VII, IX and X, so for each of these tables we display two versions, with ORIGINAL and UPDATED labels in the respective table captions.

Many of these errors are liable to have occurred through use of intermediate dataset and analysis files that were not fully corrected before production of the original paper. Some further discrepancies occur in calculations performed outside of the major tables.

We reproduced footnotes from the original tables, in full or in part, where these contained important explanations for result interpretation or when we identified a discrepancy in the footnote itself. Many of the original tables had extensive footnotes describing methodological approaches and ancillary findings; we refer the reader to the original paper where we did not reproduce them in full.

4.2 Table I: 1998 average pupil and school characteristics, pretreatment (pp.166–167)

These are simple descriptive statistics relating to pupils (Panel A and B) and schools (Panel C) that were measured in all 75 schools at the start of the study in 1998.

In the calculation of results for this table, we make allowance for the clustered nature of these data by collapsing the individual-level data to 75 observations (representing the schools), and we give each school a weight according to pupil population. In later tables, including all regressions, we perform analyses at the level of the individual while making allowance for the clustering of data when calculating standard errors. These are both valid methods for accounting for clustering in the data.

- Panel A: pupil baseline characteristics. This is for all grades (age groups), with approximately 11,500 pupils per group. We found no discrepancies between the results presented and those from reanalysis. Year of birth was the only available information for determining age (in other words, month and day were not available), which is unsurprising in this setting.
- Panel B: pupil baseline characteristics. This is for grades 3–8 with approximately 4,400 pupils per group. We found no discrepancies between the results presented and those from reanalysis. This panel includes self-reported, fieldworker-observed and measured characteristics. There are two errors in annotation of statistical significance, which are likely to be unimportant to overall findings.
- Panel C: school baseline characteristics. This is for 25 schools per group. Almost all of the results in the last four rows of this panel are affected by an error in the creation of local population densities; this is described in more detail on p.15 of this report. The discrepancies arising from this error range from modest (for example, average number of Group 1 pupils within three kilometres, for Group 1 school: ORIGINAL = 461.1, UPDATED = 430.4) to substantial (for example, average total number of primary school pupils within three to six kilometres, for Group 1 schools: ORIGINAL = 2,370.7, UPDATED = 3,431.3). These alterations in local populations lead to changes in Tables VII, IX and X, as described later.

Across all of the panels in this table, we performed significance tests to look for differences between groups. Given that the original authors allocated the schools to these groups by a quasi-randomised process, the medical literature would generally not consider it appropriate (Senn 1994) to perform such tests in baseline comparison tables — because if the randomisation has been performed appropriately, any differences arising between the groups would be due to chance. However, we acknowledge that such a practice is the norm in social science literature and, as such, most journals for these disciplines would expect it. In this case, it seems reasonable to present these analyses here to provide reassurance that the quasi-randomisation process was successful, which it broadly does appear to have been (see section 4).

Table 2: ORIGINAL Table I: Average pupil and school characteristics, pretreatment[a]

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)	Group 1– Group 3	Group 2– Group 3
Panel A: Preschool to Grade 8					
Male	0.53	0.51	0.52	0.01	−0.01
				(0.02)	(0.02)
Proportion girls < 13 years and all boys	0.89	0.89	0.88	0.00	0.01
				(0.01)	(0.01)
Grade progression (= Grade – (Age – 6))	−2.1	−1.9	−2.1	−0.0	0.1
				(0.1)	(0.1)
Year of birth	1986.2	1986.5	1985.8	0.4**	0.8***
				(0.2)	(0.2)
Panel B: Grades 3–8					
Attendance recorded in school registers (during the four weeks prior to the pupil survey)	0.973	0.963	0.969	0.003	−0.006
				(0.004)	(0.004)
Access to latrine at home	0.82	0.81	0.82	0.00	−0.01
				(0.03)	(0.03)
Have livestock (cows, goats, pigs, sheep) at home	0.66	0.67	0.66	−0.00	0.01
				(0.03)	(0.03)
Weight-for-age Z-score (low scores denote undernutrition)	−1.39	−1.40	−1.44	0.05	0.04
				(0.05)	(0.05)
Blood in stool (self-reported)	0.26	0.22	0.19	0.07**	0.03
				(0.03)	(0.03)
Sick often (self-reported)	0.10	0.10	0.08	0.02**[d]	0.02**[e]
				(0.01)	(0.01)
Malaria/fever in past week (self-reported)	0.37	0.38	0.40	−0.03	−0.02
				(0.03)	(0.03)
Clean (observed by fieldworkers)	0.60	0.66	0.67	−0.07**	−0.01
				(0.03)	(0.03)
Panel C: School characteristics					
District exam scores 1996, grades 5–8[b]	−0.10	0.09	0.01	−0.11	0.08
				(0.12)	(0.12)
Distance to Lake Victoria	10.0	9.9	9.5	0.6	0.5
				(1.9)	(1.9)
Pupil population	392.7	403.8	375.9	16.8	27.9
				(57.6)	(57.6)
School latrines per pupil	0.007	0.006	0.007	0.001	−0.000
				(0.001)	(0.001)
Proportion moderate–heavy infections in zone	0.37	0.37	0.36	0.01	0.01
				(0.03)	(0.03)
Group 1 pupils within 3 km[c]	461.1[f]	408.3[f]	344.5[f]	116.6[f]	63.8[f]
				(120.3)	(120.3)
Group 1 pupils within 3–6 km	844.5[f]	652.0[f]	869.7[f]	−25.1[f]	217.6[f]
				(140.9)	(140.9)
Total primary school pupils within 3 km	1229.1[f]	1364.3[f]	1151.9[f]	77.2[f]	212.4[f]
				(205.5)	(205.5)
Total primary school pupils within 3–6 km	2370.7[f]	2324.2[f]	2401.7[f]	−31.1[f]	−77.6[f]
				(209.5)	(209.5)

[a,b,c] These are the footnotes from the original paper. They are fully reproduced in table 3.

[d] These results are not statistically significant in the replication study.

[e] These results are statistically significant at the 90 (*) per cent confidence level in the replication study.

[f] All of these replication-study results change due to an author-identified coding error explained in section 4.8.

Table 3: UPDATED Table I: Average pupil and school characteristics, pretreatment[a]

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)	Group 1– Group 3	Group 2– Group 3
Panel A: Preschool to grade 8					
Male	0.53	0.51	0.52	0.01	−0.01
				(0.02)	(0.02)
Proportion girls < 13 years and all boys	0.89	0.89	0.88	0.00	0.01
				(0.01)	(0.01)
Grade progression (= grade−(age−6))	−2.1	−1.9	−2.1	−0.0	0.1
				(0.1)	(0.1)
Year of birth	1986.2	1986.5	1985.8	0.4**	0.8***
				(0.2)	(0.2)
Panel B: Grades 3–8					
Attendance recorded in school registers (during the four weeks prior to the pupil survey)	0.973	0.963	0.969	0.003	−0.006
				(0.004)	(0.004)
Access to latrine at home	0.82	0.81	0.82	0.00	−0.01
				(0.03)	(0.03)
Have livestock (cows, goats, pigs, sheep) at home	0.66	0.67	0.66	−0.00	0.01
				(0.03)	(0.03)
Weight-for-age z-score (low scores denote undernutrition)	−1.39	−1.40	−1.44	0.05	0.04
				(0.05)	(0.05)
Blood in stool (self-reported)	0.26	0.22	0.19	0.07**	0.03
				(0.03)	(0.03)
Sick often (self-reported)	0.10	0.10	0.08	0.02	0.02*
				(0.01)	(0.01)
Malaria/fever in past week (self-reported)	0.37	0.38	0.40	−0.03	−0.02
				(0.03)	(0.03)
Clean (observed by fieldworkers)	0.60	0.66	0.67	−0.07**	−0.01
				(0.03)	(0.03)
Panel C: School characteristics					
District exam scores 1996, grades 5–8[b]	−0.10	0.09	0.01	−0.11	0.08
				(0.12)	(0.12)
Distance to Lake Victoria	10.0	9.9	9.5	0.6	0.5
				(1.9)	(1.9)
Pupil population	392.7	403.8	375.9	16.8	27.9
				(57.6)	(57.6)
School latrines per pupil	0.007	0.006	0.007	0.001	−0.000
				(0.001)	(0.001)
Proportion moderate–heavy infections in zone	0.37	0.37	0.36	0.01	0.01
				(0.03)	(0.03)
Group 1 pupils within 3 km[c]	430.4	433.2	344.5	85.9	88.7
				(116.2)	(116.2)
Group 1 pupils within 3–6 km	1157.6	1043.0	1297.3	−139.7	−254.4
				(199.3)	(199.3)
Total primary school pupils within 3 km	1272.7	1369.1	1151.9	120.8	217.2
				(208.1)	(208.1)
Total primary school pupils within 3–6 km	3431.3	3259.8	3502.1	−70.8	−242.3
				(366.0)	(366.0)

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

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

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

Note: the original authors acknowledged the differences identified in table 2 and table 3 as part of their internal replication in 2007.

4.3 Table II: January 1998 helminth infections, pretreatment, Group 1 schools (p.168)

These data are proportions of pupils with different types of worm infections. Two readers performed egg counts for different worm infections (per 50 milligrams of stool), with values combined. The authors reported this to be a random sample of 1,894 pupils from Group 1 schools, with 15 pupils per grade from grades 3–8. It is not clear from the original report how the ICS performed this randomisation. Thresholds the authors used for moderate-to-heavy infection with hookworm, whipworm and schistosomiasis are different from those that the World Health Organization suggests (see Appendix 1 of this document). The supporting reference provided makes a case for and uses locally defined thresholds for heavy infection (Brooker *et al*. 2000) but does not mention moderate-to-heavy infection thresholds. Therefore, a more appropriate description of how the authors selected these thresholds for moderate-to-heavy infection would be 'personal communication with Dr Simon Brooker and Professor Donald Bundy', according to the authors' own report of how they actually performed this.

The right-hand column in this table shows average infection intensity, but the authors included children without infection (zero eggs seen on stool microscopy) in this calculation. The labelling here is thus misleading: this column actually represents the average worm-infection load across the entire population. Egg-count data are highly skewed, so presentation of the arithmetic means (as shown here) is not an appropriate summary of the distribution of values.

The last two rows of the table show data representing the prevalence of children with exactly n infections. The actual prevalence of ≥2 and ≥3 infections are 0.65 and 0.34, respectively. We found no other substantial discrepancies in the results presented; one unimportant difference appears to be a rounding error.

Table 4: Table II: January 1998 helminth infections, pretreatment, Group 1 schools[a]

	Prevalence of infection	Prevalence of moderate–heavy infection	Average infection intensity, in eggs per gram (se)
Hookworm	0.77	0.15	426[b] (1055)
Roundworm	0.42	0.16	2337[b] (5156)
Schistosomiasis, all schools	0.22	0.07	91[b] (413)
Schistosomiasis, schools <5 km from Lake Victoria	0.80	0.39	487[b] (879)
Whipworm	0.55	0.10	161[b] (470)
At least one infection	0.92	0.37	--
Born since 1985	0.92 ↑	0.40	--
Born before 1985	0.91	0.34	--
Female	0.91	0.34	--
Male	0.93	0.38	--
At least two infections	0.31[a]	0.10	--
At least three infections	0.28[a]	0.01	--

[a] The replication study used only data presented for 'exactly n infections'.

[b] The replication study calculated means count by including children without detected infection.

↑ The replication results coefficients are 0.01 higher than original results.

The footnotes from the original table are not reproduced here.

4.4 Table III: Proportion of pupils receiving deworming treatment in PSDP (p.170)

Discrepancies between the results presented and those from reanalysis were noted in 16 out of the 72 proportions reported in this table; these all appear to be rounding errors within 0.01 of values from reanalysis. These do not form the basis of later calculations. (PSDP = Primary-school deworming programme).

Table 5: Table III: Proportion of pupils receiving deworming treatment in PSDP

	Group 1		Group 2		Group 3	
	Girls < 13 years and all boys	Girls ≥ 13 years	Girls < 13 years and all boys	Girls ≥ 13 years	Girls < 13 years and all boys	Girls ≥ 13 years
Any medical treatment in 1998 (For grades 1–8 in early 1998)	0.78 ↓	0.19 ↑	0	0	0	0
Round 1 (March–April 1998) Albendazole	0.69 ↓	0.11	0	0	0	0
Round 1 (March–April 1998) Pranziquantel[b]	0.64	0.34	0	0	0	0
Round 2 (October–November 1998), Albendazole	0.56	0.07	0	0	0	0
Any medical treatment in 1999 (For grades 1–8 in early 1998)	0.59 ↓	0.07	0.55 ↓	0.10 ↓	0.01	0
Round 1 (March–April 1999) Albendazole	0.44	0.06	0.35	0.06 ↓	0.01	0
Round 1 (March–April 1999) Pranziquantel[b]	0.47	0.06	0.38	0.06	0.01 ↓	0
Round 2 (October–November 1999), Albendazole	0.53 ↓	0.06	0.51 ↓	0.08 ↓	0.01	0
Any medical treatment in 1999 (For grades 1–8 in early 1998)	0.73	0.10	0.71	0.13 ↑	0.02	0
Among pupils enrolled in 1999						
Round 1 (March–April 1999) Albendazole	0.55	0.08	0.46	0.08	0.01	0
Round 1 (March–April 1999) Pranziquantel[b]	0.53 ↑	0.07 ↑	0.45 ↑	0.07	0.01 ↓	0
Round 2 (October–November 1999), Albendazole	0.65	0.09	0.66	0.11	0.01	0

[a] Original footnote: Data for grades 1–8. Since month of birth information is missing for most pupils, precise assignment of treatment-eligibility status for girls born during the 'threshold' year is often impossible; all girls who turn 13 during a given year are counted as 12-year-olds (eligible for deworming treatment) throughout for consistency.

[b] Original footnote: Praziquantel figures in Table III refer only to children in schools meeting the schistosomiasis-treatment threshold (30 per cent prevalence) in that year.

↓ The replication-results coefficients are up to 0.01 lower than original results.

↑ The replication-results coefficients are up to 0.01 higher than original results.

4.5 Table IV: Proportion of pupil transfers across schools (p.172). The authors used analysis of attendance records to determine if there were movements between schools during the course of the study.

In the table, all children in the study are described simultaneously. For example, the top-left cell indicates that out of all children in the study, 0.5 per cent made a transfer from a Group 1 school to a different Group 1 school in 1998. The denominator is therefore the same for all cells in the table (n = 34,792).

Table 6: Table IV: Proportion of pupil transfers across schools

	1998 transfer to a			1999 transfer to a		
School in early 1998 (pretreatment)	Group 1 school	Group 2 school	Group 3 school	Group 1 school	Group 2 school	Group 3 school
Group 1	0.005	0.007	0.007	0.032	0.026	0.027
Group 2	0.006	0.007	0.008	0.026	0.033	0.027
Group 3	0.01	0.01	0.006	0.022	0.036	0.022
Total transfers	0.021 ↓	0.024	0.021 ↓	0.08	0.095	0.076

↓ The replication-results coefficients are up to 0.01 lower than original results.

No footnotes were used in the original table.

We noted two rounding errors in this table. As noted in the text, the transfer rates for 1999 are much higher, as the authors counted changes of school between academic years as occurring in the later year. The authors performed no statistical tests to look for differential transfer rates at the level of groups or schools, but on gross inspection, it does seem reasonable to conclude that, at least at the level of the groups, these are similar. The authors counted a maximum of one transfer per pupil per year; they did this as a simplification. We note that it was not possible to detect when children transferred to a school outside of the study area or to a school within the study area that was not involved in the project.

4.6 Table V: January to March 1999, health and health-behaviour differences between Group 1 (1998 Treatment) and Group 2 (1998 comparison) schools (p.173)

This table reports various health-related outcomes for Groups 1 and 2, mainly with data from the beginning of 1999, before Group 2 had received treatment. There is no baseline parasitological data from Group 2 in 1998, as this information was not collected for ethical reasons. In this table, there appear to be five rounding errors and six errors in the denominators described in the table footnotes; these do not affect major results.

Table 7: Table V: January to March 1999, health and health-behaviour differences between Group 1 (1998 treatment) and Group 2 (1998 comparison) schools[a]

	Group 1	Group 2	Group 1–Group 2
Panel A: *Helminth infection rates*			
Any moderate–heavy infection, January–March 1998	0.38	--	--
Any moderate–heavy infection, 1999	0.27	0.52	−0.25***
			(0.06)
Hookworm moderate–heavy infection, 1999	0.06	0.22	−0.16***
			(0.03)
Roundworm moderate–heavy infection, 1999	0.09	0.24	−0.15***
			(0.04)
Schistosomiasis moderate–heavy infection, 1999	0.08	0.18	−0.10* [c]
			(0.06)
Whipworm moderate–heavy infection, 1999	0.13	0.17	−0.04
			(0.05)
Panel B: *Other nutritional and health outcomes*			
Sick in past week (self-reported), 1999	0.41 ↓	0.45	−0.04** ↑
			(0.02)
Sick often (self-reported), 1999	0.12	0.15	−0.03**
		5	(0.01)
Height-for-age z-scores, 1999 (low scores denote undernutrition)	−1.13	−1.22	0.09* ↓
			(0.05)
Weight-for-age z-scores, 1999 (low scores denote undernutrition)	−1.25	−1.25	−0.00
			(0.04)
Haemoglobin concentration (g/L), 1999	124.8 ↑	123.2 ↑	1.6
			(1.4)
Proportion anemic (Hb < 100g/L), 1999	0.02	0.04	−0.02** [c]
			(0.01)
Panel C: *Worm-prevention behaviours*			
Clean (observed by fieldworker), 1999	0.59	0.60	−0.01
			(0.02)
Wears shoes (observed by fieldworker), 1999	0.24	0.26	−0.02
			(0.03)
Days contact with fresh water in past week (self-reported), 1999	2.4	2.2	0.2
			(0.03)

[a] Original footnote: These averages of individual-level data for grade 3–8 pupils; disturbance terms are clustered within schools. Robust standard errors in parentheses. Significantly different than zero at 99 (***), 95(**) and 90 (*) per cent confidence.

[b] Excerpt from original footnote: [original/replication] observations for parasitological results: 2,328 (862 Group 1, [1,467/1,466] Group 2); observations for haemoglobin results: 778 ([292/290] Group 1, [486/479] Group 2); observations for 1999 Pupil Questionnaire health outcomes: [9,102/9,039] ([3,562/3,545] Group 1, [5,540/5,497] Group 2 and Group 3).

[c] The results are not statistically significant in the replication study.

↓ The replication-results coefficients are up to 0.01 lower than original results.

↑ The replication-results coefficients are up to 0.01 higher than original results.

In this table, there are two discrepancies in the annotation of significance of results. For the first of these, the original authors recorded the difference in schistosomiasis-infection proportion as being significant at 90 per cent significance; however, as p = 0.101, this actually falls outside of the limit of significance. For the second of these, the difference in prevalence of anaemia (Hb<100 g/L; Group 1 0.021, Group 2 0.040, difference = −0.019, se 0.014) has a p-value of 0.194. The original authors recorded this as being significant at 95 per cent confidence (with**), when, in fact, this is a non-significant result.

As shown in the footnotes to the table, there are substantially more observations available for results derived from the Pupil Questionnaire (top four variables in Panel B, all of Panel C, approximately 9,000 pupils) than for the parasitological (approximately 2,300) and haematological (approximately 800) investigations. The column of results headed ‘Group 2’ does actually contain some observations based on pupils in Group 3, but as neither group received the intervention in year 1 (1998), this has little bearing on the interpretation.

In terms of pupil self-reported and fieldworker-observed results, the results from the start of 1999 (as shown in this table) suggest that pupils in Group 1 (Treatment) schools perceived themselves to be ‘sick’ less frequently (both ‘in past week’ and ‘often’) than pupils in Group 2 (control in year 1). However, for fieldworker observations (‘clean’ and ‘wearing shoes’), we detected no differences between these groups. At the start of the study in 1998 (see Table I results), the findings had been different: at that time, fieldworker observations suggested pupils in Group 2 and 3 schools were more frequently ‘clean’, but pupils themselves reported no differences in their self-perceived health status. Overall, these self-reported and fieldworker-observed results provide some mixed findings; there may have been some modest alterations in pupils’ self-perceived health and behaviours at different points in the study, but these were only measured at two time points in the two-year study period. As this was an unblinded study, Hawthorne effects may have played a role in self-reported outcomes and knowledge of intervention status may have affected fieldworker observations.

The authors calculated the standard errors of comparisons between groups in this table to include allowance for clustering within schools, based on the school the child was in at the start of the study. Furthermore, they made an additional adjustment for robust standard errors. Technically, they achieved this by performing individual-level linear regressions for the relevant observations of the particular outcome in STATA with a single independent variable as an indicator parameter (for example, Group 1 or Group 2) and specifying adjustment of the confidence intervals for clustering and for robust standard errors. The generic format of the STATA commands used was thus:

```
regress outcome_variable indicator_variable_of_group_status, robust
cluster(sch_id_variable);
```

We believe that this format of command makes an appropriate allowance for the clustered nature of the data for these comparisons. The authors used commands of this format for all comparisons between groups in Tables I, V, VI and VIII.

4.7 Table VI: Deworming health externalities within schools, January to March 1999 (p.179)

This table presents various results, with the comparisons it makes (and the numbers of individuals on which they are based) varying through the progress of the table. The headings of the columns are not entirely clear: for example, ‘Group 2, treated in 1999’ means ‘Children who will go on to receive drug treatment in 1999 but have not received it yet’. Furthermore, denominators vary from panel to panel. Panel A is largely composed of baseline comparisons between Groups 1 and 2 in 1998; fieldworkers had only collected parasitological data in Group 1 at this point, so no comparisons were possible. Panel B compares parasitological outcomes in Groups 1 and 2 at the start of 1999, when Group 1 had received one year of the intervention and Group 2 have not yet received the intervention. Panel C makes a crude comparison of the school attendance rate between groups over year 1 of the study (1998).

Overall, the calculations performed in Table VI represent a set of crude comparisons after just the first year of the study. For all the parasitological outcomes, there is a lack of baseline infection data for Group 2 pupils that means secular trends in parasite burden may affect these interpretations — hence, these results are not a comparison of ‘like with like’. The original authors were aware of the limitations of these crude comparisons and place no great emphasis on these results in their interpretation, concentrating rather on results of regressions in Tables VII, IX and X that combine data from all groups across both years of the study.

Table 8: Table VI: Deworming health externalities within schools, January to March 1999[a]

	Group 1, treated in 1998	Group 1, untreated in 1998	Group 2, treated in 1999	Group 2, untreated in 1999	(Group 1, treated 1998)–(Group 2, treated 1999)	(Group 1, Untreated 1998) – (Group 2, Untreated 1999)
Panel A: Selection into treatment						
Any moderate–heavy infection, 1998	0.39	0.44	–	–	–	–
Proportion of 1998 parasitological sample tracked to 1999[b]	0.36	0.36 ↓	–	–	–	–
Access to latrine at home, 1998	0.84 ↑	0.8	0.81	0.86	0.03 (0.04)	−0.06 (0.05)
Grade progression (=grade−[age−6]), 1998	−2.0	−1.8	−1.8	−1.8	−0.2**[d] (0.1)	−0.0 (0.2)
Weight-for-age (z-score), 1998 (low scores denote undernutrition)	−1.58	−1.52	−1.57	−1.46	−0.01 (0.06)	−0.06 (0.11)
Malaria/fever in past week (self-reported), 1998	0.37	0.41	0.4	0.39	−0.03 (0.04)	−0.01 ↑ (0.06)
Clean (observed by fieldworker), 1998	0.53	0.59	0.6	0.66	−0.07 (0.05)	−0.07 (0.1)
Panel B: Health Outcomes Girls <13 years and all boys						
Any moderate–heavy infection, 1999	0.24	0.34	0.51 6	0.55	−0.27*** (0.06)	−0.21** (0.1)
Hookworm moderate–heavy infection, 1999	0.04	0.11	0.22	0.2	−0.19*** (0.03)	−0.09* ↑ (0.05)
Roundworm moderate–heavy infection, 1999	0.08	0.12	0.22	0.3	−0.14*** (0.04)	−0.18** (0.07)
Schistosomiasis moderate–heavy infection, 1999	0.09	0.08	0.2	0.13	−0.11* (0.06)	−0.05 (0.06)
Whipworm moderate–heavy infection, 1999	0.12	0.16	0.16	0.2	−0.04 [(0.16)/(0.05)][f]	−0.05 (0.09)
Girls ≥ 13 years						
Any moderate–heavy infection, 1998	0.31	0.28 ↑	–	–	–	–
Any moderate–heavy infection, 1999	0.27	0.43 ↑	0.32	0.54	−0.05 (0.17)	−0.10 ↓ (0.09) (0.09)
Panel C: School Participation						
School participation rate, May 1998 to March 1999[c]	0.872	[0.764/ 0.774][f]	0.808	[0.684/ 0.690][f]	0.064**[e] (0.032) ↑	[0.080**/ 0.084**][f] (0.039) ↓

↓ The replication-results coefficients are up to 0.02 lower than original results.
↑ The replication-results coefficients are up to 0.02 higher than original results.
[a] Excerpt from original footnote: [original/replication] Observations for parasitological results: [670/669] Group 1 treated in 1998, [77/76] Group 1 untreated in 1998, [873/874] Group 2 treated 1999, [352/349] Group 2 untreated in 1999.
[b] Original footnote: We attempted to track a random sample of half the original 1998 parasitological sample. Because some pupils were absent, had dropped out or had graduated, we were only able to resurvey 72 per cent of this subsample.
[c] See original paper.
[d] The results are not statistically significant in the replication study.
[e] The results are significant at 90 (*) per cent confidence level in the replication study.
[f] The original results are within the brackets, followed by a backslash and then the replication findings.

In this table, there appear to be 20 differences between the results presented and those obtained from reanalysis, including two errors in annotation of significance of results; these differences are largely unimportant. We also found the labelling for the second row in this table somewhat unclear: the author obtained these values (0.36) by multiplying the proportion of the population resampled ('half' = 0.5, as described in footnote b) by the proportion of the group located at follow up (= 0.72).

4.8 Coding local school and pupil population densities

Local school and pupil population densities are important in the estimation of between-school indirect effects in this analysis. Averages of these values are presented in Table I and used in regressions in Tables VII, IX and X. At this point, it is pertinent to explain some coding issues that occurred in calculation of these values in the original analysis.

The authors described to us that there were two coding errors present in the steps determining the original local population-density figures.

The original code resulting in this error was as follows:

```
matrix CLOSE_D = J([_N], 12, 1000)
```

which should have been written as (difference shaded)

```
matrix CLOSE_D = J([_N], 75, 1000)
```

This code was problematic, as it erroneously limited the number of schools that could be included in this matrix calculation to 12, rather than allowing up to 75 as intended.

In addition, there were six further instances where 12 was written instead of 75 in similar lines of code. The effect of this coding error was to truncate the number of schools counted in the school and population densities to 12, rather than allowing all 75 schools to be included in this count. Since there were never more than 12 schools located at distances within three kilometres from any given PSDP school, this coding error did not affect school- and population-density figures in the published paper for distances of 1–3 kilometres. However, it affected density figures for distances of 3–6 kilometres.

A second coding error was present that miscalculated local density figures for three of the schools: school number 108 (in Group 1), 109 (in Group 2) and 115 (in Group 3). The code was as follows:

```
if (wgrp[`x', 1] == `i') {;
        matrix S_TEMP1[`i',`j'] = -1;
        matrix S_TEMP2[`i',`j'] = -1;
        matrix S_TEMP3[`i',`j'] = -1;
                }
```

This code was problematic as it erroneously assigned these three schools into a '-1' category, where it ignored their populations when calculating the local densities.

These errors affected some of the results in Table I and virtually all of the parameter values in Table VII, IX and X in the original paper. In addition to rectifying these issues, the authors have also subsequently employed a Geographic Information System computer package to convert the GPS information into more precise measures of distance. The original variables were generated manually using a simple mathematical formula to covert GPS coordinates into distances. The original authors provided the updated versions of Tables I, VII, IX and X displayed in this document to us; these are indicated as UPDATED.

4.9 Table VII: Deworming health externalities within and across schools, January to March 1999 (p.185)

Tables VII, IX and X present the result of different linear regression models, with the dependent variable relating to worm infections (Table VII), school attendance (Table IX) and exam performance (Table X). The underlying equation specifying each of these regressions is explained on pp.175–176 of the original report. Each column of these tables represents a different (numbered) regression model and the values in the column represent the coefficient estimates for different variables in each model. Where the tables give no values for a particular row in the regression, this indicates that the regression did not include that term. The models also include other parameters (beyond those whose coefficient values are shown); their presence is indicated in the row above the number of observations.

Table VII describes the within-school indirect benefits of deworming treatment on worm infections and provides data that the authors later used to calculate the between-school indirect effect and overall effect of the intervention on deworming.

The coefficient estimate of the Group 1 indicator variable in regression (2) (annotation c) is the value used to obtain the within-school indirect effect of the intervention on worm infection; this value represents a major finding of the study. Originally, the authors reported this as being 12 per cent, though the result was actually 13 per cent after correction of a rounding error. They used two parameter values in regression (1) (annotation b) to calculate the average between-school indirect effect of the intervention across all schools. Note that the value labelled ‘naïve effect’ is not a perfect representation of this quantity: ideally, the naïve effect is obtained from a regression that does not include any parameters relating to externalities. However, in this case, the difference arising from inclusion of the other parameters is negligible, so the values shown in the ORIGINAL and UPDATED tables correctly represent the magnitude of the naïve effect to two decimal places. We found two rounding errors in the table, as shown below.

When we corrected the errors in local population-density calculations and applied the authors’ modifications of methods to calculate distances, we obtained the following revised table of results. All parameter estimates are affected.

Table 9: ORIGINAL Table VII: Deworming health externalities within and across schools, January to March 1999

	Any moderate–heavy helminth infection, 1999			Moderate–heavy schistosomiasis infection, 1999			Moderate–heavy geohelminth infection, 1999		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Indicator for Group 1 (1998 treatment) School	−0.25***[a]	−0.12* ↑[c]	−0.09	−0.03	−0.02	−0.07	−0.20***	−0.11**	−0.03
	(0.05)	(0.07)	(0.11)	(0.03)	(0.04)	(0.06)	(0.04)	(0.05)	(0.09)
Group 1 pupils within 3 km (per 1000 pupils)	−0.26***[b]	−0.26***	−0.11[d]	−0.12***	−0.12***	−0.11**	−0.12*	−0.12*	−0.01
	(0.09)[b]	(0.09)	(0.13)[d]	(0.04)	(0.04)	(0.05)	(0.06)	(0.07)	(0.07)
Group 1 pupils within 3–6 km (per 1000 pupils)	−0.14**[b]	−0.13**	−0.07[d]	−0.18***	−0.18***	−0.27***	0.04	0.04	0.16
	(0.06)[b]	(0.06)	(0.14)[d]	(0.03)	(0.03)	(0.06)	(0.06)	(0.06)	(0.1)
Total pupils within 3 km (per 1000 pupils)	0.11***	0.11***	0.10**	0.11***	0.11***	0.13***	0.03	0.04	0.02
	(0.04)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Total pupils within 3–6 km (per 1000 pupils)	0.13**	0.13**	0.12*	0.12***	0.12***	0.16***	0.04	0.04	0.01
	(0.06)	(0.06)	(0.07)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)		−0.06*			0.03**			−0.04**	
		(0.03)			(0.02) ↓			(0.02)	
(Group 1 indicator) ∗ Received treatment, when offered		−0.14*			−0.02			−0.10***	
		(0.07)			(0.04)			(0.04)	
(Group 1 indicator) ∗ Group 1 pupils within 3 km (per 1000 pupils)			−0.25*			−0.04			−0.18**
			(0.14)			(0.07)			(0.08)
(Group 1 indicator) ∗ Group 1 pupils within 3–6 km (per 1000 pupils)			−0.09			0.11			−0.15
			(0.13)			(0.07)			(0.1)
Grade indicators, school assistance controls, district exam score control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	2328	2328	2328	2328	2328	2328	2328	2328	2328
Mean of dependent variable	0.41	0.41	0.41	0.16	0.16	0.16	0.32	0.32	0.32

[a] This coefficient is the approximate estimation of naïve effect in the replication study.
[b] Values used to calculate between-school indirect effect in the replication study.
[c] Values used to calculate within-school indirect effect in the replication study.
[d] Values used to calculate overall effect in the replication study.
↓ The replication-results coefficients are up to 0.02 lower than original results.
↑ The replication-results coefficients are up to 0.02 higher than original results.
The footnotes from the original table are not reproduced here.

Table 10: UPDATED Table VII: Deworming health externalities within and across schools, January to March 1999

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

[a] This coefficient is the approximate estimation of naïve effect in the replication study.

[b] Values used to calculate between-school indirect effect in the replication study.

[c] Values used to calculate within-school indirect effect in the replication study.

[d] Values used to calculate overall effect in the replication study.

Note: the original authors acknowledged these differences identified in Table VII as part of their internal replication in 2007.

Based on these fully correct parameter values obtained using the authors' methods, the within-school indirect effect of the intervention on any moderate-to-heavy worm infection is now 18 per cent (se 7 per cent; see annotation c, UPDATED Table VII), which is greater than the original estimate of 12 per cent — this is an upward revision of a major finding of the original study. Similarly, the naïve effect of the study on worm infections is revised upwards from 25 per cent to 31 per cent.

For the between-school indirect effect on having any moderate-to-heavy worm infection, the reasoning the authors used to calculate this was as follows (excerpt from p.187):

> To see this, note that the average spillover gain is the average number of Group 1 pupils located within three kilometres divided by 1000 ($\bar{N}^{T}_{03}$) times the average effect of an additional 1000 Group 1 pupils located within three kilometres on infection rates (γ_{03}), plus the analogous spillover effect due to schools located between three to six kilometres away from the school (refer to equation (1)). Based on the externality estimates in Table VII, regression 1, this implies the estimated average cross-school externality reduction in moderate-to-heavy helminth infections is $[\gamma_{03} * \bar{N}^{T}_{03,1} + \gamma_{36} * \bar{N}^{T}_{36,1}] = [0.26*454+0.14*802]/1000 = 0.23$.

Values for $N^{-T}_{03,1}$ and $N^{-T}_{36,1}$ — the average number of Group 1 pupils living in a radius area of 0–3 kilometres or 3–6 kilometres around all schools, respectively — are necessary for this calculation. The original authors found these values to be 454 and 802 children, respectively. In reanalysis, we found the values for $N^{-T}_{03,1}$ and $N^{-T}_{36,1}$ to be 448 and 1,108 children, respectively. When we apply the corrected parameter estimates from regression (1) (annotation b in UPDATED Table VII) in this calculation, using the corrected values for numbers of Group 1 pupils, we obtain the following result:

> [0.21*448 + 0.05*1108]/1000 = 0.15 = 15 per cent improvement (= between-school indirect effect).

The authors calculated the standard error associated with this result from a linear combination of variances and covariances via matrix algebra; they found that this result has a standard error of 0.11. The original analysis files did not show this calculation. Rather, the authors kindly provided it upon request. As the 15 per cent improvement is only 1.4 standard error intervals from zero, this should now be considered a non-significant result.

On pp.187–188 of the original report, the authors calculate the 'true reduction in moderate-to-heavy infections' for pupils in Group 1 schools, which incorporates both direct and indirect effects. This was originally reported as a 35 per cent reduction (se 9 per cent). This result was obtained by addition of the naïve result in regression (1) to an average between-school indirect benefit for control schools, derived from multiplication of parameters from regression (3) (annotation d, Table VII) against corresponding average population values around control schools. The values for the between-school indirect benefit calculation in the UPDATED results are (496*0.10 + 1027*0.08)/1,000 = 0.132 = 13 per cent. Therefore, an overall result of 31 per cent (naïve effect) plus 13 per cent (average cross-school externality for control schools) is a 44 per cent overall deworming benefit, considerably larger than the original findings. The original study authors calculated a 12 per cent standard error associated with this updated result, making the result highly statistically significant ($p<0.001$). The code to perform these calculations of the overall results on worm infection was not present in the original analysis files, but the authors also provided it on request.

We note there is also a coding error in the creation of the 'any moderate/heavy geohelminth infection' variable in this table. This only affects regressions (7) to (9) in this table, so does not affect the major outcome variable used for worm infections ('any moderate/heavy worm infection'). The revised results following correction of this issue are shown in the UPDATED version of Table VII.

4.10 Coding for weighting of clusters in Tables VIII and IX

In Tables I, VI, VII and VIII, the authors state that the clusters are school averages weighted according to pupil population. In Tables IX and X, there is no explicit description in the original report of how weighting was performed. The footnotes for these tables state that the regression is performed at the level of individual; weighting would thus be considered unnecessary. In fact, the authors weighted Tables IX and X by the number of pupil observations in a school and not by the number of pupils. This is a subtle but important difference.

The technical description of the processing steps the researchers performed for calculations presented in Table VIII is as follows:

The analysis file called 'namelist' holds data in a long format such that each row in the dataset describes one observation of one pupil at a fieldworker visit. One pupil, defined by the pupil id (pupid), has many observations. For example, in year 1 (1998), the pupil with pupid 1330065 has three rows relating to observations of their attendance and is in school number 182.

At the start of processing for Table VIII, one can collapse the data with the following command:

```
collapse prs wgrp* (count) np = pupid, by(sch98v1)
```

This command creates a dataset with one row representing each school, such that there are 75 rows. In each row, there is a summary variable called np that describes the number of times that each specific pupil id number within that school has occurred. Crucially, the format of this 'collapse' command counts repeated appearances of the same pupil id as separate instances, such that each individual pupil-observation contributes to the np total. Returning to the previous example, pupil number 1330065 contributes three to the overall count in school 182 in 1998 (total of 11,709 pupil observations).

For subsequent commands used in calculating data for Table VIII, the authors use the value obtained for np to weight the schools in relation to their size, as follows:

```
summ prs [aw = np]
```

As the np value represents the number of school-attendance observations, it is more correct to describe the calculations in Table VIII as having been weighted by pupil observations than by the school population.

The technical description of the same process in Table IX is as follows: the analysis again starts with the 'namelist' dataset in the long format, with each row representing one observation of one pupil. After various data-processing steps, the

following command is used to collapse the data to represent one row per pupil (text within square brackets added):

```
collapse (mean) sch98v1 prs [... various other variables ....] wgrp (sum) obs
        if ([...various conditions... ]), by(pupid yr[1998 or 1999])
```

In these data, the variable obs represents the number of pupil observations that were performed for that pupil. In subsequent regressions, this obs variable is used to weight the regressions. See, for example, the following:

```
regress prs t_any elg98 p1 mk96_s Y98sap* sap* lstd* lsem* [aw = obs],
robust cluster(sch98v1)
```

This means that the authors also weighted the regressions in Table IX by the number of pupil observations.

This technical issue of weighting regressions is thus incorrectly described in Tables VIII and IX. Both of these tables present analyses that are weighted by the number of pupil observations rather than the number of pupils. We return to discussion of the issues arising from weighting regressions by the number of observations in the 'Statistical and scientific replication' (Davey *et al*. 2014).

4.11 Table VIII: School participation, school-level data (p.191)

This table describes average school-attendance proportions for different categories of pupils in 1998 and 1999. In one row of the table, there appears to be a minor calculation error affecting the entire row, though the results are not altered substantially. There is also one error in annotation of significance. These do not affect the major findings of the paper.

We also note that the values shown in this table are based on an intermediate version of a data file called 'namelist', which subsequently underwent some minor corrections before the final version was produced. Thus, many of the results in the original version of Table VIII have slight inaccuracies due to the erroneous use of this intermediate namelist file. For completeness, we show a corrected version of Table VIII below. Many of the results for this table are slightly altered from the original version, but the substantive interpretation remains the same throughout.

Table 11: ORIGINAL Table VIII: School participation, school-level data

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)		
Panel A:					
First year post-treatment (May 1998 to March 1999)	*1st year treatment*	*Comparison*	*Comparison*	*Group 1 – (Groups 2 & 3)*	*Group 2 – Group 3*
Girls <13 years and all boys	0.841	0.731	0.767	0.093***	−0.037
				(0.031)	(0.036)
Girls ≥13 years	0.864	0.803	0.811	0.057**[a]	−0.008
				(0.029)	(0.034)
Preschool, Grade 1, Grade 2 in early 1998	0.795 ↑	0.688 ↑	0.703 ↑	0.100***	−0.018
				(0.037)	(0.043)
Grade 3, Grade 4, Grade 5 in early 1998	0.88	0.789	0.831	0.070***	−0.043
				(0.024)	(0.029)
Grade 6, Grade 7, Grade 8 in early 1998	0.934	0.858	0.892	0.059***	−0.034
				(0.021)	(0.026)
Recorded as 'dropped out' in early 1998	0.064	0.05	0.03	0.022	0.02
				(0.018)	(0.017)
Females	0.855	0.771	0.789	0.076***	−0.018
				(0.027)	(0.032)
Males	0.844	0.736	0.78	0.088***	−0.044
				(0.031)	(0.037)
Panel B:					
Second year post-treatment (March to November 1999)	*2nd year treatment*	*1st year treatment*	*Comparison*	*Group 1 – Group 3*	*Group 2 – Group 3*
Girls <13 years and all boys	0.713	0.717	0.663	0.050*	0.055*
				(0.028)	(0.028)
Girls ≥14 years	0.627	0.649	0.588	0.039	0.061*
				(0.035)	(0.035)
Preschool, Grade 1, Grade 2 in early 1998	0.692	0.726	0.641	0.051	0.085**
				(0.034)	(0.034)
Grade 3, Grade 4, Grade 5 in early 1998	0.75	0.774	0.725	0.025	0.049**
				(0.023)	(0.023)
Grade 6, Grade 7, Grade 8 in early 1998	0.77	0.777	0.751	0.02	0.026
				(0.027)	(0.028)
Recorded as 'dropped out' in early 1998	0.176	0.129	0.056	0.120*	0.073
				(0.063)	(0.053)
Females	0.716	0.746	0.648	0.067**	0.098***
				(0.027)	(0.027)
Males	0.698	0.695	0.655	0.043	0.041
				(0.028)	(0.029)

[a] The results are significant at the 90(*) per cent confidence level in the replication study.
↑ The final digits are 0.01 to 0.004 higher in replication study. The row should read (0.797 0.689 0.707).
The footnotes from the original table are not reproduced here.

Table 12: UPDATED Table VIII: School participation, school-level data

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)		
Panel A:					
First year post-treatment (May 1998 to March 1999)	*1st year treatment*	*Comparison*	*Comparison*	*Group 1 – (Groups 2 & 3)*	*Group 2 – Group 3*
Girls <13 years and all boys	0.841	0.731	0.766	0.093***	−0.035
				(0.030)	(0.035)
Girls ≥13 years	0.868	0.804	0.82	0.056*	−0.016
				(0.031)	(0.036)
Preschool, Grade 1, Grade 2 in early 1998	0.797	0.689	0.707	0.100***	−0.019
				(0.037)	(0.043)
Grade 3, Grade 4, Grade 5 in early 1998	0.877	0.788	0.827	0.071***	−0.039
				(0.024)	(0.029)
Grade 6, Grade 7, Grade 8 in early 1998	0.934	0.859	0.891	0.058***	−0.032
				(0.021)	(0.025)
Recorded as 'dropped out' in early 1998	0.066	0.051	0.03	0.024	0.022
				(0.018)	(0.017)
Females	0.855	0.771	0.789	0.076***	−0.018
				(0.027)	(0.032)
Males	0.844	0.736	0.78	0.088***	−0.044
				(0.088)	(0.037)
Panel B:					
Second year post-treatment (March to November 1999)	*2nd year treatment*	*1st year treatment*	*Comparison*	*Group 1 – Groups 3*	*Group 2 – Group 3*
Girls <13 years and all boys	0.716	0.718	0.664	0.051*	0.054*
				(0.027)	(0.027)
Girls ≥14 years	0.627	0.649	0.588	0.039	0.061
				(0.035)	(0.035)
Preschool, Grade 1, Grade 2 in early 1998	0.692	0.725	0.641	0.051	0.084**
				(0.034)	(0.034)
Grade 3, Grade 4, Grade 5 in early 1998	0.749	0.766	0.720	0.029	0.046**
				(0.022)	(0.023)
Grade 6, Grade 7, Grade 8 in early 1998	0.781	0.790	0.754	0.027	0.036
				(0.025)	(0.026)
Recorded as 'dropped out' in early 1998	0.188	0.130	0.062	0.126*	0.068
				(0.066)	(0.056)
Females	0.716	0.746	0.649	0.067**	0.097***
				(0.027)	(0.027)
Males	0.698	0.695	0.655	0.043	0.040
				(0.028)	(0.029)

Note: The original authors acknowledged these differences identified in Table VIII as part of their internal replication in 2007.

4.12 Table IX: School participation, direct effects and externalities (p.194)

This table describes the major results of the analysis on school participation, in the form of parameter estimates from different regressions. Regression (1) calculates the naïve benefits of the intervention (annotation a), and the authors use the values obtained in regression (3) later in calculations to determine the overall effect and the between-school indirect effect. The authors obtained the within-schools indirect benefit in regression (5) (annotation c).

Table 13: ORIGINAL Table IX: School participation, direct effects and externalities (dependent variable: average individual school participation per year)

	OLS (1)	OLS (2)	OLS[b] (3)	OLS (4)	OLS (5)	OLS (6)	IV-2SLS (7)
				May 98–March 99	May 98–March 99	May 98–March 99	May 98–March 99
Moderate–heavy infection, early 1999						−0.028***	−0.203* ↑
						(0.01)	(0.094)
Treatment school (T)	0.051***[a,d]						
	(0.022)[a]						
First year as treatment school (T1)		0.062***	0.060***	0.062* ↑[d]	0.056***[c]		
		(0.015)	(0.015)	(0.022)	(0.02)[c]		
Second year as treatment school (T2)		0.040*	0.034* ↑				
		(0.021)	(0.021)				
Treatment-school pupils within 3 kms (per 1,000 pupils)			0.044**		0.023		
			(0.022)		(0.036)		
Treatment-school pupils within 3–6 kms (per 1,000 pupils)			−0.014		−0.041		
			(0.015)		(0.027)		
Total pupils within 3 kilometres (per 1,000 pupils)			−0.033**		−0.035*	0.018	0.021
			(0.013)		(0.019)	(0.021)	(0.019)
Total pupils within 3–6 kilometres (per 1,000 pupils)			−0.010		0.022	−0.010	−0.021
			(0.012)		(0.027)	(0.012)	(0.015)
Indicator received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)					0.100***		
					(0.014)		
(First year as treatment-school indicator)∗ (Received treatment, when offered)					−0.012		
					(0.02)		
1996 district exam score, school average	0.063***	0.071***	0.063***	0.058[d]	0.091** ↑	0.021	0.003 ↑
	(0.021)	(0.02)	(0.02)	(0.032)	(0.038)	(0.026)	(0.023) ↓
Grade indicators, school-assistance controls and time controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.23	0.23	0.24	0.33	0.36	0.28	–
Root MSE [original/replication]	0.273	0.272	0.272	0.223 ↑	0.219	0.15	[0.073/0.069]
Number of observations	56487	56487	56487	18264	18264	2327	49(schools)
Mean of dependent variable [original/replication]	0.747	0.747	0.747	[0.784/ 0.793]	[0.784/ 0.793]	0.884	0.884 ↓

[a] This coefficient is the approximate estimation of naïve effect in the replication study.
[b] Regression (3) is affected by an error in variable creation. The values are used for later calculations in the replication study.
[c] The values are used to calculate within-school indirect effect in the replication study.
[d] These results present errors in their significance levels. See updated Table IX for the correct level of significance in the replication study.
↓ The replication-results coefficients are up to 0.02 lower than original results.
↑ The replication-results coefficients are up to 0.02 higher than original results.
The footnotes from the original table are not reproduced here.

In the original version of Table IX, which represents the major findings on school-attendance regressions, there were various rounding errors, errors in annotation of significances and other calculation errors. Most of these do not affect the major results of the study. In regression (3) there is a further coding error in the creation of one of the variables that is used for these calculations. The original line of code was as follows:

```
gen pop_36k_original = pop1_3km_original;
```

This should actually have been written as follows (difference shaded):

```
gen pop_36k_original = pop1_36k_original;
```

This error meant that the authors used the number of treatment (Group 1) students within 0–3 kilometres to create a variable that was intended to describe the population of these students within 3–6 kilometres. When we corrected both this error in variable creation (which was only used within regression (3) in Table IX) and also the error relating to local populations densities described earlier, we obtained the values for Table IX that are shown in Table 14 below.

The correction of errors largely does not affect the estimates for the naïve effect of the intervention across both years of the study (annotation a, both versions of Table IX) and the size of this effect is slightly increased (originally 5.1 per cent [se 2.2 per cent]; in reanalysis, 5.7 per cent [se 1.4 per cent]). The replication more precisely estimates the result, which is now significant at 99 per cent confidence, whereas it was previously significant at 95 per cent confidence. This result represented a major finding of the original study, and the various coding errors present in the original analysis largely do not affect it. The calculated effect is marginally greater than previously determined and now has a higher degree of statistical significance.

Table 14: UPDATED Table IX: School participation, direct effects and externalities (dependent variable: average individual school participation per year)

	OLS (1)	OLS (2)	OLS (3)	OLS (4) May 98–March 99	OLS (5) May 98–March 99	OLS (6) May 98–March 99	IV-2SLS (7) May 98–March 99
Moderate–heavy infection, early 1999						−0.025**	−0.195**
						(0.010)	(0.096)
Treatment school (T)	0.057***[a]						
	−0.014[a]						
First year as treatment school (T1)		0.063***	0.062***[b]	0.062***	0.056***[c]		
		(0.015)	(0.014)[b]	(0.022)	(0.020)[c]		
Second year as treatment school (T2)		0.039*	0.033[b]				
		(0.021)	(0.021)[b]				
Treatment-school pupils within 3 kilometres			0.040*[b]		0.022		
(per 1,000 pupils)			(0.022)[b]		(0.032)		
			−0.024[b]		−0.067***		
Treatment-school pupils within 3–6 kilometres			(0.015)[b]		(0.020)		
(per 1,000 pupils)			−0.031**[b]		−0.040**	0.014	−0.029*
			(0.012)[b]		(0.016)	(0.014)	(0.016)
Total pupils within 3 kilometres			0.012[b]		0.035***	0.016*	0.008
(per 1,000 pupils)			(0.009)[b]		(0.011)	(0.034)	(0.009)
Total pupils within 3–6 kilometres (per 1,000 pupils)							
Indicator received first year of deworming treatment, when offered (1998 for Group 1, 1999 for Group 2)					0.0104***		
					(0.014)		
(First year as treatment-school indicator) *	0.071***	0.070***	0.077***	0.58*	−0.013	0.020	−0.000
(Received treatment, when offered)	(0.021)	(0.021)	(0.022)	(0.032)	(0.020)	(0.024)	(0.022)
1996 district exam score, school average					0.106***		
					(0.034)		
Grade indicators, school-assistant controls, and time controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.22	0.22	0.22	0.33	0.37	0.29	--
Root MSE	0.279	0.279	0.278	0.223	0.217	0.150	0.069
Number of observations	56496	56496	56496	18215	18215	2327	49 (schools)
Mean of dependent variable	0.747	0.747	0.747	0.793	0.793	0.884	0.882

[a] This coefficient is the approximate estimation of naïve effect in the replication study.
[b] The replication study uses these values for later calculations of overall and between-school indirect effects.
[c] The replication study uses these values to calculate within-school indirect effects.
Note: the original authors acknowledged the differences identified in Table XI as part of their internal replication in 2007.

The authors used parameter estimates calculated in regression (3) in Table IX for a series of further calculations to determine effects on school attendance, both the overall effect on children who receive treatment and also the between-school indirect effect. The values for these effects on school participation according to the authors' methods are as follows:

Combined effect across 98–99	Absolute improvement in school attendance (%)			
	Original result p.196 (%)	Standard error	Result from fully corrected reanalysis (%)	Standard error
Average between-school indirect effect	2.0	1.3	−1.7	3.0
Average overall effect	7.5	2.7	3.9	3.2

The results on the effects on school attendance represent the key findings of the paper. In this corrected reanalysis, the direct effect of the intervention has been revised upwards, from 5.1 per cent to 5.7 per cent. However, the between-school indirect effect is reversed entirely (from +2.0 to −1.7 per cent), and we find an increased standard error interval (from 1.3 to 3.0 per cent). The original authors reported this result (2.0 per cent, se 1.3 per cent) as being 'marginally statistically significant', but it is now within one standard error interval of zero and thus is clearly non-significant.

The overall effect on school attendance is also substantially reduced (from 7.5 per cent to 3.9 per cent absolute improvement) and is now only slightly more than one standard error interval away from zero — thus, it is also non-significant.

We note that the original authors did not acknowledge the alterations in these results in their own 2007 internal replication, although their data-recording (log) files from that time show the results as found above.

The effect that we calculated for within-school indirect effect (5.6 per cent, se 2.0 per cent) in regression (5) (annotation c, both versions of Table IX) remains the same in reanalysis as originally reported. This within-school indirect effect is similar to the direct effect of the intervention — 6.2 per cent (se 2.2 per cent), from regression (4) — but note that both of these results represent data from the first year of the study only. This prompts the question: If all children in treatment schools saw fairly similar benefits in terms of school attendance, regardless of whether or not they received the drug treatment for deworming, then how much of the effect on school attendance is attributable to drug administration? We discuss this issue as part of a wider discussion of the evidence for causality in the subsequent 'Statistical and scientific replication' (Davey *et al*. 2014).

4.13 Table X: Academic examinations, individual-level data (p.199)

This table describes the major results of the analysis on exam performance, in the form of parameters from regressions. In the results presented, there appear to be two rounding errors and errors in the representation of the denominators for regressions (1) and (2). These make no difference to the overall interpretation of results from this table.

Table 15: ORIGINAL Table X: Academic examinations, individual-level data

	Dependent variable: ICS exam score (normalised by standard)		
	(1)	(2)	(3) Among those who filled in the 1998 pupil survey
Average school participation (during the year of the exam)	7 0.63*** (0.07)		
First year as treatment school (T1)		−0.032 (0.046)	−0.030 (0.049)
Second year as treatment school (T2)		0.001 (0.073) ↓	0.009 ↓ (0.081)
1996 district exam score, school average	0.74*** (0.07)	0.71*** (0.07)	0.75*** (0.07)
Grade indicators, school-assistance controls and local pupil-density controls	Yes	Yes	Yes
R^2	0.14	0.13	0.15
Root MSE	0.919	0.923	0.916
Number of observations [original/replication]	[24,958/24,979]	[24,958/24,979]	19,072
Mean of dependent variable	0.02	0.02	0.039

↓ The replication-results coefficients are up to 0.02 lower than original results.
The footnotes from the original table are not reproduced here.

When we corrected the error relating to local population densities, we obtained the following table of results; all values in regressions (2) and (3) are affected.

Table 16: UPDATED Table X: Academic examinations, individual-level data

	Dependent variable: ICS exam score (normalised by standard)		
	(1)	(2)	(3) Among those who filled in the 1998 pupil survey
Average school participation (during the year of the exam)	0.63*** (0.07)		
First year as treatment school (T1)		−0.035 (0.047)	−0.036 (0.049)
Second year as treatment school (T2)		−0.015 (0.073)	0.013 (0.088)
1996 district exam score, school average	0.74*** (0.07)	0.72*** (0.07)	0.75*** (0.07)
Grade indicators, school-assistance controls and local pupil-density controls	Yes	Yes	Yes
R^2	0.14	0.13	0.15
Root MSE	0.919	0.923	0.916
Number of observations	24,979	24,979	19,072
Mean of dependent variable	0.02	0.02	0.039

Note: The original authors acknowledged the differences identified in Table X as part of their internal replication in 2007.

The changes in results on academic performance in ICS exams that occurred as a direct result of the intervention are obtained from this table (regression (2) in both versions of Table X).

Direct effect on exam performance	Change in exam performance (standard deviations)			
	Original result p.199–201	Standard error	Result from fully corrected reanalysis	Standard error
1998 – Year 1	−0.032	0.046	−0.035	0.047
1999 – Year 2	0.001	0.073	−0.015	0.079

These corrections do not lead to any change in the interpretation of these findings. The calculations for determining the between-school indirect effect on exam performance (original result −0.049 standard deviations change, standard error 0.052, not statistically significant) were not present in the original analysis files, but the authors kindly provided them, noting that the original standard error was actually 0.051. Using these same calculation methods and applying the corrected local population-density estimates, we found that the revised between-school indirect effect on exam performance changed by 0.006 standard deviations, with a standard error of 0.059.

Appendices. We have not replicated any of these tables. We do not believe any modification of results in the appendices could affect the major findings of the study, as these do not contain any of the major results of the analysis.

5. Discussion

There are a number of issues that we identified in the pure replication that represent decisions on how the study presented the data. These do not represent inaccuracies but may influence the interpretation of the results. Many of these issues apply to several different tables, but we have illustrated them with reference to tables where the effects of these decisions are particularly noticeable.

5.1 Randomisation of schools to different treatment groups

The description in the paper on how the authors allocated schools to different treatment groups (alternating assignment from a sorted list) does not meet a strict definition of randomisation. A more accurate description might be systematic allocation resulting in 'quasi-randomisation'. The authors provided us with a more detailed description of the allocation process they used than they gave in the original manuscript. In practice, as far as can be determined from inspection of the baseline characteristics presented in Table I, the quasi-randomisation process appears to have been broadly successful in terms of the balance of the parameters displayed. There are significant differences in the year of birth between the year groups, but this may arise from differences in the extent of missing data between the groups, as discussed below.

5.2 Presentation of variability within and between groups

Table I describes various baseline characteristics of pupils and schools. For the important characteristic of pupil population in schools, it shows the average school size (in terms of number of pupils) for each group. The extent of variation in school sizes is expressed by showing the standard error of the difference in the mean school sizes across groups. This gives exactly the same value (57.6) for both Groups 1 and 2 when they are each in turn compared to Group 3 this is a correct interpretation of the calculations performed but does not convey the variation in school sizes clearly. The following table contains some other parameters that illustrate the range of school sizes within each group. The spread of sizes is indeed broadly similar between the groups, but there are some extreme sizes in Group 2.

Group (n = 25)	Mean	Median	IQR	Range
1	392.7	352	272 to 489	148 to 748
2	403.8	342	254 to 559	14 to 1,351
3	375.9	360	244 to 505	95 to 675

5.3 Presentation of denominator values

The original paper generally uses earlier tables to present proportions of individuals in a group with a particular characteristic but does not show the raw numerator and denominator data. Further, within tables, there are often changes in the denominators the authors use between different sections of a table. For example, Table 3 displays proportions of pupils receiving treatment in the different study groups in columns. In each of these columns, the denominators of pupils involved changes as the table progresses through the different stages of the study. For Group 1, there are 10,052 pupils eligible for treatment in 1998 in the top panel, there are 9,097 pupils in mid-table

(due to exit of pupils in the oldest age group from follow-up) and there are only 7,059 pupils in the lowest part of the table (ignoring pupils never present at a visit in 1999). The table does not show these changes in the denominator values. There are also minor differences between tables in the total number of pupils in the entire study: the values used in Table I (Panel A) and Table IV (34,792 pupils) differ slightly from Table III (34,810 pupils), although this difference is less than 0.1 per cent of the total.

5.4 Presentation of missing data

Throughout the paper, the authors limit the presentation of the extent of missing data — that is, measurements where no result is recorded. This is particularly evident in Table I. In Panel A of this table, there is a large amount of missing data for year of birth; this information is missing for 17 per cent of children in Group 1, 19 per cent of children in Group 2 and 31 per cent of children in Group 3. The extent of this missing data is not described in the table or the accompanying text. From the children whose age is known, there are highly significant differences in the calculated average age between pupils in the different groups, meaning that age appears to be significantly different at baseline. We cannot retrospectively determine whether these differences in the extent of missing data are due to systematic differences in data collection or due to an actual difference in the average age. The authors make no mention of this possible baseline imbalance in the accompanying text. The extent of missing data varies enormously in this study. Within Table I, for example, in the 'year of birth' variable in Panel A, 22 per cent of the values for individual children are missing, but in the weight-for-age z-score variable in Panel B, less than 0.1 per cent of data are missing. This is perplexing, as the researchers presumably knew the ages of all children for whom weight-for-age was calculated — so why did they not use this age data to complete missing age values elsewhere?

5.5 Annotation of statistically significant results

All the tables in the study make use of a standard approach of annotating the results significant at 99 (***), 95 (**) and 90 (*) per cent confidence, rather than displaying actual p-values. Whilst there is certainly merit in not restricting scientific investigation to using a significance level of 95 per cent (or α of 0.05) one should bear in mind that results at 90 per cent confidence could have occurred by chance on up to 1 in 10 occasions. Thus, for example, the result for the difference in height-for-age z-scores found in Table V (correct results are Group 1 −1.13, Group 2 −1.22, difference 0.08, se 0.05) has a p-value of 0.092. This is correctly annotated as having significance at 90 per cent confidence (*) but can be considered of questionable importance. A p-value of 0.09 means that if there were no true difference, the observed effect (or one more extreme) would occur by chance 9 times in 100.

On a wider level, there is the question of what would have been the most appropriate level of significance to use in these analyses where the authors were performing multiple comparisons. Given that the authors were studying not any single prespecified question but, rather, multiple exploratory comparisons across many different categories of outcome, it would seem prudent to have considered use of a Bonferroni correction for multiple testing or at least to have provided an explanation as to why this was not needed. The authors should have interpreted results at the lower levels of significance in this paper with extreme caution.

5.6 Generalisability of between-school indirect benefit (positive externality) results

One interesting question that arises from the examination of the between-school indirect effects is how generalisable these are to other low-income settings. It seems reasonable to suppose that there are specific geographic factors relating to the physical environment (including rainfall, soil type, average proximity of homes to schools, access to latrines and overall population density) that might play a part in determining the size of these between-school indirect effects of a school-based deworming programme. Given that many (or all) of these factors are highly variable in both time and place in African countries, it is certainly possible that the indirect between-school effects could be substantially greater or smaller in a different location. This would mean that the 'overall' result as reported in this study might be less generalisable than the naïve effect, as the latter does not try to take account of the between-school indirect effect that might have such wide geographical variation.

Overall, we are aware that many of these issues in presentation and interpretation of results represent disciplinary conventions within econometrics and that some of these might have no bearing on the findings of the study. However, the presentation and handing of missing data is considered to be highly important in the analysis of randomised trials, according to the CONSORT framework (Begg *et al*. 1996). In the statistical and scientific replication that follows this pure replication, we will examine the handling of missing data in greater detail. The issue of generalisability is important in any scientific study, and we feel that the naïve results should have assumed greater importance in discussion of results of this study, as they represent the most easily generalisable findings. As noted in section 3.5 of this replication report, the naïve results represent the type of effect that would normally be considered in a pragmatic evaluation of a cluster-randomised trial, though these would not normally be described as naïve.

6 Conclusion

The original results that the authors described were as shown in Table 17 (repeat of Table 1 as earlier). Effects that the original authors interpreted as being beneficial and statistically significant are shaded.

Table 17: Summary of results from original paper

Measure		Direct effect	Indirect effect: within school	Naïve effect	Indirect effect: between school	Overall effect
Health	Worm infection (any mod/hvy inf)	−14% (se 7%)	−12% (se 7%)	−25% (se 5%)	−23% (se 7%)	−35% (se 9%)
	Anaemia (Hb<100g/L)	Not reported	Not reported	−2% absolute prop'n (se 1%)	Not reported	Not reported
	Nutritional status	Not reported	Not reported	WAZ: −0.00 (se 0.04) HAZ: 0.09 (se 0.05)	Not reported	Not reported
School attendance (% increase)		+6.2%† (se 2.2%)	+5.6%† (se 2.0%)	+5.1% (se 2.2%)	+2.0% (se 1.3%)	+7.5% (se 2.7%)
Exam performance (average difference)		Not reported	Non-significant result, data not shown	Not reported	−0.049 sd (se 0.052)	Yr 1 −0.032 (se 0.046) Yr 2 0.001 (se 0.073)

Note: Abbreviations: sd = standard deviation; se = standard error; Hb = haemoglobin; WAZ = weight-for-age z-score; HAZ = height-for-age z-score. Page numbers refer to original study. † = year 1 data only.

Based on this reanalysis using the authors' original approaches, the revised findings are as follows:

Table 18: Summary of results from pure replication

Measure		Direct effect	Indirect effect: within school	Naïve effect	Indirect effect: between school	Overall effect
Health	Worm infection (any mod/hvy inf)	−15% (se 6%)	−18% (se 7%)	−31% (se 6%)	−15% (se 11%)	−44% (se 12%)
	Anaemia (Hb<100g/L)	Not reported	Not reported	−2% absolute prop'n (se 1%)	Not reported	Not reported
	Nutritional status	Not reported	Not reported	WAZ: −0.00 (se 0.04) HAZ: 0.08 (se 0.05)	Not reported	Not reported
School attendance (% increase)		+6.2%† (se 2.2%)	+5.6%† (se 2.0%)	+5.7% (se 1.4%)	−1.7% (se 3.0%)	+3.9% (se 3.2%)
Exam performance (average difference)		Not reported	Not reported	Not reported	0.006 sd (se 0.059)	Yr 1 −0.035 (se 0.047) Yr 2 −0.015 (se 0.079)

Note: Abbreviations: sd = standard deviation; se = standard error; Hb = haemoglobin; WAZ = weight-for-age z-score; HAZ = height-for-age z-score. † = year 1 data only.

For the outcome of worm infection, our pure replication study found beneficial effects similar to or greater than those originally reported, though the between-school indirect effect is more modest than the original authors found and is now not statistically significant. In contrast to the original report, our pure replication found little evidence of

non-worm-related health-related benefits: the intervention did not significantly affect the prevalence of anaemia, and the reported improvement in height-for-age z-score (HAZ) was of very modest size (0.08 standard error intervals of improvement) and had $p = 0.09$.

In relation to school attendance, our pure replication found beneficial effects similar to or stronger than those originally reported for the direct, indirect-within-school and naïve effects. However, after correcting coding errors in the original analysis files, some measurements of effect size were substantially different. The between-school indirect effect of the intervention on school attendance had shifted in direction and was much less precisely estimated; there was now no evidence of a statistically significant effect of this kind. The overall effect on school attendance resulting from the intervention was more modest and less precisely estimated than previously reported; this was also now not statistically significant. As in the original study, our pure replication results on examination performance showed no benefit associated with the deworming intervention. We detected a number of other discrepancies in the results originally presented, but these did not appear systematic in any way and did not have effects on the major findings of the study, other than those already described.

7 Progression to statistical and scientific replication

We aim to apply alternative statistical approaches and scientific theories of change to the effect of the deworming intervention on improving school attendance and exam performance in treatment schools, using the same raw data as in the original study reanalysed here. Specifically, we aim to conduct an analysis in accordance with the CONSORT criteria relevant to cluster-randomised trials (Begg *et al*. 1996), as far as possible. Our investigation will therefore focus on the naïve effects of the original intervention — principally, on the outcome of school attendance. We report this 'statistical and scientific replication' in a separate document (Davey *et al*. 2014).

References

2012. *London Declaration on Neglected Tropical Diseases*. London: Uniting to Combat Neglected Tropical Diseases.

2012. 'Outcomes: The Expert Panel Findings'. *Copenhagen Consensus* Center. Available from: <http://www.copenhagenconsensus.com/Projects/CC12/Outcome.aspx>. [19 July 2012].

Begg, C, Cho, M, Eastwood, S, Horton, R, Moher, D, Olkin, I, Pitkin, R, Rennie, D, Schulz, KF, Simel, D and Stroup, DF, 1996. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA*, 276(8)**,** pp.637–639.

Brooker, S, Miguel, EA, Moulin, S, Luoba, AI, Bundy, DA and Kremer, M, 2000. Epidemiology of single and multiple species of helminth infections among school children in Busia District, Kenya. *East African Medical Journal*, 77(3)**,** pp.157–161.

Davey, C, Aiken, AM, Hayes, RJ and Hargreaves, J, 2014. *Reanalysis 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).

Halloran, ME and Struchiner, CJ, 1991. Study designs for dependent happenings. *Epidemiology*, 2**,** pp.331–338.

Hamermesh, DS, 2007. Replication in Economics. *National Bureau of Economic Research Working Paper Series*, No. 13026.

Hawkes, N, 2012. Deworming debunked. *BMJ*, 346, e8558.

Kremer, M, Moulin, S and Namunyu, R, 2002. *Unbalanced Decentralization: Results of a Randomized School Supplies Provision Program in Kenya*. Washington, DC: Mimeo Brookings Institution.

Miguel, E and Kremer, M, 2004. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), pp.159–217.

Montresor, A, Crompton, David WT, Gyorkos, TW and Savioli, L, 2002. *Helminth control in school-age children: A guide for managers of control programmes*. Geneva: World Health Organization. Available from: <http://www.who.int/iris/handle/10665/42473#sthash.SP61ZOSR.dpuf>. [16 September 2014].

Senn, S, 1994. Testing for baseline balance in clinical trials. *Stat Med*, 13(17), pp.1,715–1,726.

Taylor-Robinson, DC, Maayan, N, Soares-Weiser, K, Donegan, S and Garner, P, 2007. Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin and school performance. *Cochrane Database of Systematic Reviews*, 7, no.CD000371. DOI: 10.1002/14651858.CD000371.pub3.

Appendix 1: Biology of worm infections under analysis

Worm type	Life cycle	Lifespan (all lengths are approximate)	Thresholds for moderate infection (eggs/g faeces)	Treatment used in study
Geohelminths (soil-mediated)				
Hookworm (*N. americanus*)	egg in faeces → **soil** → larvae hatches → skin contact → gut	adult lives 1–10 years in gut, larvae live in moist soil for 2 years	2,000 (WHO) 750 (Brooker *et al*. 2000)*	albendazole 600 mg ('98) / 400 mg ('99) given every 6 months
Roundworm (*A. lumbricoides*)	egg in faeces → **soil** → oral intake of egg → gut	adult lives 1 year in gut eggs last 1–3 years in soil	5,000 (WHO)*	
Whipworm (*T. trichiura*)	egg in faeces → **soil** → oral intake of egg → gut	adult lives up to 5 years in human gut	1,000 (WHO) 400 (Brooker *et al*. 2000)*	
Schistosomiasis (water-mediated)				
Schistosomiasis (*S. mansoni*)	egg in faeces → **freshwater** → snail → cercariae → skin contact → veins around gut	adult lives 4+ years in veins, lifespan in freshwater snail 1 year	100 (WHO) 250 (Brooker *et al*. 2000)*	praziquantel if ≥30% prevalence in school

Note: All information in the above table is drawn from *Manson's Tropical Diseases* (22nd edition, 2009), unless otherwise referenced.

*= threshold for moderate infection, as described in the original study (p.167). The citation provided by '(Brooker *et al*. 2000)' does not describe use of these thresholds, which the authors actually developed (according to the authors) following personal communication with Dr Simon Brooker and Professor Donald Bundy. The World Health Organization describes these standard thresholds (Montresor *et al.* 2002 p.25).

Publications in the 3ie Replication Paper Series

The following papers are available from http://www.3ieimpact.org/en/publications/3ie-replication-paper-series/:

Quality evidence for policymaking: I'll believe it when I see the replication, 3ie Replication Paper 1. Brown, AN, Cameron, DB, and Wood, BDK (2014)

TV, female empowerment and demographic change in rural India, 3ie Replication Paper 2. Iversen, V and Palmer-Jones, R (2014).

Reanalysis of health and educational impacts of a school-based deworming program in western Kenya: a pure replication, 3ie Replication Paper 3, part 1. Aiken, AM, Davey, C, Hargreaves, JR and Hayes, RJ (2014)

Replication Paper Series

International Initiative for Impact Evaluation
1625 Massachusetts Ave., NW
Suite 450
Washington, DC 20036
USA

replication@3ieimpact.org
Tel: +1 202 629 3939

www.3ieimpact.org