Calum Davey
Alexander M Aiken
Richard J Hayes
James R Hargreaves

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

Part 2: alternative analyses

December 2014

Replication
Paper 3
Part 2

Health and education

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: Davey, C, Aiken, AM, Hayes, RJ and Hargreaves, JR, 2014. *Reanalysis of health and educational impacts of a school-based deworming program in western Kenya: Part 2, alternative analyses, 3ie Replication Paper 3, part 2.* 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 2: alternative analyses

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

Alexander M Aiken
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

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

3ie Replication Paper 3, Part 2
December 2014

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.

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 school deworming programmes result in improvements in school attendance and educational achievement.

Methods: We reanalysed data from a stepped-wedge cluster-quasi-randomised trial of a drug-treatment and health-education intervention conducted in 75 schools in western Kenya. We specified two coprimary outcomes: school attendance and examination performance. These were measured by unannounced fieldworker visits and independently administered examinations. We used worm infection and nutritional parameters as secondary outcomes. We estimated effects on primary outcomes using year-stratified cluster-summary analysis and observation-level random-effects regression. We combined years with a model that accounted for secular trends.

Results: Three quasi-randomised groups of 25 schools were similar at baseline, with slight differences in mean age. There was a substantial amount of missing data. We found unexpected patterns in the school-attendance data, including a correlation between the amount of attendance data from a school and the level of attendance. In cluster-summary analysis, neither school attendance nor examination performance differed between arms in either study year. (School-attendance risk differences: 1998 5.48, 95 per cent CI −1.48—12.44, p = 0.121; 1999 2.16, 95 per cent CI −3.39—8.27, p = 0.483. Examination-performance risk difference: 1998 −0.109, 95 per cent CI −0.332—0.115, p = 0.336; 1999 −0.028, 95 per cent CI −0.228—0.171, p = 0.777.) We found some evidence of improvement in age-adjusted regression models for each year (adjusted OR 1998 1.48, 95 per cent CI 0.88–2.52, p = 0.15; aOR 1999 1.23, 95 per cent CI 1.01-1.51, p = 0.04) but not for examination performance. When we combined data from both study years in an observation-level model, the effect on school attendance was stronger than in either year (aOR 1998+1999 1.82, 95 per cent CI 1.74–1.91, p<0.001), but it had no effect on examination performance. We found evidence of reduction in hookworm and roundworm infections but not in schistosomiasis or whipworm. We found no evidence of improvement in nutritional parameters.

Discussion: We found that the evidence that the intervention improved school attendance differed according to how we analysed the data. The patterns in the data may explain this sensitivity. Our inability to review the sampling strategy guiding data collection and the potential for bias in measurement procedures necessitate caution in interpreting these results.

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

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.
Replication	A reanalysis of a study, without collection of new data. Different formats of replication include (according to 3ie terminology) pure, statistical and scientific.
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
CI	Confidence interval
CONSORT	Consolidated Standards of Reporting Trials
DVBD	Kenya Ministry of Health Division of Vector Borne Diseases
HAZ	Height-for-age z-score
Hb	Haemoglobin
ICC	Intra-cluster correlation coefficient
ICS	Internationaal Christelijk Steunfonds
ITT	Intention to treat
K	Inter-cluster coefficient of variation
PAP	Pre-analysis plan
SAP	School Assistance Programme
sd	Standard deviation
se	Standard error
WAZ	Weight-for-age z-score
WHO	World Health Organization

Background

This is a report on the replication (reanalysis) of work by Miguel and Kremer (2004) describing the impact of a school-based deworming intervention in Kenya on the health, school attendance and academic performance of school pupils. This report follows the pure replication report (Aiken *et al*. 2014; Hamermesh 2007). The analysis in this report comprises an 'internal statistical replication' and an 'internal scientific replication'. We use the term 'internal statistical replication' to mean a reanalysis of the study's original hypotheses using different handling of the same raw data (for example, different variable constructs, different data handling). We use the term 'internal scientific replication' to mean the introduction of a (different) explicit causal framework to guide analysis and interpretation of the statistical results, similar to the 'theory of change' process (Vogel 2012). We have used the qualifier 'internal' to differentiate the statistical and scientific replication analyses in this report from replication work involving collection of new data. Hammermesh (2007) uses these terms without the 'internal' qualifier to describe what we would describe as 'external replication', which uses new samples or data on different populations.

As the starting point for this replication, we have taken the results reached from the pure replication stage of this replication process, where we reproduced the study's original methodologies. The original results referred to five distinct effects on particular groups of individuals, as follows:

- **Direct effect**: this is the effect of treatment on the pupils treated with deworming drugs. This is the difference, or ratio, between the outcome distribution in pupils who received treatment and the distribution had they not received treatment.
- **Within-school indirect effect**: this is the indirect effect on all children in treatment schools arising from the treatment of children within those schools. This applies to both treated and untreated children in treatment schools. This is the difference, or ratio, between the outcome distribution in the pupils in schools where other pupils received treatment and the distribution had they not been in a school where fellow pupils received treatment.
- **Naïve effect**: this is the effect found by comparing all children in treatment schools with all children in control schools, irrespective of whether or not the children themselves received treatment. 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 medical literature would typically analyse in the pragmatic evaluation of a cluster-randomised trial. This is the difference, or ratio, between the outcome distribution in pupils in schools allocated to the intervention and the distribution had they not been allocated to the intervention.
- **Between-school indirect effect**: this is 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, whilst children in treatment schools accrue this in addition to the other effects. This is the difference, or ratio, between the outcome distribution in pupils attending schools within specified distances of schools where pupils were treated and the distribution where there were not those treated schools within that specified distance.
- **Overall effect**: this is the combination of the 'direct effect' (which applies to treated pupils only) with 'within-school indirect effect' (which applies to treatment

schools only) and the 'between-school indirect effect' (which applies to all schools). In other words, this is an estimate of the total effect of the intervention applied to a child who receives all different effects in comparison to a child who receives none of the effects. This is the difference, or ratio, between the outcome distribution in all pupils who received any of the effects and the distribution had the intervention not been allocated.

It should be noted that this terminology to describe effects is not standard in epidemiology, though it broadly corresponds with some widely used descriptions (Halloran and Struchiner 1991). Through knowledge of the causal-reasoning literature, some readers may recognise 'indirect' and 'direct' effects as referring to whether the effects are mediated by an intermediate factor (VanderWeele and Vansteelandt 2010). In this report we use the terms according to the taxonomy above and not in the manner pertaining to mediation.

The results from the pure replication are summarised as follows, with results derived from both study years unless otherwise annotated. Effects that we found to be beneficial and significant in the pure replication are shaded.

Table 1: 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: se = standard error; Hb = haemoglobin; WAZ = weight-for-age z-score; HAZ = height-for-age z-score. Examination performance is measured as a z-score. † = year 1 data only

For the statistical replication that follows, we have confined our investigation to the outcomes shown in the 'naïve effect' column above, reflecting our assessment that this is the most relevant format of analysis for these data. We have not examined the effect on anaemia for the reason stated in our pre-analysis plan: the numbers of pupils tested are too small. We have not examined the indirect between-school effect (or the overall effect that includes this) in this statistical and scientific report for the following reason: in our pre-analysis plan, we stated that we would investigate the between-school indirect effects using 'the same analytic approach as described in the original paper' (Aiken *et al*. 2013). In our pure replication report (Aiken *et al.* 2014), we reanalysed the between-school indirect effects according to precisely the methods used in the original study, and we recorded our results there, reproduced in the column headed 'indirect effect: between school'. Therefore, we have already fulfilled our stated intentions with regards to these types of effects and have not pursued them further. For the scientific

replication, we make use of a causal framework to aid our interpretation of the relationship between the intervention and the study outcomes; this is based on a diagram shown in our pre-analysis plan (Aiken *et al*. 2013). We also discuss the alternative causal relationships that might account for the findings of the analysis. The introduction of a causal diagram constitutes a statement of our 'theory of change' that links the intervention to the primary outcomes. As this was not used in the original paper, doing so comprises a 'scientific' replication. Since the original authors did not make their theory of change explicit, we cannot know whether our theory of change differs from theirs.

We developed our planned reanalysis of these data in several distinct phases. First, in response to a funding call, we read the original paper and developed an analysis plan. This comprised three components: a pure replication, a statistical replication and a scientific replication. Second, we received both the data from the original authors and detailed comments on our proposed analysis plan from an external reviewer. We considered these comments and submitted an analysis plan, which 3ie subsequently published on its website (Aiken *et al*. 2013). First, we undertook the pure replication, and during this work we had several phone calls with the original authors to clarify our understanding of the conduct of the trial and the structure of the data. As this work drew to a close we shared a copy of our pure replication report with the original authors and again received detailed comments in both verbal and written form. We considered these, made edits to our report and submitted this to our funder, 3ie. Next, we undertook our statistical and scientific replication work. We reapproached the analysis of the trial for this phase from the beginning in line with our original protocol, as we would the analysis of a registered, public-health trial with the intention of reporting in line with CONSORT guidelines. This led us to further queries about the conduct of the trial, about which we again corresponded with the original authors. Where we were unable to uncover answers to our questions, or where we were concerned about the implications of answers we received, we considered alterations to our analysis plan. We made a small number of these, as detailed below. These conversations also influenced our interpretation of the data. Finally, as we approached the end of the statistical and scientific replication work, we shared a copy of our report with the original authors. We once again received detailed input and edited our report in light of these. At this stage, we chose not to engage in further private correspondence with the authors so as to push transparent discussions about differences in interpretation of the data between our groups into the public realm. Nevertheless, we made one final addition to this report following a final correspondence from the original authors just prior to release of this report. At this stage, we added the sensitivity analyses described in the report and shown in Appendix 7.

As part of our approach to this work, the report reanalysing the raw data from the original study is given in a format consistent with the CONSORT reporting guidelines; currently, researchers in the field of public health widely adhere to these. What follows is thus an attempt to, as far as possible, describe the components of the original study using the format, layout and language normally adopted in reporting of such trials in biomedical journals. Due to the complex nature of the study, it has not been possible to write this report within the length normally available in a biomedical journal. We note that the original authors conducted the study before adherence to CONSORT guidelines became a standard practice, and our attempt to adhere to modern reporting standards is intended merely to provide a contrast to the original paper's format and should not be

interpreted as criticism. We also note that trial registration and protocol publishing, although common practice now, were by no means universal at the time the original authors conducted the study.

1. Introduction

Helminth (worm) infections are a major and persistent public-health problem, concentrated in the poorest areas of the world (Bethony *et al*. 2006; Steinmann *et al*. 2006). Associated morbidities can include anaemia, malnutrition and growth impairment. Helminth infections are relatively simple to treat with low-cost medications. A major area of debate has been the extent to which helminth infections also reduce school attendance and educational achievement and, thus, in the longer term, economic outcomes. Consequently, there has been great interest in the extent to which helminth-control programmes might have benefits in these areas, and substantial resources have already been mobilised on the basis of existing research. However, a recent Cochrane Review (Taylor-Robinson *et al*. 2012) 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.'

We reanalysed data from a cluster-quasi-randomised trial conducted in western Kenya in 1998–1999 (Miguel and Kremer 2004), which has been central to the debate over health and educational benefits associated with deworming children.

2. Methods

2.1 The intervention

The Internationaal Christelijk Steunfonds (ICS), a Dutch charitable organisation, delivered the intervention, which comprised periodic drug deworming treatment for all eligible pupils (girls 13 years old and younger and boys of all ages) in eligible schools and health education. Schools were eligible for mass drug treatment every six months with albendazole if the prevalence of geohelminth infection prior to the administration of the intervention was over 50 per cent and annual mass treatment with praziquantel treatment if schistosomiasis prevalence was over 30 per cent. The Kenya Ministry of Health Division of Vector Borne Diseases (DVBD) conducted parasitological surveys in advance of delivering the drug treatments. DVBD nurses and public-health officers delivered the drugs. Girls over thirteen years of age were not eligible for drug treatment because of potential teratogenicity concerns. For albendazole, the dosage was 600 milligrams per round in year 1 (1998) and 400 milligrams per round in year 2 (1999). For praziquantel, the dosage was 40 milligrams in year 1 (1998) and year 2 (1999). In schools with schistosomiasis prevalence of less than 30 per cent, only pupils with detected schistosomiasis infection received praziquantel.

ICS delivered health education through regular public lectures, wall charts and teacher training and focused on encouraging behaviours that prevented the transmission of worm infections such as hand washing, wearing shoes and avoiding contact with fresh water. DVBD staff trained one teacher per school for a day on delivering health messages. Several times a year, 10–15 minute presentations on worm prevention from the primary-school deworming programme project coordinator and assistant project coordinator supplemented the health lessons from teachers. Schools in the control arm did not receive drug treatment or the educational components of the intervention.

2.2 Trial design

This study was conducted in 1998–1999 in primary schools in the Funyula and Budalangi divisions of the Busia District in western Kenya (map of the area is shown in Appendix 2). Primary-school pupils in this region were typically 6–17 years old at the time of the study. In January 1998, there were a total of 92 primary schools — 61 in Funyula and 31 in Budalangi — spread across eight geographic zones. Of these 92 schools, 75 were selected to participate in the trial — 51 in Funyula and 24 in Budalangi. Twenty schools were originally excluded. Four schools were excluded because they charged considerably higher fees than other local schools. Four schools in the Budalangi division were excluded because of geographic isolation: three were on islands in Lake Victoria and a marsh separated the fourth from the rest of Budalangi. One school was excluded because it had been the pilot site for the intervention in late 1997. One school was excluded because it opened in 1998 and had few pupils in higher grades. Three additional schools (Runyu, Nangina Mixed and Kabwodo) were excluded for similar reasons to the aforementioned excluded schools, but at the point of allocation, these schools were added back in to increase the sample size. Seven schools were excluded that were receiving a different health and sponsorship intervention started in 1994–1995, and the criteria that the 1994–1995 intervention used to select these schools are not known. Seventeen schools were excluded in total (see Figure 2: CONSORT diagram).

2.3 Allocation to intervention groups

The clusters for this study were the 75 primary schools. Schools were 'quasi-randomised' into three groups (25 schools per group) and introduced the intervention in stages over two years, as shown in Figure 1: 25 Group 1 schools were in the intervention arm in both years (1998–1999), 25 Group 2 schools were in the control arm in year 1 (1998) and in the intervention arm in year 2 (1999), and 25 Group 3 schools were in the control arm in both years. This phased introduction across randomised or quasi-randomised groups is known as a 'stepped-wedge' cluster design. In this study, the researchers analysed just two steps, though the intervention was later extended to Group 3 schools.

Table 2: Schematic of stepped-wedge intervention rollout

Schools	**Year 1** (1998)	**Year 2** (1999)
Group 1 (n = 25)	**Intervention**	**Intervention**
Group 2 (n = 25)	**Control**	**Intervention**
Group 3 (n = 25)	**Control**	**Control**

Note: Stepped-wedge design is shown in schematic form. Each column represents a year of the study, and each row represents a quasi-randomly allocated group of 25 schools. The intervention was rolled out in 'steps', with Group 1 receiving the intervention in year 1, Group 2 in year 2 and Group 3 in the year after the study.

For the quasi-randomisation, schools were stratified by division and zone (Budalangi Division: Bunyala Central, Bunyala North, Bunyala South; Funyula Division: Agenga/Nanguba, Bwiri, Funyula, Namboboto, Nambuku), and the zones were listed alphabetically within each division. Within each zone, the schools were listed in increasing order of pupil enrollment as of February 1997 for grades 3–8. The three

schools that reincluded after being initially excluded were added to the bottom of this list. The allocation to Groups 1, 2 and 3 was done by allocating the first school in the list to Group 1, the second to Group 2, the third to Group 3, the fourth to Group 1 and so forth to the end of the list. This process is sometimes known as 'systematic allocation'.

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

We inferred from the original paper, in the absence of a protocol, that the combined educational and drug-treatment intervention package was intended to be delivered from the start of each year. This inference was based on the statement that, 'Due to ICS's administrative and financial constraints, the health intervention was phased in over several years. Group 1 schools received free deworming treatment in both 1998 and 1999, Group 2 schools in 1999, while Group 3 schools began receiving treatment in 2001. Thus in 1998, Group 1 schools were treatment schools, while Group 2 and Group 3 schools were comparison schools, and in 1999, Group 1 and Group 2 schools were treatment schools and Group 3 schools were comparison schools' (p.165). The paper also states, 'In what follows, "treatment" schools refer to all twenty-five Group 1 schools in 1998, and all fifty Group 1 and Group 2 schools in 1999' (p.170). We note that when reporting the results of the analysis, '1998' was operationalised as May 1998–March 1999, while '1999' was operationalised as March 1999–November 1999 (p.191), April 1999–November 1999 (p.193) or May 1999–November 1999 (p.195).

2.4 The School Assistance Programme (SAP)

Concurrently with the deworming trial, ICS also implemented a number of other interventions to attempt to improve educational outcomes. These programmes were active in 27 of the 75 deworming trial schools (7 in Group 1, 12 in Group 2, 8 in Group 2). Each of these 27 schools received (or would go on to receive) two SAP interventions. Firstly, they received one of the four following interventions:

- Seven schools received donations of textbooks in 1996 (A)
- Seven schools received financial grants in 1997 (B)
- Six schools received financial grants in 1998 (C)
- Seven schools would receive financial grants in 2000 (D)

Secondly, they received one out of the following two interventions:

1. Teacher incentives programme (13 schools). During 1998–1999, schools received an assistance programme targeting older children. The programme gave prizes to the upper primary (grades 3–8) teachers from the schools scoring highest for ICS-administered examinations.

2. Early-childhood development programme (14 schools). During 1998–1999, schools received teacher training, classroom materials, teaching manuals and salary bonuses for the nursery-school or preprimary classes.

The allocation to each of these interventions was done according to the ICS identification number (ID) of the school. The SAP schools totalled 100 and were divided by ascending school ID: 100–124 to group A, 125–149 to group B, 150–174 to group C and 175–199 to group D. The even-numbered schools received the early-childhood development programme while the odd-numbered schools received the teacher incentives programme. This allocation was done before, and independent of, the deworming trial. We are not aware of the procedure to determine the school IDs. The quasi-randomisation procedure for the deworming trial did not ensure that there was equal balance in the number of SAP schools in each group.

2.5 Outcome measurement

The authors assessed primary and secondary outcomes among a cohort of pupils who were registered in the 75 schools in grades 1–8 at the start of year 1 (1998): these pupils in all three groups were enrolled at the start of the study. The cohort was 'closed', in that pupils who joined the schools after this time were not included in the study. Outcome data were censored among pupils for whom records suggested that they had moved schools; data were included up to the point at which the students moved schools.

In accordance with our interpretation of the intention to treat, school-attendance observations of pupils in year 1 (1998) were interpreted as corresponding to the treatment condition in Group 1 and the control condition in Groups 2 and 3, and in year 2 (1999) observations were interpreted as corresponding to the treatment condition in Groups 1 and 2 and the control condition in Group 3.

Pupils in grades 1 and 2 of the cohort were not eligible for measurement of examination performance or nutritional parameters (weight-for-age and height-for-age). The assessment of helminth infection was undertaken in a subsample of pupils, as described below.

2.6 Primary outcomes

2.6.1 *School attendance.* ICS fieldworkers assessed school attendance by making unannounced school visits to check the presence of pupils. The schedule of school visits was predetermined for this trial and also for the concurrent SAP study in corresponding schools. For each year, there were eight possible visit periods. The fieldworkers scheduled schools only taking part in this trial (non-SAP) to be visited four times out of eight in year 1 (1998) and schools also in the SAP study to be visited six times in year 1 (1998) (see Figure 1). In year 2 (1999), they scheduled non-SAP schools to be visited five times and SAP schools to be visited six times. They did not share plans for school visits outside of the field team, the plans were frequently updated and they ordered the schools differently in each visit period to prevent anyone outside of the research team from learning about or predicting the day of a visit. We analysed the school-attendance data as a binary outcome: whether or not an individual pupil was present at a particular fieldworker visit.

2.6.2 *Examination performance.* ICS administered examinations in English, maths and science-agriculture for pupils in grades 3–8 in all schools. For each pupil assessed in each examination, the individual mark was transformed into a measure of deviation from the examination-specific mean (z-score). We examined the average of the results across all three examinations as a single continuous outcome.

2.7 Secondary outcomes

2.7.1 *Worm infections.* At the start of each year of the study, worm-infection rates were assessed among subsamples of pupils from intervention schools for that year. Thus in year 1 (1998), a sample was drawn from pupils across all grades in Group 1 schools prior to the drug treatment. In year 2 (1999), pupils from both Group 1 (after one year of intervention) and Group 2 (1999) (pre-intervention) schools were selected. No testing was performed for Group 2 in year 1 (1998) or Group 3 in either year, as it was felt to be unethical to test for parasite infection without an immediate plan to deliver treatment. It is unclear exactly how these subsamples of pupils were selected for parasitological investigation, although the stated intention was that these pupils should be selected at random and that the visits to the schools were not pre-announced. In Group 1, a 'representative subset' of the pupils tested in year 1 (1998) were sought out for testing again in year 2 (1999). The following procedure, quoted from an original study document, was used to collect and test for worm infections (emphasis original):

- *Stool samples were collected by ICS fieldworkers and samples were analysed by two independent readers at the Kenyan Ministry of Health. Egg counts for four different types of worm infection (hookworm, roundworm, whipworm and schistosomiasis) were enumerated by the Kato-Katz method.* ***90 stool samples are to be taken from each school****, 15 samples from each standard 3 through 8. The 15 pupils in each standard are to be randomly selected from the List of Names of all pupils.*
- *That same day, the samples are to be tested for Hookworm in the DVBD lab. The following day, the slides will be tested for Ascaris, Trichuris, and Schistosomiasis mansoni.*
- ***The Kato smear technique will be used for sample preparation. A fully quantitative method of egg counts will be employed for Hookworm, as well as for Ascaris, Trichuris, and Schistosomiasis mansoni****. Two slide series – A and B – will be created from each stool sample to ensure accurate egg counts.*
- *After egg counts are completed, the slides will be brought to the ICS in Busia. Quality control of 10% of the slides will be conducted during the course of the study, at the DVBD lab in Kisumu.*

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

2.7.2 *Weight-for-age (WAZ) and height-for-age (HAZ).* ICS staff administered questionnaires to all pupils in grades 3–8 in early 1998 and early 1999. They administered the questionnaire on a pre-announced day and only to pupils who were present. They collected weight and age in both years and height only in year 2 (1999). A

single enumerator read the scales and took height measurements for all of the pupils at a visit. DVBD staff and Busia District Hospital staff trained the enumerators in anthropometry, with external supervision at the project launch in January 1998. They measured weight with commercially available bathroom scales and height with height poles. They asked pupils their age, and ICS staff were encouraged to crosscheck against school records. For both WAZ and HAZ, we used z-score values that the original authors calculated.

2.8 Sample size

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

2.9 Allocation concealment

Leaders of the data-collection teams, the original authors and members of ICS performed the quasi-randomisation. They did not share the sequence publicly until the beginning of 1998, and they described the procedure was described to Kenyan Ministry of Health partners in early 1998. We assume that they intended Group 1 to receive the intervention first, followed by Group 2 and lastly Group 3.

2.10 Blinding

The authors made no attempt to conceal the school-intervention status from the fieldworkers collecting primary and secondary outcome data. The communities and the pupils were aware of their intervention status, and the researchers did not placebo-control the drug administration.

2.11 Ethics and consent

Ethical clearance for the drug-treatment protocol was obtained from the Ethics Committee of the Kenya Ministry of Health and Busia District Medical Officer of Health. We are not able to provide information about Institutional Review Board clearance of the protocol for the trial design or data collection, although the original authors have tried to locate these documents. The researchers gave us assurance that the trial was reviewed appropriately at the Massachusetts Institute of Technology (correspondence with Edward Miguel).

ICS obtained community consent from intervention schools in year 1 (1998) (in other words, Group 1). In both years, intervention schools held community and parent meetings immediately prior to delivery of the intervention. In these meetings, they described and discussed the project. In year 1 (1998), parents who did not wish their children to receive the drug treatment were asked to inform their school headmaster. In year 2 (1999), under recommendation from the Kenyan Ministry of Health, ICS was required to collect written consent from parents for children to receive drug treatment. Pupils in all arms were asked for their consent to take part in the questionnaire survey.

It is unclear what informed-consent procedures were carried out for attendance observations in schools. Personal communication from the original authors states, 'There were community meetings in all school communities explaining the data collection procedure and surveys. This was standard practice before launching all ICS projects. These meetings were carried out in Group 1, Group 2 and Group 3 schools. These are in addition to the deworming treatment meetings' (correspondence with Edward Miguel). We have not received copies of any documents further describing these meetings or any other aspect of the process of consent at individual or community level. We note that this study took place 15 years ago, and we accept that many of the documents relating to the study may now be difficult to locate.

2.12 Pre-analysis data handling

2.12.1 *Missing data for age and sex.* A substantial proportion of age data were missing from the original main dataset (6,646 missing age observations for 31,445 pupils; 21.1 per cent). As this is an important covariate, we used a simple method of imputation: first, where age or sex data were missing from the main dataset, we used values given in the pupil questionnaire. If age was still missing, we calculated the mean age for each grade and applied to the age of children with grade data but missing age data. Although such single imputation can be subject to problems, in this case the age range in each grade was narrow and the process very transparent. This reduced the proportion of missing age data to an acceptable level for the purposes of our analyses (702 missing observations for 31,445 pupils; 2.2 per cent). For purposes of later analyses, we categorised age into quintiles and handled age as a categorical variable. Although there were also substantial amounts of missing data for sex (3,399 missing sex observations for 31,445 pupils; 10.8 per cent), there was no straightforward way to impute these data, so we did not use sex in adjusted analyses.

2.12.2 *School-attendance data.* The data describing school attendance were measured from February to November in 1998 (year 1) and from January to November in 1999 (year 2). Pupils observed to be present at a visit were recorded as 'in attendance'. Pupils were recorded as being in 'nonattendance' if they were not present during the visit or if those observation data were missing and an entry in the 'grade' record indicated the student had dropped out of school.
We handled missingness in the outcome data on pupil attendance by applying the following steps sequentially. First, we removed from the dataset any data that had been collected during a visit that was not scheduled according to the visit plan. We did this to try to increase the likelihood that the data used were prespecified. Second, we removed entries for pupils whose observation data were missing at every visit, since the original authors thought that such students were probably incorrectly enumerated at baseline, though this was not possible to identify directly from the data (correspondence with Edward Miguel). Thirdly, we removed all data for a visit to a school where the amount of missing data per visit to the school was more than 70 per cent of all of the pupils in the school, since the original authors suggested that these occurrences in the data represented scheduled visits that did not happen because of bad weather or other logistical constraints, though again this was not possible to identify directly from the data (correspondence with Edward Miguel). To note, when pupils transferred between schools, occasionally their attendance observations were still (erroneously) assigned to

the original school. This resulted in some school visits that in reality did not take place, apparently having a very small number of pupils observed in that visit.

2.12.3 *Examination-performance data.* We used all examination data when available. We only considered examination data to be missing if they were not recorded for grade 3–8 pupils: only these grades were expected to take the examinations.

2.12.4 *Worm-infection data.* Standard WHO thresholds for moderate infection were applied to the raw egg counts (WHO 2002). We calculated the arithmetic mean of egg counts within each group, in line with standard practice for analysis of egg counts. We only considered WAZ and HAZ data to be missing if they were not recorded for grade 3–8 pupils: we only expected these grades to have anthropometric measurements taken.

2.12.5 *Nutritional data.* Weight-for-age (WAZ) and height-for-age (HAZ) data were converted to z-scores. We have used the anthropometric measures calculated by the original authors. We coded the data as missing when a pupil was recorded to be in Standard 3–8 but had missing data.

2.12.6 *Separate analyses by treatment type.* In our pre-analysis plan, we stated that we aimed to perform separate analyses by treatment type. In this report, we have not conducted this analysis. There were no data describing the eligibility for praziquantel treatment in Group 3 schools, and we felt that a separate analysis restricted to only the small number of schools (six in Group 1, treated in both years; 10 in Group 2, treated in year 2 only) that did receive praziquantel treatment in addition to albendazole treatment would add little to this report. Furthermore, in the presence of a secular trend in worm burden, it would not be not possible to determine which Group 2 schools would have been eligible in year 1 (1998).

2.13 Statistical analysis

We inspected the baseline characteristics of the pupils and schools in each arm. For each arm, we calculated cluster-summary means and confidence intervals. We used Monte Carlo simulation and calculated Moran's I (a measure of spatial autocorrelation) to test the spatial association within groups over the five nearest neighbours to assess whether quasi-randomisation had produced three groups that were geographically randomly distributed.

We carried out primary and secondary analyses out on pupils eligible for the drug treatment, specifically excluding girls over the age of 13. We performed all analyses according to the originally assigned group (intention-to-treat analysis). We also report the results for 'all pupils' in appendix tables, as we considered this to be an important format of analysis, as all children received the educational components of the intervention.

In the pre-analysis plan (PAP), we stated, 'Our analysis will initially look within each year, i.e. for 1998 and then for 1999. We will then combine the estimates of effect from the two years, accounting for the correlation in outcomes between the years due to the

fact that the same children are measured in each year' (Aiken *et al*. 2013). To carry this out, we conducted the overall primary analysis in four steps, reflecting the cluster-allocated stepped-wedge design of the trial, which increased progressively in complexity. The four steps we described in the PAP were as follows:

1. 'Summarise and display the outcomes clearly for each intervention arm in each year — for example, the proportion of children absent in the 25 schools in each group in 1998 and in 1999'.

We summarised the outcomes by calculating the mean of the school-level summary measures for each group and for each intervention arm in each year. We calculated the school-level summary attendance figures from the observations without first summarising pupil-level attendances. We compared the summary measures for each intervention arm within years using a t-test. This approach is in accordance with the vertical conceptualisation of the stepped-wedge design referred to in the PAP, although the PAP did not prespecify the use of a statistical test. The cluster-summary approach accounts for the correlation between repeat observations and within schools but does not weight according to the precision of the cluster-summary estimates.

2. 'Perform an individual-level analysis of the effect of the intervention status within a given year on the outcomes using regression models with random effects to account for clustering. We will report odds ratios and regression coefficients for intervention effect'.

We used regression models with random effects for school cluster (logistic for the binary attendance outcome at each observation and linear for examination performance) to examine the association between the intervention and the primary outcomes, stratified by year; this is a 'vertical' approach to the stepped-wedge design. We did not intend for the odds ratios calculated using logistic regression to approximate the prevalence ratio. The logistic and linear models had a similar structure. For the logistic model, supposing that π_{ijk} is the true probability of being present for the *k*th pupil in the *j*th cluster in the *i*th treatment arm, that β_i represents the intervention effect, that X is an indicator vector for intervention status and that γ_l represents the effects of covariates Z = $(z_{ijk1} \dots z_{ijkL})$, then the model fitted, for each year, was as follows:

$$\text{logit}(\pi_{ijk}) = \alpha + \beta_i \text{X} + \sum \gamma_l \text{Z}_{ijkl} + u_{ij}$$

In the above model, u_{ij} is a normally distributed random effect with mean 0 and variance as empirical variance in the school-specific mean proportions present.

In accordance with the PAP (page 10), we modelled each pupil observation of attendance as a binary outcome in this 'individual analysis' and did not aggregate observations by pupil. For the examination performance outcome, each pupil had a maximum of one measure per year. We calculated p-values for logistic regression using likelihood ratios. While at this stage in the analysis the models were not adjusted for imbalances at baseline, the regression models included terms for the population size of the school and the zone of the school, as these were used to stratify the quasi-randomisation. Therefore, the vector of covariates Z included school size and indicators for zones 1–8.

The random effect for school was used to take account of within-school clustering as well as the repeat observation of individual pupils. This is a valid, if not statistically optimal,

approach. Fitting a random effect for each pupil would have been problematic because of the small number of binary observations per pupil.

3. 'Combine the estimates of effect across the two years, accounting for correlation'.

Regression models with random effects for school and a fixed effect for year were used to examine the association between the intervention and primary outcomes in both years combined. The model was as for Step 2 but included data from both years and an indicator in Z for year 2. The effect that the regression model estimated included a comparison between year 1 (1998) and year 2 (1999) for Group 2, using Groups 1 and 3 to account for secular trends. This adds a 'horizontal' comparison into the analysis — in other words, a comparison between the same Group 2 pupils before and after the introduction of the intervention (Hussey and Hughes 2007).

4. 'Report results of any adjustment by covariates that are imbalanced at baseline. We will make adjustment for covariates if preliminary inspection of the data suggests that there is imbalance between the arms. We will include covariates in the regression models and report the adjusted estimates'.

Variables that showed imbalance between groups at baseline were included in the regression models (in other words, in the vector Z) for each year (step 2) and for combined years (step 3).

For tests of interaction — by which we mean effect measure modification — the PAP stated, 'We will test for interactions of any detected effects … by both age group and sex only'. We assessed interaction by age, but we did not investigate interaction with sex because of extensive missing data for this variable. We also performed tests for interaction by school SAP status. The PAP did not prespecify this, and we conducted these tests in addition to our preplanned analyses due to unexpected findings in the data. After preliminary examination of the data, we assumed that there was no cumulative effect of the intervention in Group 1, so we assumed the intervention status of Group 1 and Group 2 in year 2 (1998) to be the same.

We calculated the between-cluster coefficient of variation (k) for school attendance in the 50 control schools in year 1 (1998) and in the 50 intervention schools in year 2 (1999). Examination performances were normalised, so we calculated the intra-cluster coefficient of variation (ICC), using the same schools as for k.

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

We performed the secondary analyses on worm and nutrition outcomes by calculating differences in arithmetic means between groups. For all secondary outcomes, the

clustered nature of the data was accounted for by calculating the summary statistics and confidence intervals using cluster means. We compared arms using t-tests.

We made no adjustment to the p-values or confidence intervals to account for the multiple comparisons made in this analysis. The analytical plan did not prespecify an adjustment and stated all comparisons in advance, including which were primary and which were secondary.

We investigated the sensitivity of our school-attendance results to the decision about which school-attendance observations corresponded with pupils being in treatment condition and which corresponded with the control condition. This analysis was not preplanned but was undertaken following a final correspondence with the original authors in October 2014. We investigated two scenarios, based on the suggestion from the original authors that the first school visits occurred before the drug treatment was delivered in year 1 (1998) and that the drug treatment was delivered in Group 2 only after the second visit period in year 2 (1999). As described above, we do not have information from a protocol about when the drug treatment was intended to be delivered, about when the original authors made the decision to consider these first visits in each year as under control conditions or about when the educational component of the intervention was intended to be or was actually delivered. We did not have sufficient data to perform this analysis according to calendar dates. We also did not have information to explain how the timing of visits and deworming were linked.

In scenario one, we excluded observations of attendance in the first visit period in year 1 (1998) and excluded the observations in the first two visit periods in year 2 (1999). This scenario avoided comparing observations of pupils in different years in the year-specific analyses and is our preferred method of accommodating the reported dates of drug treatment.

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

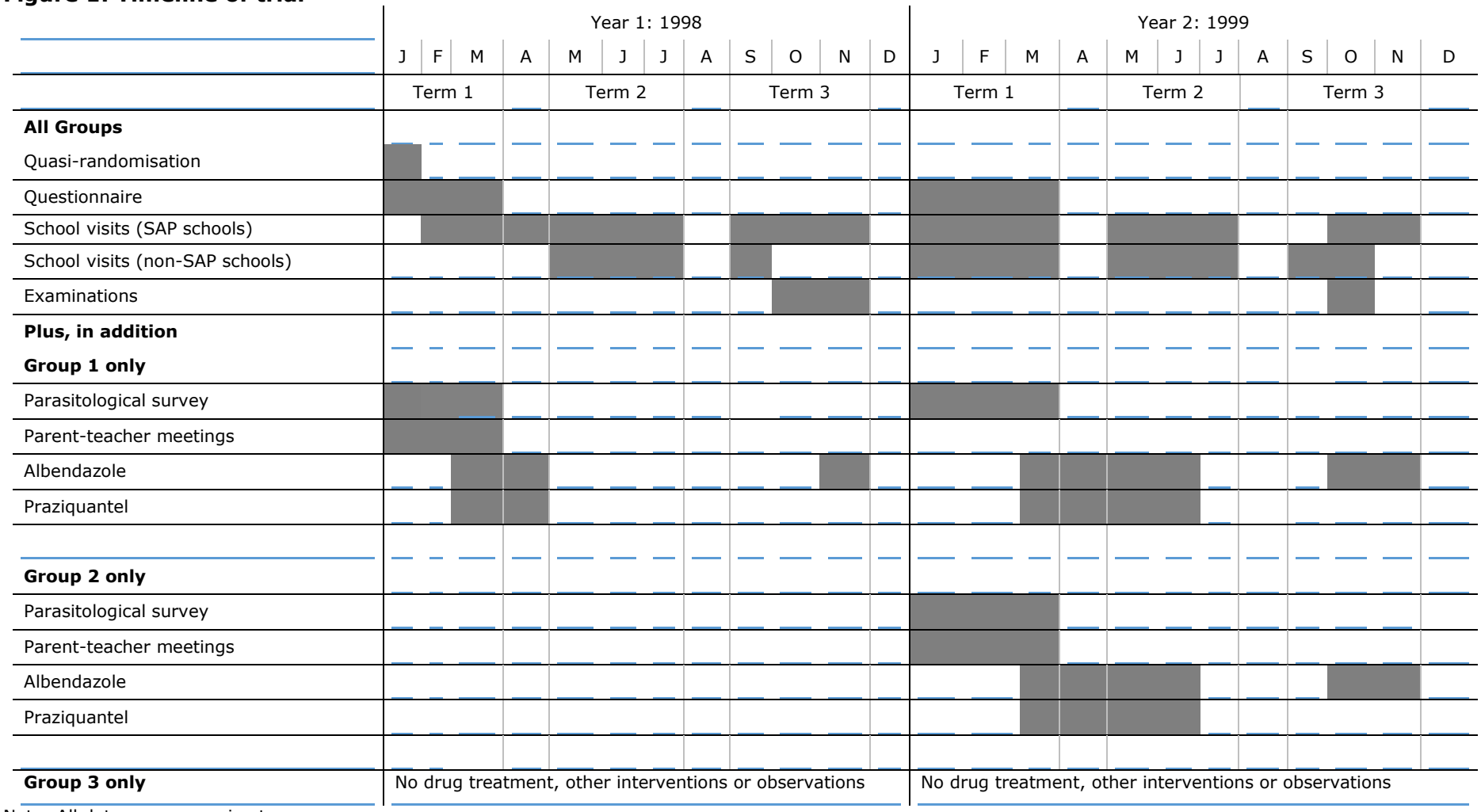
3. Results

3.1 Timeline

The trial took place between January 1998 and December 1999. Figure 1 shows the approximate dates of events in the trial. All 75 schools agreed to take part, and none dropped out. In eligible schools in the intervention arms, albendazole was administered

to eligible pupils in March–April 1998, November 1998, March–June 1999 and October–November 1999. Praziquantel was administered in March–April 1998 and March–June 1999. We do not have precise dates of drug administration or school visits. We have limited information about the timeline of the educational components, especially in year 2 (1999).

Figure 1: Timeline of trial



Note: All dates are approximate.

3.2 Baseline characteristics

Approximately 10,500 pupils were enrolled in each of the groups of 25 schools (Table 1). The numbers eligible for the drug treatment were similar across groups, with approximately 2 per cent missing eligibility data. Slightly more students were male than female in all groups. Although the sex ratio was balanced between groups, data on sex were more often missing in Group 2 (9.8 per cent) and Group 3 (17.7 per cent) compared to Group 1 (5.3 per cent). Students in Group 1 were 0.4 of a year older than those in Group 2. The mean WAZ, number of latrines per 1,000 pupils and distance to Lake Victoria were balanced across the groups. There was substantial missingness for WAZ data in all three groups. Many of the schools were quite large, with an arithmetic mean of approximately 400 students in each group. The mean number of pupils per school was similar in the three groups, but the range was much larger for Group 2 (min 37; max 1,392). A higher proportion of Group 2 schools were also enrolled in the SAP (12 out of 25). The quasi-randomisation achieved three groups dispersed with a geographical distribution indistinguishable from random (Moran's $I = -0.062$; $p = 0.14$). Appendix 2 shows a map of school locations in the study area.

Table 3: Baseline characteristics

	Group 1	Group 2	Group 3
Pupil-level characteristics			
Number of pupils	10,612	10,752	10,081
Female ≤ 13 and all male: year 1 (1998)	9,180	9,299	8,660
Female ≤ 13 and all male: year 2 (1999)	8,661	8,749	8,172
Missing eligibility data (%)	187 (1.8)	303 (2.8)	250 (2.5)
Proportion male (95% CI)	0.53 (0.52–0.54)	0.51 (0.47–0.55)	0.52 (0.50–0.53)
Missing sex (%)	562 (5.3%)	1,053 (9.8%)	1,784 (17.7%)
Mean age (95% CI) After imputation (95% CI)	11.8 (11.6–12.0) 11.4 (11.2–11.6)	11.4 (11.0–11.7) 11.0 (10.8–11.2)	12.3(12.1–12.6) 11.2 (11.0–11.4)
Missing age (%) After imputation (%)	1,662 (15.7) 155 (1.5)	1,929 (17.9) 334 (3.1)	3,055 (30.3) 213 (2.1)
Mean WAZ	−1.38 (-1.44–1.33)	−1.45 (-1.53–1.36)	−1.44 (-1.52–1.36)
Missing WAZ n/N (3–8th standard 1998) (%)	1,792/6,233 (28.8)	1,740/5,672 (30.7)	1,382/5,498 (25.1)
School-level characteristics			
Number of schools	25	25	25
School Assistance Programme (SAP)	7	12	8
Received textbooks in 1996	2	4	1
Received grants in 1997	2	3	2
Received grants in 1998	2	2	2
Early-childhood development	2	7	5
Teacher incentives	5	5	3
Latrines per 1,000 pupils (95% CI)	7.4 (6.1–8.8)	6.2 (4.7–7.7)	6.6 (5.2–7.9)
Km to Lake Victoria (95% CI)	10.0 (7.9–12.2)	9.9 (6.7–13.2)	9.5 (6.9–12.0)
Pupils per school (mean [min–max])	424 (168–772)	430 (37–1,392)	403 (103–752)

Note: All pupils who were enrolled or registered in school at the start of year 1 (1998) are included in the denominator. We calculated point estimates and confidence intervals for pupil characteristics using the means of cluster-means summary measures.

3.3 Drug treatment

In all schools tested, geohelminth infections were identified in over 50 per cent of the pupils, and therefore albendazole was offered to all eligible pupils in intervention schools. In year 1 (1998), 6,616 pupils in Group 1 received any drug treatment (72.1 per cent of those eligible), while none received treatment in Group 2 or 3. In year 2 (1999), 4,516 Group 1 pupils (52.1 per cent) and 4,159 Group 2 eligible pupils (47.5 per cent) received any drug treatment, as well as 91 pupils in Group 3.

3.4 Educational component of intervention

Field documents from 1999 include the following information about the delivery of the education components of the intervention to Group 1 schools in year 1 (1998):

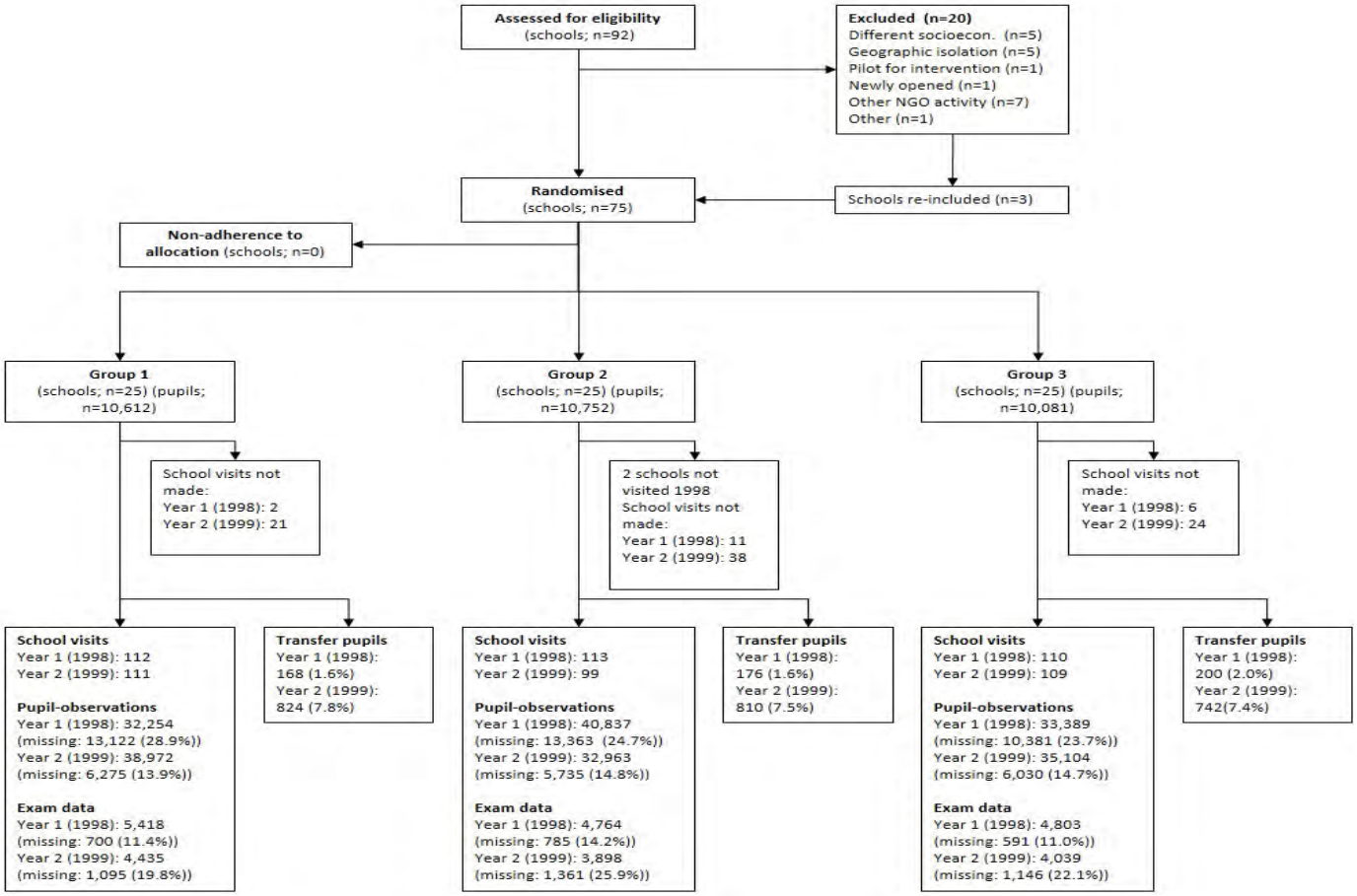
> *Health messages about worms [were] disseminated to teachers and pupils. The messages focused on the types of worms common in the region, mode of infection or transmission and preventive measures. The exercise was carried out from September 21 - 25, 1998. The education was for 25 Group W_1 schools that received treatment. Health Education Charts were used in the teaching and a*

copy of the chart and a guide book was left in the school. A contact teacher in each school was trained on how to utilize the charts and continue with the health education activity in the school.

The original authors have not yet provided us with similar documentation from the second year of the study, but we have no reason at this time to believe that the educational components of the intervention were handled differently.

3.5 Primary outcomes

Figure 2: CONSORT diagram



Note: Study participants were enumerated at the start of the study if they were registered at the school in grades 1–8 at the start of 1998. Follow up is shown for primary outcome data. 'Pupil observations' refers to the number of times that any pupils were observed in the group, excluding observations after transferring schools. Missingness for examination data is based on the number of pupils in standards 3–8 who had not moved schools.

3.5.1 *Trial profile: School openings and closures.* In year 1 (1998), two schools temporarily closed (school numbers 133 and 134 in Group 2), and other local schools absorbed pupils from these schools. These two schools reopened in year 2 (1999), and we chose to include these two schools in our analyses, including making use of a small number of observations recorded at school 134 in year 1 (1998). In addition, one school in group 3 (number 271) had no attendance observations recorded in year 1 (1998), and one school in Group 2 (number 133) had no examination scores results in year 2 (1999). It is unclear why these data were unavailable. The authors did not report that any schools closed in year 2 (1999), but one school in Group 2 (number 133) had no pupils recorded as present at any of the visits performed in that year.

3.5.2 *Trial Profile: School-attendance measurements.* In year 1 (1998), Group 1 schools were visited 112 times, Group 2 schools 113 times and Group 3 schools 110 times (Figure 2). A total of 19 planned school visits were not made, with the majority (11) of these missed visits in Group 2. In year 2 (1999), there were 111 visits to Group 1 schools, 99 to Group 2 and 109 to Group 3. A total of 83 planned school visits were not made in year 2 (1999), substantially more than in the first year of the study: 21 non-visits in Group 1, 38 in Group 2 and 24 in Group 3. Group 2 had substantially higher numbers of non-visits in both years of the study, and in year 2 (1999), more than a quarter of planned visits (38 out of 137; 28 per cent) were not conducted. Examining the visits that were successfully conducted, data were available for approximately 74 per cent of the pupils in these visits in year 1 (1998) and approximately 86 per cent in year 2 (1999). Within each year, there were broadly similar proportions of missing data across the three groups for attendance observations in visits that were successfully conducted.

3.5.3 *Trial profile: examination-performance measurements.* Examination data were available for approximately 5,000 pupils in each arm in year 1 (1998) and approximately 4,000 pupils in each arm in year 2 (1999). In year 1 (1998), there was a moderate amount of missing examination data: 11.4 per cent in Group 1, 14.2 per cent in Group 2 and 11.0 per cent in Group 3. In year 2 (1999), there was a higher proportion of missing examination data: 19.8 per cent in Group 1, 25.9 per cent in Group 2 and 22.1 per cent in Group 3. Examination data are recorded as missing when children were in grades 3–8, had no ICS examination score and had not moved schools in the previous year.

3.5.4 *Trial profile: movements between schools.* A total of 544 (1.7 per cent) pupils had moved schools by the end of year 1 (1998), and 2,376 (7.6 per cent) pupils had moved by the end of year 2 (1999): movements occurring between school years are included in the total for year 2 (1999). The proportions that moved are similar in each arm, and there did not appear to be asymmetrical movements into or out of any study arm. During year 1 (1998), 168 (1.6 per cent) Group 1 pupils moved to a different school, as compared with 176 pupils (1.6 per cent) in Group 2 and 200 pupils (2.0 per cent) in Group 3. By the end of year 2 (1999), 824 (7.8 per cent) Group 1 pupils moved to a different school, as compared with 810 (7.5 per cent) Group 2 pupils and 742 (7.4 per cent) Group 3 pupils.

3.6 Between-cluster coefficient of variation

The measured coefficient of variation for school attendance was 0.17 in year 1 (1998) and 0.11 in year 2 (1999). For examination performance, the values of ICC were 0.20 in year 1 (1998) and 0.16 in year 2 (1999).

3.7 Primary outcomes

Table 2 shows the results of the cluster-level and individual-level analyses comparing trial arms.

Step 1 results. The table shows the means of the cluster summaries for each group and intervention status (control, intervention). In all three groups and in both intervention conditions, the mean attendance is higher in year 1 (1998) than in year 2 (1999), but the difference in examination performance between the years differs by group. In year 1 (1998), intervention schools had a mean attendance of 5.48 per cent (95 per cent CI −1.48–12.44) higher than control schools, although this was not statistically significant (t-test p-value 0.12). In year 2 (1999), the intervention schools had a 2.16 per cent (95 per cent CI −3.39–8.27) greater mean attendance than control schools, but there was no statistical evidence of a difference (t-test p-value 0.48). These risk differences correspond to odds ratios of 1.78 and 1.21, respectively. In year 1 (1998) and year 2 (1999), there was no evidence of an association between intervention and examination performance in the cluster-means analysis.

Step 2 results. In the random-effects logistic regression for year 1 (1998), there was limited statistical evidence of an association between the intervention and attendance (OR 1.77; 95 per cent CI 0.91–3.44; $p = 0.097$). There was some evidence of an effect of the intervention on attendance in the individual-level analysis in year 2 (1999) (OR 1.23, 95 per cent CI 1.01–1.51, p-value 0.047), although with a smaller point estimate of effect. The corresponding regression models for exam performance found no evidence of effect of the intervention in either year.

Step 3 results. When we combined both years, we found strong evidence of an effect on school attendance (OR 1.78, 95 per cent CI 1.70–1.87; $p < 0.001$). There remained no evidence of an effect on examination performance.

Step 4 results. We adjusted for age and SAP because we considered these to be unbalanced across the groups at baseline (Table 1). In the adjusted analysis, we found no evidence for an effect in year 1 (1998) (aOR 1.48, 95 per cent CI 0.88–2.52; $p = 0.15$) and evidence for an effect in year 2 (1999) (aOR 1.23, 95 per cent CI 1.01–1.51; $p = 0.044$). In the analysis combining both years, the adjusted-odds ratio of 1.82 (95 per cent CI 1.74–1.91; $p < 0.001$) was approximately equivalent to the unadjusted analysis. There was no evidence for an effect of the intervention on examination performance after adjustment.

We tested the effect of the intervention on attendance in both years combined (that is, Step 3) for interaction with age and SAP status. There was strong evidence of an interaction with age ($p < 0.001$), with a stronger effect for younger age groups than for older:

≤ 7 yrs:	aOR 2.43; 95% CI 2.27–2.61
8–9 yrs:	aOR 1.87; 95% CI 1.76–2.00
10–11 yrs:	aOR 1.93; 95% CI 1.82–2.05
12–13 yrs:	aOR 1.53; 95% CI 1.44–1.64
≥ 14 yrs:	aOR 1.42; 95% CI 1.32–1.52

There was evidence of an interaction with SAP status (p = 0.045), with the effect in SAP schools being aOR 1.88 (95 per cent CI 1.78–2.00) and in non-SAP schools aOR 1.74 (95 per cent CI 1.63–1.86).

For school-attendance and examination-performance outcomes, we also performed analyses applied to all pupils. The results for these two primary outcomes were similar if we included all pupils in schools (results shown in Appendix 4), as opposed to drug-eligible pupils, as shown in the main analysis above.

We plotted the proportion of all pupil observations recorded as 'in attendance' in each school against the number of observations made in that school, and then we stratified the results by year and by Group (Figure 3) and fitted ordinary-least-squares regression lines. The table shows the cluster summaries by intervention status, with 48 control and 25 intervention schools in year 1 (1998) and 25 control schools and 50 intervention schools in year 2 (1999). In year 1 (1998), there were several schools in all of the groups that had more than 95 per cent attendance; in year 2 (1999), no schools had such high levels of attendance. All of the schools with attendance above 95 per cent did not participate in the SAP. In year 1 (1998) in Group 2, there was one school that was an outlier in terms of the number of observations, with more than 6,000 observations; this was the school that had absorbed additional pupils from the schools that were temporarily closed (Figure 3, data point not shown). In Group 2 in year 2 (1999), there was one school with very few observations and no children recorded present at the school (Figure 3, data point not shown).

3.8 Sensitivity analysis

The results of the analysis exploring the sensitivity of the school-attendance results to the handling of the treatment condition are shown in Appendix 7. In scenario one, 11,588 observations at the start of year 1 (1998) were excluded, as well as 31,404 observations during the first two visit periods in year 2 (1999). In comparison with our prespecified analysis, the year-specific results were approximately unchanged, with only the cluster summary mean difference in year 2 (1999) being slightly larger (3.57, 95 per cent CI −1.33–8.47, p-value = 0.150). The logistic regression results for year 2 (1999) were unchanged. For the combined-year logistic regression analysis, the unadjusted and adjusted ORs were larger than in the prespecified analysis (OR 2.08, 95 per cent CI 1.98–2.19, p-value < 0.001; aOR 2.13, 95 per cent CI 2.02–2.25, p-value < 0.001).

In scenario two, 11,588 attendance observations performed at the start of year 1 (1998) were excluded, and 31,404 observations occurring during the first two visit periods in year 2 (1999) were handled as year 1 observations. In comparison with our prespecified analysis, we found that the cluster summary mean differences were larger in year 1 (7.38, 95 per cent CI −0.18–14.94, p = 0.056) and had smaller p-value. In the unadjusted logistic regression analyses, there were 104,213 observations in year 1. In the unadjusted and adjusted regression models, the ORs for year 1 were closer to the null (year 1 OR 1.64, 95 per cent CI 1.06–2.55, p-value = 0.030; year 1 aOR 1.44, 95

per cent CI 1.03–2.01, p-value = 0.036), but the p-values were smaller. The results for year 2 (1999) were the same as for scenario one. The unadjusted and adjusted combined-year logistic regression ORs were larger than in our prespecified analysis (OR 1.89, 95 per cent CI 1.80–1.98, p-value < 0.001; aOR 1.92, 95 per cent CI 1.82–2.01, p-value < 0.001).

Table 4: Primary outcomes

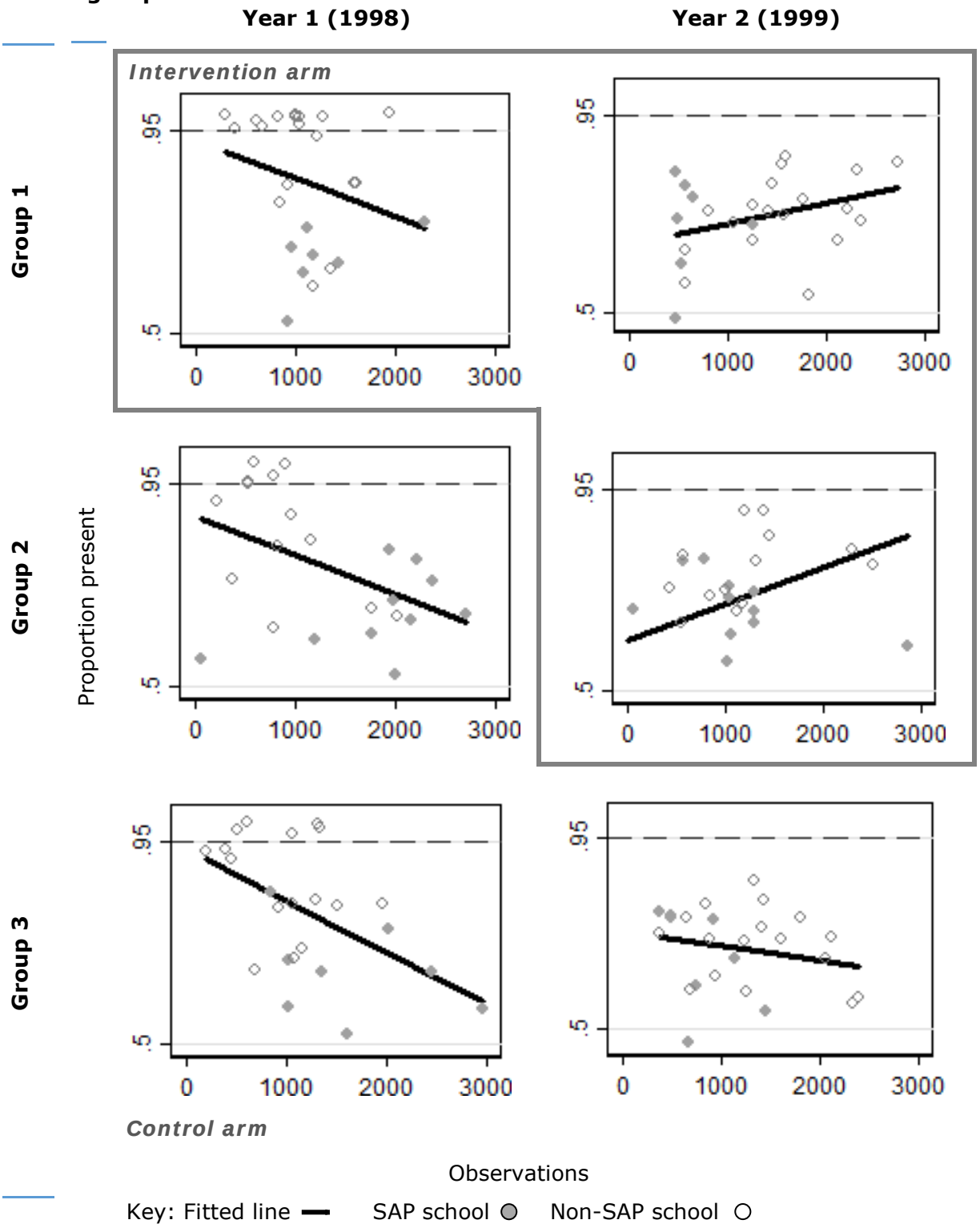
Yr(s)	Cluster summaries (Step 1)							Individual-level random effects (adjusted for pupil population size and zone) (Steps 2–4)			
	Attendance (%)							Unadjusted (N 1998 = 90,571; N 1999 = 89,170)		Adjusted for age, SAP (N 1998 = 89,540; N 1999 = 88,735)	
	Grp 1 % (N)	Grp 2 % (N)	Grp 3 % (N)	Intervention % (N)	Control % (N)	Risk difference (95% CI)	P-val	Odds ratio	P-val (LR test)	Adjusted odds ratio	P-val (LR test)
1998	84.2 (25)	77.0 (24)	80.4 (24)	84.2 (25)	78.7 (48)	5.48 (−1.48–12.44)*	0.121	1.77 (0.91–3.44)	0.097	1.48 (0.88–2.52)	0.150
1999	72.5 (25)	70.9 (25)	69.5 (25)	71.7 (50)	69.5 (25)	2.16 (−3.39–8.27)**	0.483	1.23 (1.01–1.51)	0.047	1.23 (1.01–1.51)	0.044
1998 + 1999				-	-	-	-	1.78 (1.70–1.87)	< 0.001	1.82 (1.74–1.91)	< 0.001
	Mean ICS Examination score							Unadjusted (N 1998 = 12,011; N 1999 = 9,830)		Adjusted for age, SAP (N 1998 = 11,999; N 1999 = 9,826)	
	Grp 1 mean (N)	Grp 2 mean (N)	Grp 3 mean (N)	Intervention mean (N)	Control mean (N)	Difference (95% CI)		Treatment coefficient	P-val	Treatment coefficient	P-val
1998	−0.031 (25)	0.037 (25)	0.118 (25)	−0.031 (25)	0.077 (50)	−0.109 (−0.332–0.115)	0.336	−0.131 (−0.321–0.058)	0.173	−0.135 (-0.323–0.054)	0.161
1999	−0.033 (25)	0.082 (24)	0.052 (25)	0.023 (49)	0.052 (25)	−0.028 (−0.228–0.171)	0.777	−0.015 (−0.199–0.168)	0.870	−0.017 (-0.201–0.166)	0.854
1998 + 1999				-	-	-	-	−0.117 (−0.292–0.058)	0.191	−0.121 (-0.293–0.052)	0.169

Note: Cluster summaries are the unweighted mean of school-level summaries. The individual analyses are adjusted for the variables used in stratifying the randomisation. The total units of analysis for the combined year 1 and year 2 analyses are the sums of the yearly totals. The random effect is fitted for the school. We excluded transfer pupils after they had moved schools.

* This risk difference corresponds to an odds ratio of 1.78.

** This risk difference corresponds to an odds ratio of 1.21.

Figure 3: Scatter plot of proportion present against number of observations, by year and group



Note: Scatter plots of proportion are presented against the number of observations in each school by year and by allocation group. The dotted line indicates 95 per cent attendance. We excluded two schools in Group 2 from the charts to preserve the scale: one in year 2 (1999) where no pupils are recorded present and one in year 1 (1998) with a disproportionately large number of observations (approximately 6,000).

3.9 Secondary outcomes

3.9.1 *Worm infections.* Table 3 shows results for comparisons of parasitological testing. At the start of year 2 (1999), substantially more pupils were tested in Group 2, the control group, than in Group 1, the intervention group. For both hookworm and roundworm, there was strong evidence ($p \leq 0.01$) of substantially lower worm-infection rates (both mean egg count and proportion with moderate infection) in Group 1 than in Group 2. However, for whipworm and schistosomiasis, there was little statistical evidence ($p > 0.1$) of any difference in either egg count or proportion with moderate infection, although the absolute difference in schistosomiasis burden is substantial.

Table 5: Secondary outcomes: worm infections at the start of year 2 (1999)

Type of worm infection	Group 2 pupils pre-intervention	Group 1 pupils with one year of intervention	Difference (95% CI)	P-value
Pupils tested (n)	1,233	746		
Average egg count (eggs/g; arithmetic mean)				
Hookworm	694	151	−543(−744 to −342)	< 0.001
Roundworm	4,283	1,289	−2,994 (−4,540 to −1,448)	< 0.001
Whipworm	374	254	−120 (−386 to 146)	0.367
Schistosomiasis	245	115	−130 (−316 to 56)	0.165
Proportion with moderate infection (WHO thresholds)				
Hookworm	7.8%	1.8%	−6.0% (−8.7 to −3.2)	< 0.001
Roundworm	23.6%	7.8%	−15.7% (−23.7 to −7.7)	< 0.001
Whipworm	7.6%	6.6%	−1.0% (−5.2 to 7.3)	0.747
Schistosomiasis	17.1%	8.0%	−9.1% (−20.2 to 2.0)	0.107

Note: We accounted for the clustered nature of the data by calculating the 95 per cent confidence intervals around the mean of the cluster means. There were missing data for individual Group 1 pupils tested for hookworm (n = 2) and whipworm (n = 4) and in Group 2 for hookworm (n = 1), whipworm (n = 6) and schistosomiasis (n = 3).

In other comparisons of worm infections between groups, results suggested that some secular changes in worm-infection rates occurred between 1998 and 1999. We show and discuss these in Appendix 3. We also conducted analyses including all pupils (in other words, not excluding girls ≥13yrs) – these results are very similar to those presented above and are shown in Appendix 5.

3.9.2 *WAZ and HAZ results.* After one year of intervention, there is no evidence of a difference in WAZ between either control group alone (Group 2 or Group 3) and Group 1 or when we combine the data from the control schools (Table 4). For HAZ, there is some evidence of a lower mean HAZ in Group 3 relative to Group 1 (p-value = 0.06). When we combine control schools (Group 2 and 3), there is very limited evidence (p-value = 0.55) of a difference in HAZ from the Group 1 schools that had received the intervention at this time.

Table 6: Secondary outcomes: WAZ and HAZ at the start of year 2 (1999)

Group(s)	Number tested	Intervention status	Mean z-score	Difference from Group 1 (95%CI)	P-value
Weight-for-age z-score (WAZ)					
1	2,982	Received	−1.329	ref	-
2	2,097	None	−1.282	0.047 (−0.094 to 0.188)	0.507
3	2,195	None	−1.380	−0.051 (−0.190 to 0.088)	0.469
2+3	4,212	None	−1.332	−0.003 (−0.119 to 0.125)	0.962
Height-for-age z-score (HAZ)					
1	2,982	Received	−1.231	ref	-
2	2,098	None	−1.129	0.102 (-0.133 to 0.337)	0.390
3	2,196	None	−1.453	−0.192 (−0.455 to −0.010)	0.061
2+3	4,214	None	−1.294	0.064 (−0.149 to 0.173)	0.552

Note: We accounted for the clustered nature of the data by calculating the 95 per cent confidence intervals around the mean of the cluster means. WAZ data were unavailable for 2,128 (41.7 per cent) Group 1 pupils, 2,722 (56.5 per cent) Group 2 pupils and 2,448 (52.7 per cent) Group 3 pupils, all in grades 3–8. HAZ data were unavailable for 2,127 (41.6 per cent) Group 1 pupils, 2,721 (56.5 per cent) Group 2 pupils and 2,447 (52.7 per cent) Group 3 pupils, all in grades 3–8.

We also conducted similar analyses in all pupils, rather than restricted to drug-eligible pupils, as shown above. The results are broadly similar; we show them in Appendix 6.

4. Discussion of statistical replication

Our statistical reanalysis of data from a cluster-quasi-randomised stepped-wedge trial conducted in western Kenya in 1998 and 1999 examined whether a combined education and drug-treatment intervention for deworming children improved school attendance or examination performance. In a fully adjusted logistic regression model making maximum use of the data available, there appeared to be strong evidence of an improvement in school attendance. However, the size of the point estimates and the strength of the evidence were not consistent in the analytic steps progressively building up to this fully adjusted model. That is, we found no evidence of effect with cluster summaries, some evidence with individual analysis stratified by year and a larger point estimate of effect when both years were combined than we found in either individual year. This inconsistency, as well as other concerns related to the quality of data and an unexpected pattern of correlations in the observations, raises uncertainty about the reliability of the fully adjusted result. By contrast, throughout all the progressive steps of analysis, there was consistently no evidence of an effect on examination performance; this result is consistent with the original analysis. There was some evidence that the intervention reduced worm burden in intervention schools, especially for hookworm and roundworm, but there is no evidence that either WAZ or HAZ were improved. There was evidence of interaction of the effect on school attendance by age and also by whether schools were involved in another intervention programme (the School Assistance Programme) operating concurrently in a subset of these schools.

4.1 Overview of strengths and limitations

The trial had several strengths. It was a large and innovative study and remains, to our knowledge, the only cluster-randomised trial to investigate the potential impact of school-based deworming on school attendance. The data on attendance were collected using direct observations of a very large sample of pupils, rather than using school registers or relying on recall, and was able to track pupils who moved between study schools over two years.

Viewed from the perspective of current practice in biomedical research, the trial had several procedural limitations. These included the absence of clearly prespecified plans for sampling, data collection, data management and analysis. While these practices were first devised for trials of pharmaceutical products, they are now routinely applied in evaluations of public-health, social and behavioural interventions published in journals from these disciplines. Such practices were, to our knowledge, much less common for randomised trials in the economics literature at the time the original trial was first published and only became standard practice in the medical literature around that time. More recently, there has been a growing debate within economics on some of the benefits of these approaches. As epidemiologists, we fully support the importance of such practices. The absence of a clearly specified protocol for collection of these data initially compromised our confidence in the results relating to school attendance. We therefore approached the data cautiously by starting with simpler but arguably more robust and transparent analyses and progressively building up to more complex forms of analysis. During this process, we found several discrepancies in the data and results, which we explore in more detail below.

4.2 Sensitivity to weighting of data

We found inconsistent results from our analysis of the effect of the intervention on school attendance. In view of the stepped-wedge cluster-randomised design, we started with simple unweighted cluster summaries before proceeding to individual-level analyses. In further analysis, we observed that there was a relationship between the number of attendance observations performed in a school and the overall rate of attendance in that school (Figure 2). The association between the number of pupil observations and the overall attendance in schools was noticeably different by intervention status; these were directly related in two out of three intervention-group years but were inversely related in all of the control-group years. In Group 2, which changed from control to intervention status between study years, the direction of this association switched between years.

Weighting is commonly used to increase statistical efficiency in cluster-randomised trials by giving greater weight to clusters that provide more precise estimates. In step one in our analysis, we use an unweighted analysis that should be robust and give an unbiased effect estimate but may not be statistically optimal. In steps two to four, we use random-effects regression methods that weight according to the precision of each cluster estimate, which depend on *both* the numbers of observations *and* the intracluster correlation (ICC). This method should provide greater power and precision. The approach used in the original paper of weighting directly by number of observations without taking account of the ICC does not maximise precision. Furthermore, this approach increases the risk of bias if there are underlying correlations between cluster size and outcome. In light of the correlations illustrated in Figure 3, we therefore do not think it would be advisable to perform a weighted analysis based on the number of pupil observations, as in the original analysis, and we think that an analysis partly weighted by number of pupil observations, such as our logistic regression models, may also be biased (Hayes and Moulton 2009).

4.3 Combining years of study for results on school attendance

The stepped-wedge rollout of the intervention implies that this study can be analysed as a stepped-wedge trial. Combining data over two years requires some accounting for the potential for secular trends to influence the comparison between schools that are in different arms in each year (in other words, Group 2). In our primary analysis, we adjust for year by including an indicator term, which makes maximum use of the data and may improve the power of the study because of the comparison between the same schools when in different arms (Hussey and Hughes 2007). When examining the process of combining results on school attendance across the two different years of this study, several separate but inter-related issues become apparent:

4.3.1 *Closed cohort.* This is a closed cohort, so the study population in year 2 (1999) is a different population to that in year 1 (1998). They are on average one year older, some pupils have dropped out and some have aged out (in other words, left school after completing Grade 8). Due to limitations in the data collection, we did not attempt to censor pupil records, other than when there were transfers between schools. Thus, during the study, a pupil observed to be 'absent' on a particular study visit was progressively more likely to represent a pupil who had permanently left school, either as

a dropout or as an aged-out pupil, rather than being likely to represent an instance of sporadic absenteeism. The original authors also did not censor pupils who had dropped out or aged out and also made no particular allowances for this feature of the dataset.

4.3.2 *Schools with high attendance in year 1 (1998) only.* When we examined a scatter plot of school attendance (Figure 3), we were surprised that there were schools with over 95 per cent attendance in year 1 (1998) but none in year 2 (1999). An explanation for this change between the years might be that a substantial number of pupils dropped out or aged out of the schools by year 2 (1999) and were thus only contributing 'absent' data in the second year of the study. However, if this were the case, we would expect dropouts and age-outs to occur in broadly similar proportions in all schools, such that all points in the scatter plot would shift down to a similar degree in year 2 (1999). In fact, all the schools with very high attendance rates in year 1 (1998) had much lower attendance in year 2 (1999), whereas other schools were largely unaffected. It is unclear to us why the changes in attendance between study years differed so greatly.

The schools with very high (> 95 per cent) attendance in year 1 were all non-SAP schools, as shown in Figure 3. In year 2, SAP and non-SAP schools appeared to be much more similar to each other, in terms of school-attendance patterns. We note that in year 1 (1998), fieldworker visit schedules were different for SAP and non-SAP schools (visits in different months, fewer visits for non-SAP schools; see Figure 1), but in year 2, visit schedules had less variation between schools. It therefore seems possible that school SAP status indirectly affected the measured level of attendance in schools due to systematic differences in the way the visits were conducted in different years of the study. This could lead to bias in analysis of these data, especially when combining results across the two study years.

4.3.3 *Year-stratified results do not combine to give similar point estimates.* We found an unexpected discrepancy between the year-stratified logistic regression results and the combined-years models for the school-attendance outcome, for both unadjusted and adjusted individual-level results. In our analysis, the combined-years model showed a stronger effect (aOR = 1.82), substantially higher than, rather than being an approximate average of, the two year-specific effects (aOR 1.48 and 1.23). The inclusion of a within-group comparison for Group 2 schools, which were control schools in year 1 (1998) and intervention schools in year 2 (1999), could explain this counterintuitive finding if the effect of the intervention was very strong when estimated by this horizontal comparison of Group 2. The Group 2 comparison across years also reduces the level of between-cluster variation and may therefore have greater statistical power. The increase in power ordinarily represents an advantage of the stepped-wedge design.

We are concerned about the reliability of this combined estimate of effect across the two study years, because it depends strongly on the 'horizontal' comparison of outcomes between year 1 (1998) and year 2 (1999) in Group 2. Figure 3 shows that there was probably a bias towards more pupil observations in schools with low attendance in year 1 (1998) (control condition), while we saw the opposite bias in year 2 (1999) (intervention condition). This would potentially lead to overestimation of the effect of the intervention on attendance, particularly in an analysis weighted, in part, by the number of observations.

4.4 Why are confidence intervals in 1998 substantially wider than for 1999?

In the adjusted model, the 95 per cent confidence interval for the effect size in year 1 (1998) is wide (0.88 to 2.52) when compared to the corresponding figure for year 2 (1999; 1.01 to 1.51). This is surprising, given that these results are based on similar numbers of pupil observations in each year (89,540 pupil observations in year 1, 88,735 in year 2). Furthermore, the width of the confidence intervals are very similar in the cluster-summary and observation-level model for year 1 (1998) but are considerably narrower in the logistic regression model in year 2 (1999). Part of the reason for this may be the very high attendance (near to 100 per cent) for some schools in year 1 (1998), especially in Group 1 and in non-SAP schools. Near-perfect attendance in these schools means there was very high ICC for the observations in these schools, which would lead to a large design effect and, hence, lower power and less precision for the outcome and effect measures. In year 1 (1999), no schools had these very high levels of attendance; as such, ICC did not so substantially reduce the precision of the estimates.

4.5 Assumption of no cumulative effect of intervention

We assumed that the effect of the intervention was the same in both years, which the finding of lower attendance rates in Group 1 schools in year 2 (1999) of the study supported. Given that only two years of data are available, it is hard to examine this assumption further. We do not feel that this assumption is likely to have a substantial influence on the interpretation of the study.

Overall, these various issues lead us to have uncertainty about the validity of estimates arising from combined-year analyses of these data, and such results should therefore be interpreted with caution.

4.6 Missing data

Some of the patterns of missing data would be best understood by comparing how actual data collection differed from data collection planned in a prespecified protocol and the reasons for any deviations; this was not possible. It is not clear to us why there were pupils listed in the dataset who did not have any observation data. We cannot rule out the possibility that in some instances missing attendance data indicated absence from school rather than data not being recorded. If this differed by study arm in this unblinded study, then the primary analysis would be at risk of bias. As the extent of missingness in attendance data was similar in each of the groups, we believe that this risk is low. The rule we applied to identify visits that were scheduled but did not take place had an arbitrary cut-off of 70 per cent missing data, which was not informed by knowledge of the data-collection process, but we believe it is unlikely that we have excluded any 'real' school visits. In addition to issues with the school-attendance data, there were substantial missing data for the important covariates of age and sex; it is unclear why this was not collected. The need to impute large amounts of data for age means that this variable is liable to have reduced accuracy throughout the analysis, possibly biasing any calculations performed with this variable. For sex, missingness was so extensive that we did not attempt to adjust for this important covariate in our analyses.

4.7 Sensitivity analysis

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

We prefer the year-specific results from scenario one for considering the sensitivity of the results to school-attendance observations, which were conducted prior to actual treatment. Scenario one does not combine data from years 1998 and 1999 in the analyses for year 1. In this scenario, the year-specific results were very similar to the main results. The combined-year effect estimates were somewhat stronger but, as outlined above, we advise that this combined-year result should be treated with caution.

In the second scenario, in year 1, the statistical evidence for an effect was stronger. The inclusion of additional observations in the year 1 analysis probably contributed to the narrowing of the confidence intervals. The difference in the mean cluster-summary analysis for year 1 was larger. However, this result should be treated with caution. In this scenario, the year 1 analysis includes observations from periods in 1998 and 1999, and we are uncertain about the implications of combining data from the two years (see above).

In the absence of a protocol, it has not been possible to conduct either a true intention-to-treat or per-protocol analysis. Our interpretation of the study was that it intended to deliver the whole intervention package in each calendar year, which we stated in our pre-analysis plan – as such, this remains our primary format of analysis. However, we recognise that there is value in exploring effects of deworming treatment according to the reported timing of deworming treatment. Such analyses can demonstrate sensitivity to analytical decisions or approximate a per-protocol analysis in the absence of a protocol.

4.8 Interactions of intervention effect

We found strong evidence of an interaction of the effect of the intervention with pupil age. The effect was strongest amongst the youngest children. It seems plausible that pupils' ages might influence the effect of the intervention on school attendance.

4.9 Blinding and the Hawthorne effect

As the study was unblinded, ideally, we would want to be sure that fieldworker data-collection practices were the same in all schools. In practice, this is hard to verify retrospectively, so it is a possibility that there were, consciously or unconsciously, variations in data collection between groups.

The trial design is also vulnerable to the 'Hawthorne effect': this occurs when study participants know they are taking part in an experiment and change their behaviours. Parent associations and teachers in all schools were informed about the project early in 1998 and parents in intervention schools were invited to engage with the school when the intervention was about to start. Pupils, parents and teachers in all schools would therefore have been aware of the trial's intentions and may have adjusted their behaviours accordingly.

4.10 Deworming results

There was evidence of a substantial deworming effect for hookworm and roundworm but no change for whipworm or schistosomiasis. Schistosomiasis treatment was administered in fewer schools; therefore, it is difficult to interpret the large point-estimate differences with such wide confidence intervals. Without more information about how the sampling was performed, and the degree of success in relocating the subsample in Group 1, the substantial difference in the number of pupils sampled in Group 1 and Group 2 at the start of year 2 (1999) raises concerns that the samples may not be comparable.

Although the faecal egg counts are highly negatively skewed (in other words, a small number of individuals have very high egg-excretion rates), the recommended practice in evaluation of differences in eggs-per-gram counts is to compare arithmetic means (WHO 2002). Recent research has suggested that revisions are necessary to this approach (Levecke *et al*. 2011), but this discussion is outside of the scope of this report. As the sample sizes in all groups analysed here are large (> 600 pupils tested), this approach is unlikely to seriously affect the qualitative conclusions of this analysis. Furthermore, the results for the effects on moderate infection are similar to the egg counts, supporting the validity of the averaging approach.

4.11 WAZ and HAZ results

We found no effect of the intervention on WAZ. The weak evidence of effect ($p = 0.06$) on HAZ when we compared Group 1 and Group 3 is unlikely to be meaningful, as when Groups 2 and 3 are combined (neither of these Groups had received intervention at the start of year 2 [1999]), there is no evidence of a difference from Group 1 ($p = 0.55$). There is also a substantial amount of missing data for both WAZ and HAZ outcomes, possibly because data were not collected from children who were absent on the day of the survey. Since this day was pre-announced in schools, these data are potentially subject to bias if children who were absent on that day were systematically different from those who had anthropometric testing performed. Furthermore, the degree of missingness varies by group, with more pupils tested in Group 1. Group 1 pupils may have been more likely to attend school than pupils in other groups, for various reasons; as such, different pupils may have been sampled, biasing the comparisons.

In the absence of information about sampling frames and refusal rates, we estimated the missingness indirectly by comparing the sample sizes in each group to the total pupils eligible for the questionnaire. To our knowledge, no sampling fraction was used for the anthropometric testing; if in fact one was used, we will have overestimated the extent of

missingness amongst these data. However, the proportion of missing data varies by group; this would not be explained by use of a sampling fraction. On the basis of these findings, we find no clear evidence to support changes in WAZ or HAZ associated with the intervention.

5. Scientific replication

Notwithstanding these various methodological issues, we sought to map our findings against a prespecified conceptual hierarchy reflecting our theory of change (Figure 4) in order to further consider the strength of evidence provided by the trial and make recommendations for further research and policy. Outcomes on the diagram are in boxes: those for which there is evidence of a difference are shaded in grey (including the implementation of the intervention itself). Since we made no adjustment for multiple comparisons, effects with moderately small p-values should be interpreted cautiously. We note that odds ratios will be further from one than prevalence ratios, because the prevalence of attendance is high. The major outcomes in the diagram are as follows, starting from the top-left:

- There is good evidence that the intervention was allocated to the schools designated as Group 1 in years 1 and 2 (1998 and 1999) and also to those designated Group 2 in year 2 (1999). This is recorded in the report on the randomisation procedure, and consultations took place with the schools.
- There is good evidence that all of the schools in the intervention condition were eligible for drug treatment and that many of the pupils in those schools received treatment. There is good evidence that few pupils in the control condition schools received deworming treatment.
- There is moderate evidence from study records that the educational component of the intervention was administered in the intervention schools in a variety of modalities during the course of the study.
- There is some mixed evidence of behavioural changes, as previously discussed in the pure replication report (Aiken *et al*. 2014). We did not re-examine findings from the pupil questionnaire in the statistical replication, as we do not consider self-reported health status in primary-school children in an unblinded study to be a reliable measure.
- There is strong evidence to suggest that intervention-condition schools had a lower average worm-infection burden for hookworm and roundworm, though the evidence is limited for whipworm and schistosomiasis.
- There is no evidence of differences in non-worm health outcomes, operationalised as WAZ and HAZ. In the pure replication, we had also found no evidence of changes in haemoglobin level or prevalence of anaemia (Aiken *et al*. 2014).
- The strength of evidence of a difference in school attendance by intervention condition is sensitive to choices about analysis approach and is weakened by unexplained patterns in the outcome data and extensive missingness.
- There is consistently no evidence of a difference in examination scores.

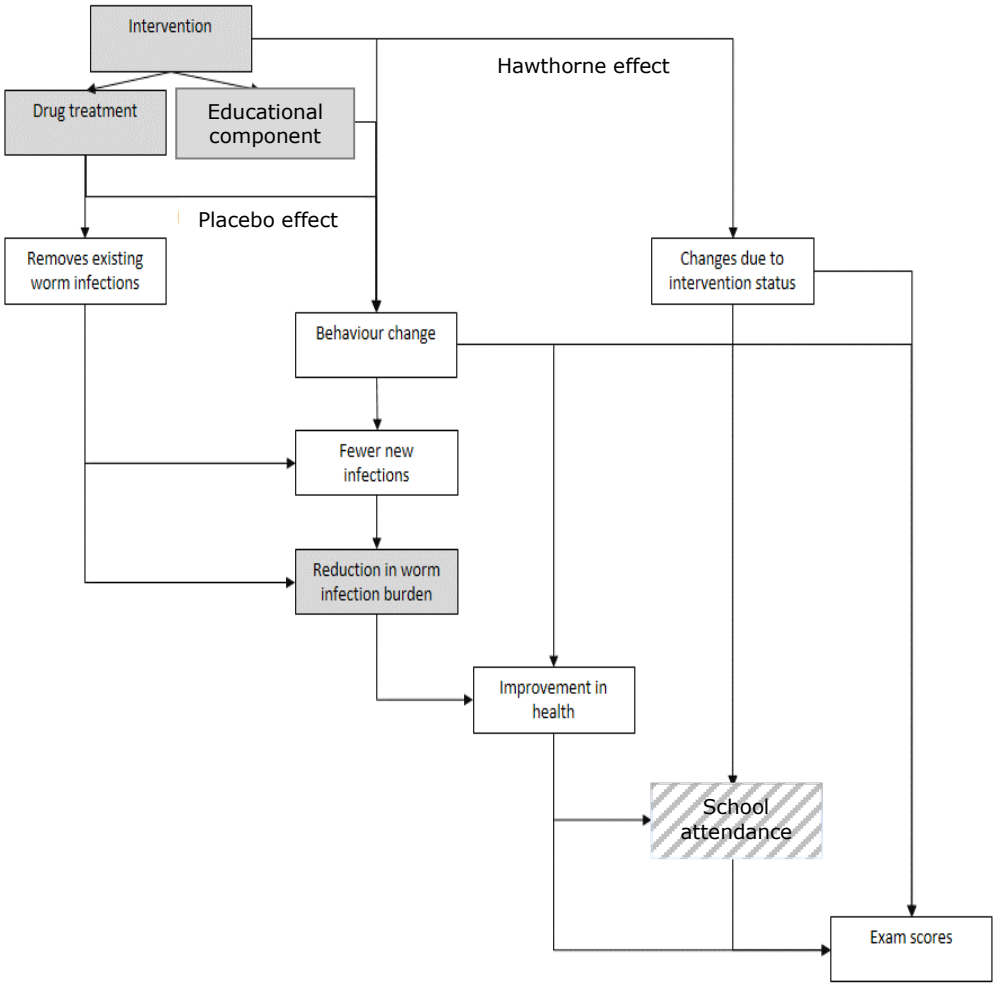
The evidence of a reduction in worm-infection burden in two of the four worm types supports the assertion that the drug component of the intervention was delivered and suggests that either the intervention removed existing infections or that there were fewer new infections — or both. Without good evidence of a difference in any connecting health outcome (in other words, excluding reduction in worm burden itself), it is uncertain whether health changes provide a causal link between the reduction in roundworm and hookworm burden and the change in school attendance. Furthermore, the pure replication found improvements in school attendance to be similar whether or not children had received drug treatment, and in this analysis, all results were similar whether applied to drug-eligible pupils (main analysis) or all school pupils (see Appendix

4). This further undermines confidence in a causal relationship between drug administration and changes in school attendance.

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

Finally, the possible improvement in school attendance did not appear to lead to an improvement in exam performance. This result is consistent with the findings of the original analysis. While we note that examinations represent only one approach to measuring educational attainment, an improvement in school attendance without a demonstrable improvement in educational attainment would be of uncertain benefit to the pupils.

Figure 4: Theory of change diagram



Note: We show the outcomes in boxes that are linked by arrows representing hypothesised causal pathways. Outcomes for which there is evidence of a difference between arms are shaded. The striped box for 'school attendance' indicates that the evidence for a difference between the arms of the study is mixed.

We had limited data on many of the outcomes along the causal chain. In particular, there are limited data on exactly why it was that pupils were not going to school and the possible behavioural changes that could alter attendance. If pupils in intervention schools did indeed attend school more frequently, as suggested by some of our analyses, we would ideally like to know more about those students who changed attendance behaviours and why this occurred. It would also be important to consider the quality of attendance change: was improved attendance sporadic, sustained or did it fall off over time from the intervention? Qualitative data might have informed these questions.

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

What would constitute a similar setting? This study was conducted in rural western Kenya in 1998–1999, and the researchers found that all schools tested had > 50 per cent baseline prevalence of worm infection. This suggests that Busia District was a 'high worm burden' setting at that time. For the results of this trial to be applied to other settings, there would have to be a similarly high burden. As the nature of the causal pathway operating here is uncertain, it is unclear what other aspects of the setting would

need to be similar for the intervention to work in the same way. For example, poverty and gender bias are two other factors that almost certainly impact school attendance, but these effects operate in complex ways that vary substantially from place to place, which might alter the effects of this intervention. A recent high-profile publication reported from a large trial looking at the effect of deworming and vitamin A supplementation on preschool mortality in north India found that deworming had no effect in this lightly infected area (Awasthi *et al*. 2013).

As ever, more research is needed to answer these questions. To our knowledge, the original study remains the largest and most influential study ever to have examined the educational impacts of deworming school children. Questions over the impacts of deworming children in low-income countries clearly remain of great scientific interest.

6. Conclusion

The results from the pure replication, according to a fully corrected repeat of the authors' original methods, were as follows; effects that the pure replication found to be both beneficial and significant are shaded.

Table 7: 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: se = standard error; Hb = haemoglobin; WAZ = weight-for-age z-score; HAZ = height-for-age z-score. Effects that were found to be beneficial and significant in the pure replication are shaded. Examination performance is measured as a z-score.

Our statistical replication has concentrated on the cells within the bold borders, as shown below. Cells where the data are most consistent with a beneficial effect are shaded. The hatched lines for the school-attendance outcome indicate that we conclude the evidence for this outcome is mixed.

Table 8: Summary of results from statistical and scientific replication

Measure		Direct effect	Indirect effect: Within school	Naïve effect	Indirect effect: Between school	Overall effect
Health	Worm infection (any mod/hvy inf)	Not examined	Not examined	Difference for roundworm and hookworm only	Not examined	Not examined
	Anaemia	Not examined	Not examined	Not examined	Not examined	Not examined
	Nutritional status	Not examined	Not examined	WAZ: no difference HAZ: no difference	Not examined	Not examined
School attendance (adjusted odds ratio)		Not examined	Not examined	aOR 1998+1999 1.81 (95% CI 1.74–1.90). Results sensitive to analytic choices	Not examined	Not examined
Examination performance (average difference)		Not examined	Not examined	−0.103 (−0.274–0.067)	Not examined	Not examined

Note: aOR = adjusted odds ratio, 95 per cent CI = 95 per cent confidence interval.

We found that the quasi-randomisation approach used did adequately balance schools on important baseline parameters. Our statistical reanalysis found that the strength of

evidence that a combined education and drug-treatment intervention delivered at the school level was associated with improved school attendance differed depending on how we analysed the data. The dataset had substantial amounts of missing data and some hard-to-explain patterns in school attendance. While there was some evidence that combined deworming and educational intervention has an effect on school attendance, there was considerable uncertainty about the extent of this effect. There was no apparent improvement in intervention schools for examination performance.

For worm infection, we found there is good evidence that the intervention for hookworm and roundworm infections has a beneficial effect, but there is no evidence for whipworm or schistosomiasis. As we know little about the sampling, and without baseline data from control groups, we cannot precisely enumerate these effects. We found no evidence of benefits for nutritional parameters (WAZ and HAZ).

In our scientific replication, we mapped our results against a prespecified theory of change. We found limited evidence for intermediate steps on a causal pathway linking reduction in worm infections to school attendance and examination performance via changes in health. Even if the result showing the strongest effect of the intervention on school attendance were accepted, we suggest that various pathways of change are plausible. One possible explanation is that behavioural changes unrelated to drug treatment occurred in this unblinded study that led to the observed changes in school attendance. With reference to this framework, we conclude that while these data provide some evidence of an effect of a combined educational and drug-treatment intervention on school attendance, the results are dependent on analytic choices and are at risk of bias. We found no evidence of an effect on examination performance.

7. Registration number and name of trial registry

This study was not, to our knowledge, registered as a trial in advance of being conducted.

8. Protocol

The pre-analysis plan for this reanalysis can be found in reference (Aiken *et al*. 2013) and is freely available online.

9. Funding and conflicts of interest

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 authors of this reanalysis report have conducted this work as consultants for 3ie. The funders have had no role in design, conduct or reporting of this reanalysis, other than facilitating contact with the original study authors. We have no conflicts of interest to declare.

ICS Africa provided funding for the intervention. The World Bank Research Department, the Partnership for Child Development, the MacArthur Foundation and the University of California, Berkeley, provided funding for the evaluation of the effects of the intervention.

References

Aiken, AM, Davey, D, Hargreaves, JR and Hayes, RJ, 2013. *Deworming schoolchildren in Kenya: Replication plan* [Online]. Washington, DC: International Initiative for Impact Evaluation (3ie). Available from: <http://www.3ieimpact.org/media/filer_public/2013/05/14/aiken_replication_plan_final.pdf> [Accessed 16 September 2014].

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)

Awasthi, S, Peto, R, Read, S, Richards, SM, Pande, V, Bundy, D and DEVTA team, 2013. Population deworming every 6 months with albendazole in 1 million pre-school children in North India: DEVTA, a cluster-randomised trial. *Lancet*, 381(9876), pp.1,478–1,486.

Bethony, J, Brooker, S, Albonico, M, Geiger, SM, Loukas, A, Diemert, D and Hotez, PJ, 2006. Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. *Lancet*, 367(9521), pp.1,521–1,532.

Halloran, ME and Struchiner, CJ, 1991. Study designs for dependent happenings. *Epidemiology*, 2(5), pp.331–338.

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

Hayes, RJ and Moulton, LH, 2009. *Cluster Randomised Trials*. Chapman&Hall/CRC.

Hussey, MA and Hughes, JP, 2007. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28(2), pp.182–191.

Levecke, B, Speybroeck, N, Dobson, RJ, Vercruysse, J and Charlier, J, 2011. Novel Insights in the Fecal Egg Count Reduction Test for Monitoring Drug Efficacy against Soil-Transmitted Helminths in Large-Scale Treatment Programs. *PLoS Neglected Tropical Diseases*, 5(12), e1427. DOI: 10.1371/journal.pntd.0001427

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.

Steinmann, P, Keiser, J, Steinmann, P, Keiser, J, Bos, R, Tanner, M and Utzinger, J, 2006. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *Lancet Infectious Diseases*, 6(7), pp.411–425.

Taylor-Robinson, DC, Maayan, N, Soares-Weiser, K, Donegan, S and Garner, P, 2012. 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

VanderWeele, TJ and Vansteelandt, S, 2010. Odds Ratios for Mediation Analysis for a Dichotomous Outcome. *American Journal of Epidemiology*, 172(12), pp.1,339–1,348.

Vogel, I, 2012. Review of the use of 'Theory of Change' in international development [Online]. Glasgow, UK: UK Department for International Development (DFID). Available from: <http://r4d.dfid.gov.uk/pdf/outputs/mis_spc/DFID_ToC_Review_VogelV7.pdf> [Accessed 16 September 2014].

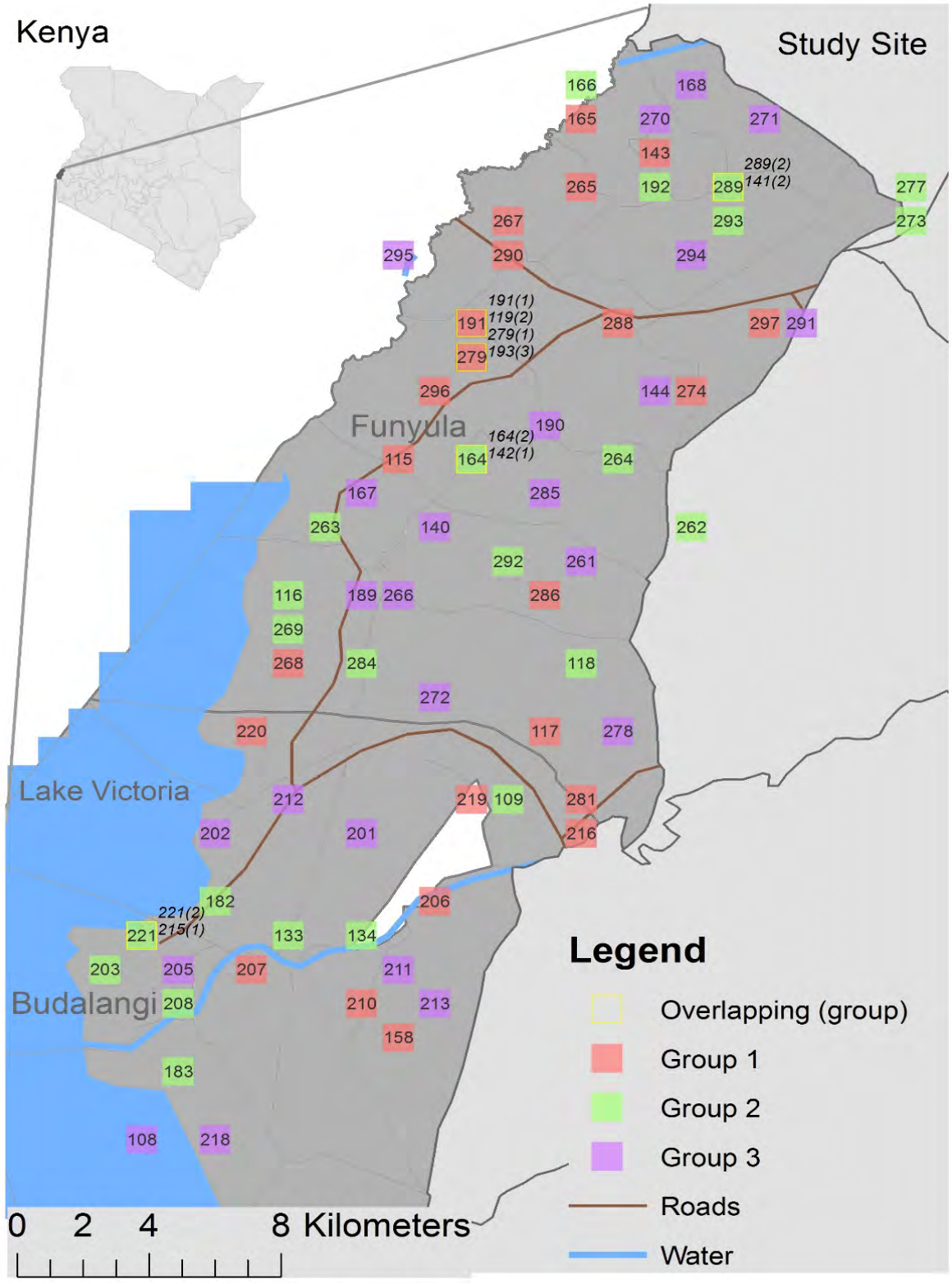
Appendix 1: Sample-size calculations

Before performing this reanalysis, we estimated what size of effects on attendance could plausibly be detected (power to detect = $1-\beta \geq 80$ per cent) with this study size at different significance levels (α). We based these estimates on the assumption of complete data for attendance outcomes in schools, with a harmonic mean of 400 pupils per school, 25 schools per group, a baseline rate of attendance of 72 per cent (in other words, 28 per cent absent, as reported in the original paper) and a coefficient of variation between schools (k) of 0.25, which represents a moderate degree of between-school variation. This also included an estimated adjustment for the stepped-wedge design (Hayes and Moulton 2009). The original analysis estimated an approximate 5 per cent improvement in attendance associated with the naïve effect of the intervention. On this basis, we calculated that the trial had approximately 70 per cent power to detect a true effect size of similar magnitude to the observed effect in the original study (the original authors found naïve effect on school attendance to be 5.1 per cent, revised to 5.7 per cent in the pure replication) at the 5 per cent significance level.

	Power to detect effect at differing significance level (α)			
% improvement in school attendance	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
3%	< 70%	< 70%	< 70%	< 70%
5%	~ 80%	~ 70%	< 70%	< 70%
7%	> 90%	> 90%	80–90%	~ 70%
9%	> 90%	> 90%	> 90%	~ 80%
11%	> 90%	> 90%	> 90%	> 90%

Note: This table represents the power to detect different intervention effects on the attendance outcome, calculated for the between-cluster coefficient of variation of k = 0.25.

Appendix 2: Map of study area



Note: This map represents the study region and schools. Locations are approximate because the researchers truncated the GPS location data gathered at the time of the study. The map displays schools with the same approximate location with a yellow square with the school ID label to the top right of the location and with the intervention group in brackets.

Appendix 3: Parasitological trends over time

Parasitological data were available for Group 1 and Group 2 only. No data were collected from Group 2 in year 1 (1998). Therefore, there were three possible comparisons:

A. Group 1 in year 1 (1998) against Group 1 in year 2 (1999)
B. Group 1 in year 1 (1998) against Group 2 in year 2 (1999)
C. Group 1 in year 2 (1999) against Group 2 in year 2 (1999) (see Table 3)

The table below shows the results for the first two of these comparisons, both in terms of (arithmetic mean) egg counts and proportions with moderate (or heavy) infection, according to WHO-defined thresholds.

Comparison A is a crude before-after comparison in Group 1. There is evidence for reduction in both the egg count and the proportion with moderate infection for both hookworm and roundworm between 1998 and 1999 ($p < 0.05$ for all these comparisons). However, there is no evidence of difference in either egg count or proportion with moderate infection for whipworm and schistosomiasis. Strictly, we should not have analysed the Group 1 effects as two separate samples, because some of these were paired observations. Restricting the analysis to just these paired samples within Group 1 shows direct evidence of a decrease in worm burden for pupils receiving treatment (results not shown), with the same qualitative conclusions as reached with the analyses shown here.

Comparison B assesses the secular trend by comparing Group 1 in year 1 (1998) with Group 2 in year 2 (1999). For both groups, these data were collected prior to the participants first receiving the intervention. Although all four types of worm infections show an increase in both mean egg count and proportion with moderate infection, there is only strong statistical evidence for hookworm (egg count and proportion moderately infected) and roundworm (egg count only).

Comparison A	Group 1 pupils pre-intervention	Group 1 pupils with one year of intervention	Difference (95% CI)	P-value
Year sampled	1998	1999		
Pupils tested (n)	1,801	861		
	Average egg count (eggs/g; arithmetic mean)			
Hookworm	427	233	−194 (−70 to −318)	0.003
Roundworm	2,410	1,548	−862 (−2,049 to 325)	0.151
Whipworm	169	265	95 (−49 to 240)	0.190
Schistosomiasis	92	112	20 (−104 to 144)	0.750
	Proportion with moderate infection (WHO thresholds)			
Hookworm	4.8%	2.3%	−2.5% (−4.7 to −0.4)	0.022
Roundworm	16.2%	8.9%	−7.4% (−14.5 to −0.2)	0.044
Whipworm	4.2%	6.7%	2.5% (−1.3 to 6.3)	0.195
Schistosomiasis	7.2%	7.9%	0.7% (−8.0 to 9.4)	0.878
Comparison B	**Group 1 pupils pre-intervention**	**Group 2 pupils pre-intervention**	**Difference (95% CI)**	**P-value**
Year sampled	1998	1999		
Pupils tested (n)	1,801	1,477		
	Average egg count (eggs/g; arithmetic mean)			
Hookworm	427	690	263 (57 to 469)	0.014
Roundworm	2,410	4,216	1,806 (166 to 3,447)	0.032
Whipworm	169	369	200 (−45 to 445)	0.108
Schistosomiasis	92	226	134 (−27 to 295)	0.101
	Proportion with moderate infection (WHO thresholds)			
Hookworm	4.8%	7.6%	2.8% (−0.2 to 5.8)	0.065
Roundworm	16.2%	23.6%	7.4% (−1.7 to 16.4)	0.108
Whipworm	4.2%	7.2%	3.0% (−2.1 to 8.2)	0.237
Schistosomiasis	7.2%	16.7%	9.5 (−1.6 to 20.6)	0.092

Note: There were no missing data in 1998. In 1999, there were missing data for Group 1 pupils tested for hookworm (n = 2) and whipworm (n = 4) and in Group 2 for roundworm (n = 2), whipworm (n = 5) and schistosomiasis (n = 2).

Appendix 4: Primary analyses, including all nontransferring pupils

Yr(s)	Cluster summaries							Individual-level random effects (adjusted for pupil population size and zone)			
				Attendance (%)				Unadjusted (N 1998 = 106,480; N 1999 = 107,039)		Adjusted for age, SAP (N 1998 = 104,198; N 1999 = 106,078)	
	Grp 1 % (N)	Grp 2 % (N)	Grp 3 % (N)	Intervention % (N)	Control % (N)	Difference (95% CI)	P-val	Odds ratio	P-val (LR test)	Adjusted odds ratio	P-val (LR test)
1998	84.0 (25)	77.4 (24)	80.5 (24)	84.0 (25)	79.0 (48)	5.00 (−1.92–11.89)	0.154	1.63 (0.90–2.95)	0.110	1.38 (0.88–2.15)	0.163
1999	71.2 (25)	69.7 (25)	68.3 (25)	70.4 (50)	68.3 (25)	2.17 (−3.71–8.05)	0.464	1.23 (1.02–1.48)	0.034	1.23 (1.02–1.48)	0.032
1998 and 1999				-	-	-	-	1.74 (1.66–1.82)	< 0.001	1.81 (1.74–1.90)	< 0.001
				Mean ICS Examination score				Unadjusted (N 1998 = 14,985; N 1999 = 12,372)		Adjusted for age, SAP (N 1998 = 14,961; N 1999 = 12,36)	
	Grp 1 mean (N)	Grp 2 mean (N)	Grp 3 mean (N)	Intervention mean (N)	Control mean (N)	Difference (95% CI)	P-val	Treatment coefficient	P-val	Treatment coefficient	P-val
1998	−0.071 (25)	−0.006 (25)	0.050 (25)	-0.071 (25)	0.022 (50)	−0.093 (−0.310–0.124)	0.393	−0.111 (−0.292–0.070)	0.230	−0.113 (−0.299–0.073)	0.235
1999	−0.058 (25)	0.055 (24)	0.011 (25)	-0.002 (49)	0.011 (25)	−0.013 (−0.206–0.180)	0.893	−0.005 (−0.182–0.172)	0.956	0.008 (−0.187–0.171)	0.930
1998 and 1999				-	-	-	-	−0.102 (−0.270–0.067)	0.237	−0.104 (-0.275–0.067)	0.233

Note: Cluster summaries are the unweighted means of school-level summaries. We adjusted the individual analyses for the variables used in stratifying the randomisation. The total units of analysis for the combined year 1 and year 2 analyses are the sums of the yearly totals. The random effect is fitted for the school. Cluster summaries could not be compared across years because secular trends would influence the result. We excluded transfer pupils after they had moved schools.

Appendix 5: Secondary outcome: worm infections in all pupils tested

Type of worm infection	Group 2 (pre-intervention)	Group 1 (after one year of intervention)	Difference (95% CI)	P-value
Pupils tested (n)	1,477	861		
Average egg count (eggs/g; arithmetic mean)				
Hookworm	690	233	−457 (−657 to −257)	< 0.001
Roundworm	4,216	1,548	−2,668 (−4,244 to −1,092)	0.001
Whipworm	369	265	−104 (−376 to 168)	0.445
Schistosomiasis	226	112	−114 (−284 to 56)	0.183
Proportion with moderate infection (WHO thresholds)				
Hookworm	7.6%	2.3%	−5.3% (−8.0 to −2.7)	< 0.001
Roundworm	23.6%	8.9%	−14.8% (−23.0 to −6.5)	0.001
Whipworm	7.2%	6.7%	−0.5% (−6.4 to 5.3)	0.853
Schistosomiasis	16.7%	7.9%	−8.8% (−19.8 to 2.1)	0.112

Note: Data shown represent all tested individual children in the year that they were tested, regardless of eligibility for drug treatment. Missing data: There were missing data for Group 1 pupils tested for hookworm (n = 2) and whipworm (n = 4) and in Group 2 for roundworm (n = 2), whipworm (n = 5) and schistosomiasis (n = 2).

Appendix 6: Secondary outcomes: WAZ, HAZ in 1999, in all pupils tested

Group(s)	Number tested	Intervention status	Mean z-score	Difference from Group 1 (95% CI)	P-value
Weight-for-age z-score (WAZ)					
1	3,425	Received	−1.262	ref	-
2	2,459	None	−1.213	0.049 (−0.091 to 0.189)	0.486
3	2,601	None	−1.302	−0.040 (−0.179 to 0.098)	0.561
2 and 3	5,060	None	−1.259	−0.003 (−0.124 to 0.118)	0.956
Height-for-age z-score (HAZ)					
1	3,426	Received	−1.147	ref	-
2	2,460	None	−1.046	0.101 (−0.126 to 0.329)	0.378
3	2,602	None	−1.339	−0.192 (−0.417 to −0.034)	0.094
2 and 3	5,062	None	−1.196	−0.048 (−0.156 to 0.252)	0.639

Note: We accounted for the clustered nature of the data by calculating the 95 per cent confidence intervals around the mean of the cluster means. WAZ data were unavailable for 2,517 (42.4 per cent) Group 1 pupils, 3,192 (56.5 per cent) Group 2 pupils and 2,885 (52.6 per cent) Group 3 pupils, all in grades 3–8. HAZ data were unavailable for 2,516 (42.3 per cent) Group 1 pupils, 3,191 (56.5 per cent) Group 2 pupils and 2,884 (52.6 per cent) Group 3 pupils, also all in grades 3–8.

Appendix 7: Sensitivity analysis

Scenario 1: Dropping first visits in 1998 (11,588 dropped), dropping first and second visits in 1999 (31,404 dropped)

Cluster summaries								Individual-level random effects (adjusted for pupil population size and zone)			
Yr(s)	Attendance (%)							Unadjusted (N 1998 = 81,985; N 1999 = 68,228)		Adjusted for age, SAP (N 1998 = 81,242; N 1999 = 67,945)	
	Grp1 % (N)	Grp2 % (N)	Grp3 % (N)	Intervention % (N)	Control % (N)	Difference (95% CI)	P-val	Odds Ratio	P-val (LR test)	Adjusted Odds Ratio	P-val (LR test)
Year 1 Visits 98(2) – 98(8)	84.0 (25)	76.1 (24)	80.0 (24)	84.0 (25)	78.1 (48)	5.91 (-1.35 — 13.17)	0.109	1.80 (0.92 — 3.54)	0.090	1.49 (0.88 — 2.54)	0.143
Year 2 Visits 99(3) – 99(8)*	70.5 (25)	73.3 (24)	68.3 (25)	71.8 (49)	68.3 (25)	3.57 (-1.33 — 8.47)	0.150	1.22 (1.00 — 1.52)	0.079	1.22 (0.97 — 1.52)	0.088
1998 + 1999								2.08 (1.98 — 2.19)	<0.001	2.13 (2.02 — 2.25)	<0.001

Scenario 2: Dropping first visits in 1998 (11,588 dropped), coding first and second visits in 1999 as year 1, inc. as 'control' for Group 2 (~31,000 changes)

Cluster summaries								Individual-level random effects (adjusted for pupil population size and zone)			
Yr(s)	Attendance (%)							Unadjusted (N 1998=104,213; N 1999=68,228)		Adjusted for age, SAP (N 1998=103,318; N 1999=67,945)	
	Grp1 % (N)	Grp2 % (N)	Grp3 % (N)	Intervention % (N)	Control % (N)	Difference (95% CI)	P-val	Odds Ratio	P-val (LR test)	Adjusted Odds Ratio	P-value (LR test)
Year 1 Visits 98(2) – 99(2)	83.0 (25)	73.1 (25)	78.2 (25)	83.0 (25)	75.6 (50)	7.38 (-0.18 — 14.94)	0.056	1.64 (1.06 — 2.55)	0.030	1.44 (1.03 — 2.00)	0.036
Year 2 Visits 99(3) – 99(8)*	70.5 (25)	73.3 (24)	68.3 (25)	71.9 (49)	68.3 (25)	3.57 (-1.33 — 8.47)	0.150	1.22 (1.00 — 1.52)	0.079	1.22 (0.97 — 1.52)	0.088
1998 + 1999								1.89 (1.80 — 1.98)	<0.001	1.92 (1.82 — 2.01)	<0.001

* Results for year 2 are the same in both scenarios.

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 Part 1: a pure replication, 3ie Replication Paper 3, part 1. 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 2: alternative analyses, 3ie Replication Paper 3, part 2. Aiken, AM, Davey, C, Hayes, RJ and Hargreaves, JR (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