Does Mass Deworming Affect Child Nutrition? Meta-analysis, Cost-Effectiveness, and Statistical Power*

Kevin Croke[1], Joan Hamory Hicks[2], Eric Hsu[2], Michael Kremer[3], and Edward Miguel[2]

[1]World Bank and Harvard T.H. Chan School of Public Health
[2]University of California, Berkeley
[3]Harvard University

July 7, 2016

*We are grateful to the Campbell Collaboration deworming team, especially Vivian Welch, for generously sharing data and information on their work. We thank Harold Alderman, Shally Awasthi, Don Bundy, Serene Joseph, Chengfang Liu, Scott Rozelle, and Walter Willett for help interpreting their studies, searching for original data, or supplying us with additional results. We thank Egor Abramov, Kamran Jamil, Spencer Ma, Cole Scanlon, Wen Wang, and Kevin Xie for research assistance. We thank Amrita Ahuja, Harold Alderman, Sarah Baird, Matthew Basilico, Peter Hotez, Macartan Humphreys, Rachael Meager, Antonio Montresor, and Dina Pomeranz for helpful comments. Michael Kremer declares that he is a former board member of Deworm the World, a US non-profit organization. He has received no funding from Deworm the World. He is also a part-time employee of USAID, which financially supports deworming activities. This paper was written in his academic capacity and USAID had no influence over the writing of this paper.

Abstract

Background

The WHO has recently debated whether to reaffirm its long-standing recommendation of mass drug administration (MDA) in areas with more than 20% prevalence of soil-transmitted helminths (hookworm, whipworm, and roundworm). There is consensus that the relevant deworming drugs are safe and effective, so the key question facing policymakers is whether the expected benefits of MDA exceed the roughly \$0.30 per treatment cost. The literature on long-run educational and economic impacts of deworming suggests that this is the case (**?**; **?**; **?**; **?**; **?**). However, a recent meta-analysis, **?** (hereafter TMSDG), disputes these findings. The authors conclude that while treatment of children known to be infected increases weight by 0.75 kg (95% CI: 0.24,1.26; $p = 0.0038$), there is "substantial evidence" that MDA has no impact on weight or other child outcomes. This has led some to question the WHO policy and the literature on long-run impacts.

Methods

We first examine statistical power in TMSDG. Next, we update the TMSDG analysis by including studies omitted from that analysis and extracting additional data from included studies. To do this, we follow procedures outlined in the *Cochrane Handbook for Systematic Reviews of Interventions* (**?**), such as deriving standard errors from p-values when the standard errors are not reported in the original article. The updated sample includes twice as many trials as analyzed by TMSDG, substantially improving statistical power.

Results

We find that TMSDG is underpowered, that is, the analysis would conclude that MDA has no effect even if the true effect were (1) large enough to be cost effective relative to other interventions in similar populations, or (2) of a size that is consistent with results from studies of children known to be infected. The hypothesis of a common zero effect of multiple-dose MDA deworming on child weight at longest follow-up is rejected at the 10% level using the TMSDG dataset, and with a p-value < 0.001 using the updated sample. Adding any one of five individual updates to the TMSDG data in isolation leads to rejection of the null hypothesis at the 5% level. Applying either of two study classification approaches used in previous Cochrane Reviews (prior to TMSDG) also leads to rejection at the 5% level.

In the full sample, including studies in environments where prevalence is low enough that the WHO does not recommend deworming, the average effect on child weight is 0.134 kg ([95% CI: 0.031,0.236], random effects estimation). Results are robust to removing any individual study. In environments with greater than 20% prevalence, where the WHO currently recommends mass treatment, the average effect on child weight is 0.148 kg ([CI: 0.039,0.258], random effects). In environments with more than 50% prevalence, where the WHO recommends multiple annual doses, the estimated effect is 0.182 kg ([CI: 0.070,0.293], random effects).

The implied average effect of MDA on infected children in the full sample (calculated by dividing estimated impact by worm prevalence for each study and applying a random effects model) is 0.301 kg. This likely reflects considerably larger effects on those with moderate to severe intensity infections and smaller effects on those with light infections. At 0.22 kg per U.S. dollar, the estimated average weight gain per dollar expenditure from deworming MDA (assuming two annual treatments) is more than 35 times that from school feeding programs as estimated in RCTs.

1 Introduction

Soil transmitted helminths (hookworm, whipworm, and roundworm) are estimated to affect 1 in 4 people in endemic countries (**?**), and at least 1.3 billion people worldwide (**?**). The intensity of infection is correlated with prevalence, and highly skewed, so while many infected individuals have light infections, only a minority have the moderate to severe intensity infections believed to account for the bulk of morbidity (**?**).

There is consensus that treatment is safe and effective. Indeed, deworming drugs are the standard of care for those known to be infected. Because individual collection and testing of stool samples prior to treatment is prohibitively expensive and logistically impractical in many low-income contexts, the World Health Organization (WHO) has long recommended mass drug administration (MDA) in endemic areas. In particular, the WHO recommends annual treatment in areas with more than 20% prevalence of soil-transmitted helminths, and multiple annual treatments where prevalence is above 50% (**?**).

Studies of the long-term impact of deworming in multiple environments suggest that even under conservative assumptions and allowing for uncertainty, the expected benefits of MDA exceed the $0.30 per treatment costs (**?**; **?**; **?**; **?**; **?**).[1] Many development organizations that have examined the evidence have judged deworming to be highly cost effective, including the Copenhagen Consensus (**?**), the Disease Control Priorities Project (**?**)[2], Givewell (**?**; **?**), the Abdul Latif Jameel Poverty Action Lab (J-PAL) (**?**), and the World Bank (**?**).

However, a recent Cochrane Review (**?**), hereafter "TMSDG," argues against this long-standing consensus. The Review expresses support for treating those known to be infected (p. 30), and

[1] Givewell calculates the cost of deworming for soil-transmitted helminths in India at $0.30 per child per treatment, which includes both drug and delivery costs, including the value of staff time (**?**).

[2] We note that while analysts at Givewell identified errors in the calculations of deworming's cost effectiveness presented in the Disease Control Priorities Project, their revised calculations also find it to be a cost effective intervention at $82.51 – $138.28 per DALY (**?**).

estimates that single dose treatment for those known to be infected (and to have moderate intensity infections on average) increases weight by 0.75 kg (95% CI: 0.24,1.26; $p = 0.0038$). However, it takes a strong stand on mass drug administration in areas where worms are endemic, arguing that there is "substantial evidence" that mass treatment has no impact on child weight or other outcomes. This has led a subset of the TMSDG authors to conclude in other work that the belief in long-term educational and economic impacts discussed above are "delusional" (**?**).

Of course, failure to reject the null hypothesis of no effect does not constitute evidence of no effect. Particularly when treatment is inexpensive and side-effects are mild, studies and meta-analyses must have adequate power in order to rule out possibly modest effect sizes that would still render treatment cost-effective. Unfortunately, underpowered meta-analyses are extremely common in many areas of health research (**?**), and it is important to examine when this is the case. In the case of deworming, the cost of MDA for STH is estimated at just $0.30 per treatment (**?**). Since the majority of those treated in MDA programs will either be uninfected or have only light intensity infections rather than the moderate to severe infections thought to account for the bulk of STH morbidity (**?**, **?**), statistical power to pick up population-wide effects is typically limited (**?**). TMSDG include studies in environments with prevalence below the 20% threshold at which the WHO recommends deworming, further weakening statistical power by including estimates from settings where there are few worm infections to begin with.

In this paper, we first assess statistical power in TMSDG, concluding that it is inadequate to rule out weight gain effects that would either (i) make MDA cost effective relative to school feeding programs aimed at similar populations, or (ii) that would be reasonable to expect in MDA programs, given the estimated impacts in populations that have been pre-screened for infection and that have moderate intensity infection on average. We then update the work of TMSDG to create a more comprehensive, and thus better-powered, meta-analysis. The update includes 22 estimates from 20 studies examining the impact of multiple-dose MDA on child weight at longest follow-up, twice as many as TMSDG. It includes four studies not identified in TMSDG and additional

data from six studies discussed by TMSDG but not included in their meta-analysis for the child weight outcome, either using procedures in the *Cochrane Handbook for Systematic Reviews of Interventions* (**?**) to extract additional data or contacting the original authors to obtain additional information.[3] Additionally, in three cases, the updated analysis includes improved estimates, for example, by obtaining information on intra-cluster correlation directly from the original study authors rather than by imputing data from other studies.

With the full data set, the hypothesis of a common zero effect of multiple-dose deworming on weight is rejected with $p < 0.001$. With the TMSDG sample, this hypothesis is rejected at the 10% level, but applying either a study classification approach used in the previous 2012 Cochrane review or a study classification approach used in earlier Cochrane reviews leads to rejection with $p < 0.01$.[4] Any one of five other individual updates to the TMSDG data leads to rejection of the null hypothesis of a common zero effect at the 5% level.

Using our updated sample, and following TMSDG in including studies from low-prevalence environments where the WHO does not recommend MDA, mass deworming is estimated to increase child weight by 0.134 kg (95% CI: 0.031,0.236; $p = 0.01$; random effects estimation). This effect remains robust (at $p < 0.05$) when any individual trial estimate is dropped from the meta-analysis. The result is also nearly unchanged even when simultaneously dropping any two of the 22 estimates: among the 231 possible combinations of two studies that could be dropped simultaneously, in 96% of cases the estimated effect remains statistically significant at $p < 0.05$, and the largest p-value is just 0.067. In areas with prevalence above 20%, where the WHO recommends MDA, the average estimated impact on child weight is 0.148 kg (CI: 0.039,0.258; $p = 0.008$; random effects). Where the WHO recommends multiple annual MDA (areas with prevalence above 50%), the average estimated weight gain is 0.182 kg (CI: 0.070,0.293; $p = 0.001$; random effects).

The implied average effect of MDA deworming on infected children in the full sample (cal-

[3] In several cases we received data from the Campbell Collaboration, who had themselves directly contacted the original authors. See Appendix A for details on the source of all new studies.

[4] See section 5.1 for more detailed discussion of these study classification issues.

culated by dividing estimated impact by worm prevalence for each study and applying a random effects model) is 0.301 kg.[5] This average effect likely conceals substantial heterogeneity. Light infections are often asymptomatic, and only between 2 and 16 percent of the population experience moderate to severe intensity infections in the studies in our sample that report this information,[6] so implied effects in the subpopulation of those with moderate to severe intensity infections are likely much larger. For more general context on the implied average effect size of 0.301 kg, the difference in weight gain for boys at the 25th versus at the 50th percentile of the weight-for-age distribution between ages 2 and 3 is 0.2 kg (**?**).[7]

Moreover, this gain comes at modest cost compared to some other common interventions. The implied weight gain per U.S. dollar of expenditure is 0.22 kg, assuming two MDA treatments per year. For comparison, the weight gain per dollar of expenditure estimated in RCTs by **?** for school feeding programs is less than 0.01 kg, suggesting that relative to school feeding, deworming is highly cost effective in increasing weight in school-age children in low-income countries. This echoes a recent epidemiological study that similarly finds deworming to be highly cost-effective (**?**).

This paper is organized as follows. Section **??** provides background information on soil-transmitted helminths, mass drug administration, and earlier literature, including TMSDG, and

[5]The implied average effect on infected children in the subsamples with 20% and 50% prevalence are similar, at 0.249 and 0.276 respectively. The confidence intervals overlap with that of the full sample.

[6] In particular, **?** reports 15% prevalence of any helminthic infection among preschool aged children at baseline in Peru, with only 1.8% of children exhibiting moderate or heavy infections, while **?** report hookworm, whipworm, and roundworm prevalences of 77%, 42%, and 55%, with 15%, 16%, and 10% prevalence of moderate to severe infection respectively. Moreover, **?** use thresholds that are lower than the WHO thresholds used by **?** for moderate to severe intensity infections, suggesting that the comparable prevalence of moderate-severe infections in the **?** sample is lower than the figures presented here.

[7]According to WHO growth charts, a boy at the 25th percentile of the weight for age distribution grows 1.9 kg between age 2 and age 3, while a boy at the 50th percentile grows 2.1 kg over that year. Note that the difference in weight gain for boys at the 25th versus at the 50th percentile of the weight-for-age distribution between ages 3 and 4 is also 0.2 kg (**?**). In our full sample, the median duration of follow up at which weight gain is measured is 1 year.

assesses whether TMSDG is adequately powered to rule out cost-effectiveness. Section **??** discusses the sample, including criteria for study inclusion, the procedure used to identify studies, and the general principles guiding data extraction and determination of which estimates to use in the meta-analysis. The appendix details the application of these principles to individual studies. Section **??** describes our hypothesis testing and estimation strategy. Section **??** replicates the TMSDG analysis, tests the hypothesis of a common zero effect of multiple dose MDA, and estimates the impact of multiple dose MDA on child weight, both in environments where the WHO recommends mass deworming and more broadly. Section **??** concludes with a discussion of implications, methodology, and directions for future research.

2 Background

The potential health consequences of worm infections are generally agreed to depend on the number of worms in the body (i.e., infection intensity), rather than a simple binary indicator of infection status. Infection intensity is highly skewed and is strongly correlated with disease prevalence: in low prevalence populations relatively few people have severe infections, while many more do in high prevalence populations (**?**). STH are spread via eggs deposited in the local environment when individuals defecate in their surroundings or do not practice proper hygiene after defecating. Due to the transmission mechanism, school-aged children are especially vulnerable to these worm infections, and also play an important role in spreading them in the local community (**?**).

There is widespread acceptance that those who are known to be infected with intestinal helminths should be treated. Indeed, this is the standard of medical care (**?**; **?**; **?**), and some consider not treating individuals known to be infected as unethical. New trials of this type are therefore typically not conducted, but TMSDG identify five trials of single-dose treatment on what the authors term "infected children,"[8] and one trial of multiple dose treatment. The TMSDG meta-analysis found a

[8]As described in more detail below, this terminology is slightly different than "screened for infection," which was the classification used in the previous Cochrane Review (**?**).

child weight gain effect of 0.75 kg across the single dose trials (95% CI: 0.24,1.26; $p = 0.0038$).[9] These trials were largely conducted in settings with considerable infection intensity: for the 3 (out of 5) single dose studies that report infection intensities as measured by eggs per gram (epg), all three report mean epg values which are equivalent to a moderate intensity infection for at least one type of worm (**?**; **?**; **?**).[10]

Because deworming treatment is inexpensive and safe but diagnosis is comparatively expensive (necessitating lab analysis of a stool sample) and logistically difficult in many contexts, the WHO recommends annual mass treatment in areas where worm infections are above 20% and multiple treatments annually where prevalence is greater than 50%. Screening for worm infections requires testing stool samples, which in turn necessitates skilled staff, laboratory facilities, and re-contacting infected individuals for treatment, which can be challenging in many contexts where worm infections are endemic. Furthermore, the Kato-Katz test - the most commonly used method for testing for worms in these regions - has an estimated specificity between 52% and 91% (**?**; **?**), suggesting that many infections would go undetected, and thus presumably untreated, even with proper screening. TMSDG note that screening for worm infections is not recommended by the WHO because the cost of screening is 4 to 10 times that of the treatment itself (**?**, p. 7). Taken together, this suggests that a policy of screened treatment for worm infections would be costly, logistically complicated, and imprecise.

Subsequent to the WHO recommendation for MDA, a social science literature emerged measuring the longer-term educational and economic impact of deworming. Four studies in three moderate to high prevalence settings – in Kenya, Uganda, and the historical southern United States – all find substantial long-run impacts of deworming on educational outcomes (**?**; **?**; **?**; **?**). Two of these studies also report economic outcomes and both find positive effects.

[9]The degree of confidence that TMSDG have in these findings is unclear. At one point in the text they note that the case for treating infected children is "obvious" (p. 30). However, they elsewhere describe the meta-analysis results as "low quality evidence" (**?**, p. 2).

[10]These are our calculations, using the infection intensity thresholds from **?**.

- **?** finds that Ugandan children exposed to a deworming program originally studied in **?** have higher math test scores nearly a decade later, with effect sizes of over 0.2 standard deviation units.

- **?** finds that infant children who lived in Kenyan communities where older school-age children were dewormed show large cognitive test score improvements ten years later, presumably due to reduced infection through beneficial spillover effects. The magnitude of the effect is 0.2 to 0.3 standard deviation units, which is equivalent to between 0.5 to 0.8 years of schooling.

- Using a difference-in-difference estimation methodology rather than a randomized design to study a deworming campaign in the U.S. South in the early 1900s, **?** finds that deworming led to increased school enrollment and attendance for children, and improved literacy and boosted income by 17% for adults who were treated as children.

- Finally, **?** estimate that a decade after treatment, males who participated in mass deworming in Kenya worked 17% more hours per week and had higher living standards, missing approximately one fewer meal per week. Females were approximately one-quarter more likely to have passed the primary-school leaving exam and attended secondary school. The estimated value of benefits, in terms of the net present value of future earnings net of increased schooling costs, exceeds the cost by more than one hundred fold.

These results suggest that the expected benefits of deworming would greatly exceed its costs even if one took a conservative approach, assuming a very low probability of effects of the measured magnitude or assuming that true effects are considerably less than the measured effects (**?**).

While TMSDG estimate positive impacts of deworming treatment on weight of children known to be infected, they argue in contrast that there is "substantial evidence" that mass deworming does not improve weight or other child outcomes. Note that this is a considerably different – and far

more demanding – statement than a claim that the null hypothesis of no effect cannot be rejected (for instance, due to a lack of statistical power), or that MDA has a positive impact in some settings but not others.

TMSDG has been controversial. **?** note substantial changes over time in the way the meta-analysis in the Cochrane Review on deworming is presented (see the evolution across **?**, **?**, **?**, and **?**). **?** also note that although the text refers to a protocol (**?**, p.29), they were unable to find a publicly available pre-specified protocol for the updated review, leading to confusion over the precise hypothesis being tested, uneasiness over how studies are grouped for analysis, and concern about which studies are included versus excluded. A number of authors (**????**) have expressed concern over lack of consideration given to the effects of different STH species, treatments, and drug distribution strategies.

2.1 Is TMSDG Adequately Powered?

A simple calculation suggests that the TMSDG analysis is underpowered to rule out the possibility of effects that would make mass drug administration cost effective relative to school feeding. As noted, the estimated weight gain in kilograms per dollar spent from school feeding is less than 0.01 (from RCT studies; **?**). Given the $0.60 per year treatment cost of deworming in environments where two annual doses are required, an effect of just 0.006 kg would make deworming cost effective relative to school feeding (setting aside, for the time being, the issue of other outcomes).

Examining the random effects meta-analysis estimator that is the focus in TMSDG and taking as given the variance of the effect across studies TMSDG estimate, and the standard errors TMSDG report for the underlying studies, the implied Minimum Detectable Effect (MDE) size at 95% confidence and 80% power is 0.28.[11] Using TMSDG's random effects estimate of 0.08 kg, there

[11]These estimates were obtained through a series of Monte Carlo simulations. On iteration k of the process, we simulate for each study i an estimated effect β_{ik} as a draw from $N(\hat{\mu}, \hat{\tau}^2 + \hat{\sigma}_i^2)$. Here, $\hat{\tau}^2$ is the estimated variance of the local effect within the RE model (note that $\hat{\tau}^2$ is 0.074 in the TMSDG sample), and $\hat{\sigma}_i^2$ is the squared standard error of the estimated effect in study i. We

is only 17% statistical power to detect a significant effect at 95% confidence.

TMSDG also appears to be underpowered to detect effects that might be reasonable to expect from MDA given the estimated effects TMSDG report in studies of those known to be infected (and to have moderate intensity infections on average), given the preponderance of individuals with no infection or light infections in most MDA samples; the average prevalence in their sample of studies is 46% (among those studies reporting prevalence). The suggestion that there are positive effects from treating individuals known to be infected, yet no impact from treating populations containing infected individuals creates an apparent paradox, since if there are effects of deworming among those known to be infected, then one would expect a smaller, but still positive average effect of MDA in endemic populations. A simple exercise suggests that this hypothesis cannot be rejected using TMSDG's data. Under the assumption that MDA does not affect weight in uninfected children,[12] the implied treatment effect on weight for infected children is given by dividing the intention-to-treat effect by prevalence. Calculating this effect for each study[13] and then applying a random effects model to the estimated effects on infected children, yields a confidence interval for the average effect on infected children whose upper end is 0.584 kg. This, in turn, is well within the confidence interval of (0.24 kg, 1.26 kg) that TMSDG estimate for the effect of single dose treatment on those children known to be infected and have typical infection intensities far greater than in the MDA sample. Thus it is impossible to reject even the hypothesis that effects on weight are proportional to prevalence, let alone the hypothesis that weight gain effects are zero for the uninfected, and small for those with light infections and substantially larger for those with

vary $\hat{\mu}$ to compute statistical power for different effect magnitudes. Next, we take the simulated β_{ik} ($i = 1,...,N$) and apply a random effects model to obtain a simulated point estimate for a given effect size as well as a standard error. We repeat this process 10,000 times. The reported power of a test is the proportion of simulations where the null hypothesis that the average effect size is zero is rejected. The reported 80% MDE is an estimate of the effect size that would deliver a test with 80% power.

[12]This assumption may not hold literally if MDA creates epidemiological externalities, but it will hold under the null hypothesis.

[13]Note that these calculations include estimated prevalence for two studies in TMSDG's sample which do not report prevalence within their study - **?** and **?**. See Section **??** for more details.

moderate to severe intensity infections.

To understand why statistical power in TMSDG is so low, note first that TMSDG include only 11 treatment effect estimates on the impact of multiple dose mass deworming on child weight at longest follow-up, three of which have fewer than 100 subjects, and one of which only has 198 individuals. Larger studies are typically cluster-randomized and while cluster randomization offers the opportunity to pick up epidemiological externalities, it also limits power. The cluster-randomized studies have between 48 and 124 clusters, so there are relatively few units of randomization in some cases.

Second, in low worm infection prevalence populations, very large samples are needed to have sufficient statistical power to pick up effects of mass drug administration. A simple statistical rule of thumb is that since statistical power improves with the square root of the sample size, detecting an effect of given size among a proportion q of a population requires a sample $1/q^2$ times as large as among a population in which everyone is infected. If effects were proportional to prevalence, picking up an effect in a population with prevalence 0.46, for instance, would require a sample 4.7 times as large as in a population in which all are infected. Since low-intensity infections are expected to have little impact, and since intensity rises non-linearly with prevalence (**?**), only a small fraction of the population in low prevalence environments will have moderate to heavy infections and the necessary increase in sample size may be much larger if weight gain effects are concentrated among those with moderate-severe intensity infections.[14]

Together, these factors imply that it will be important to use the most efficient estimators possible, for example, by including baseline values of weight when possible rather than just endline values, in order to improve statistical power. It will also require efforts to employ all the available studies and data.

[14] **?** explain that the relationship between prevalence and intensity is well described by: $Mean\ Intensity = k * (1 - Prevalence)^{-1/k} - k$, where k is a parameter representing the degree of aggregation or dispersion of the parasite in the population. Estimates of k lie in the range 0.11 to 0.81 (**?**). Taking the derivative with respect to prevalence implies that intensity is expected to rise more than linearly in prevalence.

3 Sample and data extraction procedures

This section describes trial inclusion criteria, the search procedure for identifying studies, and the procedures for extracting data from included trials.

3.1 Trial inclusion criteria

Our analysis includes randomized controlled trials of deworming MDA with multiple doses that include child body weight as an outcome. Following what TMSDG term their "main comparison" (**?**, p. 4), we consider only trials in which multiple doses of deworming treatment were administered, and include treatment effect estimates from the longest follow-up reported. We focus on child weight gain because it is an important nutritional outcome which could potentially be improved over relatively short time horizons.[15] Moreover, there is substantial evidence for this outcome available in existing studies, which we expect to lead to relatively more adequate statistical power.[16] Weight is also highlighted as one of the three primary outcomes examined in TMSDG (p.11).[17] Only trials for which a proper intention-to-treat estimate can be obtained are included. Therefore, we require that the study (or trial authors, through personal communication) report out-

[15]The studies we consider generally take place in low-income settings where child obesity is not considered a widespread issue.

[16]For many of the other outcomes they examine, TMSDG include relatively few studies in their meta-analysis. Only weight (10 studies), height (8), and hemoglobin (9) have more than three studies that are aggregated in formal meta-analysis. As height deficiencies and stunting are generally conceived in the nutrition literature as the result of cumulative undernutrition over extended periods, we considered that height was unlikely to respond to deworming over the course of relatively short run trials (the median length is 12 months). Hemoglobin is an important outcome but it is most closely associated with hookworm (hookworm infection is the third leading contributor to the global burden of anemia, whereas neither of the other two STH species rank among the top causes of anemia (**?**); see also **?**. Of the 7 trials in the TMSDG meta-analysis of hemoglobin (which produce 9 treatment estimates, since two are factorial), only one appears to have any significant hookworm prevalence (**?**); the other six either do not report hookworm prevalence or report very low values, between 1% – 11%.

[17]The other two are hemoglobin and cognition.

comes for the population assigned to treatment and comparison groups, independent of whether they received treatment or not.

When estimating the mean effect of MDA on weight, we report results both in the set of trials that take place in settings where the WHO recommends deworming (i.e., those where prevalence of either hookworm, whipworm, or roundworm is over 20%, which is the threshold for annual MDA, or 50%, which is the threshold for multiple annual MDA), and in the full sample (for completeness and comparability to TMSDG).

3.2 Search procedure

We start with the sample of studies included in TMSDG, as a well-known and oft-cited systematic review, for their analysis of the impact of multiple-dose deworming treatment of "all children living in an endemic area" (i.e. mass drug administration or MDA) at longest follow-up on child weight. We supplement this sample with additional studies we could identify that meet the trial inclusion criteria above. The Campbell Collaboration generously shared information on additional trials which they identified for use in their own forthcoming systematic review on the impacts of mass deworming. Although we do not know exactly which studies the Campbell Collaboration use in their meta-analysis (which is not yet published), all studies we include were identified by the Campbell Collaboration and any study that they identified for their analysis of the impact of multiple-dose MDA on weight was considered for inclusion in our sample.

3.3 Data extraction and choice of estimator

Given the importance of statistical power, we sought to use the most precise unbiased estimator available. We also followed guidelines in the *Cochrane Handbook for Systematic Reviews of Interventions* (**?**). We use the following principles for extraction of data and selection of estimates from included trials in our analysis below:

i. If treatment effects are presented without standard errors, standard errors are calculated using other presented data (e.g., t-statistics, p-values, or 95% confidence intervals), where possible following the formulas provided in the *Cochrane Handbook* (**?**, section 7.7.3.3).

ii. If results are reported in figures rather than in the text or in a table, *Web Plot Digitizer* software (**?**) is used to extract numerical estimates from the figures.

iii. If key information on treatment impacts is missing from a paper (and cannot be derived from what is presented), original microdata (where available) is used to obtain relevant estimates. We also obtain information from trial authors in several cases, through either direct communication or thanks to the generosity of the Campbell Collaboration research team.

iv. Where studies report multiple treatment impact estimates, we follow the standard in TMSDG and the medical literature of favoring unadjusted estimates. If studies do not report unadjusted estimates and we are unable to obtain them directly, but the studies do report treatment effect estimates adjusted with standard covariates or baseline values (such as child age and sex), these estimates are included in the analysis. (Note that since expected weight gain varies with age, including age as a covariate should generally improve precision of the estimates, and that including age or other pre-determined variables as a covariate should not induce bias).

v. When there was a choice between treatment effect estimates based on a comparison of endline differences and treatment effect estimates based on a comparison of changes from baseline to endline, the "changes" estimate is used, since it typically is more precise. Using the single difference treatment effect typically leads to a substantial loss of statistical precision: when outcomes are highly autocorrelated over time (as is the case for body weight), failure to use baseline values in measurements of treatment effects results in far less precise estimates (**??**). Estimates that take into account baseline information remain unbiased, while typically improving precision, and thus are preferable under standard statistical criteria, such as under the

goal of minimizing mean squared error. Following the *Cochrane Handbook*, when baseline and endline means and measures of variance were present but variance of the changes are missing in the original text, the standard error for changes is calculated using a correlation coefficient for the value between baseline and endline imputed from other studies (**?**, section 16.1.3.2).

vi. In the event of apparent textual contradictions about key parameter values in a trial (for example, language in the study text which reports significant effects versus reported standard deviation values which imply non-significant results), we first try to obtain the original microdata to perform the estimation ourselves. Where this is not possible, we assess which statistics were the primary focus of reporting in the text and contact the original authors for clarification.

vii. When possible, treatment effect estimates are extracted based on an Analysis of Covariance (ANCOVA) model, rather than estimates based on difference-in-difference estimators. The *Cochrane Handbook* states that since ANCOVA estimates "give the most precise and least biased estimates of treatment effects they should be included in the analysis when they are available" (**?**, section 9.4.5.2). Properly including baseline weight measures in the analysis is also critical in contexts with baseline imbalance (**?**).

The full sample includes 22 estimates from 20 studies, twice as many as TMSDG. In particular, the full sample includes four studies not identified in TMSDG or unpublished when their review was compiled (**?**; **?**; **?**; **?**); data extracted from six studies discussed by TMSDG but not included in their meta-analysis for the MDA child weight outcome (**?**; **?**; **?**; **?**; **?**; **?**); and improved estimates from three studies that are included in the TMSDG meta-analysis (**?**; **?**; **?**). Note that we classify **?** as an MDA trial; the Cochrane authors classify it in this way in their 2012 Review but change the classification system in the 2015 update; we retain the 2012 classification in this analysis.[18] See the appendix for detailed information on each individual study in the full sample.

[18]TMSDG state that "We changed the classification of **?** and **?**. Previously these trials were in the 'all children in an endemic area' category, whereas now they are classified in the 'children with

4 Hypothesis tests and estimation strategy

In light of TMSDG's conclusion that there is "substantial evidence" of no impact of deworming MDA, we first report a test of the hypothesis that the true impact of multiple-dose deworming on weight is zero in all settings. This involves testing the null hypothesis that $B = 0$ in a standard fixed effects meta-analysis estimate:

$$\hat{Y}_i = B + \sigma_i \tag{1}$$

where $\hat{Y}$ is the estimated effect in study i; B is the true deworming treatment effect, and σ_i is a random variable, representing measurement error, assumed to be distributed normally, with mean zero and a standard deviation equal to the standard error of the estimated treatment effect in each study. Rejection of this null hypothesis implies that deworming affects child weight in at least some circumstances.

We then report the estimated average impact using a random effects model:

$$\hat{Y}_i = B + \mu_i + \sigma_i \tag{2}$$

where B is the underlying true average effect, μ_i is a normally distributed random variable denoting the difference between the average effect and the effect in the particular context, and σ_i represents measurement error due to sampling variation, which is assumed to be captured in the study-specific

infection.' This decision was based on reviewing the trials with parasitologists and examining the prevalence and intensity of the infection where clearly the whole community was heavily infected" (**?** p. 154). It is worth noting that although TMSDG exclude **?**, they include **?**; the highest recorded worm baseline prevalence in **?** by STH species is 92% (for ascaris); the highest prevalence in **?** is also 92% (for whipworm). Thus this reclassification does not appear to have been done systematically by worm prevalence. In our view, assessing the merits of the WHO policy by including studies in environments with prevalence below WHO thresholds while excluding MDA studies in areas with high prevalence may lead to risk of bias. Since our goal is to examine the effect of MDA, we retain the 2012 classification. **?** is a single dose deworming trial so does not enter into the present meta-analysis but **?** has one multi-dose treatment arm, which we incorporate. See appendix A.3 for more detail.

standard errors.

Finally, under the assumption that deworming has no effect on uninfected children, one can calculate for each study an implied estimated effect on weight for infected children as the estimated intention-to-treat effect divided by prevalence. If one takes the estimated prevalence to be accurate and not subject to measurement error, then standard errors for these estimated effects are straightforward to compute. One can then apply a random effects model to estimate the average treatment effect on infected children. In a few cases, a study does not report an exact value for prevalence, but we are able to identify whether the study has below or above 50% prevalence. If prevalence is above 50%, we then compute the average prevalence among all studies in the sample reporting greater than 50% prevalence and assign that value to the study. We proceed similarly for studies with below 50% prevalence that do not report an exact value for prevalence.[19]

5 Results

Subsection **??** first replicates the results of the TMSDG subsample. Subsection **??** tests the hypothesis of a common zero effect. Finally, subsection **??** reports estimated effects of MDA on weight.

5.1 Verifying replication of results in the TMSDG Sample

We call the sample of 11 treatment effect estimates from 10 studies used in the TMSDG meta-analysis for the impact of MDA on child weight gain the "TMSDG sample." See Table **??** for a list of these studies. Figure **??** verifies that our estimation procedure yields results similar to TMSDG when using this sample. In particular, this figure presents a forest plot with TMSDG estimates; inserting the effect sizes, standard errors, and sample sizes reported for each of these studies in

[19]See appendix A.7 for more detail on how prevalence categories were assigned when STH prevalence was not reported in the study text.

TMSDG's text and figures into the relevant formulas provided in **?** using the R statistical software package replicates the TMSDG results.

Using the data presented in TMSDG, the hypothesis of a common zero effect of deworming on child weight gain is rejected at the 10% level ($p = 0.089$). Figure **??** shows the fixed effect estimate used to test this hypothesis.

The null hypothesis of a common zero effect is more strongly rejected when applying either of two study classification approaches used in previous Cochrane Reviews (prior to TMSDG).

1. The first versions of the Cochrane Review (**?**; **?**) did not create separate categories for test-and-treat (i.e. treatment only of children who have been diagnosed with STH) and MDA studies (i.e. treatment of the whole population), but rather considered all the data together. At a minimum, reporting results with the full sample before turning to the subgroup analysis seems reasonable, since if deworming has a positive effect on infected individuals, and if there is no effect on uninfected individuals, then deworming must have a positive effect on weight in a population that includes infected individuals. In the case of the TMSDG dataset, using all of the multiple dose at longest follow-up studies would only involve the addition of the one multiple dose study the authors identify but classify as separate from the MDA studies - **?**.[20]

2. The 2012 Cochrane Review changed this approach, however, introducing a distinction that effectively distinguished between test-and-treat and MDA studies (**?**). Applying this classification used in the 2012 Cochrane Review to the studies in TMSDG also leads to the inclusion of **?**, which was classified as a MDA study by **?** (or, as a "target population treated" study, using the language in that review) rather than a test-and-treat study.

Thus, applying either of these procedures from Cochrane Reviews prior to TMSDG results in the addition of **?** to the TMSDG sample. With **?** classified in this way but otherwise using the data

[20]Note that we believe that **?** should be included in the set of MDA studies, since individuals were treated without first being screened, as discussed in Appendix A.3.

set in the TMSDG study, the null hypothesis of a common zero effect is strongly rejected with a p-value of 0.009 (see Table **??**).

5.2 Testing the hypothesis of a common zero effect

We now turn to the full sample of 22 treatment effect estimates. Table **??** shows that the null hypothesis of a common zero effect is rejected in the full sample at $p < 0.001$ in areas with prevalence above the thresholds at which WHO recommends MDA (20% prevalence) and in areas where it recommends multiple annual dose MDA (50% prevalence).

This result is not reliant on the addition of any one study to the TMSDG sample: any one of six updates leads to rejection of the hypothesis of a common zero effect with p$<$ 0.05 (see Table **??**).

Figures **??** and **??** show point estimates and confidence intervals for the effect of mass deworming drug administration on weight from each of the studies included in the TMSDG and full sample, respectively, as well as the point estimates and confidence intervals from the fixed effects estimation needed to test the hypothesis of a common zero effect.

The rejection of a common fixed effect of zero implies that MDA deworming affects child weight in at least some circumstances. If effects are positive in some circumstances, then unless they are negative in other circumstances, average effects must be positive. There is no scientific reason to believe that deworming has negative side effects on weight. With only one negative estimate significant at the 5% level out of 22 estimates in the full sample (Figure **??**), the patterns in the data seem consistent with the hypothesis that the true effect of MDA on weight is never negative. In future work, we hope to more formally examine if this hypothesis can be rejected. However, for the sake of comparability, we follow TMSDG in the next section by imposing that the distribution of true effects is normal around some unknown mean.

5.3 Estimating the impact of deworming

Table **??** reports random effects estimates. In the full sample, the estimated weight gain effect is 0.134 kg [CI: 0.031,0.236; $p = 0.01$]. Of course, the full sample includes trials conducted in low infection prevalence areas where the WHO does not currently recommend mass deworming.[21] In areas with greater than 20% prevalence, where the WHO recommends MDA deworming, the estimated treatment effect is 0.148 kg [95% CI: 0.039,0.258; $p = 0.008$]. In areas with more than 50% prevalence, where the WHO recommends multiple doses annually, the estimated effect is 0.182 kg [CI: 0.070,0.293; $p = 0.001$].

Our full sample estimate has more statistical power than TMSDG. Using the same approach as above, we find an MDE of 0.15 kg in the full sample, roughly half the MDE of 0.28 kg using the TMSDG sample.

This full sample effect remains robust (at $p < 0.05$) when any individual trial estimate is dropped from the meta-analysis, as shown in Table **??**. The level of significance with which the null hypothesis that the mean effect is zero can be rejected remains high even when simultaneously dropping any two of the 22 estimates: among the 231 possible combinations of two studies that could be dropped simultaneously, in 96% of cases the estimated effect remains statistically significant at $p < 0.05$, and the largest p value is just 0.067 (not shown). The results are similarly robust to dropping any one or two studies in the subsample of studies with prevalence greater than 20%

[21] TMSDG examine outcomes by subgrouping based on infection prevalence, but they split the data into three subgroups, one containing only two studies, thus limiting statistical power for each test. In the weight gain multiple dose comparison, TMSDG analyze 5 studies from low prevalence settings (defined as less than 50% infection), 4 from medium prevalence settings (50% to 70% prevalence), and 2 from high prevalence settings (over 70% prevalence). They take this tripartite division from an earlier WHO framework (World Health Organization, 2002). Creating groupings with only a few studies makes the resulting estimates far less precise. The *Cochrane Handbook* notes that "when there are few studies or if the studies are small, a random-effects analysis will provide poor estimates of the width of the distribution of intervention effects" (**?**, section 9.5.4). Note that this tripartite division was not pre-specified in the original Cochrane pre-analysis plan (**?**). Examining all studies conducted in environments in which the WHO recommends MDA while excluding studies where the WHO does not recommend MDA would enhance policy relevance.

and those with prevalence greater than 50%.

The full sample estimate of 0.134 kg comes from studies with average worm prevalence of 51%. Assuming treatment only affects weight in the worm-infected population, this implies an average effect of roughly 0.301 kg among those with worms (calculated by dividing estimated impact by worm prevalence for each study and applying a random effects model).[22] This in turn likely represents the average of a considerably larger effect among the small proportion of those infected who have moderate or severe intensity infections and a considerably smaller effect among the majority of those infected, who have light-intensity infections.

While the average weight gain is fairly modest, it is far from negligible relative to the very low cost of deworming. Given the estimated effect of 0.134 kg and the estimated cost of $0.60 per person treated for two treatments per year (in the multiple dose context we focus on in this meta-analysis), the estimated cost per kg gained is $\frac{\$0.60}{0.134kg}$ = $4.48 per kg. For reference, it is worth comparing this to nutritional programs aimed at similar populations, in particular, school feeding programs. **?** conduct a Cochrane Review of the impact of school feeding, and **?** combine these results with data on costs to estimate the cost effectiveness of school feeding. They estimate that over a one year period, "the cost of an extra kilogram of weight ranged from $112 to $252 in the RCTs and $38 to $86 in the [controlled before-and-after studies]" (p.177-8). This suggests that just focusing on the weight outcome, deworming is highly cost effective relative to another widely implemented intervention. As noted by **?**, school feeding is implemented in over 72 countries by the World Food Programme alone. To the extent that school feeding programs aim to produce child weight gain, deworming is likely to be a highly cost effective option for policymakers who already support school feeding.[23]

[22]Note that this calculation includes estimated prevalence for three studies in the full sample which do not report prevalence within their study - **?**, **?**, **?**. See Table **??** and Section **??** for more details.

[23]Note that this comparison is imperfect for a number of reasons, including that the cost figures for deworming are for India whereas those for school feeding are based in three African countries. Moreover, both school feeding and deworming may affect other outcomes. However, it is unlikely

Of course, a complete cost-effectiveness analysis of mass deworming treatment would also need to consider effects on later educational and labor market outcomes, in addition to child weight gains, and as noted above these are often substantial (**?**; **?**; **?**; **?**; **?**).

6 Conclusion

We began with the question of whether the expected benefits of mass drug administration according to WHO guidelines for deworming exceeds the cost.

To summarize, the null hypothesis of a common effect of zero weight gain from multiple-dose mass drug administration is rejected at the 10% level using the TMSDG data. Employing either of two classification systems used in previous Cochrane Reviews would lead to rejection at $p < 0.01$. Any one of five other updates to the data leads to rejection of the null hypothesis at the 5% level. Combining all updates leads to rejection with $p < 0.001$.

Reasonable people may disagree about statistical methods for analyzing data. However, at a minimum, it seems clear that implementing MDA generates child weight gains in some circumstances. Since the null hypothesis of a common zero effect is rejected, MDA must have positive impacts in at least some environments. This implies that if one accepts the standard view that antihelminthic drugs have no substantial side effects, the expected effect of deworming on child weight is positive.

Applying standard approaches from the *Cochrane Handbook* to a larger set of studies yields an estimated average effect of deworming on weight of 0.134 kg [95% CI: 0.031,0.236], corresponding to an estimated average effect of 0.301 kg [CI: 0.071, 0.530] among those with worm infections. While this effect is modest, it is substantial relative to the cost of deworming, and suggests

that appropriate adjustments for these factors would overturn the conclusion that deworming is highly cost effective in increasing weight relative to school feeding programs, given that the cost per kg of weight gain is an order of magnitude higher for school feeding. Moreover, while school feeding may also promote school participation, deworming, too, has been found to be highly cost effective in increasing school participation (**?**).

that deworming is many times as cost effective as widely implemented school-feeding programs at improving nutrition among school age children in low-income populations. Moreover, the findings that deworming improves nutrition in at least some environments implies that the literature on the long-run educational and economic impacts of deworming cannot be dismissed on *a priori* grounds, and this literature suggests that the expected benefits of MDA greatly exceed the cost.

Our results also suggest that the data from studies of mass deworming are consistent with the data from studies of deworming of those known to be infected, given the much lower prevalence and intensity of infection in MDA studies than in test-and-treat studies, and the substantial confidence intervals around both the MDA and test-and-treat estimates.

A key lesson of this paper is that large samples are needed to have adequate statistical power to pick up a minimum detectable effect (MDE) that corresponds either to 1) what would be expected as the average effect of MDA in populations in which about half the population is uninfected and only a small fraction have moderate-severe intensity infections, or 2) what would be needed to form a policy judgment that the expected benefit of deworming is less than the cost.[24] This implies that analyses which divide up the set of available studies will likely be underpowered, limiting the scope for further subsample tests. While the results here suggest that overall, multiple-dose MDA increases child weight, they also suggest that if MDA had similar impacts on weight across drug type or worm species, a meta-analysis focusing on any one species of worm or drug may well be underpowered. We therefore think it would be appropriate for any future studies designed to explore heterogeneity, across worm species or drug type, for example, to report a test of the hypothesis that the average effect of MDA for each worm species or each drug is the same, rather than to simply test the hypothesis that the effect of any one individual drug or MDA against any one species has zero impact on weight. Beyond its relevance for health research, greater awareness

[24]The estimated effect in the full sample is about 50% greater than the estimate in the TMSDG sample, though well within the confidence interval of (-0.11 kg, 0.27 kg) found by TMSDG in their meta-analysis of the impacts of multiple-dose MDA on weight. However, the confidence interval shrinks by 47% to (0.03 kg, 0.24 kg) with the full data set. Thus, incorporation of additional information into the analysis helps address the problem of insufficient power in TMSDG.

of the limitations of under-powered meta-analyses will become increasingly important as more social scientists conduct meta-analyses (**?**).

A further methodological lesson is that it is appropriate to explicitly test the null hypothesis of a common zero effect. Finally, in evaluating policy, it is appropriate to focus on all mass-drug administration studies conducted in environments where mass deworming is recommended under WHO guidelines rather than to mix in studies from environments in which worm prevalence is sufficiently low that the WHO does not recommend mass treatment, or to selectively exclude studies of mass drug administration conducted in high-prevalence environments.

While we have argued that deworming MDA is cost effective based on its impact on child weight alone, there is evidence that deworming also leads to gains in other outcomes (**?**). TMSDG aggregate data from a more limited number of studies for outcomes other than weight, and some of these are from low-prevalence environments where the WHO does not recommend deworming, further reducing statistical power. This means that for each outcome assessed individually, it is typically impossible to reject either the hypothesis of no average effect, or the hypothesis of effects large enough to make deworming cost effective. In such a setting, it may be appropriate to consider the joint hypothesis that there is no impact on any of the child outcomes considered. We hope to do this in future work.

We have begun to explore heterogeneity of impact with covariates suggested by the scientific literature, including prevalence and intensity, age, and whether the study design captures epidemiological spillovers. One could also examine heterogeneity by helminth type, drugs used, and co-morbidities. However, given the limited number of studies and hence limited degrees of freedom, scope for examining heterogeneity while maintaining statistical power is limited. In future work, we hope to systematically examine heterogeneity, but here we simply capture heterogeneity using random effects estimation.

We follow TMSDG in assuming a normal distribution of the effect of MDA across study environments, but future work could relax the assumptions of symmetry and normality on the dis-

tribution of effects across studies and estimate these from the data. There is no reason, *a priori*, to expect that the distribution of deworming effects follows such a distribution. In fact, the underlying science naturally suggests a non-symmetric distribution, with positive effects in some cases and negligible effects in cases with low prevalence and intensity of infections.

We hope to examine a Bayesian, rather than frequentist, approach to meta-analysis for policy analysis in future work. The implicit loss function implied by requiring 95% confidence to undertake MDA without regard to the statistical power of the test is one in which there is a high cost of a false positive and a low cost of a false negative. That might be appropriate if, for example, the Food and Drug Administration were considering a drug that might have side effects. However, in the deworming context, the drugs have already been through regulatory approval, the monetary cost of deworming is low, and there is no evidence of serious side effects, while there is at least some potential that deworming has large long-run benefits (**?**). Thus, the cost of a false positive is low while the cost of a false negative is potentially quite substantial.

A Bayesian approach to estimating whether the expected benefit of MDA according to WHO guidelines exceeds the cost would start with a prior on the effect of deworming. Studies of the impact of treatment on those known to be infected provide a natural prior. It would factor in the range of benefits that have been estimated from deworming, each with an attached probability weight in order of assess whether summing across the range of potential benefits times their likelihood yields a benefit greater than the estimated cost. Since the net present value of the long-run educational and economic benefits has been estimated as more than one-hundredfold that of the cost (**?**), assessing even the subjective probability of these benefits would likely lead to the conclusion that the expected benefits of MDA exceed their cost.

Table 1: Summary of Studies

Study: Full Name	Study: Short Name	TMSDG Estimate (SE)	Full Estimate (SE)	Prevalence	Country
1. **?**	Kruger 1996	-0.38 (0.23)	(same)	0.38	South Africa
2. **?**	Watkins 1996	0.13 (0.11)	(same)	0.92	Guatemala
3. **?**	Donnen 1998	-0.45 (0.17)†	(same)	0.11	Zaire
4. **?**	Awasthi 2000	-0.05 (0.08)	(same)	0.13	India
5. **?**	Dossa 2001a	0.00 (0.27)	(same)	0.59	Benin
6. **?**	Dossa 2001b	0.00 (0.14)	(same)	0.59	Benin
7. **?**	Alderman 2006	0.15 (0.09)‡	(same)	>50%±	Uganda
8. **?**	Awasthi 1995	0.98 (0.15)	(same)	≤20%∓	India
9. **?**	Awasthi 2001	0.17 (0.34)	0.17 (0.07)	0.09	India
10. **?**	Hall 2006	0.0 (0.07)	0.05 (0.06)	0.84	Vietnam
11. **?**	Sur 2005	0.5 (0.47)	0.29 (0.09)	0.54	India
12. **?**	Willett 1979	(not included)	0.16 (0.08)	0.55	Tanzania
13. **?**	Joseph 2015	(not included)	0.04 (0.05)	0.12	Peru
14. **?**	Miguel 2004	(not included)	-0.76 (0.44)	0.77	Kenya
15. **?**	Ndibazza 2012	(not included)	0.01 (0.09)	0.03	Uganda
16. **?**	Gupta 1982a	(not included)	0.027 (0.175)	0.51	Guatemala
17. **?**	Gupta 1982b	(not included)	0.13 (0.15)	0.54	Guatemala
18. **?**	Ostwald 1984	(not included)	0.70 (0.45)	0.96	Papua New Guinea
19. **?**	Gateff 1972	(not included)	0.35 (0.13)	>50%≀	Cameroon
20. **?**	Wiria 2013	(not included)	0.19 (0.45)	0.75	Indonesia
21. **?**	Liu 2016	(not included)	0.03 (0.15)	0.31	China
22. **?**	Stephenson 1993	(not included)	0.90 (0.18)	0.92	Kenya

Notes: This table summarizes key features of the studies included in the meta-analysis. We follow the meta-analysis literature by referring to studies using the first author and year only, but report full references in the first column of this table. We were able to include Gateff 1972, Gupta 1982, Ostwald 1984, Ndibazza 2012, Wiria 2013, and Liu 2015 thanks to the generosity of the Campbell Collaboration deworming review team. Effect sizes are in kg. †The estimated effect for Donnen 1998 was taken from TMSDG, who obtained unadjusted estimates from trial authors. ‡The estimated effect size for

Table 2: Tests of the Hypothesis of Common Zero Effect, Adding Updates Individually

Study	Effect Size	P-value
1. TMSDG	**.061**	**.089***
2. TMSDG (using prior Cochrane classifications)	**.092**	**.009*****
3. Sur 2005	.092	.006***
4. Willett 1979	.077	.021**
5. Joseph 2015	.054	.066*
6. Awasthi 2001	.086	.006***
7. Ostwald 1984	.066	.069*
8. Gateff 1972	.082	.019**
9. Liu 2015	.06	.089*
10. Stephenson 1993	.092	.009***
11. Wiria 2013	.062	.084*
12. Ndibazza 2012	.054	.105
13. Gupta 1982a	.06	.09*
14. Gupta 1982b	.065	.063*
15. Hall 2006 (ANCOVA)	.073	.032**
16. Miguel 2004	.056	.121
17. Full Sample (rows 3-16)	**.111**	**<0.001*****

Notes: This table presents meta-analysis treatment effect estimates for the impact of multiple-dose mass deworming on weight using a fixed effects model, and the p-values associated with a test of the null hypothesis of a common zero effect across all studies included in the sample. Row (1) includes the TMSDG sample of studies (described in the notes of Figure **??**). Row (2) adds Stephenson 1993 (classified by TMSDG as a study of "infected children" rather than "all children living in an endemic area") to the TMSDG sample, following classification approaches used in prior Cochrane Reviews. Rows (3) through (16) make the sample changes described for the full sample (described in the notes of Figure **??**) one at a time (holding the remainder of the TMSDG sample constant), and then for the full sample altogether in Row (17). For brevity we refer to each study by the first author and year; see Table **??** for the full reference. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Estimated Weight Gain (kg) and Test Results Across Samples

	TMSDG	Full sample	Full with >20% prevalence	Full with >50% prevalence
Fixed effect estimate (s.e.)	0.061	0.111	0.142	0.157
	(0.036)	(0.022)	(0.030)	(0.031)
P-value: test for common zero effect	0.089*	<0.001***	<0.001***	<0.001***
Random effects estimate (s.e.)	0.078	0.134	0.148	0.182
	(0.098)	(0.052)	(0.056)	(0.057)
P-value: random effects estimate	0.426	0.011**	0.008***	0.001***
Number of studies	*11*	*22*	*16*	*14*

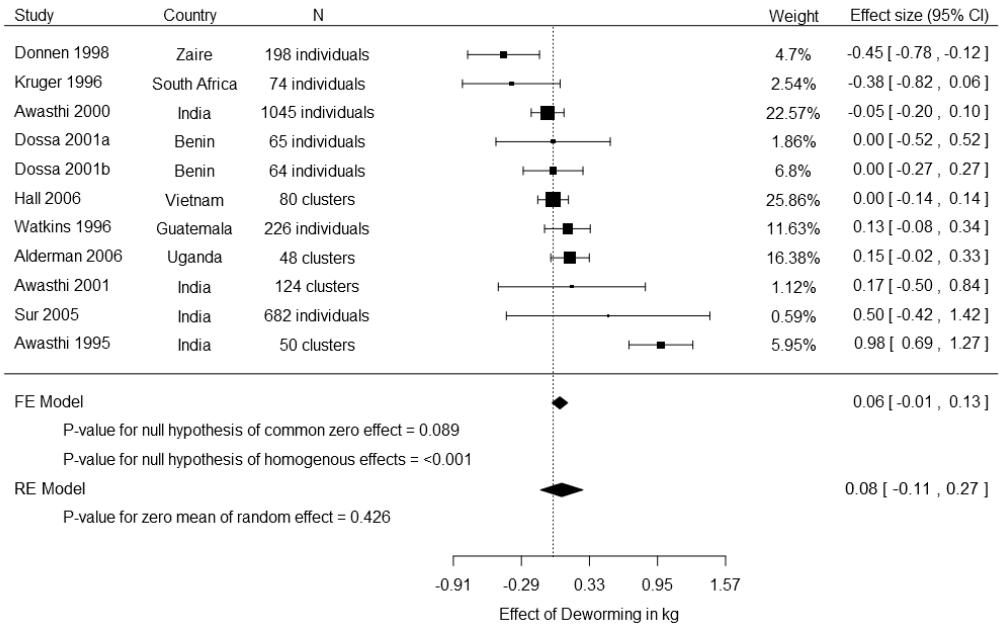
Notes: This table presents treatment effect estimates and key test results across the samples discussed in the main text. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Robustness of Random Effects Estimates to Dropping Individual Studies

Study	Effect Size	Standard Error	P-value
Full Sample	**0.134**	**(0.052)**	**0.011****
1. Kruger 1996	0.150	(0.052)	0.004***
2. Watkins 1996	0.134	(0.055)	0.016**
3. Donnen 1998	0.158	(0.051)	0.002***
4. Awasthi 2000	0.146	(0.055)	0.008***
5. Dossa 2001a	0.137	(0.054)	0.010***
6. Dossa 2001b	0.140	(0.054)	0.010***
7. Alderman 2006	0.132	(0.056)	0.018**
8. Awasthi 1995	0.095	(0.043)	0.028**
9. Awasthi 2001	0.131	(0.057)	0.021**
10. Hall 2006	0.139	(0.057)	0.016**
11. Sur 2005	0.123	(0.055)	0.024**
12. Willett 1979	0.132	(0.056)	0.018**
13. Joseph 2015	0.140	(0.058)	0.016**
14. Miguel 2004	0.144	(0.052)	0.006***
15. Ndibazza 2012	0.141	(0.055)	0.011**
16. Gupta 1982a	0.138	(0.054)	0.011**
17. Gupta 1982b	0.134	(0.054)	0.014**
18. Ostwald 1984	0.127	(0.053)	0.016**
19. Gateff 1972	0.122	(0.054)	0.023**
20. Wiria 2013	0.133	(0.053)	0.012**
21. Liu 2016	0.139	(0.054)	0.011**
22. Stephenson 1993	0.105	(0.049)	0.031**

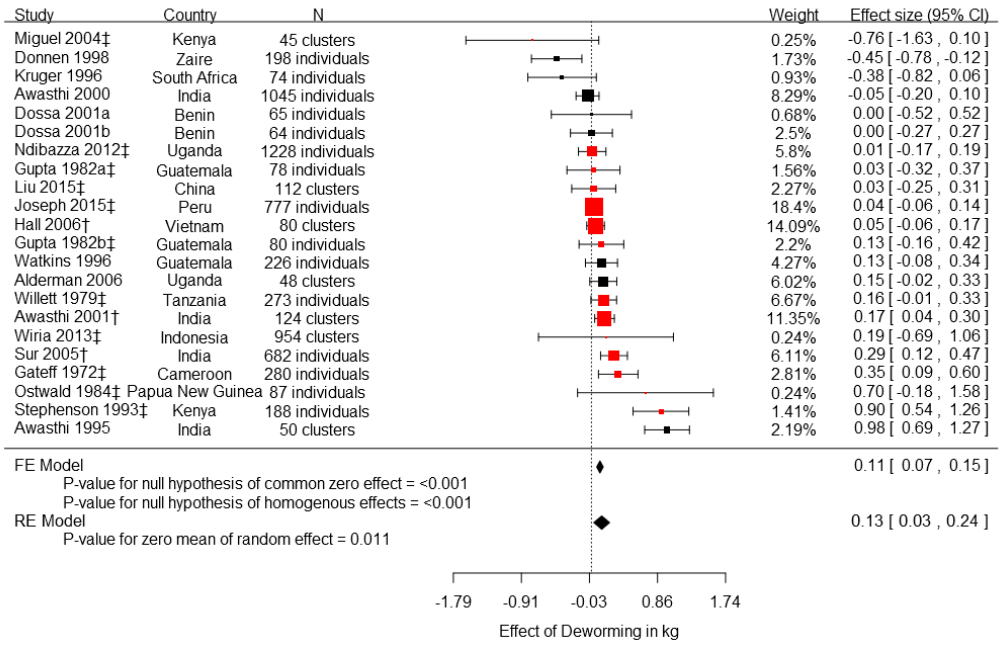
Notes: This table presents meta-analysis treatment effect estimates for the impact of multiple dose mass deworming on weight using a random effects model, dropping one study from the sample at a time (holding the remainder of the full sample constant). For brevity we refer to each study by the first author and year; see Table **??** for the full reference. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure 1: Effect of Mass Deworming on Child Weight, TMSDG Sample



Notes: This meta-analysis forest plot includes all studies in the "TMSDG sample", which is the set of studies included in **?** for the study of treating "all children living in an endemic area" with multiple doses of deworming medication at longest follow-up on the weight gain outcome, for a total of 11 effect estimates. For brevity, we refer to each study by the first author and the year; see Table **??** for the full reference. Squares denote the point estimate, and the whiskers show the 95% confidence interval. Effect sizes and standard errors are taken directly from **?**, not from the original articles. The point estimate squares are sized according to the weight each study is given in the fixed effect meta-analysis (calculated according to the precision of the study). The dotted vertical line represents zero effect. The lower panel displays the estimated effect across all studies using fixed and random effects models, the p-value associated with a test of the null hypothesis of a common zero effect across all studies, and the p-value of the random effects estimate.

Figure 2: Effect of Mass Deworming on Child Weight, Full Sample



Notes: This meta-analysis forest plot includes all studies in the "full sample", as described in the text, for a total of 22 effect estimates. For brevity, we refer to each study by the first author and the year; see Table **??** for the full reference. Squares denote the point estimate, and the whiskers show the 95% confidence interval. For studies that were included in **?**, effect sizes and standard errors are taken directly from that meta-analysis, except for the updates described to **?**, **?**, and **?**. The full sample additionally includes several studies that were not included in the TMSDG meta-analysis: **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, and **?**. The point estimate squares are sized according to the weight each study is given in the fixed effect meta-analysis (calculated according to the precision of the study). The dotted vertical line represents zero effect. † denotes a study with data updated since TMSDG. ‡ denotes a study not included in TMSDG. The lower panel displays the estimated effect across all studies using fixed and random effects models, the p-value associated with a test of the null hypothesis of a common zero effect across all studies, and the p-value of the random effects estimate.

Appendix A Details on data extraction for the full sample

This appendix describes the full sample of studies included in the primary analysis, and describes data extraction for these studies, following the general principles outlined in section **??**, which were used to generate the full sample of 22 treatment impact estimates from 20 different trials. Table **??** presents descriptive characteristics of the full sample.

?? lists the studies included from TMSDG without any updates. **??** discusses studies included in the updated analysis which were not mentioned in TMSDG, and which we presume the TMSDG authors were unaware of. **??** describes the process of incorporating studies that were mentioned in TMSDG but not included in their child weight gain analysis, for example, by using formulas in *The Cochrane Handbook* (**?**) to derive standard errors from other reported data. The subsequent sections detail adjustments to some estimates included in the TMSDG sample: **??** discusses cases in which more precise difference-in-difference estimates could be used instead of simply looking at endline differences, while **??** discusses cases in which ANCOVA estimation could be used. Section **??** describes the process for resolving conflicting information in **?**. **??** explains how studies were classified according to WHO recommendations for MDA based on helminth prevalence.

Appendix A.1 Estimates adopted from TMSDG

TMSDG include 11 treatment effect estimates from 10 different trials in their meta-analysis of the impact of multiple dose deworming of "all children living in an endemic area" on weight gain at longest follow-up. Eight of these treatment effect estimates are included without alteration in the updated sample: those from **?**, **?**, **?**, **?**, **?**, **?**, and **?**. Note that the clustered, unadjusted estimates from **?**, and the unadjusted estimates from **?**, were not contained in the published versions of the trials, but were obtained by Cochrane authors directly from the original trial authors. We use these same estimates in our sample.

Appendix A.2 Incorporating studies not mentioned in TMSDG

The full sample employed in this paper additionally incorporates four studies not mentioned in TMSDG.

? was likely not included in TMSDG simply because it was published in 2015, conceivably after the final literature review was conducted for the meta-analysis. The trial targeted children between ages 1 and 2 in rural Peruvian communities over the course of 1 year. The study presents a treatment effect and 95% confidence interval from the multiple dose treatment arm.[25] A formula provided in *The Cochrane Handbook* was used to compute the standard error (following Principle **??** in Section **??** of the main text).

?, was likely not included in TMSDG simply because it is a new, still unpublished study. This cluster randomized trial targeted school-aged children in China. The unadjusted difference-in-difference treatment impact estimate, 95% confidence interval, and associated p-value were obtained from the study authors, thanks to communication facilitated by the Campbell Collaboration. Again, we use a formula provided in *The Cochrane Handbook* to compute the standard error (following Principle **??** in Section **??** of the main text).

The Campbell Collaboration alerted us to the existence of **?** and **?**. **?** is a trial involving school-aged children in Papua New Guinea, and **?** is a study of school-aged children in rural Cameroon. It is unclear why neither of these studies are mentioned in TMSDG. Treatment effects and standard errors were calculated using information available in the published papers and formulas provided in **?** (following Principle **??** in Section **??** of the main text).

[25] Two treatment arms involved just a single dose of deworming and were not included.

Appendix A.3 Incorporating studies mentioned in TMSDG but omitted from weight gain meta-analysis

Six studies mentioned in TMSDG but not incorporated in their meta-analysis for the weight gain outcome are included in the sample.

1) **?** is acknowledged in TMSDG, but not included in their meta-analysis for weight gain, possibly (although it is not entirely clear) because the trial authors report only an adjusted treatment effect of mass deworming on weight gain or because the standard errors of the treatment effect are not directly reported in the text. Following TMSDG's preference for unadjusted treatment effect estimates, we contacted the trial authors in an attempt to obtain the microdata in order to extract unadjusted values, but after searching in his archives, Dr. Willett determined that the original data had been destroyed. We thus include what appears to be an adjusted treatment effect measure in our full sample (following Principle **??** in Section **??** of the main text). The covariates used are baseline weight, study induction date (there were two separate study intakes), and age at the time of induction. All three of these are likely to improve precision of the estimates. Although treatment impact standard errors are not directly reported in the study, information presented is used to calculate standard errors of treatment effects, following the procedure and formulas in **?** (following Principle **??** in Section **??** of the main text).

2) **?** report estimated impacts on weight-for-age z-score, but do not report estimates for the raw weight outcome. As a result, this study is not included in the TMSDG sample for meta-analysis on weight. However, the original trial data is publicly available, and we computed the estimated impact on weight using that data and an ANCOVA specification (following data extraction principles **??** and **??**).[26] Schools which received treatment for schistosomiasis (praziquantel) are dropped.

3) **?** was not included in the weight meta-analysis in TMSDG, likely because the study reports

[26] **?** corrects rounding, coding, and typographical errors in the original paper and presents updated data and results. We use these updated data and refer to updated results throughout, although we continue to reference **?** for simplicity.

only impacts on outcomes derived from weight (weight-for-age and weight-for-height), but does not present estimates for the raw weight outcomes. The data for this trial is not publicly available, but the Campbell Collaboration generously shared information on the raw weight impact from this study obtained through correspondence with the study authors, allowing inclusion of this trial in our full sample (following data extraction principle **??**).

4) **?** is classified in TMSDG as a single dose trial, but this appears to be erroneous based on our reading of the article. In their abstract, the study authors write "481 households (2022 subjects) and 473 households (1982 subjects) were assigned to receive placebo and albendazole, respectively, every three months." Furthermore, this trial does not report raw weight outcomes in the study text, although they were measured. The Campbell Collaboration authors had contacted the original authors and received from them baseline and endline measures of weight and standard deviations of those values for all study participants under age 16, and generously shared these estimates with us.[27] **?** does not report variance of changes, so a correlation coefficient is required to impute the standard error of the treatment effect. A correlation coefficient was estimated using a study with author-provided raw microdata of baseline and endline weight values (**?**). Using this estimated correlation coefficient of 0.89 yields a standard error of 0.4458 for **?**.[28] We thus incorporate this trial into our sample using data extraction principles **??** and **??**.

5) **?** was included in the 2012 Cochrane Review as a case of mass treatment and since we are examining mass treatment studies, we include this study. Prevalence at baseline was 92%, so while this is a high prevalence community, this was not a test and treat study, but an MDA study. Departing from the previous review, TMSDG classify this as a study of "infected children", and do not include it in their meta-analysis of the effect of "all children living in an endemic area". Note that in the 2015 update, the Cochrane Review changed its test-and-treat category, previously

[27] It is not entirely clear whether the values that were calculated account for clustering, but since the household clusters had so few children per cluster, additional clustering would not substantially affect standard errors.

[28] Another trial for which authors provided raw microdata, **?**, has the extremely similar baseline-endline correlation coefficient of 0.90.

called "screened for infection", to "children known to be infected." The result of this choice is that in the 2015 update, TMSDG no longer classify **?** and **?** as mass treatment programs.[29] The distinction used in the 2012 Cochrane Review, between "test and treat" and "mass treatment", corresponds more closely to the decision facing policymakers, and we preserve the original distinction. In doing so, one treatment effect and standard error from **?** is incorporated in the full sample, which measured the impact of multiple doses in an unscreened, but heavily infected, population of Kenyan schoolchildren.[30] We are able to calculate this treatment impact and standard error, following data extraction principle **??**.

6) **?** was excluded from the TMSDG analysis for reasons that are unclear (to us). TMSDG note in the "Characteristics of excluded studies" section that "[There are] only two units of allocation for relevant comparison. Children randomly divided into 4 groups, 'taking care that age distribution was similar in each group"'. The 4 groups were then allocated 1 of 4 different single treatment regimens; no details given." (p. 97). Following data extraction principle **??**, we calculate treatment effects and standard errors from the deworming versus placebo comparisons (n=78), and the deworming plus giardia treatment versus giardia treatment only comparisons (n=80) in the published paper.

Like TMSDG, the full sample excludes **?**. We received raw data generously from the study authors (via the Campbell Collaboration). However, the shared data only contained observations

[29]TMSDG state that "We changed the classification of **?** and **?**. Previously these trials were in the 'all children in an endemic area' category, whereas now they are classified in the 'children with infection.' This decision was based on reviewing the trials with parasitologists and examining the prevalence and intensity of the infection where clearly the whole community was heavily infected" (**?** p. 154). It is worth noting that although TMSDG exclude **?**, they include **?**; the highest recorded worm baseline prevalence in **?** by STH species is 92% (for ascaris); the highest prevalence in **?** is also 92% (for whipworm). Thus this reclassification does not appear to have been done systematically by worm prevalence. In our view, assessing the merits of the WHO policy by including studies in environments with prevalence below WHO thresholds while excluding MDA studies in areas with high prevalence may lead to risk of bias.

[30]**?** and the other treatment arm of **?** tested single dose deworming so are excluded from our analysis.

for children who had received the full set of intended doses of deworming medicine, rather than all who had been assigned to treatment, regardless of whether or not they received full treatment. Therefore a valid intention-to-treat analysis could not be conducted and estimates from this data were not included in the meta-analysis.

We also follow TMSDG in excluding **?** since the text indicates that the non-mortality outcomes such as weight were only measured for a subset of children from a randomly chosen cluster, but that *within* clusters, measured children were not chosen randomly.

Appendix A.4 Increasing precision using differences-in-differences estimates

? is included in the TMSDG sample using an endline-only comparison. The updated sample uses additional data from the article in order to calculate a difference-in-difference estimate, following data extraction principles **??** and **??**.

In particular, Web Plot Digitizer software (**?**) was used to extract difference-in-difference estimates for **?** from a figure. Extracting endline values and endline error bars from the graph nearly exactly reproduces in RevMan software the treatment effect of 0.5 and (abnormally large) standard error of 0.4717 reported in TMSDG (Web Plot Digitizer yields a treatment effect of 0.53 and a standard error of 0.46). Data from the figure and from p-values reported in the paper text was used to calculate the standard error of the baseline to endline change using the same software to extract data from the figure in which baseline and endline values are reported. This data was combined with the regressions of the treatment effect reported in the paper. The standard error of the change was calculated following the formulas and procedures in **?**, using information on the treatment effect, p-values, and degrees of freedom.[31] The change in weight from baseline to endline in **?** is 0.2925 (note that this is a smaller treatment effect than the 0.5 difference at endline used by

[31] As the change in weight over time is not reported in the text of the paper, the same method was used that we believe **?** used to estimate the endline difference in means, i.e. using data from Figure **??** in the article.

TMSDG). In the text of the article it explicitly states that the p-value of this change is 0.001.[32] The t statistic is calculated using the p-value and degrees of freedom. Once the t statistic is obtained, the standard error can be calculated using the following formula:

$$standard\ error = \frac{treatment\ effect}{t\ statistic}, \tag{3}$$

The *tinv* function in Excel was used to determine that, given a p-value of 0.001 and a sample size of 683 (and thus 681 degrees of freedom), the t statistic is 3.3048. This, in turn, using equation 1, implies a standard error of 0.0885. This result takes **?** from being an outlier with an extremely large standard error of 0.4717 (despite a relatively large sample of n=683) to having similar standard errors to the other study in the TMSDG sample with comparable sample size: **?**, with sample size of 1,045, has a standard error of 0.0760.[33] This revised standard error is included in the full sample.

Appendix A.5 Use of ANCOVA to account for baseline imbalance in outcome

Among all of the studies included in our meta-analysis, only one reports baseline imbalance in the weight outcome measure - **?**. Specifically, the control group is heavier, at 20.7 kg, compared to 20.5 kg in the treatment group – a difference which is statistically significant at $p = 0.01$ and thus quite unlikely to occur by chance.

As noted in section **??**, the *Cochrane Handbook* states clearly that when baseline data is available, the preferred analytical approach is to control for the baseline value of the outcome using an Analysis of Covariance (ANCOVA) specification, instead of the difference-in-difference specification used by the original authors and by TMSDG (data extraction principle **??**). A further

[32] See p. 261 and p. 265 of **?**.

[33] We contacted Dr. Sur to obtain the original micro data from the trial, in order to verify these calculations directly from the original microdata. Unfortunately, Dr. Sur is now retired and thus no longer has access to the micro data

advantage of this method is not only that ANCOVA is a more efficient estimator (**??**), but that it also reduces bias in cases of any baseline imbalance (**?**). We use microdata obtained directly from the **?** trial authors in order to estimate this ANCOVA specification (properly accounting for clustering), and obtain an effect size of 0.05 (SE 0.06).[34]

Appendix A.6 Resolving apparently conflicting reporting

In the text of **?**, the authors report conflicting treatment effect estimates, an issue that was also noted by TMSDG in their meta-analysis (**?**, p.43). In particular, the text of **?** states that deworming produced positive and significant effects on weight; the authors write that "Mean (+ SE) weight gain in Kg in control versus ABZ [i.e. treatment] areas was 3.04 (0.03) versus 3.22 (0.03), (p=0.01)" (p. 823). Later in the text, however, a similar treatment effect and level of statistical significance, but a different set of standard errors for the treatment effect, is reported: "The mean weight gain in 1.5 years in the albendazole plus vitamin A group was 5.57% greater than that in the vitamin A group alone (3.22 KG (SD: 2.03, SE: 0.26) vs. 3.05 KG (SD: 1.47 SE: 0.19) P-value=0.01)." (p. 825).

We follow data extraction principle **??** in consideration of this issue. In their meta-analysis, TMSDG use the reported treatment effect (0.17 kg), and appear to calculate the standard error using the second set of values (SE 0.26 and SE 0.19). Based on the p-values calculated from these numbers, and in contradiction to the p-value of 0.01 reported in the study, TMSDG refer to these results as not statistically significant, with a standard error of 0.341. By contrast the standard error

[34] A second issue with this trial relates to the imputation of clustered standard errors. In TMSDG, the treatment effect values (for a weight gain of 0.00) are included in the meta-analysis using the results reported in an unpublished manuscript obtained from the trial authors. TMSDG note that while some estimates were analyzed using methods to account for clustering, the main unadjusted results in the manuscript did not appear to use clustered standard errors, so they adjust the standard errors using an ICC that they obtain from **?**, which was a cluster randomized trial in Uganda. In this analysis the original trial data are used to calculate, rather than impute, the clustered standard errors.

is 0.0650 if one uses the p-value of 0.01 and treatment effect of 0.17 to back out a standard error, following, as in section **??**, the formulas and procedures in **?**, section 7.7.3.3. [35]

Three pieces of evidence were used to assess which estimate to use. First, we consulted directly with Dr. Awasthi about this issue. She expressed disagreement with TMSDG's interpretation of the results, and confirmed that she agreed with the interpretation of the study's results and calculation of the study's results and calculation of the study's standard errors using the p-values and effect sizes used here.[36] Second, the standard error for the weight outcome presented in the TMSDG analysis is 0.341, very large for the size of this large trial (124 clusters, and over 2,000 participants). In fact, this is 1.5 to 3 times larger than the weight outcome standard errors that TMSDG calculate for other trials in their original sample with only a fraction of the sample size.[37] By contrast, if the standard error is calculated using the p-value and treatment effect (SE=0.0650), this makes it comparable to the other large cluster RCTs.[38] Finally, we note that it is the (statistically significant) p-value that is reported consistently in the paper, rather than the standard error. Essentially, it is either the case that the authors entered incorrect measures of variance at one point in the paper, or one believes that the authors' interpretation of the full set of study results was incorrect. Given our correspondence with Dr. Awasthi, the evidence from the standard errors of comparable studies, and the fact that the p-value is reported consistently in the paper while the standard errors differ, the standard error derived from the p-value is incorporated into the full sample.

[35]There is yet a third possible way to calculate standard errors from data reported in this paper. This would be to use a set of standard errors reported in the abstract (0.03 for both treatment and control changes from baseline). These figures imply a still smaller standard error of 0.04.

[36]Personal communication, March 23, 2016. Dr. Awasthi also noted that the original micro data is no longer available.

[37]For instance, **?**, n=74, SE 0.2241), **?**, n=226, SE=0.1059), **?**, n=198, SE=0.1665), and the two treatment arms from **?** (n=65, SE 0.265 and n=64, SE=0.1385).

[38]For instance, **?** (40 clusters, SE 0.0599), **?** (50 clusters, SE 0.0892), **?** (50 clusters, SE 0.148)) and the large individually randomized trials (**?**, n=683, SE=0.0885, **?**, n=1,045, SE=0.076). We do note, however, that there are two large cluster RCTs in the full sample with comparably large standard errors: **?** (73 clusters, SE=0.44 and **?** (954 household clusters, SE=0.45).

Appendix A.7 Classification of studies by prevalence

Studies are classified according to WHO guidelines for MDA recommendations which are in turn based on whether helminth prevalence is greater than 20%, in which case MDA is recommended, and greater than 50%, in which case multiple dose MDA is recommended. Helminth prevalence in a study is classified based on the maximum prevalence across all worms reported in that study. Where possible, helminth prevalence level (see Table **??**) is classified based on prevalence described within the study itself, using cutoffs that are appropriate for WHO policy guidelines. One study in our sample is classified based on prevalence from an earlier study done in the area and which was used for targeting of the intervention, rather than baseline data collection within the trial itself (**?**). (Another study in our sample, **?**, does not report on prevalence at all, and is classified based on two other subsequent trials conducted in the same area of India – **?** and **?**). Finally, **?** is classified according to information from local health center statistics provided in the article, although the authors do not report baseline prevalence in their own sample.