Appendix for "Meta-Analysis and Public Policy: Reconciling the Evidence on Deworming"

March 4, 2022

For Online Publication

Appendix A: Details on data extraction for the full sample

This appendix describes the full sample of studies included in the primary analysis in this paper and the data extraction for these studies, following the general principles outlined in Section 3.2 of the main text. Section A.1 lists the studies included in our sample for which the estimates are the same as in Taylor-Robinson et al. (2015). Section A.2 discusses studies included in our sample which were not mentioned in Taylor-Robinson et al. (2015), and which in several cases the Taylor-Robinson et al. (2015) authors were unaware of at the time of that review. Section A.3 describes the process of incorporating studies that were mentioned in Taylor-Robinson et al. (2015) but not included in their analyses, for example, by using formulas in *The Cochrane Handbook* (Higgins and Green, 2011) to derive standard errors from other reported data. The subsequent sections detail adjustments to some estimates included in the Taylor-Robinson et al. (2015) sample: Section A.4 discusses cases in which more precise difference-in-differences estimators could be used instead of endline differences, while Section A.5 discusses cases in which ANCOVA estimation could be used. Section A.6 describes the process for resolving conflicting information in Awasthi and Pande (2001). Section A.7 describes the correction of a data extraction error by Taylor-Robinson et al. (2015) regarding the study by Awasthi et al. (2000). Throughout, we discuss whether our data differ from those in the updated review by Taylor-Robinson et al. (2019). Section A.8 explains how studies were classified according to WHO recommendations for MDA based on helminth prevalence. Section A.9 describes our update of Taylor-Robinson et al.'s (2015) search for trials.

A.1 Estimates adopted from Taylor-Robinson et al. (2015)

The meta-analyses by Taylor-Robinson et al. (2015) of the impact of multiple dose deworming of "all children living in an endemic area" on child weight, MUAC, height, and hemoglobin at longest follow-up include the following.

- Weight: 11 estimates from 10 studies (Alderman 2006[*]; Hall 2006; Sur 2005; Dossa 2001a[*],b[*]; Awasthi 2001; Awasthi 2000[*]; Donnen 1998[*]; Kruger 1996a[†]; Watkins 1996[*]; Awasthi 1995/2008[*])
- MUAC: 4 estimates from 3 studies (Dossa 2001a[*],b[*]; Donnen 1998[*]; Watkins 1996[*])
- Height: 8 estimates from 7 studies (Dossa 2001a[*],b[*]; Awasthi 2001; Awasthi 2000[*]; Donnen; 1998[*]; Kruger 1996a[†]; Watkins 1996[*]; Awasthi 1995/2008[*])
- Hemoglobin: 9 estimates from 7 studies (Ndibazza 2012; Kirwan 2010; Goto 2009; Le Huong 2007a[*],b[*]; Dossa 2001a[*],b[*]; Awasthi 2000; Kruger 1996a[*])

Our updated sample includes 7 estimates of the weight effect, 4 of the MUAC effect, 6 of the height effect, and 5 of the effect on hemoglobin without alteration, as well as their associated standard errors. These are marked with a star in the list above. Note that the clustered, unadjusted estimates from Alderman et al. (2006) of the weight effect, and the unadjusted estimates from Donnen et al. (1998) of the effects on weight, height, and MUAC, 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. The weight and height estimates and standard errors from Kruger (1996a), differ slightly from what we obtain based on the published data, likely due to rounding in Taylor-Robinson et al.'s (2015) data extraction procedures. These

estimates, which are marked with a dagger, have been corrected in the 2019 update, and we use the latter.

Taylor-Robinson et al. (2019) also included without alteration the estimates and standard errors marked with a star above, except for the standard error of the height effect from Dossa 2001b, which was revised from 0.64 to 1.099. Following the procedures from the Cochrane guidelines for data extraction, we obtain a standard error which corresponds to the one from the 2015 review and include this one in our meta-analysis.

A.2 Incorporating studies not mentioned in Taylor-Robinson et al. (2015)

The full sample employed in this paper additionally incorporates five studies not mentioned in Taylor-Robinson et al. (2015).

Joseph et al. (2015) was not included in Taylor-Robinson et al. (2015) because it was published in 2015, 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 treatment effect estimates on weight and height, and their corresponding p-values, from the multiple dose treatment arm.[1] The main body of the article presents estimates using imputed child weight and height data for children lost to follow-up, but ITT estimates without any imputation can be obtained from the supplementary material (Table S4, Joseph et al., 2015), which we use in our data extraction. A formula provided in *The Cochrane Handbook* was used to compute the standard error (following Principle i). Taylor Robinson et al. (2019) include this trial in their update but extract point estimates and standard errors of the weight and height effects based on data which included imputed observations (Table 4, Joseph et al., 2015). While

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

the point estimates and standard errors for the height effect differ depending on whether imputed data were used, these are the same up to two decimal points for the weight effect.

Similarly, Carmona-Fonseca and Correa-Botero (2015) was not included in Taylor-Robinson et al. (2015). Carmona-Fonseca and Correa-Botero (2015) is a factorial design of the provision of deworming treatment and vitamin A, targeting children under 15 years of age in a poor area of Colombia over the course of two years. The study collected baseline and endline data on child weight, height, and hemoglobin, among other outcomes. These data were generously shared with us by the study authors. We use these data to obtain ANCOVA estimates of the weight, height, and hemoglobin effects (following Principles iii and iv), based on the comparison of the deworming only and the placebo groups and the comparison of the deworming plus vitamin A and vitamin A only groups.[2] While Taylor-Robinson et al. (2019) note that they were aware of this trial while updating their review, they do not include this study in their update and note that they are "awaiting clarification on randomization" (p. 112). In their abstract, Carmona-Fonseca and Correa-Botero (2015) state "Methodology: Clinical, randomized, controlled trial with parallel groups intervened and evaluated each 3-4 months for 4 times, followed 12 months" (p. 10) and additional details are provided in the Materials and Methods section.

Liu et al. (2017) was not included in Taylor-Robinson et al. (2015) because it was unavailable at the time of the review. This cluster randomized trial targeted school-aged children in China. The unadjusted difference-in-differences estimates of the treatment effect on weight, height, and hemoglobin, as well as their associated 95% confidence intervals and p-values were

[2] The deworming intervention was randomized at the household level, with all children in the same household being allocated to the same treatment group. Because data identifying siblings are not available, we are unable to cluster errors to account for within household correlation in the ANCOVA model. Positive (negative) intra-cluster correlation (ICC) would lead to under-(over-) estimation of the standard errors. However, we note that time-invariant household characteristics, which are likely to induce non-zero correlation in nutritional outcomes across siblings, are partly controlled for in the ANCOVA model due to the inclusion of the baseline dependent variable.

obtained from the study authors, thanks to communication facilitated by the Campbell Collaboration. While p-values are reported with up to three decimal points, the bounds of the confidence intervals are presented only up to two decimal points. To calculate standard errors as precisely as possible, we use a formula provided in *The Cochrane Handbook* to compute standard errors (following Principle i) based on the reported p-values. Taylor-Robinson et al. (2019) also obtain data from this trial from the Campbell Collaboration. While the point estimates of the weight, height, and hemoglobin effects that they use are the same as the ones we use here, the standard errors of the weight and hemoglobin effects that they extract are larger than the ones we use. They may have calculated the standard errors from the 95% confidence intervals, the bounds of which were rounded to two decimal points. The 95% confidence intervals that they report for these outcomes do not coincide, up to two decimal points, with the ones we received from the Campbell Collaboration and which they presumably received as well.

Welch et al. (2016) included two studies of multiple-dose MDA which were not identified by Taylor-Robinson et al. (2015): Ostwald et al. (1984) and Gateff et al. (1972). Ostwald et al. (1984) is a trial involving school-aged children in Papua New Guinea and reports treatment effects on weight, height, and hemoglobin. Gateff et al. (1972) is a study of school-aged children in rural Cameroon and it includes weight as an outcome. Treatment effects and standard errors were calculated using information available in the published papers and formulas provided in Higgins and Green (2011) (following Principle i). Taylor-Robinson et al (2019) note that these two studies were missed in the literature search conducted for Taylor-Robinson et al. (2015). Taylor-Robinson et al. (2019) include these two studies in their update; the estimates and standard errors that they include in their meta-analysis are the same as the ones we use here, except for the standard error of the weight effect from Gateff et al. (1972). The paper reports the mean difference between

treatment and control and the "standard deviation of the distribution" (*l'écart type des distributions*) (Table 3, p. 107), which Taylor-Robinson et al (2019) extract as the standard error of the treatment effect. However, it is clear from other reported statistics that this is meant to be the standard deviation of the differences in outcome between each pair of children (who had the same baseline weight), not the standard error of the treatment effect estimator. Under the interpretation of Taylor-Robinson et al (2019), the implied t-statistic would be 0.25, which differs from the paper's reported t-statistic of 2.66 and the paper's clear interpretation (in several places in the text) that the treatment effect on weight was statistically significant. If the phrase "*écart type des distributions*" is interpreted as the standard deviation of the differences in outcome between each pair, then the implied t-statistic from the one-sample t-test is identical to the paper's reported t-statistic of 2.66. We extract this standard error by dividing the reported point estimate by this reported t-statistic.

A.3 Incorporating studies mentioned in Taylor-Robinson et al. (2015) but omitted from their meta-analyses of the effect of deworming MDA on child weight, MUAC, height, or hemoglobin

- Willett et al. (1979) is acknowledged in Taylor-Robinson et al. (2015) and Taylor-Robinson et al. (2019), but not included in their meta-analysis for weight gain, because the trial authors report only an adjusted treatment effect of mass deworming on weight gain and because the standard errors of the treatment effect are not directly reported in the text. Following the preference by Taylor-Robinson et al. (2015) 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 could no longer be located. We thus include what appears to be an adjusted treatment effect measure in our full sample (following Principle vi). 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 estimators. Although treatment impact standard errors are not directly reported in the study, the reported p-value is used to calculate standard errors of treatment effects, following the procedure and formulas in Higgins and Green (2011) (following Principle i). Taylor-Robinson et al. (2019) do not include this study in their update because "It is unclear from the primary data whether the effect estimate is adjusted. Furthermore, it is not possible to back calculate from the ANOVA p-value to obtain the standard error for the effect estimate. It is possible to obtain the F ratio statistic from the p value and degrees of freedom (which are known), but there are too many unknown values in the formula for the F ratio to back calculate any further." (pp. 151-152) As noted above, the controls (potentially) included in the analysis would not lead to bias. Furthermore, we disagree with the view that the standard error cannot be calculated from the p-value. Taylor-Robinson et al. (2019) acknowledge that an F ratio statistic can be calculated from the p-value which, we note, will be equivalent to the square of the t-statistic (for the two-tailed test of the hypothesis that a single coefficient is equal to zero), which can then be used to calculate the standard error using the reported point estimate (Stock and Watson, 2011). The Cochrane Handbook notes "Where exact P values are quoted alongside estimates of intervention effect, it is possible to estimate standard errors." (Higgins and Green, 2011 p. 2:25).

- Miguel and Kremer (2004) 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

Taylor-Robinson et al. (2015) 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 iii and iv).[3] Schools which received treatment for schistosomiasis (praziquantel) are dropped.[4] We shared the point estimate and standard error of the weight effect we estimated with the publicly available data from Miguel and Kremer (2004) with the authors of the Cochrane Review, and they incorporated these in their updated meta-analysis (Taylor-Robinson et al. 2019).

- Ndibazza et al. (2012) was not included in the weight or height meta-analysis by Taylor-Robinson et al. (2015), likely because the study reports only impacts on related outcomes (weight-for-age, weight-for-height, and height-for-age), but not on raw weight or height. The data for this trial is not publicly available, but the Campbell Collaboration generously shared information on the raw weight and height impacts from this study obtained through correspondence with the study authors, allowing inclusion of this trial in our full sample (following Principle iii). Taylor-Robinson et al. (2019) include estimates of the weight and height effect in their update, and these estimates and their standard errors correspond to the ones we use in this paper. Taylor-Robinson et al. (2019) also update the standard error of the hemoglobin estimate from the 2015 review, likely based on the information provided by the Campbell Collaboration; the point estimate and standard error of the hemoglobin

[3] Miguel and Kremer (2014) correct rounding, coding, and typographical errors in the original paper and present updated data and results. We use these updated data and refer to updated results throughout, although we continue to reference Miguel and Kremer (2004) for simplicity.

[4] Miguel and Kremer (2004) also collected endline height and hemoglobin data, which we do not use in our updated sample. This is because there are baseline imbalances in weight and thus plausibly other nutritional indicators. While the ANCOVA strategy can account for these baseline differences for weight, we lack the data to do the same for height or hemoglobin.

effect used by Taylor-Robinson et al. (2019) also correspond to the ones we use in this paper.

- Wiria et al. (2013) is classified in Taylor-Robinson et al. (2015) 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 reports body mass index (BMI) rather than raw weight or height outcomes in the study text. However, the Campbell Collaboration authors had contacted the original authors and received from them baseline and endline measures of weight and height, as well as standard deviations of those values for all study participants under age 16, and generously shared these estimates with us.[5] Wiria et al. (2013) does not report variance of changes, so a correlation coefficient is required to impute the standard error of the treatment effect. Correlation coefficients were estimated using a study with author-provided raw microdata of baseline and endline weight and height values (Hall et al., 2006). The estimated correlation coefficients are 0.89 for weight and 0.91 for height. Imputing these values, we estimate standard errors for Wiria et al. (2013) of 0.394 and 0.535, for weight and height, respectively.[6] We thus incorporate this trial into our sample using Principles iii and iv. Taylor-Robinson et al. (2019) do not include this trial in their update because "… there was a great deal of missing data. At the nine-month follow-up analysis, data were available for less than 16% of the 4004 individuals who were included in the trial (for both change score data and end values), and at the 21-month follow-up analysis,

[5] 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.
[6] Another trial for which authors provided raw microdata, Goto et al. (2009), has the extremely similar baseline-endline correlation coefficients of 0.90 for weight and 0.83 for height.

data were available for 13% of the 4004 individuals (end values data only)." (p. 18). However, this does not appear to be a correct interpretation of the study's attrition rate. While the total sample size of the trial is 4,004 participants, for multiple outcomes data was collected intentionally from only a subset of recipients. Table 2 of their paper shows that BMI was collected for only a subset of respondents (1770 out of 4004); the same table shows that 813 of these 1,770 individuals with BMI data were under 19. The Campbell Collaboration research team requested and received baseline and endline weight data for all children under 16, and they received data on 524 children under 16. Assuming equal distribution of respondents in each year from ages 0-19, the population under 16 is 84% of the population under 19, implying that the study had a sample of approximately 683 individuals under 16. The relevant attrition calculation therefore involves dividing the endline sample (524) by the number of children from whom data was collected from at baseline (683). By this calculation approximately 77% of individuals were retained to the end of the study, rather than the 13% reported by Taylor-Robinson et al (2019). This is also consistent with the article text, which reports that 80.1% of respondents were retained at endline. Thus under the reasonable interpretation that the reason for the smaller number of weight and height observations is that data on these outcomes (like several other key outcomes in the trial) was deliberately collected from only a subset of respondents, the attrition levels in this study are not notably different than others in the sample. Therefore we include the trial.

- Stephenson et al. (1993) 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 88%, so while this is a high prevalence community, this was

not a test-and-treat study, but an MDA study.[7] Departing from the previous review, Taylor-Robinson et al. (2015) classify this as a study of "infected children", and do not include it in their meta-analysis of the weight, MUAC, or height effect of "all children living in an endemic area". In the 2015 update, the Cochrane Review changed its "test and treat" category, previously called "screened for infection", to "children known to be infected." Thus in the 2015 update, Taylor-Robinson et al. (2015) no longer classify Stephenson et al. (1989) and Stephenson et al. (1993) as mass treatment programs.[8] 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, we incorporated in our sample additional treatment effect estimates on weight, MUAC, and height from Stephenson et al. (1993). These estimates measured the impact of multiple doses in an unscreened, but heavily infected, population of Kenyan schoolchildren.[9] We are able to calculate this treatment impact and standard error, following Principle i. Taylor-Robinson et al. (2019) again include this study in their updated review for the analysis of multiple-dose deworming. Their point estimates and standard errors correspond to the ones we use here.

[7] We calculate prevalence as the sample-size weighted average prevalence across all treatment and placebo arms.

[8] Taylor-Robinson et al. (2015) state that "We changed the classification of Stephenson et al. (1989) and Stephenson et al. (1993). 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" (Taylor-Robinson et al. (2015) p. 154). It is worth noting that although Taylor-Robinson et al. (2015) exclude Stephenson et al. (1993), they include Watkins et al. (1996); the highest recorded worm baseline prevalence in Watkins et al. (1996) by STH species is 91% (for ascaris); the highest prevalence in Stephenson et al. (1993) is 88% (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.

[9] Stephenson et al. (1989) and the other treatment arm of Stephenson et al. (1993) tested single dose deworming so are excluded from our analysis.

- Gupta and Urrutia (1982) was excluded from the meta-analysis of the weight and height effects by Taylor-Robinson et al. (2015). Regarding this study, Taylor-Robinson et al. (2015) state "[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). We note, however, that the two-step randomization (first, of children into groups and, then, of groups into experimental arms) implies that the unit of treatment allocation is the child (n=159). Following Principle i, 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=81) in the published paper. Taylor-Robinson et al. (2019) also include this study in their updated review for the analysis of multiple-dose deworming. Their point estimates and standard errors correspond to the ones we use here.
- Stoltzfus et al. (1997) was excluded from the meta-analysis of the effect of deworming MDA on weight and height in both the 2015 and 2019 Cochrane reviews; in Taylor-Robinson et al. (2019) they explain that they do not include this trial in their update because it "presented data from subgroups, selected on the basis of factors such as… frequency of treatment (Stoltzfus 1997 (Cluster))… [which] were not randomized and have not been included in meta-analysis." (p. 28) The study has two treatment arms, twice- and thrice-yearly deworming, and we combine data from these two treatment arms to calculate the estimate of twice- or thrice-yearly deworming, following the procedure and formulas in Higgins and Green (2011) and Principle vi. The authors present separate estimates for

children under and over 10 years of age. Likewise, we calculate estimates of twice- or thrice-yearly deworming on children under and over 10 years of age, separately.

- Hall et al. (2006) was not included in the meta-analysis of the effect of mass deworming on MUAC or height by Taylor-Robinson et al. (2015) or Taylor-Robinson et al. (2019) because Hall et al. (2006) had not adjusted standard errors to account for clustering and they were unable to impute an intra-cluster correlation (ICC) coefficient for these outcomes. Taylor-Robinson et al. (2015; 2019) do include this trial in their meta-analysis of the weight effect, imputing a weight ICC coefficient from Alderman et al. (2006). We use microdata, obtained from the study authors, thanks to communication facilitated by the Campbell Collaboration, to obtain ANCOVA estimates of deworming MDA on MUAC and height (following Principles iii and iv). Taylor-Robinson et al. (2019) do not include these MUAC and height estimates in their update.

  Furthermore, in a table in which Taylor-Robinson et al. (2019) respond to our 2016 version of this paper (p. 154) they make the following statement about Hall et al. (2006): "It is important that this study is published, and in the public domain. The unpublished manuscript does not provide data on baseline balance. We therefore are using the change values as provided in the manuscript, and decided against the post hoc adjustment re-analysis of Croke et al The weight gain in intervention and control group is ZERO in the unpublished manuscript; the post hoc adjusted value used in Croke et al is 139g, SE 57G, p value 0.016." This is incorrect: in the 2016 working paper version of this paper we report a treatment effect for Hall et al 2006 of 0.05 kg, SE 0.06 (p. 26). We agree with Taylor-Robinson et al. on the importance of publication of this study.

- Kruger et al. (1996) is a factorial design study of the effect of deworming MDA and of the provision of iron-fortified soup. While the meta-analyses of the effects on weight, height, and hemoglobin by Taylor-Robinson et al. (2015) incorporate the comparison of children given deworming and unfortified soup to those given placebo and unfortified soup, they exclude the comparison of children given deworming and fortified soup to those given placebo and fortified soup. We incorporate additional estimates of the weight, height, and hemoglobin effects using the latter comparison. Kruger et al. (1996) present results separately for children with low and adequate baseline iron stores. Following the procedure and formulas in Higgins and Green (2011), we combine data from children with low and adequate baseline iron stores to calculate the estimated treatment effect on child weight, height, and hemoglobin, for all children.[10] Taylor-Robinson et al. (2019) also include this comparison in their updated review for the analysis of multiple-dose deworming. Their point estimates and standard errors correspond to the ones we use here.
- Rousham and Mascie-Taylor (1994) was not included in the meta-analysis of the effect of mass deworming on MUAC by Taylor-Robinson et al. (2015) for reasons that are unclear to us, but may be because the single usable outcome (MUAC) is presented without a standard deviation. This is a cluster-randomized study (at the village level) of the effects of deworming treatment, every 2 months, targeting children aged 2 to 6 years. Taylor-Robinson et al. (2015) write "Rousham 1994 (Cluster) had 13 units of randomization (villages) containing 1476 children in total and had also not adjusted for clustering, but no outcomes from this trial were suitable for meta-analysis." (p. 16). Rousham and Mascie-Taylor (1994) present difference-in-differences estimates of the effects on weight-for-age,

[10] Taylor-Robinson et al. (2015) also followed this procedure for the one comparison they included in their meta-analyses.

height-for-age, and weight-for-height z-scores, as well as on MUAC, controlling for sex and age-squared. These controls are likely to increase the precision of the estimators. While the first three estimates are not suitable for inclusion in their meta-analysis of the weight or height effects, the MUAC estimate can be included in the meta-analysis of the MUAC effect. Furthermore, while TMCDG state that the analysis did not adjust for the clustered design of the trial, our interpretation is that the statistical analysis did account for clustering at the village level. Rousham and Mascie-Taylor (1994) write "Since the allocation of mebendazole or placebo tablets was randomized by village rather than individual, heterogeneity of nutritional status between villages was tested using a hierarchical (nested) analysis of variance." (p. 319) We extract the difference-in-differences estimate of the MUAC effect (longest follow-up), and calculate the standard error from the reported F-statistic (following Principles i and vi). Taylor-Robinson et al. (2019) do not include this study in their updated review.

Taylor-Robinson et al. (2015; 2019) exclude Goto et al. (2009) from their meta-analysis of the effect of deworming MDA on child weight and height, but include it in their meta-analysis of the effect on hemoglobin. We received raw data generously from the study authors (via the Campbell Collaboration). However, the shared data only contained observations 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-analyses. Taylor-Robinson (2019) include this study for the analysis of the effect of multiple dose deworming on hemoglobin.

We also follow Taylor-Robinson et al. (2015) and Taylor-Robinson et al. (2019) in excluding Awasthi et al. (2013) 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.

A.4 Increasing precision using difference-in-differences estimators

- Sur et al. (2005) is included in the meta-analysis of the weight effect by Taylor-Robinson et al. (2015; 2019) using an endline-only comparison. Our analysis uses additional data from the article in order to calculate a difference-in-differences estimate, following Principles ii and iv. In particular, Web Plot Digitizer software (Rohatgi, 2015) was used to extract difference-in-differences estimates for Sur et al. (2005) from the paper's Figure 1 (p. 265). Data from the figure and from p-values reported in the paper text were used to calculate the standard error of this estimator. The standard error of the change was calculated following the formulas and procedures in Higgins and Green (2011), using information on the treatment effect, p-values, and degrees of freedom.[11] The change in weight from baseline to endline in Sur et al. (2005) is 0.2925 (note that this is a smaller treatment effect than the 0.5 difference at endline used by Taylor-Robinson et al. (2015; 2019)). In the text of the article Sur et al. state that the p-value of this change is 0.001.[12] 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:

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

[11] As the change in weight over time is not reported in the text of the paper, the same method was used that we believe Taylor-Robinson et al. (2015) used to estimate the endline difference in means,
[12] See p. 261 and p. 265 of Sur et al. (2005).

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 the above formula, implies a standard error of 0.0885.[13] This revised standard error is included in our analysis. Taylor-Robinson et al. (2019) acknowledge that a difference-in-differences estimate can be obtained but argue that "…in order to calculate the standard error of this effect estimate, Croke et al. have assumed a t-test was used to generate the p-value for this difference in differences." The Cochrane Handbook notes "Where exact P values are quoted alongside estimates of intervention effect, it is possible to estimate standard errors." (Higgins and Green, 2011 p. 2:25). We follow the procedures outlined in the Cochrane handbook to extract the standard errors. We note that the assuming that a t-test was used is both consistent with the Cochrane guidelines and with Taylor-Robinson et al.'s (2019) thinking. In their discussion of Willet et al. (2019), Taylor-Robinson et al. (2019) acknowledge that an F-ratio statistic can be calculated from the p-value, and we note that this statistic is equivalent to the square of the t statistic (for the two-tailed test of the hypothesis that a single coefficient is equal to zero) (Stock and Watson, 2011).

- Kirwan et al. (2010) is included in the meta-analysis of the effect on hemoglobin by Taylor-Robinson et al. (2015) using an endline-only comparison. Our sample uses additional data from the article in order to calculate a difference-in-differences estimate, following Principles i and iv. Taylor-Robinson et al. (2019) keep the endline-only estimate in their updated review.

A.5 Increasing precision using ANCOVA

[13] 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.

We use microdata obtained directly from the Hall et al. (2006) trial authors in order to estimate the ANCOVA specification (properly accounting for clustering) for the weight outcome, following Principles iii and iv.[14] Taylor-Robinson et al. (2019) "…decided against the post hoc [ANCOVA] adjustment re-analysis of Croke et al" (p. 154). It is unclear to us whether the authors had access to the microdata which would have allowed estimation of an ANCOVA model; however if so, Cochrane guidelines favor the use of ANCOVA estimates over difference-in-differences estimates.

A.6 Resolving apparently conflicting reporting

In the text of Awasthi and Pande (2001), the authors report conflicting treatment effect estimates of the weight effect, an issue that was also noted by Taylor-Robinson et al. (2015, p.43). In particular, the text of Awasthi and Pande (2001) 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). On one hand, the latter set of standard errors (of the treatment and control means) is consistent with the reported standard deviations, assuming that the data were averaged and analyzed at the cluster level; Taylor-Robinson et al. (2019) state that this was the case, and use the latter set of standard errors. On the

[14] A second issue with this trial relates to the imputation of clustered standard errors. In Taylor-Robinson et al. (2015), 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. Taylor-Robinson et al. (2015) 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 Alderman et al. (2006), 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.

other hand, while the first set of standard errors is consistent with the p-value of 0.01, the set used by Taylor-Robinson et al. (2019) is not.[15]

We follow Principle v in consideration of this issue. Taylor-Robinson et al. (2015) use the reported treatment effect for weight (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, Taylor-Robinson et al. (2015) refer to these results as not statistically significant, with a standard error of 0.341. By contrast the standard error is 0.0650 if one uses the p-value of 0.01 and treatment effect of 0.17 to calculate a standard error, following the formulas and procedures in Higgins and Green (2011), section 7.7.3.3.[16]

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 the interpretation of the results by Taylor-Robinson et al. (2015), and confirmed that she agreed with the interpretation of the study's results and calculation of the study's standard errors using the p-values and effect sizes used here.[17] Second, the standard error for the weight outcome presented in the Taylor-Robinson et al. (2015) analysis is 0.341, which is 1.5 to 3 times larger than the weight outcome standard errors that Taylor-Robinson et al. (2015) calculate for other trials in their original sample

[15] While there is fundamental ambiguity in the text regarding the correct set of standard errors, we note that the one used by Taylor-Robinson et al. (2019) might be upward biased and that the one we use might come from an individual-level analysis, which would be downwardly biased. Averaging data at the cluster level, as may have been done for the first set of standard errors, will over-state standard errors in the absence of perfect intra-cluster correlation. By contrast, standard errors from analysis at the individual level without appropriate clustering will understate standard errors.

[16] 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.

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

with only a fraction of the sample size.[18] By contrast, if the standard error is calculated using the p-value and treatment effect (SE=0.0650), this makes it comparable to several other large cluster RCTs in the sample.[19] Finally, we note that it is the (statistically significant) p-value that is reported consistently in the paper, rather than the standard error. It is either the case that the authors entered incorrect measures of variance at one point in the paper, or else the authors' interpretation of the full set of study results was incorrect throughout the paper. 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 (following Principle i).

In their analysis of the height effect of mass deworming, Awasthi and Pande (2001) report a single set of standard errors for mean height in the treatment and control groups; however, these are not consistent with the reported p-value of the test of the hypothesis that the effect was zero. As with weight, we derive standard errors for the treatment effect estimator based on p-values. In their 2019 update, Taylor-Robinson et al. (2019) do not update the standard errors of the weight or height effects from this trial, reporting instead that they have communicated with a statistician who is familiar with the trial's analysis, and that the relationship between standard deviation and standard error in the paper's main analysis section is consistent with the trial's sample size, which they therefore believe is more accurate (p. 155).[20]

[18] For instance, Kruger et al. (1996) (n=74, SE=0.2241), Watkins et al. (1996) (n=226, SE=0.1059), Donnen et al. (1998) (n=198, SE=0.1665), and the two treatment arms from Dossa and Ategbo (2001) (n=65, SE=0.265 and n=64, SE=0.1385).

[19] For instance, Hall et al. (2006) (40 clusters, SE 0.0599), Alderman et al. (2006) (50 clusters, SE=0.0892), Awasthi et al. (2008) (50 clusters, SE=0.148) and several large individually randomized trials (Sur et al. (2005), n=683, SE=0.0885, Awasthi et al. (2000), 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: Miguel and Kremer (2004) (73 clusters, SE=0.44) and Wiria et al. (2013) (954 household clusters, SE=0.45).

[20] They write that (p. 155) "We know that this trial was analysed by cluster (Richard Peto, pers. Com). The paper abstract and main results provides differing estimates of variance for weight gain. The abstract gives a standard error

A.7 Resolving data extraction errors

The meta-analysis of the hemoglobin effect by Taylor-Robinson et al. (2015) includes the study by Awasthi et al. (2000), which reports the following standard deviations for hemoglobin (g/dl): 0.66g/dl (N=601) in the treatment group and 0.65g/dl (N=444) in the control group. The forest plots presented by Taylor-Robinson et al. (2015) shows that they extracted a s.d. of 0.65 for the treatment group and of 0.66 for the control group. They appear to have transposed these figures; we follow the original study's reporting of standard deviations.

A.8 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 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 (Alderman et al. (2006)). Another study in our sample, Awasthi et al. (2008), does not report on prevalence at all, and is classified based on two other subsequent trials conducted in the same area of India –

of 0.03 for weight gain in both groups, and the results gives a standard error of 0.26 in the intervention and 0.19 in the control. The data in the main results are analysed at the level of the cluster: using the relationship between SE and SD, we calculate n for the intervention as 61, and for the control 60. This corresponds (allowing for rounding errors) with the units randomised in the paper. We therefore used, for weight change, intervention 3.22 (SE 0.26) and control 3.05 (SE 0.19)."

Awasthi et al. (2000) and Awasthi and Pande (2001). Finally, Gateff et al. (1972) 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.

For studies that screened children for infection we calculate the population-wide worm prevalence as follows.

- Yap et al. (2014) report that 99.1% of children in their study area were infected with some worm species and that, among those infected, 94.3% were infected with whipworm (the most prevalent among all species). Therefore, we estimate prevalence at 0.991*0.943*100=93.4%.
- Sakar et al. (2002) report that roundworm prevalence in their study setting is 78.5%. The study does not report the prevalence of other worm species and we assume that roundworm prevalence is the highest among all species.
- Freij et al. (1979) report that, among children ages 1 to 4 years in their study setting, 48.9% were infected with roundworm. The study does not report the prevalence of other worm species and we assume that roundworm prevalence is the highest among all species.
- Tee et al. (2013) report that 15.3% of the children in their study population were infected with whipworm only (i.e., negative for other worm species). We contacted Dr. Lee to obtain data on worm prevalence by species (independent of co-infection), who told us that the data are no longer available. Dr. Lee also told us that he believes that whipworm is the most common infection in their setting (in Malaysia), and pointed us to the study by Huat et al. (2012), in a similar setting in Malaysia, reporting that 15% of children were infected with whipworm alone, 5% with roundworm alone, and 15% with both roundworm and

whipworm. We thus estimate whipworm infection prevalence in this setting as 0.153*(0.30/0.15)*100=30.6%, and assume it is the highest among all worm species.

For Awasthi (2008) we impute infection prevalence using the average among all trials in our sample where prevalence is under 20%. For Gateff et al (1972) and Alderman et al (2006), we impute infection prevalence using the average among all trials in our sample where prevalence is over 50%.

A.9 Updated search for trials

We updated the systematic search for trials by Taylor-Robinson et al. (2015) to identify studies published between April 14, 2015 (the Taylor-Robinson et al. search date) and June 29, 2018.

We searched for trials in the Cochrane Central Register of Controlled Trials (CENTRAL), MEDLINE, EMBASE, and LILACS, following the search procedure outlined in the Additional Table 1, "Detailed search strategies," in Taylor-Robinson et al. (2015, pp. 127-128). The search resulted in 345 titles, of which only Carmona-Fonseca and Correa-Botero (2015) and Namara et al. (2017) met the inclusion criteria. Neither of these trials reported estimates of the effect of mass deworming on the child nutrition outcomes examined in this paper. Through correspondence with the respective trial authors, we received confirmation that both trials collected data on some of these outcomes. By the closure of our study update (August 10, 2018), we had received microdata from the trial by Carmona-Fonseca and Correa-Botero (2015) but not from Namara et al. (2017). The latter study followed up children, from a trial (Ndibazza et al., 2012) which was discussed by Taylor-Robinson et al (2015), for a longer period of time. While we do not include the data from the longer follow-up by Namara et al. (2017), we do include the data from Ndibazza et al. (2012). Further, on July 12, 2018, we looked up each study that was identified through the above methods

and that met the criteria for inclusion in our study on Google Scholar. For each of these studies, we identified all the articles that cited the respective study, and that were identified by Google as meeting the following additional search criteria "albendazole OR mebendazole OR piperazine OR levamisole OR pyrantel OR tiabendazole OR thiabendazole." We added all these titles to a Google Library and listed all articles published in 2015 or later, that met the additional search criteria "weight OR height OR MUAC OR "arm circumference" OR hemoglobin OR haemoglobin." We identified 143 titles, of which only Bhattacharyya et al. (2018) met our inclusion criteria. This trial reports effects on child weight, but is not included in our analysis as there are inconsistencies with the statistics that are reported (e.g., the confidence interval for the weight gain in the treatment group does not include the point estimate). We have contacted the trial authors for clarification but have not received a response (August 10, 2018).

Table A.1: Differences in the study inclusion and data extraction between Taylor-Robinson et al. (2019) and this paper

Paper	TMDCG inclusion or exclusion rationale	Our inclusion or exclusion rationale	TMCDG data extraction procedure	Our data extraction procedure
Carmona-Fonseca and Correa-Botero (2015)	**Exclude** Listed as "awaiting assessment"; "awaiting clarification on randomization"	**Include** Details for randomization are provided in the Methods section	Study not included	Obtain ANCOVA estimates using author-shared data following Principles iii and iv
Goto et al. (2009)	**Include**	**Exclude** Study text states that analysis was based on whether children received treatment or not, not on whether they were randomized into treatment or control (i.e. not intention-to-treat)	Calculate treatment effect based on comparison of endline values	Study not included
Rousham and Mascie-Taylor (1994)	**Exclude** States that analysis did not account for cluster randomization; also states that no relevant outcomes were reported.	**Include** Study text reports use of hierarchical analysis to account for clustered design; MUAC outcomes are reported.	Study not included	Calculate the standard error from the reported F-statistic and difference-in-differences estimate following Principles i and vi
Stoltzfus et al. (1997)	**Exclude** Treatment effects presented by subgroup rather than full sample	**Include** Treatment effect for full sample can be calculated from subgroup treatment effects	Study not included	Calculate standard errors from the reported p-value following Higgins and Green (2011) and Principle i
Willett et al. (1979)	**Exclude** Standard errors not presented and cannot be calculated from p-value and treatment effect; unclear whether treatment effect was adjusted for covariates	**Include** Adjusted effect can be calculated from treatment effect and p-value; include adjusted effect following Principles i and vi	Study not included	Include adjusted treatment effect following Principle vi; calculate standard error from the reported p-value following Higgins and Green (2011) and Principle i
Wiria et al. (2013)	**Exclude** Exclude based on large number of missing values, calculating attrition by comparing final weight/height	**Include** Include because attrition should be calculated based on weight/height subsample; not on full study sample	Study not included	Estimate standard error using baseline-endline correlation coefficient from comparable studies using author-provided data following Principles iii and iv

	subsample to total study population			
Sur et al. (2005)	Included by both	Included by both	Calculate treatment effect based on comparison of end line values	Extract difference-in-differences treatment effect from graph using WebPlotDigitzer following Principle ii; calculate standard errors from reported p-values following Principle iv
Gateff et al. (1972)	Included by both	Included by both	Calculate treatment effect based on difference-in-differences; use reported standard deviation to calculate standard error.	Calculate treatment effect based on difference-in-differences following Principle iv; use study text reports of statistical significance to calculate standard error consistent with reported t statistic and p-value, following Principle v.
Kirwan et al. (2010)	Included by both	Included by both	Calculate treatment effect based on comparison of endline values	Calculate treatment effect based on difference-in-differences, following Principle iv.
Awasthi and Pande (2001)	Included by both	Included by both	Extract measures of variance from paper's results section.	Extract measures of variance from paper abstract; interpretation based on correspondence with author, following Principle v.
Liu et al. (2017)	Included by both	Included by both	Calculate treatment effect based on difference-in-differences; use 95% confidence intervals to calculate standard errors	Calculate treatment effect based on difference-in-differences; use treatment effects and p-values to calculate standard errors, following Principle i.
Hall et al (2006)	Included by both	Included by both	Calculate treatment effect for weight gain based on difference-in-differences as reported in manuscript; impute ICC using value from Alderman et al. (2006)	Calculate treatment effects for weight, height, and MUAC gain with ANCOVA specification using microdata provided by authors, following Principle iii.
Dossa et al. (2001)	Included by both	Included by both	Calculate treatment effect based on difference-in-differences	Calculate treatment effect based on difference-in-differences (calculated standard error of height differs between TMCDG and this study)
Joseph et al (2015)	Included by both	Included by both	Calculate treatment effect based on difference-in-	Calculate treatment effect based on difference-in-

			differences using sample with imputed values for missing values	differences using sample without imputation

Table A.2: Statistical power to detect effects that render deworming MDA cost-effective relative to alternative policies

	Taylor-Robinson et al. (2015)				Taylor-Robinson et al. (2019)						Welch et al. (2016)	
	Weight (kg)	MUAC (cm)	Height (cm)	Hb (g/dl)	Weight (kg)	Weight (kg) *Before 2000*	Weight (kg) *Since 2000*	MUAC (cm)	Height (cm)	Hb (g/dl)	Weight (s.d)	Height (s.d)
	1	2	3	4	5	6	7	8	9	10	11	12
Panel A: Estimates of the child nutrition effects of deworming MDA												
Point estimate	0.078	-0.034	0.015	-0.017	0.114	0.258	0.022	0.068	0.023	0.012	0.049	0.031
s.e.	0.098	0.109	0.081	0.030	0.065	0.164	0.029	0.130	0.057	0.031	0.034	0.027
p-val	0.426	0.757	0.852	0.572	0.079	0.116	0.449	0.603	0.680	0.698	0.150	0.255
Average Doses	3	2.5	2.9	4.4	3.8	3.3	4.3	2.4	4.1	3.8	2.2	2
Panel B: Minimum detectable effect (in absolute units) to reject the hypothesis that the effect of deworming MDA is zero at the 95% confidence level, with 80% power												
MDE	0.276	0.306	0.227	0.085	0.181	0.461	0.083	0.365	0.159	0.088	~0.294 kg (0.095 s.d.)	~0.195 cm (0.075 s.d.)
Panel C: Minimum average effect (in absolute units) that renders deworming MDA cost effective relative to												
School feeding	0.006	0.003	0.006	--	0.008	0.007	0.009	0.003	0.008	--	0.005	0.004
Preschool feeding	0.005	NA	0.011	0.002	0.006	0.006	0.007	NA	0.015	0.002	0.004	0.008

Appendix B: Cost effectiveness calculations

We use the estimate from Kristjansson et al. (2016) of the average per capita cost of providing school-based food supplementation for a school year of 200 days (10 months) and for a fixed daily ration of 401 kcal (Kristjansson et al., 2016), specifically, the average caloric content in the school feeding programs included in an unpublished update of the Kristjansson et al. (2007) review; this average cost is $41. Following Kristjansson et al. (2016), and assuming that the cost per calorie delivered in school and preschool feeding programs are the same, we estimate the per capita cost of a one calendar year provision of a daily ration of 391 kcal, the average in the preschool feeding programs reviewed by Kristjansson et al. (2015) at $41*1.2*(397/401) = $48.70.

Appendix C: Testing if the mean differences, between MDA and test-and-treat trials, of the impact of deworming of infected children on weight, MUAC, height, and Hb are jointly zero

We also conduct a Wald test of the hypothesis that the mean differences, between MDA and test-and-treat trials, of the impact of deworming of infected children on weight, MUAC, height, and Hb are jointly zero. Following the methods of subsection 4.1, and assuming a common correlation coefficient for the estimators of these differences, we are unable to reject the null at the 95% confidence level for correlation coefficients between -0.20 and 0.35. For values of the correlation coefficient between -0.33 and -0.21 and between 0.36 and 0.99, we reject the null at the 95% confidence level. These results are based on conducting Wald tests for every value of the correlation coefficient between -1 and 1 in 0.01 increments.

Appendix D: Discussion of Welch et al. (2016)

Welch et al. (2016) identify many of the missing studies mentioned above, but are underpowered primarily because they subdivide deworming studies based on the type of drugs used, the frequency of treatment, and whether the trial compared deworming to pure placebo versus trials in which deworming plus an additional intervention is compared to the additional intervention alone. Thus instead of reporting a single meta-analysis which aggregates a large number of studies, they conduct multiple small sample meta-analyses.[21] When one relaxes the narrow category of analysis from the main comparison in Welch et al. (2016), for example to include trials where approved drugs aside from albendazole were used, or where deworming was done more or less frequently than twice per year, one obtains statistically significant estimates of the effect of deworming on weight, and in some cases for height (Table D.1, columns 2-4).

[21] For the weight effect, Welch et al. (2016, p. 145) also presented "…an analysis of any mass deworming treatment (any drug, any frequency) compared to placebo, and found an effect size of 0.03 SMD for weight (95% CI: 0.00 to 0.07, 30 trials, I2=46%) with a total of 59, 691 participants. This corresponds to a difference of 0.05 kg." This analysis is not emphasized by the authors (e.g., not presented in the "Summary of findings table"). We also note that Welch et al. (2016) do not exclude studies from settings with <20% infection prevalence or that administered a single dose of deworming treatment. As discussed above, treatment effect estimates from low prevalence settings are expected to be small as there are few infected children to begin with. In addition, estimates are also expected to be small in single dose trials as the length of follow-up is typically shorter than in multiple-dose trials and nutritional gains take time to occur. While we are uncertain about which trials were included in the broad-category weight analysis, in other analyses, Welch et al. (2016) include Jinabhai et al. (2001) and Nga et al. (2009), both single dose trials with a length of follow-up of 4 months.

Table D.1: Relaxing the main category of analysis in Welch et al. (2016)

	Main Comparison	Including additional studies				
	Alb. (std.)	Other STH drugs (std.)	Alb. (other freq.)	Vitamin A; Hygiene educ.	Koroma 1996; Stephenson 1989	Other STH drugs (all freq.); Alb. (other freq.);
	1	2	3	4	5	6
Panel A: Weight or WAZ						
Point estimate	0.049	0.061	0.046	0.044	0.250	0.123
s.e.	0.034	0.027	0.027	0.021	0.072	0.035
p-val	0.150	0.026	0.088	0.036	0.001	0.000
N	11	14	15	14	13	28
Panel B: Height or HAZ						
Point estimate	0.031	0.026	0.044	0.054	0.180	0.094
s.e.	0.027	0.021	0.023	0.030	0.079	0.036
p-val	0.255	0.215	0.055	0.069	0.022	0.009
N	9	10	13	10	11	23

Notes: Estimation method is random-effects meta-analysis of standardized mean differences. Column 1 presents estimates of the main comparison of Welch et al. (2017): albendazole twice per year (i.e., standard frequency) versus placebo. We expand the sample by including: trials with other deworming drugs at standard frequency (column 2); albendazole trials at other frequencies, when estimates for the standard frequency are not available (column 3); trials with vitamin A as a co-intervention, or where the treatment included hygiene education (column 4); trials excluded by Welch et al. (2016) because they increased cross-trial heterogeneity (column 5). Column 6 expands the main comparison by considering all deworming drugs at all frequencies, as well as the additional studies from columns 4 and 5.

A second reason for limited power is that Welch et al. (2016) exclude two studies from their main comparison (Koroma et al., 1996; Stephenson et al., 1989) because they led to high heterogeneity in their meta-analyses. The authors noted that heterogeneity measured by the I^2 statistic was reduced from 93% to 61% after removing both of these studies. They scrutinized these trials and identified baseline imbalance and unclear allocation concealment, and thus decided to exclude them from their meta-analyses (p.73). Stephenson et al. (1989) was balanced on roundworm and whipworm prevalence and baseline height and weight, but the treatment group had higher hookworm prevalence and intensity, and the placebo group had higher ascaris intensity. The treatment group in Koroma et al. (1996) had lower mean weight-for-age and height-for-age z

scores, relative to the control group. However, because both studies report difference-in-differences estimates, baseline imbalances should not lead to bias unless they are correlated with changes in child weight and height. Note that the estimated weight and height effects in Koroma et al. (1996) and Stephenson et al. (1989) are both positive and the largest among those considered in the main analyses by Welch et al. (2016).

Excluding studies because they increase measures of cross-study treatment heterogeneity will mechanically lead to downward biased estimates of cross-study variance. Further, when the distribution of true effects is right-skewed, excluding estimates at the right tail will lead to downward bias in estimation of average effects. In subsection 4.2 of the article, we discuss the reasons why one would naturally expect the distribution of effects of MDA on child nutrition to be right-skewed. When one extends Welch et al.'s (2016) main comparison to include studies excluded because they increased cross-study variance, one rejects the hypothesis of zero average effect of MDA on weight (column 5, panel A) at the 99% confidence level and on height (column 5, panel B) at the 95% level. Finally, when one relaxes the category of the main analysis to include: (i) deworming trials with any drug, at any frequency, (ii) trials with vitamin A as a co-intervention, or where hygiene education was part of the treatment, and (iii) trials excluded by Welch et al. (2016) because they increased cross-study heterogeneity, one rejects the null hypothesis of a zero average effect on weight and height at the 99% confidence level (column 6).

Appendix E: Robustness of random effects estimates to dropping individual study estimates and pairs of estimates

The finding of a significant effect of deworming MDA on child weight, MUAC, and height appear robust in settings with over 20% worm infection prevalence (see Table 4). One can reject the null hypothesis that the mean effect of deworming MDA on child weight is zero at the 95% confidence level after dropping any one study estimate and after dropping any pair of estimates, among 210 possible combinations. In these same settings, after dropping one estimate at a time, one can reject the null hypothesis of a zero mean effect on MUAC 67% of the time; the highest p-value is 0.056. Dropping one pair of estimates at a time, one can also reject the null for 8 pairs out of 15 possible combinations. In addition, after dropping one study estimate at a time, one can reject the null hypothesis of no average effect of MDA on height 94% of the time, and when dropping pairs of estimates, one can reject the null 79% of the time.

Table E.1: Robustness

	Weight	MUAC	Height
Panel A: Dropping one study estimate at a time			
Number of times RE estimate remains significant at 95% confidence level	21/21 (100%)	4/6 (67%)	15/16 (94%)
Highest p-val	0.002	0.056	0.214
Panel B: Dropping two study estimates at a time			
Number of times RE estimate remains significant at 95% confidence level	210/210 (100%)	8/15 (53%)	96/120 (80%)
Highest p-val	0.005	0.198	0.430

Notes: Robustness of the finding of a significant average effect of deworming MDA across settings (μ) on child weight, MUAC, and height, based on the random-effects estimator, to dropping any one or two studies. Sample is comprised of settings with over 20% worm infection prevalence.

Appendix F: Additional figures and tables

Table F.1: Test of qualitative interaction

Weight (W)	MUAC (M)	Height (H)	Hemoglobin (Hb)
1	2	3	4
Panel A: Full Sample of MDA trials			
1. H0: All treatment effects are non-negative (p-value)			
0.361	0.230	0.711	0.967
2. H0: All treatment effects are non-positive (p-value)			
<0.001	<0.001	0.057	0.132
Panel B: Pooling all MDA and test-and-treat trials			
1. H0: All treatment effects are non-negative (p-value)			
0.493	0.381	0.823	0.930
2. H0: All treatment effects are non-positive (p-value)			
<0.001	<0.001	<0.001	0.160

Notes: The test statistics used for testing these hypotheses are described in subsection 4.1.

Table F.2: Cost-effectiveness analysis

	Deworming MDA (full sample)		
	Average effect [average no. doses]	Average effect per 2 doses = 2*((1) / av. no. doses)	Gain per $1,000 spent = (2)*(1,000 / cost of 2 treatments)†
	1	2	3
Weight (kg)	0.140 [3.81]	0.074	108.3 [47.8, 193.9]
MUAC (cm)	0.127 [3.57]	0.071	104.7 [46.2, 187.3]
Height (cm)	0.064 [4.00]	0.032	47.3 [20.9, 84.7]
Hb (g/dl)	0.028 [3.92]	0.014	21.1 [9.3, 37.7]

Notes: Estimates of the average child nutrition effects of deworming MDA correspond to our random effects estimates. † We assume a per capita cost of $0.34 for one deworming treatment. This is the current cost estimate for India (GiveWell, 2017), and it incorporates an estimate of the opportunity cost of the time that teachers spend in deworming programs, based on their wages. In square brackets we show a lower and upper bound of the outcome gain per $1,000 spent, using the higher cost per treatment of $0.77 that GiveWell (2017) estimates for African countries (also inclusive of the time of teachers) and the lower cost per treatment of $0.19 in India, if one values the opportunity cost of the time of teachers at one quarter of their wage, respectively. (Muralidharan and Sundararaman (2015) show that public school teachers in India obtain large rents, with average wages being over four times as large as those of private sector teachers.)

Table F.3: Pessimism of Priors

	MDA (>20% prevalence)				School feeding			Preschool feeding		
	Gain per $1,000 spent†	Standard error of RE estimate	Informational value of average study	Posterior precision with improper prior = 1/(2)^2	Gain per $1,000 spent	Minimum pessimism (MDA) that leads to indifference = ((1)/(4) - 1)*(3)	Minimum relative pessimism (MDA) that leads to indifference‡ = ((1)/(4) - 1)	Gain per $1,000 spent	Minimum pessimism (MDA) that leads to indifference = ((1)/(7) - 1)*(3)	Minimum relative pessimism (MDA) that leads to indifference‡ = ((1)/(7) - 1)
	1	2		3	4	5	6	7	8	9
Weight (kg)	144.6	0.043		529.6	6.2	11786.4	22.3	4.9	15010.6	28.3
MUAC (cm)	166.5	0.086		134.0	3.3	6653.1	49.6	NA	NA	NA
Height (cm)	80.0	0.038		676.1	6.1	8251.6	12.2	11.1	4202.2	6.2
Hb (g/dl)	82.6	0.038		675.2	--	--	--	1.4	38109.8	56.4

Notes: Estimates of the average child nutrition effects of deworming MDA correspond to our random effects estimates. † We assume a per capita cost of $0.34 for one deworming treatment. This is the current cost estimate for India (GiveWell, 2017), and it incorporates an estimate of the opportunity cost of the time that teachers spend in deworming programs, based on their wages. Estimates of the child nutrition effects of school feeding programs in low and middle income countries come from the systematic review by Kristjansson et al. (2007). Estimates for weight and height correspond to random effect estimates. Estimates for MUAC and hemoglobin come from a single study in Kenya (Neumann, 2003). Estimates of the child nutrition effects of preschool-feeding programs in low and middle income countries come from the systematic review by Kristjansson et al. (2015a). Estimates for weight, height, and hemoglobin correspond to random effect estimates, no estimate of the effect on MUAC is provided in the review. $41 is the per capita cost estimate of the daily provision of a ration of 401kcal for a 200-day school year, and $48.7 is the per capita cost estimate of the daily provision of a ration of 397kcal for a calendar year (Kristjansson et al., 2016). ‡ Measures the minimum level of pessimism about the mean effect of MDA that would lead to indifference with feeding programs, relative to the posterior precision of the mean MDA effect that obtains with an improper prior.

Table F.4: Robustness to exclusion of additional studies

	Weight	MUAC	Height	Hb
	1	2	3	4
Panel A: MDA trials with ≥20% prevalence				
Point estimate	0.154	0.198	0.087	0.069
Se.	0.044	0.086	0.039	0.038
p-val	<0.001	0.022	0.024	0.073
N	21	6	16	11
Panel B: Robustness Analysis				
B1. To dropping studies not mentioned by Taylor-Robinson et al. (2015)				
Point estimate	0.146	0.198	0.095	0.097
Se.	0.053	0.086	0.041	0.057
p-val	0.005	0.022	0.021	0.088
N	16	6	12	7
B2. To dropping studies erroneously not classified as MDA by Taylor-Robinson et al. (2015)				
Point estimate	0.131	0.108	0.091	0.069
Se.	0.035	0.056	0.040	0.038
p-val	<0.001	0.056	0.022	0.073
N	19	5	14	11
B3. To dropping studies not included in Taylor-Robinson et al. (2015) likely because standard errors are not directly reported or because only adjusted estimates are available				
Point estimate	0.149	0.198	0.066	0.069
Se.	0.053	0.086	0.048	0.038
p-val	0.005	0.022	0.166	0.073
N	18	6	14	11
B4. To dropping studies not included in Taylor-Robinson et al. (2015) likely because only estimates on "related" measures were presented (e.g. weight-for-age Z score instead of weight)				
Point estimate	0.168	0.198	0.087	0.069
Se.	0.041	0.086	0.039	0.038
p-val	<0.001	0.022	0.024	0.073
N	20	6	16	11

Notes: Panel B.1 excludes Joseph et al. (2015), Carmona-Fonseca and Correa-Botero (2015a,b), Liu et al (2017), Ostwald et al. (1984) and Gateff et al. (1972). Panel B.2 excludes Wiria et al. (2013) and Stephenson et al. (1993). Panel B.3 excludes Willett et al. (1979) and Stoltzfus et al. (1997a,b). Panel B.4 excludes Miguel and Kremer (2004) and Ndibazza et al. (2012).

Appendix G: Bayesian analysis of the mean effect of deworming MDA with skewed distributions

In this section, we turn to a Bayesian approach aimed at estimating the mean effect of deworming MDA across settings, while allowing for skewness in the distribution of true effects. In particular, we consider the half-normal and the and skew-normal distributions. While the former imposes right-skewness, the latter generalizes the normal distribution and allows for both a right and left skewness (it also allows for zero skewness).[22]

For the model with a half-normal distribution, we use the following prior for the scale parameter: $\tau_k \sim U(0,10)$.[23] For the model with a skew-normal distribution we use a prior of $U(-5,5)$ for the location parameter, of $U(0,10)$ for the scale parameter, and of $N(0, 10^2)$ for the shape or skewness parameter.[24]

Appendix Table G.1, panel A, presents posterior mean and standard deviation for the mean effect of deworming MDA. Both the half-normal and the skew-normal distribution produce posterior means that are similar to or slightly larger than our random-effects estimates. For instance, the posterior means for the effect size on weight gain range from 0.171 kg (skew-normal distribution) to 0.214 kg (half-normal distribution).

[22] We use the STAN platform to estimate posterior distributions using Markov Chain Monte Carlo (MCMC) methods. All results use 12 chains with 10000 iterations for each chain (5000 used for warmup).
[23] Results are similar when we instead use a prior of $\tau_k \sim U(0,5)$.
[24] Results are similar when we instead use $U(-10,10)$ as the prior for the location parameter.

Table G.1: Mean effect of deworming MDA with skewed distributions

	Weight (kg)	MUAC (cm)	Height (cm)	Hb (g/dl)
Panel A: Posterior mean effect across settings (s.d.)				
Half-normal	0.215	0.230	0.069	0.049
	(0.043)	(0.109)	(0.037)	(0.035)
Skew-normal	0.172	0.139	0.065	0.041
	(0.058)	(0.199)	(0.046)	(0.041)
Panel B: Posterior mean for skewness parameter (s.d.)				
Skew-normal	8.681	-0.002	-0.008	1.365
	(5.935)	(10.036)	(10.352)	(10.280)

Notes: For models in which the true average MDA effect is drawn from a half-normal distribution, we used a uniform prior for the parameter with support from zero to ten. For models in which the true average effect of MDA deworming is drawn from a skew-normal distribution, we used uniform priors for the location and scale parameters with support from -5 to 5 and 0 to 10, respectively. For the shape parameter we used a normal prior centered at zero with variance 100.

Panel B presents posterior distributions for the skewness parameter when using the skew-normal distribution. For all outcomes, the mean is positive, suggesting right-skewed distributions of effect sizes. For weight gain, very little mass in the posterior is below zero, which strongly supports a right-skewed distribution for effect sizes. For other outcomes, the posterior distribution has a meaningful amount of mass in the negative range of the support, suggesting somewhat weaker support for a right-skewed distribution.

Appendix H: Publication bias in the deworming literature

Andrews and Kasy (2019) estimate that negative estimates of the effect of MDA on weight are ten times less likely to be published than positive ones, suggesting that estimates of deworming's impact based on published estimates are upward biased. However, their method assumes that true effects of MDA are independent of the standard errors. Because the effect of MDA depends on infection intensity, which is expected to be a non-linear function of infection prevalence (Anderson and May, 1985), we can examine whether true effects are independent of standard errors by examining the correlation between prevalence and standard errors. For our full sample of weight

effects, the correlation is large, at 0.342. Thus, Andrews and Kasy's (2019) method for estimating selective publication probabilities may be unwarranted in the case of deworming. In general, we note that a positive correlation between true effects and standard errors will arise whenever researchers have an unbiased prior of the effect of the intervention they seek to evaluate (e.g., by conducting a pilot or by obtaining a signal) and design their interventions to obtain adequate power (e.g., the standard of 80%) to detect a significant effect at the 95% confidence level, as suggested by most guidelines on experimental design. The positive correlation arises as smaller standard errors are needed to detect smaller effects.

Appendix References

Alderman, Harold, Joseph Konde-Lule, Isaac Sebuliba, Donald Bundy and Andrew Hall. 2006. "Effect on weight gain of routinely giving albendazole to preschool children during child health days in Uganda: a cluster randomised controlled trial." BMJ: British Medical Journal 333(7559):122–124.
URL: http://www.jstor.org/stable/40699371

Awasthi, S, R Peto, VK Pande, R Fletcher, S Read and Bundy DA. 2008. "Effects of deworming on malnourished preschool children in India: an open-labelled, cluster-randomized trial." PLoS Neglected Tropical Diseases 2(4):e223.

Awasthi, Shally, Richard Peto, S Read, SM Richards, V Pande, Donald A Bundy and DEVTA team. 2013. "Population de-worming with 6-monthly albendazole: DEVTA, a cluster-randomised trial among 1 million preschool children in North India." The Lancet 381(9876):1478–1486.

Awasthi, Shally and Vinod Kumar Pande. 2001. "Six-monthly de-worming in infants to study effects on growth." Indian Journal of Pediatrics 68(9):823–7.

Awasthi, Shally, Vinod Kumar Pande and Robert H. Fletcher. 2000. "Periodic deworming with albendazole and its impact on growth status and diarrhoeal incidence among children in an urban slum of India." Indian Pediatrics 37:19–29.

Carmona-Fonseca, Jaime, and Adriana Correa-Botero. "EFECTO DEL ALBENDAZOL Y LA VITAMINA A PERIÓDICOS SOBRE HELMINTOS INTESTINALES Y ANEMIA EN NIÑOS DEL URABÁ ANTIOQUEÑO (COLOMBIA)." Biosalud 14.1 (2015): 9-25.

Donnen, P, D Brasseur, M Dramaix, F Vertongen and et al. 1998. "Vitamin A Supplementation but not deworming improves growth of malnourished preschool children in Eastern Zaire." Journal of Nutrition 128(8):1320–7.

Dossa, R. A. M. and E. A. D. Ategbo. 2001. "Impact of iron supplementation and deworming on growth performance in preschool Beninese children." European Journal of Clinical Nutrition 55(4):223.

Frison, Lars and Stuart J. Pocock. 1992. "Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design." Statistics in Medicine 11(13):1685–1704.
URL: http://dx.doi.org/10.1002/sim.4780111304

Freij, L., Meeuwisse, G.W., Berg, N.O., Wall, S. and Gebre-Medhin, M., 1979. Ascariasis and malnutrition. A study in urban Ethiopian children. The American journal of clinical nutrition, 32(7), pp.1545-1553.

Gateff, C, G Lemarinier and R Labusquiere. 1972. “Chimiotherapie antihelminthique systematique au thiabendazole en milieu scolaire africain.” Annales de la Societe Belge de Medicine Tropicale 52(2):103–112.

Goto, Rie, C.G. Nicholas Mascie-Taylor and Peter G. Lunn. 2009. “Impact of anti-Giardia and anthelminthic treatment on infant growth and intestinal permeability in rural Bangladesh: a randomized double-blind controlled study.” Transactions of The Royal Society of Tropical Medicine and Hygiene 103(5):520–529.

Gupta, M C and J J Urrutia. 1982. “Effect of periodic antiascaris and antigiardia treatment on nutritional status of preschool children.” The American Journal of Clinical Nutrition 36(1):79–86.
URL: http://ajcn.nutrition.org/content/36/1/79.abstract

Hall, Andrew, L Nguyen Bao Khanh, Don Bundy, N Quan Dung, T Son Hong and R Lansdown. 2006. “A randomized trial of six monthly deworming on the growth and educational achievements of Vietnamese school children.” Unpublished manuscript.

Higgins, Julian PT and Sally Green. 2011. Cochrane Handbook for Systematic Reviews of Interventions. Cochrane.

Huat, L.B., Mitra, A.K., Jamil, N.I.N., Dam, P.C., Mohamed, H.J.J. and Muda, W.A.M.W., 2012. Prevalence and risk factors of intestinal helminth infection among rural Malay children. Journal of global infectious diseases, *4*(1), p.10.

Joseph, Serene A., Martin Casapia, Antonio Montresor, Elham Rahme, Brian J. Ward, Grace S. Marquis, Lidsky Pezo, Brittany Blouin, Mathieu Maheu-Giroux and TheresaW. Gyorkos. 2015. “The Effect of Deworming on Growth in One-Year-Old Children Living in a Soil-Transmitted Helminth-Endemic Area of Peru: A Randomized Controlled Trial.” PLoS Negl Trop Dis 9.

Kahneman, D. and Tversky, A., 1979. Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), pp.263-292.

Kerwin, Jason. 2015. “The Effect of HIV Infection Risk Beliefs on Risky Sexual Behaviors: Scared Straight or Scared to Death?” Working paper.

Kirwan, P., Asaolu, S.O., Molloy, S.F., Abiona, T.C., Jackson, A.L. and Holland, C.V., 2009. Patterns of soil-transmitted helminth infection and impact of four-monthly albendazole treatments in preschool children from semi-urban communities in Nigeria: a double-blind placebo-controlled randomised trial. BMC infectious diseases, 9(1), p.20.

Kruger, M, CJ Badenhorst, EPG Mansvelt, JA Laubscher and AJS Benade. 1996. “The effect of iron fortification in a school feeding scheme and anthelminthic therapy on the iron status and growth of 6-8 year old school children.” Food and Nutrition Bulletin 17(1).

Le, Huong Thi, et al. "The effect of iron fortification and de-worming on anaemia and iron status of Vietnamese schoolchildren." British Journal of Nutrition 97.5 (2007): 955-962.

Liu, C., Lu, L., Zhang, L., Luo, R., Sylvia, S., Medina, A., Rozelle, S., Smith, D.S., Chen, Y. and Zhu, T., 2017. Effect of deworming on indices of health, cognition, and education among schoolchildren in rural China: a cluster-randomized controlled trial. The American journal of tropical medicine and hygiene, 96(6), pp.1478-1489.

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." Journal of Development Economics 99(2):210 – 221.

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

Miguel, Edward and Michael Kremer. 2014. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, Guide to Replication of Miguel and Kremer (2004).".
URL: http://emiguel.econ.berkeley.edu/assets/miguel research/46/PSDP-REP 2014-11.pdf

Ndibazza, Juliet, Harriet Mpairwe, Emily L. Webb, Patrice A. Mawa, Margaret Nampijja, Lawrence Muhangi, Macklyn Kihembo, Swaib A. Lule, Diana Rutebarika, Barbara Apule, Florence Akello, Hellen Akurut, Gloria Oduru, Peter Naniima, Dennison Kizito, Moses Kizza, Robert Kizindo, Robert Tweyongere, Katherine J. Alcock, Moses Muwanga and Alison M. Elliott. 2012. "Impact of Anthelminthic Treatment in Pregnancy and Childhood on Immunisations, Infections and Eczema in Childhood: A Randomised Controlled Trial." PLoS ONE 7(12):e50325.

Ostwald, Rosemarie, Mark Fitch, Rainer Arnhold, Jennifer Shield, Louie Dexter, Jan Kilner and Richard Kimber. 1984. "The effect of intestinal parasites on nutritional status in well-nourished school age children in Papua New Guinea." Nutrition Reports International 30(6):1409–1421.

Rohatgi, Ankit. 2015. "WebPlotDigitizer.".
URL: http://arohatgi.info/WebPlotDigitizer

Sarkar, N.R., Anwar, K.S., Biswas, K.B. and Mannan, M.A., 2002. Effect of deworming on nutritional status of ascaris infested slum children of Dhaka, Bangladesh. Indian pediatrics, *39*(11), pp.1021-1026.

Stephenson, L. S., M. C. Latham, E. J. Adams, S. N. Kinoti and A. Pertet. 1993. "Physical Fitness, Growth and Appetite of Kenyan School Boys with Hookworm, Trichuris trichiura and Ascaris lumbricoides Infections Are Improved Four Months after a Single Dose of Albendazole." Journal of Nutrition 123(6):103646.

Stephenson, LS, MC Latham, KM Kurz, SN Kinoti and H. Brigham. 1989. "Treatment with a single dose of albendazole improves growth of Kenyan schoolchildren with hookworm,

Trichuris trichiura, and Ascaris lumbricoides infections." American Journal of Tropical Medicine and Hygiene 41(1).

Stock, J., Watson M. 2011. "Introduction to Econometrics (3rd edition)". Addison Wesley Longman.

Stoltzfus, R.J., Albonico, M., Tielsch, J.M., Chwaya, H.M. and Savioli, L., 1997. School-based deworming program yields small improvement in growth of Zanzibari school children after one year. The Journal of nutrition, 127(11), pp.2187-2193.

Sur, D, D.R. Sahan, B Manna, K Rajendran and S.K Bhattacharya. 2005. "Periodic deworming with albendazole and its impact on growth status and diarrhoeal incidence among children in an urban slum of India." Transactions of the Royal Society of Tropical Medicine and Hygiene 99(4):261–7.

Taylor-Robinson, David C, Nicola Maayan, Karla Soares-Weiser, Sarah Donegan and Paul Garner. 2015. Cochrane Database of Systematic Reviews 7.

Tee, M.H., Lee, Y.Y., Majid, N.A., Noori, N.M. and Raj, S.M., 2013. Growth reduction among primary school-children with light trichuriasis in Malaysia treated with albendazole. Southeast Asian J Trop Med Public Health, *44*(1), pp.19-24.

Watkins, William E., Jos R. Cruz and Ernesto Pollitt. 1996. "The effects of deworming on indicators of school performance in Guatemala." Transactions of The Royal Society of Tropical Medicine and Hygiene 90(2):156–161.

Willett, Walter, WL Kilama and CM Kihamia. 1979. "Ascaris and Growth Rates: A Randomized Trial of Treatment." American Journal of Public Health 69(10):987–991.

Wiria, Aprilianto E., Firdaus Hamid, Linda J. Wammes, Maria M. M. Kaisar, Linda May, Margaretta A. Prasetyani, Sitti Wahyuni, Yenny Djuardi, Iwan Ariawan, Heri Wibowo, Bertrand Lell, Robert Sauerwein, Gary T. Brice, Inge Sutanto, Lisette van Lieshout, Anton J. M. de Craen, Ronald van Ree, Jaco J. Verweij, Roula Tsonaka, Jeanine J. Houwing-Duistermaat, Adrian J. F. Luty, Erliyani Sartono, Taniawati Supali and Maria Yazdanbakhsh. 2013. "The Effect of Three-Monthly Albendazole Treatment on Malarial Parasitemaia and Allergy: A Household-Based Cluster-Randomized, Double-Blind, Placebo-Controlled Trial." PLOS One 8(3):e57899.

Yap, P., Wu, F.W., Du, Z.W., Hattendorf, J., Chen, R., Jiang, J.Y., Kriemler, S., Krauth, S.J., Zhou, X.N., Utzinger, J. and Steinmann, P., 2014. Effect of deworming on physical fitness of school-aged children in Yunnan, China: a double-blind, randomized, placebo-controlled trial. PLoS neglected tropical diseases, *8*(7), p.e2983.