Supplementary Appendix

Commentary

Assessing Long-Run Deworming Impacts on Education and Economic Outcomes: A Comment on Jullien, Sinclair and Garner (2016)

Sarah Baird[1], Joan Hamory Hicks[2], Michael Kremer[3], Edward Miguel[4*]

9 November 2016

[1] Milken Institute School of Public Health, Department of Global Health, The George Washington University, Washington, District of Columbia, USA, [2] University of California, Center for Effective Global Action, Berkeley California, USA, [3] Department of Economics, Harvard University and NBER, Cambridge, Massachusetts, USA, [4] Department of Economics, University of California, Berkeley and NBER, Berkeley, California, USA.

*Corresponding author. E-mail: emiguel@berkeley.edu.

Supplementary Appendix

Jullien, Sinclair and Garner (2016; hereafter, JSG)[1] seek to appraise the methods of three recent papers that document long-term (roughly 7 to 10 year) impacts of child deworming interventions in Africa, namely, Baird, Hicks, Kremer and Miguel (2016a; hereafter, Baird)[2] in Kenya, Croke (2014)[3] in Uganda, and Ozier (2016)[4] in Kenya. Beyond our partial response in the Commentary, where we faced strict word limits, here we present a number of additional reactions.

One broad point is worth making up front. As no observer (to our knowledge) has suggested that deworming drugs create significant side effects, the relevant question for policymakers seeking to assess whether to implement mass deworming programs is whether the expected benefit of deworming (taking into account the uncertainty of that benefit) exceeds the cost. We would not claim that there is zero uncertainty about the benefits of deworming. However, the benefits estimated in Baird[2] exceed deworming's (minimal) costs by more than 100-fold, so deworming would be a good investment even if there were substantial uncertainty about its benefits. The fact that all three of these studies of long-term benefits, as well as Bleakley (2007)[5], find meaningful positive impacts, suggests that it would be difficult to reach the conclusion that the cost of mass drug administration exceeds the expected benefit in areas with high prevalence of worms. As discussed below, the specific issues raised by JSG[1] do not invalidate this conclusion.

(1) No research study is perfect along all dimensions, and an assessment of which studies to include in a meta-analysis or systematic review should take into account both studies' weaknesses as well as strengths. Existing meta-analyses, such as Taylor-Robinson et al (2015)[6], for example, include studies of mass drug administration in areas with worm prevalence far below the WHO threshold for mass drug administration, which can be seen as an important weakness of those studies. JSG[1] emphasize what they see as weaknesses of the three studies, and largely ignore their strengths. All three studies provide policy-relevant evidence for mass deworming in areas of high infection prevalence, where the current WHO policy would

recommend mass deworming drug administration. All three evaluate cluster-randomized trials, and thus have the potential to capture epidemiological spillovers. The three studies examine real-world programs in two different countries. In addition, they all successfully obtained long-run data 7 to 10 years after the deworming intervention, with rich educational and economic outcomes of direct interest to policymakers, and balanced follow-up rates between the treatment and control groups, itself a substantial feat. Many of the studies included in existing reviews of the deworming literature, such as Taylor-Robinson et al (2015)[6], do not have these useful characteristics. It seems far from clear to us that it would be appropriate to exclude these three studies from systematic reviews and meta-analyses while including other studies, for example, studies conducted in areas below the WHO threshold for mass drug administration.

(2) An observer might legitimately place lower weight on the results of a study if one believed that many outcomes had been examined and only those with positive outcomes were reported. In a research field in which pre-analysis plans are standard practice, one might also think there was a high likelihood of selective reporting in the absence of such a plan. JSG[1] criticize all three papers for not registering pre-analysis plans, but until recently, these plans were unknown in Economics and Political Science (the home disciplines of the authors on the three studies), with the American Economic Association RCT registry only established in 2013. It is worth noting that pre-analysis plans and study registration only became the norm in clinical trials around 2000, and only become ubiquitous in health research in the last decade. Thus much influential existing epidemiological research also did not feature a priori analytic plans. The use of pre-analysis plans certainly confers some advantages, but it would be inappropriate to treat the absence of such a plan as indicating bad intent on the part of researchers in a context in which such plans are not the norm, or to dismiss evidence on the basis of the absence of a pre-analysis plan in such a context.

Moreover, social science research has developed several alternative approaches to improving the reliability of empirical results in the absence of pre-analysis plans. In lieu of pre-

analysis plans, most social science empirical studies typically feature multiple robustness checks and alternative specifications in an attempt to assess the validity of results, and all of the three critiqued studies pursue this strategy (and this can explain some of the increase in hypotheses tested in Baird[2] in subsequent versions of the working paper). This is a strength of these studies, but it is not described as such by JSG[1].

Of course, there could also be unconscious researcher bias, and this could potentially lead to the choice of particular statistical specifications or covariates that would influence the results. Under the null hypothesis of no effect, a team of researchers interested in finding positive results could have a probability of finding effects greater than 5% in the presence of such unconscious bias, and analogously, a team of researchers interested in finding no effect would have a higher than 20% probability of finding no effect in a study with 80% power, say, if they were subject to unconscious bias in choosing among specifications, covariates or outcome variables. It is striking here that JSG[1] do not suggest that alternative specifications or covariates would meaningfully change the results in Baird[2]. Note also that if there were a, say, 10% rather than 5% chance that the results in Baird[2] would arise under the null hypothesis of no treatment effect, the fundamental conclusion that the expected benefit of deworming vastly exceeds its cost would remain valid under any reasonable set of statistical priors.

There are also major differences in norms in the social sciences versus public health regarding the circulation (and online posting) of the working paper versions of studies, which are sometimes called "discussion papers" for this reason. Baird[2] is a good example of this case, given that it took several years for the authors to assemble their results, receive feedback from colleagues, and more time to work through the journal publication process. Economics researchers typically circulate incomplete working paper versions quite widely, and post working paper versions in public series even before submission for publication, in the hope of getting comments from fellow scholars to improve the quality of their statistical analysis. This is the normal state of affairs in our disciplines, but it is not portrayed as such by JSG[1]. They inaccurately characterize the initial working paper versions as the authors' complete intended analysis (without checking with the authors if this was in fact the case), and effectively consider

all subsequent versions of the paper as post-hoc data mining. However, economics researchers often circulate working papers that are incomplete, with additional analyses planned (e.g., analysis of results broken down by gender in Baird[2], as we discuss in our Commentary).

JSG[1] could have usefully provided a more balanced discussion of the fact that there have been multiple versions of Baird[2] circulated, rather than implying that the existence of these working paper versions is "proof" of selective presentation of results. Quite the opposite: pre-publication results in economics papers are discussed, vetted, and improved over time through sharing within the research community, and typically additional robustness checks and specification checks are introduced which can increase, rather than decrease, the scientific community's confidence in the results.

(3) JSG[1] present "risk of bias" assessments, based on the earlier Cochrane report[6], in their Table 2. We have a number of concerns with this table, and discuss a few of these here.

First, in the case both of Miguel and Kremer (2004)[7] and Alderman et al. (2006)[8] (two cluster RCTs), one of the main drivers behind their assessment of "high risk" of bias is the fact that neither of these studies was double-blinded. However, it is logistically impossible to carry out a truly double-blinded cluster randomized deworming trial: deworming generates minor side effects (mainly transient gastrointestinal distress) in roughly 10% of those who take the pills, so community members in a placebo community would quickly deduce that they were in fact not receiving real deworming drugs if there are no local cases of minor side effects. Yet cluster randomized trials have many scientific advantages, including their ability to capture spillover effects, and the fact that they more closely resemble how a large-scale mass deworming policy would be carried out. Frankly, the fact that evidence from unblinded cluster randomized control trials is downgraded for this reason seems to us to indicate that the "risk of bias" tools used in epidemiology and public health are problematic in this context.

Second, a concern JSG[1] appear to have about lack of blinding among participants is that treatment effects might work mainly through placebo effects. There are several ways to explore

these issues in the data from Miguel and Kremer (2004)[7]. First, there are sizeable numbers of students in treatment schools who did not receive deworming treatment either due to absence on the day of deworming or because they were adolescent girls (who were meant to be excluded from treatment due to potential drug side effects). If the effect were mainly driven by placebo effects, rather than real deworming impacts, then there would be no meaningful effects on those students who did not themselves take the deworming pills. This would be true both for the untreated within treatment schools, and for control school students located within 3 km of a treatment school. Yet these are populations that both Miguel and Kremer (2004)[7] and Aiken et al. (2015)[9] found experienced reduced worm infection burden, presumably due to epidemiological externalities, and also show gains in school participation (even though a placebo effect is not plausible for them). This point is discussed in greater detail in the appendix to Hicks et al. (2015)[10]. Placebo effects also cannot plausibly explain the externality findings in Miguel and Kremer (2004)[7], or in Ozier (2016)[4], for that matter.

Unfortunately, despite our earlier communications with them on this topic, JSG[1] did not acknowledge these points, and thus gloss over the methodological strengths of the two trials upon which Baird[2], Croke (2014)[3] and Ozier (2016)[4] are based. Rather they label them as "high risk of bias" for using a cluster design that cannot plausibly be blinded to participants.

Third, JSG claim that Miguel and Kremer (2004)[7] is at high risk of bias due to the fact that there was simultaneous health (worm prevention) education carried out in the treatment schools at the same time as drug administration. This raises the possibility that it was health education rather than the deworming drugs themselves that drove impacts.

While the health education intervention limits the ability to study the impact of deworming pills alone, it does carry other important advantages. Note that worm prevention education is carried out in tandem with drug treatment in many real-world mass deworming programs, and thus studying the combined impact of these two complementary interventions (both of which aim to reduce worm infection) is important for understanding public policy. Moreover, as discussed in Miguel and Kremer (2004)[7], there is no evidence of worm prevention

behavioral change in their deworming treatment schools, strongly suggesting that the deworming drugs per se were the key drivers of any impacts.

Fourth, JSG claim that a major weakness of Alderman et al. (2006)[8] is that there was a non-trivial rate of treatment in the control group. Despite this, Alderman et al.[8] apparently have sufficient statistical power to estimate moderate positive deworming treatment effects, all of which should be thought of as lower bounds on true treatment effects in the absence of this contamination – making the results of Alderman et al. (2006)[8] and Croke (2014, the follow-up)[3] all the more striking. This important scientific point is ignored in JSG's discussion.

Taken together, JSG[1] appear to apply quite a narrow approach to considering the scientific value of the studies they critique. Many study features that appear to have limitations according to simple rule-of-thumb criteria also have important scientific strengths, or advantages for drawing implications for public policy, as noted above.

(4) Related to the narrowness of their approach, JSG mention, but do not discuss, the Bleakley (2007)[5] findings on long-run impacts of deworming in the US. JSG[1] are, of course, free to choose their inclusion criteria, but a policymaker interested in the best informed estimate of the likely impact of deworming would put some weight on the findings in Bleakley (2007)[5], given its uniquely long time frame (across multiple decades), credible statistical design, massive administrative dataset, and the fact that it is the real-world follow-up of an actual mass deworming program, albeit one undertaken in the last century. JSG's approach effectively puts zero weight on the results of this study.

(5) Another main issue for JSG[1] regarding selective reporting is the number of hypothesis tests presented in Baird[2]. They write that "There is a high risk of false positive results with the number of hypotheses tested for statistical significance increased from 228 in Baird 2011a to 650 in Baird 2016." But JSG[1] never clearly discuss why an increase in the number of tests is

evidence for selective reporting, or any other form of bias. We have no indication of what exact hypothesis JSG[1] are proposing to test here.

There are actually straightforward reasons why the number of tests increases across versions of Baird[2]. The number of hypothesis tests increases in large part because later versions included a breakdown of results by gender, roughly tripling the number of tests (i.e., tests on the full sample, for females, and for males). There is also an increase over time as the authors included additional robustness checks suggested to them by conference participants or journal referees, and some reduction in other tests as the authors refined the outcomes presented in the main tables (although many of these alternative outcomes were retained in the Appendix).

In fact, a close look at JSG[1] Appendix 3 provides evidence running counter to JSG's broad claims about selective reporting. For instance, if the data were all "noise" and there was no relationship between deworming and later outcomes, one might expect that the proportion of tests yielding statistically significant results at the $p<0.05$ level would be roughly 5%. Yet the proportions of significant results are much higher than this and nearly unchanged over time as additional hypotheses are added, ranging between 24% and 30%, according to JSG[1].

A few other patterns are noteworthy. If Baird[2] had been systematically cherry-picking statistically significant results over time (and no longer reporting null results) as they "mined" their dataset, they might have shown no (or only very few) non-significant results, and/or the proportion of tests with $p<0.05$ would presumably have risen sharply over time. Yet, as noted, it remains stable at between 24 to 30% across all versions of the paper. So that pattern, too, does not appear consistent with selective reporting.

If the true relationships were all "noise", and Baird[2] were simply presenting the significant findings, then that would imply that there would have to be approximately 4 outcomes not shown in the Baird[2] tables for each outcome that is shown. Why? In order to drive the proportion of significant ($p<0.05$) results from roughly 25% down to the pure noise level of 5%, many outcomes would simply need to "disappear" from the analysis. But this again is implausible. Baird[2] already show the standard educational and labor market outcomes (e.g., school enrollment, attainment, wages, hours worked, sector, meal consumption), including

some that are not statistically significant. Everybody is subject to some unconscious bias, so we cannot completely rule out the possibility that there were some mistaken judgement calls. However, we do know that we attempted to be as even-handed as possible in the presentation of the results, and we can certainly rule out the possibility that we excluded four fifths of potentially relevant results.

The same issues arise in JSG's presentation in their Table 4. There they claim that, among the 30 main hypothesis tests they consider, 17 of these are statistically significant at 95% confidence, while 13 are not. Seventeen of 30 is greater than half, and an order of magnitude larger than the 5% we would expect if the results were pure "noise". (Incidentally, there appear to be "a" and "b" superscripts in this table from an earlier version in which the authors considered significance in a more nuanced fashion, at both $p<0.05$ and for $0.05<p<0.1$; they have apparently abandoned this nuance in the published version, for reasons that are unclear.)

JSG[1] make strong insinuations about selective reporting in Baird (2016a)[2], related to the number of hypotheses tested, but a closer look at the actual pattern of results is not consistent with the most plausible forms of selective reporting. Given the seriousness of the issue of selective reporting in the research community, JSG[1] should marshal more convincing evidence before making these claims, in our opinion. We do not find their "narrative analysis" convincing.

(6) One of the most serious charges of selective reporting are made when JSG[1] write that "Some outcomes reported in early versions were dropped. It is not made clear to the reader why, but it is likely to be due to the failure to demonstrate an effect (for example, cognitive test results reported in 2011a, but absent in 2016; with no apparent effect on Ravens matrices or English vocabulary)."

While JSG[1] make some rather inflammatory claims about selective reporting of non-significant results in Baird[2] here, JSG[1] themselves "cherry-pick" one or two outcomes that support their narrative of selective reporting in Baird[2], while misrepresenting broader patterns.

First of all, the cognitive test results (Ravens matrices, English vocabulary) are reported in the appendix of the Baird (2016a)[2] paper: "We also have test score information for KLPS-2 respondents, from an English vocabulary test and a Raven's Matrices test. There are no statistically significant treatment effects of deworming on either of these outcome measures (normalized within the sample), using the standard regression specification, either in the full sample or broken out by gender (not shown)." A text search on "test" yields this information.

There were also multiple statistically significant impacts reported in early versions of Baird[2] that were then not reported in later versions, as the authors refined their analysis and made the main tables a bit more compact. It is not the case that only null results disappeared; some statistically significant results did, too. For instance, the 2011 versions of Baird[2] present evidence that the days of work missed due to poor health in the past month among wage earners was lower in the treatment group ($p<0.05$). This interesting and statistically significant result was cut from subsequent versions, but this fact is not mentioned in JSG[1].

(7) As we discuss in our Commentary, JSG[1] contains a discussion of the multiple testing adjustments in Baird[2]. Specifically, JSG[1] write: "In their 2016 abstract, Baird et al. state that men stayed enrolled for more years of primary school, and women were approximately one quarter more likely to have attended secondary school. These statements are supported by 'statistically significant' results within the text, but … neither result is robust to the authors' own adjustments for multiple inferences."

This discussion, and the related information in JSG's Table 6 is quite misleading, in our view. First of all, in the main text and in the supplementary materials, Baird[2] present all multiple testing adjustments, and discuss how statistical significance levels change with the adjustments.

As discussed above, including all of this material in a 100 to 250 word Economics abstract is literally impossible.

More importantly, where there are changes in significance levels with the multiple testing adjustments, they are typically minor in Baird[2], as a glance at the False Discovery Rate (FDR) adjusted q-values in the supplementary appendix tables (and Table 1 in our Commentary) indicates. For instance, in the case of these two outcomes that JSG[1] focus on above, namely, men's primary school enrollment and women's secondary school enrollment, both results remain statistically significant at 90% confidence after adjustment (with q-values of 0.07 and 0.08, respectively). The labor supply results for men also remain significant at 90% after adjustment. Given an estimated deworming benefit/cost ratio of more than 100, moving from a 5% chance that the results would arise by chance under the null hypothesis of no effect to a 10% chance that they would arise under the null hypothesis does not change the conclusion that the expected benefits of mass deworming greatly exceed the cost under any reasonable set of priors on the null hypothesis. However, as we note in our main Commentary, JSG's discussion does not present the actual adjusted p-values.

(8) The discussion of whether effects are consistent across related outcomes (the final column JSG's table 6) could also be improved. The goal of this exercise appears to be for JSG[1] to find an outcome within the set of educational or labor market outcomes on which there was no treatment effect. If so, they list it here as "evidence" of selective reporting in Baird[2].

This is an inappropriate strategy, in our view, and creates the potential for unconscious bias on the part of JSG[1] to influence the assessment. These are outcomes that Baird[2] themselves present to the reader; they are not being hidden. And the multiple testing corrections discussed above account for the pattern of significance levels across a set of related outcomes (e.g., educational outcomes) to determine whether the significant ones were likely to have been produced by chance or sampling variation. As noted above, in most cases statistically significant results remain robust at either the 95% or 90% level post-adjustment for multiple testing.

In our view, this column is superfluous if the multiple inference adjustment p-value results are presented appropriately (i.e., the adjusted p-value rather than just “yes” and “no”).

Moreover, we have made the replication data and materials available to the scholarly community (Baird et al 2016b)[11], so JSG[1] and other scholars can implement alternative multiple testing adjustments if they prefer, or seek to assess the robustness of the results in other ways.

(9) Table 5 in JSG[1] appears to selectively ignore several important outcome variables in Baird[2], including the meal consumption measure, and the sectoral employment variables (including manufacturing employment).

References (for Supplementary Appendix)

1. Jullien S, Sinclair D, Garner P. International Journal of Epidemiology, 2016; XX, XX-XX.
2. Baird S, Hicks JH, Kremer M, Miguel E. Worms at work: Long-run impacts of a child health investment. Quarterly Journal of Economics. 2016a; 131(4): 1637-1680.
3. Croke K. The long run effects of early childhood deworming on literacy and numeracy: Evidence from Uganda. 2014. Available from: http://scholar.harvard.edu/files/kcroke/files/ug_lr_deworming_071714.pdf.
4. Ozier O. Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming. 2016. Available from: http://economics.ozier.com/owen/papers/ozier_early_deworming_20160727.pdf.
5. Bleakley H. Disease and Development: Evidence from Hookworm Eradication in the American South. Quarterly Journal of Economics. 2007; 122(1): 73–117. Available from: doi: 10.1162/qjec.121.1.73.
6. Taylor-Robinson D, Maayan N, Soares-Weiser K, Donegan S, Garner P. Deworming drugs for soil transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance. Cochrane Database of Systematic Reviews. 2015(7). Available from: doi: 10.1002/14651858.CD000371.pub6.
7. Miguel E, Kremer M. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Econometrica. 2004; 72(1): 159–217.
8. Alderman H, Konde-Lule J, Sebuliba I,Bundy D, Hall A. Effect on weight gain of routinely giving albendazole to preschool children during child health days in Uganda: cluster randomized controlled trial. British Medical Journal. 2006. Available from: doi: 10.1136/bmj.38877.393530.7C.
9. Aiken AM, Davey C, Hargreaves JR, Hayes RJ. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. International Journal of Epidemiology. 2015;1-9. doi: 10.1093/ije/dyv127.
10. Hicks JH, Kremer M, Miguel E. Commentary: Deworming externalities and schooling impacts in Kenya: a comment on Aitken et al. (2015) and Davey et al. (2015). International Journal of Epidemiology. 2015; 1-4. doi: 10.1093/ije/dyv129.

11. Baird S, Hicks JH, Kremer M, Miguel E. Replication data for: Worms at work: long-run Impacts of a child health investment. 2016b. Harvard Dataverse, doi: 10.7910/DVN/ZNSY5O.