Corrigenda for "Disease and Development: Evidence from Hookworm Eradication in the American South"*

Hoyt Bleakley[†]

First version: December 1, 2017
This version: January 5, 2018

This document corrects some errors and/or ambiguities in the published version of the study referenced in the title.[1] More than a decade after its publication and over 15 years after doing the main analysis, I had occasion to re-read this article in the fall of 2017. With fresh eyes, I noted some inconsistencies, omissions, and ambiguities that were not apparent to me back then in the process of multiple revisions of the paper.

This note is written with the assumption that the reader is familiar with the original paper, otherwise many of the references below will make little sense. Readers unfamiliar with the original paper are encouraged to read it before reading this document. Recall that the RSC is the "Rockefeller Sanitary Commission," which operated the anti-hookworm campaign analyzed, and SEA is a "State Economic Area", a grouping of contiguous and relatively homogeneous counties. SCS are "Sequential Cross Sections", which refer to the use of pooled census data for various years to analyze the flow of school attendance, *inter alia*. RC is a "retrospective cohort" analysis, in which individuals are tracked by their state and year of birth, and their income is compared with their potential exposure to the RSC. "Long-term RC" refers to the exercise that pools various census years to construct cohort-level averages of income proxies (as contrasted with the use of single census year).

Many of these ambiguities resulted from what turned out to be lossy compression when reducing the very long manuscript and various associated results into a version that was short enough for publication. Therefore, I wish to also bring the reader's attention to a 2006 working paper: Stigler Center Working Paper #205.[2] I released this in part as an effort to have a longer version of the paper in the pre-publication record. The working paper is itself

*Comments and questions welcome. I thank David Roodman, whose queries identified several ambiguities in the published paper. Any errors are my own.

[†] Department of Economics and Population Studies Center, University of Michigan. Postal address: 611 Tappan St, Ann Arbor, MI 48109-1220. Electronic mail: hoytb at umich dot edu.

[1] "Disease and Development: Evidence from Hookworm Eradication in the American South." *Quarterly Journal of Economics,* February 2007, 122:1. `http://www.jstor.org/stable/25098838`

[2] `http://www.chicagobooth.edu/assests/stigler/205.pdf`

also a compressed version of the project, of course, but it is long enough to be informative in regard to certain ambiguities in the published version.

To be clear, I do not believe that any of these items referenced below affect the basic conclusions of the project. They might, however, confuse a reader or potential replicator as to the path along which I arrived at such conclusions. **In any case, I accept full responsibility for these errors.**

These corrigenda are listed in the order of appearance in the paper.

- page 73, footnote from the title: while this paper started as the first essay of my dissertation, in the sense of people writing three essays and referring to them as 'the' chapters of their dissertation, it was in fact the second chapter of my disseration. The MIT dissertation formatting rules required an introduction to the thesis, which was itself, in fact, the first chapter of my thesis.

- page 85: the 1930 census data from IPUMS was marked as 'preliminary' at the time when these results were generated. It was therefore did not include 1930 in the data set. The omission of 1930 was stated in the appendix, but should have been at this point in the text as well.

- page 85: "number" should read "numbers."

- page 88: Figure II reports 95% confidence intervals, a fact that should have been stated in the figure notes. My long-standing habit is to copy the two right-most columns from Stata regression output, which reports 95% confidence intervals.

- page 88: Figure II reports coefficients on pre-RSC hookworm for regressions of school enrollment in the years 1900–1950, and a similar figure has appeared since the dawn of the project, going back to drafts of my dissertation. Nevertheless, the Spence (Growth) Commission later re-published something essentially similar to the published version, albeit with a Figure III that includes 1870–80 estimates, and a version of that figure also appears in the Stigler Center working paper. See Figure 1 of this document for a comparison of the two graphs.

  I wish to provide some explanation for this discrepancy.

  At the outset of the project, I used data from 1900-50 only. I was dissuaded from going farther back in part by the missing 1890 census and in part by the lack of digitized information on 19th-century county boundaries. (I believe that the NHGIS files were not publicly available at that time. If they were, I was not aware of them.)

  Very late into the project, I became aware of the NHGIS data and that the SEAs had been coded for 1870–80. Consequently, I spent some time reviewing the historical county boundary changes using the newly available NHGIS. It seemed to me that using SEA was an acceptable harmonization strategy for the 1870–80 counties, albeit less good than for the 20th century. At that time, I decided to produce some results (seen in the figure in the versions produced in the Spence Commission report and in the Stigler Center working paper) that incorporated the 1870–80 data.

The paper was relatively near to publication at that time, and I was a bit reluctant to open the can of worms associated with such a dramatic change in the sample. For example, this could involve redoing all of the tables, justifying the harmonization strategy going farther back in time, and perhaps provoking an additional round with the referees. I recall mentioning this issue in some of the correspondence with the editor, but I think that the subject got lost in the shuffle of all the other things that had been added to the paper in response to referee comments.

So, I did not pursue the issue of including 1870–80 in the SCS.

- page 89: "panel F" should read "row E."

- page 90: In the text describing Panel B of Table III, I neglected to mention that the estimated coefficients on the added control variables × post-RSC can be seen in Stigler Center WP#205, Table IV. (Please note that Table IV in the working paper has two panels labelled "Panel C" by accident. These are two separate specifications.) Note further that the parental background controls enter the specification in the third row of Panel B simply as a main effects (i.e., without an interaction with post or with year). This is described in the working paper, with reference to Table IV therein, as follows.

  > In Panel D, I include controls for parental background, but these do not materially affect the estimated hookworm coefficient. I proxy for each parents income with the occupational income score, and include a binary indicator for whether that parent is present (missing parents getting an imputed income of zero). While these parental SES variables are generally highly statistically sig-nificant, their inclusion results in changes of the hookworm coefficient that are less than a standard error. Similar results are obtained when using literacy or Duncans socioeconomic indicator as the measures of parental background, as well as if allowing these parental variables to vary across year.
  >
  > Finally, Panel E of Table IV presents the estimated coefficient on hookworm × Post$_t$ when all of the above controls are included in the specification. The estimates are similar to the baseline.

  There is no exact equivalent of the working paper's Table IV Panel D, but Panel E in that same table corresponds to the third row of Panel B in the published paper's Table III.

- page 94, final paragraph: I write that

  > I consider a simple parameterization of the cross-cohort comparison: the number of childhood years potentially exposed to the anti-hookworm campaign, times the pre-eradication hookworm intensity in the state of birth. (page 94)

I wish to elaborate on this here. 'Simple' was in the sense of 'stylized' and perhaps 'simplified.' In a summary paper[3] on this and other work, I state that

> The dashed line measures the approximate number of years of potential childhood exposure to the hookworm-eradication activities in the South. (page 220)

The phrase 'approximate numbers of years' is not meant simply as a rhetorical hedge. Instead, it is meant to convey uncertainty about the exact functional form, with specific reference to the scope and slope for the partially exposed cohorts.

While the simple functional form for exposure suggests the presence of testable kinks in the hookworm/income coefficients at two specific year-of-birth cohorts, there are at least three reasons why such a test would be inappropriate.

- The campaign was not instantaneous.
- Infection was not uniform across ages in childhood.[4]
- The reduction in human capital from a given intensity of infection might not be of equal magnitude across ages in childhood.

One could use these features to adjust the measure of childhood exposure, although this would interject more discretion and more unknowns into the functional form. The decision that I made in pursuing this research agenda was to opt for a transparent and comparatively inflexible and nondiscretionary approach. I understand that it is an approximation, but it is a simple one that also ties the hand of the researcher.

The published paper states the limits of this approximation when compared with more flexible functional forms.

- page 99: In the first full paragraph, which describes the data construction briefly, I neglected to mention that *only white males* were included in the underlying sample for the long-term RC analysis. That only whites are included is mentioned in the Data Appendix, however. The exclusion of females from the sample was lost on the cutting-room floor when revising the working paper for publication.

    - Failing to include a justification in the paper for focusing on whites in the long term retrospective/cohort analysis was an unfortunate omission. I am grateful to David Roodman for noticing this and finding a brief justification for this choice in another of my papers ("Malaria Eradication in the Americas[...]", *AEJ: Applied*, 2010). Footnote 7 (on page 11) of that paper, which reads in part as follows.

[3] "Economic Effects of Childhood Exposure to Tropical Disease," *American Economic Review: Papers & Proceedings* 2009, 99:2, 218223. `http://www.aeaweb.org/articles.php?doi=10.1257/aer.99.2.218`

[4] Smillie, Wilson G., and Donald L. Augustine. "Intensity of Hookworm Infection in Alabama: Its Relationship to Residence, Occupation, Age, Sex, and Race." *Journal of the American Medical Association.* 85 (1925), 1958–1963.

> I focus on US whites for several reasons. First, only a small proportion of blacks lived outside of the most malarious states among the earlier cohorts, which means that they make for an imprecisely measured point of comparison. Second and more importantly, that same population of blacks was less likely to have been enslaved, which means that they make for an inappropriate control group for those blacks born into slavery in the malarious [S]outh. [...]

The argument about regional aspects of malaria and black population applies equally well to hookworm. I think that I chose the "have been enslaved" language deliberately in that case. This would include the so-called freedmen in the 1880 census, for example, who were born into slavery, even if they were emancipated by the time we observe them in the 1880 census, for example.

Taking a step back for a moment, I recognize that there is a tradeoff here in the long-term analysis. Those people who were born only a decade or two apart are more likely to be comparable to each other, but unlikely to be useful in sorting out the cross-cohort timing of income convergence. I made the judgment call that this comparability problem was too severe in the case of blacks because of enslavement at the outset of the sample, their distinct regional distribution over time, and later the effects of the increasing integration of blacks into the mainstream economy.

– I did not include women in the long-term RC results also for reasons of comparability over the stretch of 150 birth years. I did, however, neglect to document this choice in the published paper. Omitting from the published version the text indicating that I only included males in the long-term RC sample was a serious expositional error.

  This important piece of information seems to have ended up on the cutting-room floor when shrinking the draft paper down to a publishable size. On page 26 of the 2006 working paper and page 98 of the published paper, notice that the sections are broadly similar, and indeed some paragraphs are close to identical. However, the third paragraph of the working paper, in which I mention that the sample starts with males in the microdata, was cut in half. This dropped any mention of males in the text, and instead referred the reader to the Appendix for details on the data construction. Unfortunately, I failed to notice that the appendix did not mention the exclusion of women from the base sample for the very long-term results. Such a change also happened in the figure that presents the cohort-specific coefficients on hookworm (figure 5 in the working paper in figure 3 in the published paper). That the sample consists of males is mentioned in the working paper, but was cut for space in the published paper, apparently in the hope of using the appendix as a backstop to describe that and features of the sample. That males form the basis of the sample also appears in Tables 11 and 12 in the working paper. Table 11 corresponds to Table 6 in the published paper, and again the table note is heavily compressed in part by omitting a detailed description of the sample. Table 12 in the working paper has no equivalent in the published paper, although the results are mentioned in the text. Again, text that would've indicated males are the base sample for these long-term analyses ended up on the cutting room floor.

That being said, I would defend the choice of looking exclusively at males in the long-term analysis. (Or, if not exclusively, at least firstly and separately.) The 150 years of birth cohorts covers a span of time that experiences fairly substantial changes in both female labor-force participation and in the character of that participation in labor market.

- page 101, footnote 25: I wrote that "I have experimented with higher-order polynomial trends and found no estimates of exposure that are statistically significant for $n \leq 5$." I wish to clarify the meaning of this statement here. I hope that no one reads this to mean that I toyed with cohort polynomials of order 6+ and found robust results for exposure. Instead, this is meant to say that *results for exposure are not robust to controlling for higher-order trends.* Table VI displays results controlling for up to a quadratic in year of birth and AR(2) in the cohort-specific hookworm coefficient itself.

- page 107, section I.A, lines 3–4: The reported access date for the IPUMS data is incomplete. The first time that I accessed the data would have been September 2001, when I began work on this project as a chapter in my PhD thesis. It is probable that I downloaded a few additional variables after May 2003 as well. (The parental occupation variables do not appear in early versions of the paper, for example.) I did not appreciate at the time that the IPUMS was a developing resource, and thus I did not adequately document these versioning issues.

- page 108, 3rd line of final paragraph: I failed to note my reason for excluding the 2000 census from the long-term analysis across cohorts using occupational income scores. I did so because of the following warning on the IPUMS website (emphasis below is mine).

> User Caution: The translation of occupation codes into the 1950 classification is particularly problematic for 1980-2000 censuses, the ACS and the PRCS. Significant reorganizations of the occupational classification scheme by the Census Bureau in 1980 and again in 2000 mean that occupation scores in this period will be more distorted than for earlier decades. *The difficulty is most acute for the 2000 census, the ACS and the PRCS, because the classification into the 1950 system was performed solely on the basis of the occupational titles without the benefit of supporting technical documentation.* (IPUMS website, accessed December 1, 2017)[5]

In the Stigler Center WP#205, Table XII, I show that the parameter estimate on childhood exposure to hookworm is somewhat sensitive to including 2000 or to excluding

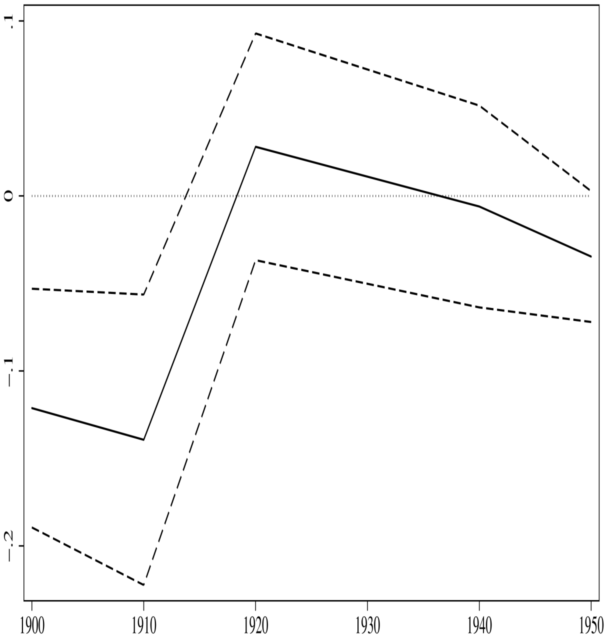
[5] `http://usa.ipums.org/usa-action/variables/OCCSCORE#comparability_section`

1980–90 censuses. I did not produce graphs of the $\hat{\beta}_k$ for these alternative samples, however.

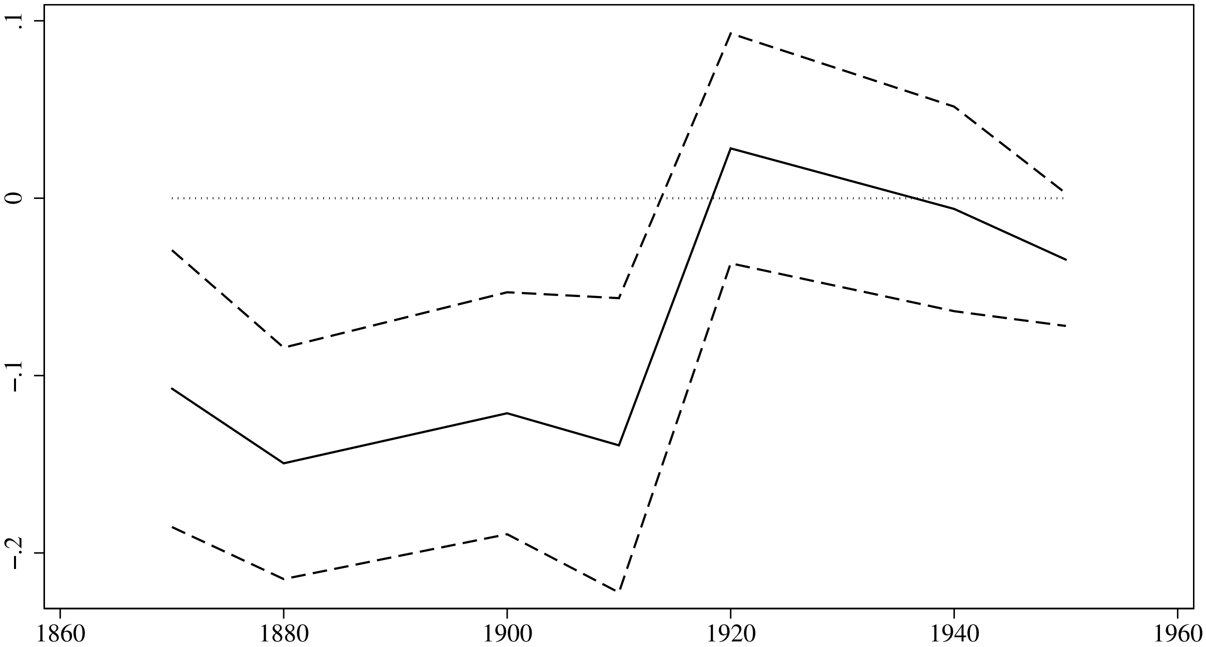
- page 110, 4th line down: "dependent" should read "independent."
- page 110, final paragraph: the paragraph should also state that the SEAs (county groups) were used as a strategy for managing missing data. Because the counties within an SEA were relatively homogeneous, I used the non-missing data from the other counties in the SEA to construct the SEA variable.
- page 111, first full paragraph: Several of the aggregate variables that (i) were added at a late stage of the project and that (ii) were attributed to ICPSR study #3 may in fact have been drawn from ICPSR study #2896, which effectively superceded study #3.
- page 112: In the listing for county spending on health, there is no mention of natural logarithms, nor should any use of natural logarithms be inferred. There are counties with zero spending on health, especially in the earlier years, so taking logs would result in many missing values. (The works by John Ferrell and cited in the published paper discuss health-spending patterns in the region at that time.) This variable is simply the change in the level of per-capita spending, from 1902 to 1932. The use of changes, rather than levels, was clear in the working papers (e.g., Table IV, Panel B, of the Stigler Center WP#205), but such mention was unfortunately omitted when shinking the paper for final publication.
- Stigler Center WP#205, page 17, lines 7–8: In describing control variables for Panel B of Table IV, I stated that the results presented there had variables for "changes in health and sanitation spending by county governments over various intervals." The "over various intervals" part is incorrect; the variable indicated is for the change in the level of per-capita spending from 1902–32. This language was inadvertently included from a previous version of the paper, in which I included controls for such changes from 1902–32, 1902–12, and 1913–32.

Figure 1: Comparison of Figure II ("Hookworm Eradication and School Attendance, 1900–50") in published paper with version that includes 1870–80

Panel A: Figure II from published paper, adjusted to match scale in Panel B



Panel B: Figure III from Stigler Center Working Paper #205, which includes estimates from 1870–80


.1
0
-.1
-.2
1860
1880
1900
1920
1940
1960