Fetal Iodine Deficiency and Schooling: A Replication of Field, Robles and Torero (2009)[*]

Niklas Bengtsson

Uppsala University, Uppsala, SE-75236, Sweden

niklas.bengtsson@nek.uu.se

Fredrik Sävje

Yale University, New Haven, CT 06520, USA

fredrik.savje@yale.edu

Stefan Swartling Peterson[†]

Uppsala University, Uppsala, SE-75236, Sweden

stefan.peterson@kbh.uu.se

Abstract

Scholars have theorized that congenital health endowment is an important determinant of economic outcomes later in a person's life. Field, Robles and Torero [2009, *American Economic Journal: Applied Economics*, 1(4), 140–169] find large increases in educational attainment caused by a reduction of fetal iodine deficiency following a set of iodine supplementation programs in Tanzania. We revisit the Tanzanian iodine programs with a narrow and wide replication of the study by Field et al. We are able to exactly replicate the original results. We find, however, that the findings are sensitive to alternative specification choices and sample restrictions. We try to address some of these concerns in the wide replication; we increase the sample size fourfold and improve the precision of the treatment variable by incorporating new institutional and medical insights. Despite the improvements, no effect is found. We conclude that the available data do not provide sufficient power to detect a possible effect since treatment assignment cannot be measured with sufficient precision.

Keywords: Education; fetal origins hypothesis; iodine deficiency; prenatal exposure; replication

JEL classification: I12; I21; J16; O15

[*] We extend our gratitude to Douglas Almond, Sebastian Axbard, Per Engström, Rita Ginja, Linna Martén, Eva Mörk, three anonymous referees and seminar participants at Uppsala University and EEA Gothenburg 2013 for comments and suggestions. Bengtsson and Sävje acknowledge financial support from the Swedish International Development Cooperation Agency through grant Uforsk SWE-2012-109.

[†] Also affiliated with UNICEF.

I Introduction

How does improved health *in utero* affect educational outcomes later in life? Field, Robles and Torero (2009, hereafter FRT) seek to shine light on this question by estimating the effect of iodized oil capsule (IOC) distribution programs launched 1986 in Tanzania on educational attainment. The programs are noteworthy for their combined size. The targeted districts contain a quarter of the Tanzanian population, and a total of 6 million capsules were distributed (Peterson et al., 1999). FRT use the lagged roll-out of the programs for identification. They find large effects. The intent-to-treat estimates indicate that being protected from iodine deficiency in utero increased educational attainment by 0.35 years on average. The estimated effect is particularly large for girls, with an estimated intent-to-treat effect of 0.59 years. Using the coverage rates of the programs to derive the hypothetical case of achieving full coverage, the authors calculate that "the expected increase in grade attainment for a child protected from fetal iodine deficiency is a minimum of 0.73 years."

FRT conclude that countries with similar iodine supplementation programs have experienced a 4.8% increase in school participation as an effect of the supplementation. The results provide evidence of a causal link between the geographical health environment and economic development. In particular, FRT's study is one of the first to establish a link between fetal health and outcomes later in life using quasi-experimental methods (see, e.g., Almond and Currie, 2011 for a review of this literature). Since the publication of the original article, the empirical strategy and data have been used to address other aspects of child health and household behavior (Adhvaryu and Nyshadham, 2016).

We revisit the Tanzanian experience by replicating FRT's study. We first attempt to exactly reproduce FRT's results using the original data. The exercise is successful, but it highlights a series of sample restrictions and other specification choices that warrant further investigation. We examine the robustness of the results with respect to these choices and find that the large estimates rely on the exact specification used in the original study. The estimated effects are smaller and not statistically significant at conventional levels with alternative specifications that are well-motivated with respect to both identification and statistical efficiency.

We investigate seven aspects of FRT's specification. We are, however, primarily concerned about four choices for which we fail to find a clear motivation. First, FRT use within-household birth order as a control variable, but their code does not accurately sort the respondents by birth date. The resulting birth order is essentially random. Second, they use the education level of the spouse of the household head as a control variable. The variable is coded as missing in households where the head does not have a spouse. As a result, all single-parent households are dropped from FRT's analysis. The dropped observations constitute approximately 21% of the sample. Third, FRT exclude 14-year-olds from the analysis. This cohort was targeted by the IOC programs, and FRT include 14-year-olds in their supplementary analyses. Including the 14-year-olds increases the sample size by 22%. Fourth, treatment is defined as whether the respondent was protected from iodine deficiency during pregnancy. The respondent's birth date is therefore of great importance when deriving the treatment variable. In their main analysis (using the Tanzanian Household Budget Survey that started surveying in 2000), FRT derive the birth year as 2000 minus the reported age. This approach ignores the timing of the interviews (which is reported at the monthly level), and, in particular, it disregards that a quarter of the sample was surveyed in 2001. FRT's results are fairly robust to each of these issues addressed in isolation. Combined, however, the four issues produce estimates that are 3.5 times higher than a specification that only addresses these aspects of the analysis. With such a specification, the estimated effects are close to zero and not statistically significant despite lower estimated standard errors.

Prior studies have documented both cognitive and health effects of fetal iodine deficiency, and it is plausible that some of these effects carry on to schooling outcomes. A relevant critique of the null result in our narrow replication is that it would be possible to detect positive effects with better data. In particular, the actually assigned treatments are never observed in the original study, and FRT's independent variable is the probability of treatment as approximated by a model. Based on evidence in the medical literature, their model assumes that the first trimester of the pregnancy is particularly sensitive to iodine deficiency, and they exploit that the IOC programs provide variation in protection from such deficiencies. The programs are, however, poorly documented; their start dates and lengths are not known, and no program was able to reach all targeted people. Furthermore, the existing survey data do not provide accurate information of when and where the respondents were born. It is therefore not known when the respondents were in utero. Finally, exact biological and medical details about iodine intake and depletion are needed for an accurate model of the protection from deficiency, but such details are not known. These uncertainties introduce measurement error that could lead to imprecision and attenuation bias.

We attempt to mitigate these concerns in a wide replication of the original study. We keep the fixed-effects approach from FRT but use new medical and institutional insights to improve the model used to derive treatment probabilities. We also extend the data with four additional data sets. The sample is almost four times as large as in FRT. The changes should improve our ability to detect non-zero effects, but the estimates are close to zero and not statistically significant at conventional levels. We conclude that the existing data do not provide evidence of whether the fetal origins hypothesis holds.

The rest of the paper is structured as follows. We provide a short background on iodine deficiency disorders and the IOC programs in the next section. We also discuss the identification strategy used in FRT and their results in light of the current medical literature on iodine deficiency. Section 3 presents the results from our narrow replication. In Section 4, we discuss how to estimate the probability of protection from iodine deficiency using the available data, and we report the results from the wide replication. Section 5 concludes.

II Background

Iodine is a chemical element and a micronutrient important for the synthesis of thyroid hormones. Thyroid hormones play a vital role in the regulation of metabolism and are essential for growth and development in humans. The conditions resulting from low levels of thyroid hormones due to iodine deficiency are referred to as *iodine deficiency disorders*. The disorders are varied and involve conditions with serious consequences for the well-being of those affected. The focus in FRT is disorders resulting from iodine deficiency during pregnancy, which may affect fetal development.

The negative effects of severe iodine deficiency were discovered early in the 20th century. It was established that significant deficiency during pregnancy leads to *cretinism*, which entails both severe health problems and cognitive disabilities for the child. A string of studies towards the end of the last century suggested that milder forms of iodine deficiency during pregnancy could also be associated with hindered cognitive development (see, e.g., Dugbartey, 1998; Pop et al., 1999; Haddow et al., 1999; Lavado-Autric et al., 2003; Pop et al., 2003). The evidence is, however, scant, and there is an active discussion concerning the exact period of the pregnancy that is sensitive to mild deficiency, which skills are affected and how persistent the effects are (Zoeller and Rovet, 2004). No experimental evidence exists on the long-term consequences of milder forms of deficiency. To the best of our knowledge, FRT was the first study to investigate such effects using rigorous quasi-experimental methods. This made the study an important contribution to both economics and medicine.

FRT's identification strategy exploits the fact that the early pregnancy appears particularly sensitive to iodine deficiency (Pharoah et al., 1971; Cao et al., 1994). In particular, during the first trimester, the fetus cannot itself synthesize thyroid hormones and is completely reliant on the mother's hormone production. FRT use the delayed roll-out of IOC programs in Tanzania to exploit variation in protection from iodine deficiency in utero. Following the increased awareness of the benefits of deficiency prevention, several large-scale IOC supplementation programs were introduced in the late 1980s in Tanzania. The purpose of the programs was to supply IOCs to the most affected populations until universal salt iodization began in the early 1990s (Assey et al., 2009). In total, 27 districts were targeted.

The intended structure of the programs was to distribute IOCs every second year starting in 1986. For each distribution round, all people aged from 2 to 45 years were to be given an IOC containing 400 mg of iodine, and children aged from 12 to 23 months were to be given a dose of 200 mg of iodine (Peterson et al., 1999). Delays in both initial and repeated distribution rounds were common due to administrative problems, and only 10 districts received their initial round in 1988 or earlier. The average coverage rate was 64%, and full coverage was never achieved in any district. The programs nevertheless reached a substantial number of individuals. A conservative estimate is that the programs provided 12 million person-years of protection from iodine deficiency (Peterson et al., 1999).

By the early 1990s, the salt iodization programs had started.[1] During this period, the focus of the IOC programs shifted from districts with high levels of iodine deficiency to districts not yet reached by the salt iodization programs, namely districts where fewer than 75% of households had access to iodized salt (Peterson et al., 1999). Thus, after 1990, the absence of an IOC program does not necessary indicate that the population is unprotected from iodine deficiency (Assey et al., 2009).

Empirical strategy

FRT's data consist of children surveyed at school-going age who potentially benefited from the IOC programs in utero. The data set in their main analysis is constructed from the *Tanzanian Household Budget Survey* conducted in 2000 and 2001 (hereafter, THBS 2000). They exploit the lagged roll-out of the IOC programs to identify the causal effect of protection from iodine deficiency in utero, effectively comparing treated and untreated children in the targeted cohorts. They adopt a fixed-effects approach where treatment is considered to be as-if randomly assigned conditional on district and birth date effects. Their exact regression specification is:

$$Y_{idt} = \beta_0 + \beta_1 T_{dt} + \boldsymbol{\beta}_2 \mathbf{X}_{idt} + \mu_d + \lambda_t + \varepsilon_{idt}, \tag{1}$$

where Y_{idt} denotes educational attainment as measured by the number of completed grades for respondent i born in district d in year t. The treatment variable T_{dt} is the calculated probability of in utero protection from iodine deficiency for an individual born in district d in year t. Thus, β_1 is the coefficient of interest. μ_d and λ_t are district and birth year fixed effects, and $\mathbf{X}_{idt}$ is a vector of control variables measured at either the individual or household level.[2] FRT allow the error term ε_{idt} to be correlated within each cohort in a district by clustering the standard errors at the district-birth year level.

The treatment variable T_{dt} is the probability of being protected from iodine deficiency in utero. Ideally, this variable would be an indicator taking value one if a respondent was protected during the relevant part of the pregnancy, zero otherwise. Such an indicator requires detailed information on the timing, coverage and intensity of the IOC programs; the time and place of birth of the respondent; and the need for and depletion of iodine during pregnancy. None of these factors are known. Instead, FRT's treatment variable is an approximation of the probability of protection based on the limited data available.

[1] Data on the exact dates of the salt iodization programs are, to our knowledge, not available.

[2] In detail, FRT include a "correction factor" intended to capture misspecification in their treatment model; an indicator of whether the respondent's mother was 23 years old or younger at the time of birth; indicators for the respondent's age, gender and birth order; the number of children in the household; the household head's and spouse's education level; indicators for the month of the interview; indicators of whether the household resides in an urban setting, whether the household owns their home, whether the main dwelling has a grass roof and whether the household experienced food problems; and distance to nearest secondary school and nearest health clinic.

Original results

FRT find that the deficiency protection the IOC programs offered increased educational attainment by 0.35 years on average. Given that the programs did not reach all targeted people, this is best seen as an intent-to-treat effect and, thus, a lower bound of the true effect. Scaling up the point estimate by the programs' coverage rate, FRT conclude that iodine deficiency protection in utero increases educational attainment by 0.73 years. This figure is, however, likely also an underestimated effect. First, the remaining imprecisions in the treatment variable would lead to effect attenuation not accounted for by FRT's adjustments (see the discussion in subsequent sections). Second, FRT's strategy presumes that children not protected during the first trimester did not benefit from the IOC programs at all. While the first trimester appears to be particularly sensitive to iodine deficiency, existing studies do not allow us to rule out that protection has an effect later in the pregnancy or even after birth (Zoeller and Rovet, 2004 provide an overview). Treatment is, thus, likely to spill over into the control group and further attenuate the estimates.

FRT highlight two aspects of their results. First, they find no effects on health outcomes. They interpret this as an indication that the estimated reduced-form effect primarily affects educational attainment through an effect on cognitive ability. This is in line with the discussion in the medical literature, where mild iodine deficiency during pregnancy has been found to primarily affect cognitive development. However, the magnitude of the implied per-protocol effects reported by FRT resembles those for severe deficiency, and severe deficiency would also affect health outcomes (see, e.g., Pharoah and Connolly, 1987). It is, however, possible that small biological effects interact with the social environment in the long term and lead to large effect on outcomes such as educational attainment.

Second, FRT find considerably stronger effects for girls (point estimate of 0.59 years) than for boys (0.19 years). In fact, the effect is only statistically significant for girls. FRT cite two laboratory studies on rodents as support for their findings of gender differentials. Chan et al. (2005) report different responses between male and female guinea pigs in the regulation of thyroid hormone receptors, and Friedhoff et al. (2000) observe behavioral differences between male and female rats when their mothers' thyroid glands were completely removed. To our knowledge, no study on humans prior to FRT had found large gender differences of either mild or severe iodine deficiency.[3]

[3] The evidence from studies following FRT is mixed. Politi (2010) investigates the effect of iodine deficiency on educational attainment in humans in Switzerland and does not find support for a stronger effect for females. Murcia et al. (2011) find gender differences for one-year-olds in Spain (the study, however, lacks a clear and credible identification strategy). Both Adhvaryu et al. (2018) and Feyrer et al. (2017) investigate the effects of salt iodization programs in the U.S. using fixed-effects methods similar to FRT. The former study finds large effects for women but no effects for men. The latter does not investigate the effect for women due to data limitations but finds large effects for men. Finally, Deng and Lindeboom (2018) find large effects of salt iodization on the educational attainment of women (but not for men) in China.

III Narrow replication

Our aim in this section is to replicate the findings of FRT using the original data sources. In a first part, we attempt to reproduce the exact results of FRT without changes to the original specification. In a second part of the narrow replication, we make small changes to the specification to address issues discovered during the replication effort.

FRT share the assembled data files and the code for the main analysis through the publishing journal's website. However, documentation on how the data were cleaned and how variables were defined is not publicly available. We managed to replicate the undocumented parts of the analysis by trial and error. Our recreated data set was compared observation by observation to the original to detect any differences.

We were able to replicate the FRT data set using the original data sources except for the following variables:

1. We did not attempt to replicate the outcome variables in the supplementary analyses. That is, we did not replicate the sickness variables presented in FRT's Table 6 or the alternative schooling measures presented in FRT's online appendix.

2. We did not attempt to replicate the matched total goiter rate variable that FRT use in the supplementary analysis in their Table 4.

3. We attempted but were unable to to fully replicate the program coverage rate variable used in FRT's main analysis. We could not find the procedure they used to assign program coverage when a respondent was affected by several overlapping IOC programs. We found a procedure that replicated the coverage rate variable for children aged 10-14 (i.e., including FRT's main sample of 10-13-year-olds). However, this procedure does not replicate the variable for the extended sample not used in this replication.

4. We could not replicate FRT's birth seasonality information, which is used to calculate treatment probabilities. We therefore use the partially assembled data set provided by FRT to construct the birth seasonality variable for the narrow replication. This is the only variable we use in the replication that is not derived from original data sources.[4]

5. Our calculated treatment probabilities differ slightly from those in FRT's data. The discrepancy is very small (the maximum difference is at the eighth decimal place) and is likely due to differences in floating point arithmetic in the computer architecture used for the calculations.

In sum, although we failed to recreate some parts of the FRT sample, our reconstructed data set is sufficient for a full replication of FRT's main analyses. Table 1 presents the replicated results, which correspond to the first two parts of Table 3 in FRT. The results are replicated exactly.

[4] In the wide replication, we follow FRT's stated approach and estimate birth seasonality from the original data using one of the additional data sets (DHS 2004) that contains detailed birth date information.

Table 1: Replication of the main results in FRT

	(1)	(2)	(3)	(4)
All	0.347**	0.246**	0.559***	0.632**
	(0.148)	(0.114)	(0.197)	(0.283)
	[1395]	[1395]	[1395]	[690]
Females	0.594***	0.429***	0.824***	1.611***
	(0.170)	(0.135)	(0.262)	(0.461)
	[678]	[678]	[678]	[192]
Males	0.190	0.134	0.384	1.045*
	(0.160)	(0.136)	(0.240)	(0.548)
	[717]	[717]	[717]	[208]

Notes: The table replicates the analysis presented in parts 1 and 2 in Table 3 in FRT. Each cell presents the result from a separate regression. The outcome in all columns is the number of completed grades. The first column presents the results from FRT's main specification where the treatment variable is the estimated probability of being protected from iodine deficiency. The second column changes the treatment variable to FRT's binary treatment indicator. The third column interacts the treatment variable from the first specification with the coverage rate of the IOC programs to adjust for incomplete coverage. The last column retains the interacted treatment specification from column three but uses household fixed effects rather than district-level indicators. Each cell reports point estimates, estimated standard errors clustered at the district-cohort level within parentheses and number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

Potential issues discovered during the replication

Several aspects of FRT's data compilation and analysis struck us as potentially problematic during the replication effort. Only one of these aspects is unambiguously an error (see issue 1 below). The remaining aspects are, to varying degrees, judgement calls. Some of these issues became apparent only in light of new institutional and medical insights that presumably were unavailable to FRT. We want to address all these potential concerns, but we do so in two separate sections. In this narrow replication, we focus on the issues we consider to have less of a judgement call character. These are aspects of the analysis that we believe many economists would consider to be "empirical Pareto improvements" in the sense that they improve some aspect of the analysis without making any other aspect of the analysis worse off (this precludes, e.g., all changes that involve bias-variance trade-offs). They are also aspects that can be easily changed without requiring large changes to other parts of FRT's analysis (this precludes, e.g., changes to the treatment model). Aspects of FRT's analysis that involve trade-offs or are more clearly judgement calls are discussed in our wide replication in Section IV.

We investigate seven aspects of FRT's analysis in the narrow replication. The first four (labeled 1-4 below) are not discussed by FRT, and we fail to find motivations for these aspects given the identification strategy. The two subsequent specification choices (labeled 5 and 6 below) are documented by FRT, but we still fail to find (or understand) how the identification strategy provides motivation for them (especially considering the large reduction in sample size they entail). The final issue is a purely econometric issue concerning the estimation of standard errors and p-values.

Below, we have ordered the potential issues from least to most contentious based on how we believe most economists would view them. Apart from investigating the issues in isolation and together, we also investigate how the estimated effects change as we cumulatively address increasingly contentious issues.

The final issue (number seven) does not follow this pattern because it is, arguably, one of the least contentious issues. The consequence of this issue is, however, quite predictable (i.e., higher standard errors and p-values), and we opt to present it last for that reason.

In detail, the issues we address are as follows:

1. When deriving birth order, FRT use the `sort` command in Stata (StataCorp, 2007) to sort observations in ascending order by household ID and descending order by age. This command is, however, not programmed to accurately sort data in descending order, and the variable is therefore constructed from observations sorted in a pseudo-random way. We use the `gsort` command, which yields the correct birth order.

2. FRT add the educational level of the spouse of the household head as a covariate. Spouse education is, however, not reported when there is no spouse in the household. As the statistical software used by FRT silently drops observations with missing values, FRT's specification implicitly excludes all single-parent households. Approximately 21% of the children in the sample live in such households. We could not find the motivation for this restriction in FRT. We add the excluded children and include an indicator for missing spouses as a control variable.

3. FRT only include children aged between 10 and 13 years in their main analysis. Children aged 14 years are excluded.[5] FRT do not discuss why 14-year-olds are dropped. The general motivation for the age restriction is that the chosen cohorts were potentially affected by the IOC program in utero and that the age span represents "the modal age of enrollment." As we discuss below, it appears that the 14-year-olds satisfy both these criteria. Moreover, FRT include children aged 10-14 in their supplementary investigation using a different data set (Part 3 of their Table 3).

   Since the THBS survey was conducted in 2000 and 2001, children with a reported age of 14 were born either in 1986 or 1987. These children would have benefited from iodine deficiency protection in the counterfactual setting where all programs started on time. In practice, IOC programs had started only in a few districts when this cohort was in utero, and the estimated treatment probabilities are generally small. However, this does not constitute a threat to identification. The low estimated treatment probabilities could in fact be beneficial for identification. Programs were frequently repeated in some of the districts, and all children aged 10-13 have treatment probabilities close to one in these districts (see Figure S1 in the online appendix). As a result, there is no effective control group in these districts, and the effects are estimated by extrapolation from children with high, but not full, protection. Similar to a classical difference-in-difference approach, if children aged 14 have low treatment probabilities, their inclusion will lead to more credible identification by making the estimates more precise and less model-dependent.

[5] FRT investigate the effects when including younger cohorts in their robustness checks, but they never include children aged 14 when investigating the THBS 2000 data.

FRT's second motivation—the modal age of enrollment—suggests that there is insufficient variation in the outcome variable for children aged 14 to warrant their inclusion. This appears not to be the case: for children aged 13, average grade attainment is 3.2 grades with a standard deviation of 1.8. For children aged 14, the average is 4.1 grades with a standard deviation of 2.2.

We include children aged 14 in the analysis. This increases the sample size by approximately 22%. As FRT calculate the birth order variable after imposing the age restriction, we recalculate birth order with the new cohort included.

4. FRT do not use all available data when estimating the respondents' birth date. They derive birth year for the THBS 2000 survey as 2000 minus the reported age in years at the time of the interview. This procedure disregard both the reported survey year and month (the survey took place from May 2000 to May 2001). To illustrate the problem, consider three 12-year-old respondents. The first is interviewed in May 2000, the second in October 2000 and the third in May 2001. Possible birth months are: May 1987–May 1988 for the first; October 1987–October 1988 for the second; and May 1988–May 1989 for the third. FRT's approach assigns the average treatment probability during the period January 1988–Dec 1988 to all three respondents. This is a poor approximation of the actual intervals and may exacerbate the measurement error in treatment. We exploit all information on the timing of the survey to calculate treatment probabilities as the average over the possible birth months (see the wide replication for additional details).[6]

5. FRT drop all children that cannot be linked to a unique mother in the household. The main motivation for this restriction is to avoid orphan children, which arguably are more mobile than other children in the household (making such observations more prone to measurement error due to inter-district migration). Orphanhood is, however, reported in the data, and these children can easily be excluded from the sample without matching them to mothers. A second motivation for linking children to unique mothers is to calculate the mothers' age at birth, which FRT use as a control variable. However, as we discuss in the wide replication, this control variable is not necessary for identification and lacks clear medical and institutional motivation. A consequence of imposing this restriction is that nearly all children in polygamous households are excluded.

   We include children that cannot be linked to a unique mother. The "young mother" covariate is missing for these children, so we include an indicator for missingness as a control variable.

[6] We keep the birth year fixed effect specification FRT use, and the birth year is estimated using the same approach as in the wide replication (also accounting for birth seasonality on the regional level). An alternative approach is to use the interaction of age, survey year and month as fixed effects. This would more flexibly adjust for potential selection effects by cohort, but the specification is further from FRT. It also raises concerns how to define other variables relying on the birth year estimate, such as the clustering variable. We do not take issue with FRT's argument that assignment is as-if random given district and birth year fixed effects, and in an effort to remain as close as possible to their original specification, we opt not to change the fixed effects. In the online appendix, we report the results presented in Table 2 when using the interacted cohort fixed effects (Table S1). We also report the results when, as in FRT, assigning treatment probabilities on yearly basis, but improving the estimation of birth year (Tables S2 and S3). The qualitative conclusions are unchanged.

6. Motivated by that nonresident children in the household do not have schooling outcomes recorded in the survey, FRT exclude all children that are not sons or daughters of the household head or spouse. The restriction unnecessarily excludes large groups of children permanently living in the household. Grandchildren of the household head is the largest group that is excluded. The necessary variables are readily available for all of these children. We include children that are grandchildren of the household head.

7. Finally, FRT cluster their standard error estimators at the district-cohort level. As FRT's treatment variable is the same for respondents in the same cohort and district, such clustering is reasonable. However, FRT's treatment model estimates that protection from iodine deficiency continued for several years after an IOC program ended. In other words, the treatment variable is also correlated across different cohorts in the same district. As discussed by Bertrand et al. (2004) and others, such serial correlation will lead to an underestimation of the standard errors. This increases the risk of a Type I error (i.e., the rejection of the null hypothesis of no effect when it is, in fact, true). Furthermore, the roll-out of the IOC programs exhibits some geographical correlation between districts. In the same way, such spatial dependencies will also lead to downward bias in the estimated standard errors.

   It is beyond the scope of this replication study to address possible spatial dependencies, but we attempt to address concerns of possible serial correlation. We cluster the standard error estimators at the district level rather than the district-cohort level. This approach has the disadvantage that the number of clusters is greatly reduced (to 25).[7] This makes us concerned that the asymptotic approximation of the null distribution is inappropriate given the Wald-type test used to derive significance levels. To address the issue, we derive the null distribution using wild cluster bootstrap. We use the procedure outlined in Appendix B in Cameron et al. (2008) and impose the null hypothesis of zero effect when deriving the null distribution, which we estimate using 1,000 bootstrap iterations.

Table 2 presents the estimated effects of the IOC programs when these issues are addressed. The first panel presents the effect of addressing each issue in isolation. The left-most column (labeled "FRT") is the estimated effects using FRT's original specification (i.e., it is exactly the first column in Table 1 in this paper or the first panel in Table 3 in FRT). Each of the remaining columns correspond to the seven issues discussed above (the labels are the number in the list). The results are fairly robust to the issues when addressed in isolation. The one exception is when 14-year-olds are included (column "3"), which reduces the effect for girls by almost half.

The second panel of Table 2 explores the estimated effects when several issues are addressed together. The left-most column (labeled "All issues") reports the estimates when all issues are addressed. We see

[7] Following FRT, we drop 2 of the 27 districts where the IOC programs started after 1990.

Table 2: Addressing potential issues found in FRT's analysis

Panel 1: Each issue addressed in isolation

	FRT	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All	0.347**	0.340**	0.262**	0.244*	0.278*	0.329**	0.434***	0.347*
	(0.148)	(0.145)	(0.119)	(0.127)	(0.150)	(0.139)	(0.135)	(0.176)
	[1395]	[1395]	[1691]	[1705]	[1395]	[1558]	[1500]	[1395]
Females	0.594***	0.627***	0.521***	0.328**	0.556***	0.573***	0.687***	0.594**
	(0.170)	(0.171)	(0.157)	(0.160)	(0.157)	(0.183)	(0.172)	(0.185)
	[678]	[678]	[819]	[839]	[678]	[760]	[730]	[678]
Males	0.190	0.170	0.113	0.195	0.063	0.156	0.222	0.190
	(0.160)	(0.156)	(0.147)	(0.163)	(0.190)	(0.151)	(0.149)	(0.182)
	[717]	[717]	[872]	[866]	[717]	[798]	[770]	[717]

Panel 2: All issues addressed except one

	All issues	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All	0.082	0.092	0.200	0.259	0.107	0.098	0.112	0.082
	(0.135)	(0.135)	(0.133)	(0.151)	(0.102)	(0.136)	(0.135)	(0.116)
	[2610]	[2610]	[2036]	[2139]	[2610]	[2056]	[2333]	[2610]
Females	0.074	0.082	0.328*	0.479**	0.122	0.191	0.169	0.074
	(0.149)	(0.153)	(0.158)	(0.168)	(0.157)	(0.141)	(0.146)	(0.157)
	[1286]	[1286]	[1000]	[1045]	[1286]	[1009]	[1143]	[1286]
Males	0.079	0.091	0.090	0.065	0.087	0.017	0.056	0.079
	(0.162)	(0.150)	(0.162)	(0.172)	(0.127)	(0.180)	(0.168)	(0.150)
	[1324]	[1324]	[1036]	[1094]	[1324]	[1047]	[1190]	[1324]

Panel 3: Addressing issues cumulatively

	FRT	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All	0.347**	0.340**	0.250**	0.155	0.098	0.112	0.082	0.082
	(0.148)	(0.145)	(0.118)	(0.112)	(0.122)	(0.118)	(0.116)	(0.135)
	[1395]	[1395]	[1691]	[2056]	[2056]	[2333]	[2610]	[2610]
Females	0.594***	0.627***	0.542***	0.258*	0.191	0.169	0.074	0.074
	(0.170)	(0.171)	(0.156)	(0.147)	(0.148)	(0.155)	(0.157)	(0.149)
	[678]	[678]	[819]	[1009]	[1009]	[1143]	[1286]	[1286]
Males	0.190	0.170	0.088	0.086	0.017	0.056	0.079	0.079
	(0.160)	(0.156)	(0.144)	(0.150)	(0.164)	(0.152)	(0.150)	(0.162)
	[717]	[717]	[872]	[1047]	[1047]	[1190]	[1324]	[1324]

Notes: The table presents the results from the main specification of FRT when the issues discovered in the narrow replication are addressed. In the first panel, each issue is addressed individually. That is, in each column we address only the corresponding issue and leave all other aspects as in FRT's original specification. The first column (labeled "FRT") presents the estimates from FRT (i.e., the first column in Table 1). The subsequent columns change one aspect of the specification, where the numbers correspond to the numbered issues in the text. The second panel presents the results when all issues except one are addressed. The first column (labeled "All") addresses all seven issues discussed in the text. The subsequent columns address six of the seven issues, where the omitted issue is indicated by the column number. The third panel addresses the issues cumulatively. The first column (labeled "FRT") is FRT's exact specification. Cumulatively, each subsequent column addresses one issue discussed in the text. The final column (7) addresses all issues, and it is exactly the results presented in the first column of the second panel. In this way, the third panel connects the first two panels. In summary, the issues discussed in the text are as follows: (1) using `sort` to derive birth order; (2) discarding children in single-parent households; (3) discarding children that are 14 years old; (4) not using all birth date information when calculating treatment probabilities; (5) discarding children that cannot be linked to a unique mother; (6) discarding grandchildren of household heads; and (7) ignoring serial correlation by clustering the standard errors at the district-cohort level. Each cell reports point estimates, estimated standard errors within parentheses and the number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

that the effects are small and not statistically significant at conventional levels. The remaining seven columns present the estimates when all but one of the issues are addressed. The label of the column indicates the issue that is not addressed (e.g., column "4" addresses issues 1-3 and 5-7). While the estimates vary from column to column, the picture is qualitatively the same: the estimated effects are considerably closer to zero. The lack of robustness is not driven by a single issue, but Column 3 stands out with more modest reductions. The inclusion of 14-year-olds appears to be the most consequential aspect of FRT's specification.

The final panel presents the estimates from specifications that cumulatively address the issues. The left-most column (labeled "FRT") is the estimated effects using FRT's original specification. The subsequent columns each address one of the potential issues in order. For example, column 3 shows the estimates when addressing issues 1, 2 and 3. We see that the first issue (the use of the `sort` command) is largely inconsequential for the results. The next two issues, however, lead to larger changes. When we include children living in households with a single parent (column "2") and children aged 14 (column "3"), the estimates are less than half of those reported by FRT and not statistically different from zero at conventional significance levels. Models 4-6 reduce the point estimate further. The last issue (accounting for serial correlation between cohorts) leads to small changes in the estimated standard errors, and perhaps unexpectedly, a decrease for girls. The point estimates are, however, already insignificant at this point, and the changes do not affect the qualitative conclusions.

Another way of presenting the combined issues is to categorize them by type. Table S4 in the online appendix presents the results when addressing the issues discussed above in groups. We consider all combinations of changes to the sample restrictions (issues 2, 3, 5 and 6); changes to the definition of variables used in the analysis (issues 1 and 4); and changes to the estimation procedure (issue 7). The analysis shows that all three types of specification issues affect the results, but FRT's sample restrictions appear to be most consequential.

We conclude that FRT's findings rely on a set of specific specification choices. FRT do not use all cohorts affected by the IOC programs when estimating the effect of the IOC program. Furthermore, among the cohorts they include, they restrict the study to children of household heads in non-polygamous families where both parents are present. We fail to find medical, institutional or econometric reasons to restrict the study to this subsample. When the sample is broadened, the large effects estimated by FRT disappear.

While our replication of FRT's study does not lend support to the conclusion that the IOC programs had a positive effect, we also cannot conclude the opposite; it is possible that the programs had a fairly large effect that we are unable to detect. There are two main issues. The first issue is the lack of statistical power due to the low sample size. The second issue is measurement error. We do not observe which respondents were treated; at best, we can derive a rough estimate of treatment. This introduces measurement error that may bias the estimates towards zero and reduces our ability to detect non-zero effects. FRT discuss one such source—that the program did not reach everyone in a targeted district—and inflate their baseline result by 60% to account for this. However, this is just one among a host of uncertainties that introduce errors in the treatment variable. In the wide replication, we address these problems by improving the measurement of the treatment variable and increasing the sample size.

IV Wide replication

Our wide replication extends FRT's analysis in two directions. First, we pool data from five surveys containing information on educational attainment in Tanzania. This is in an effort to maximize sample size and, thereby, precision.[8] Second, we make additional changes to FRT's specification not explored in the narrow replication. Our main concern with FRT's analysis is the potentially large measurement error in the treatment variable. We attempt to improve the treatment model by incorporating new institutional and medical insights.

Additional data sources

We investigate the effects of protection from iodine deficiency using data on the cohorts born between 1986 and 1990 from a collection of five household surveys conducted in Tanzania between 1999 and 2010. Three are the *Demographic and Health Surveys* (DHS) for the years 1999, 2004-2005 and 2009-2010 (hereafter labeled DHS 1999, DHS 2004 and DHS 2010). The fourth is the first wave of the *National Panel Survey* conducted in 2008-2009 (labeled NPS 2008), and the fifth is the *Tanzanian Household Budget Survey* of 2000-2001 (THBS 2000). FRT use the THBS 2000 data set in their main analysis and the DHS 2004 data set in their supplementary analysis.

The surveys differ somewhat in their construction and focus, but they are all representative of the whole of Tanzania and contain the relevant data needed for the analysis.[9] All surveys were conducted by the National Bureau of Statistics in Tanzania with the involvement of international organizations.

We investigate each survey separately, as well as all surveys collected in a pooled analysis. The pooled analysis benefits from a maximized sample size, increasing our ability to detect possible effects. However, as the surveys are conducted in different years, the respondents will be of different ages (typically, older) when all data are pooled compared to the respondents in FRT analysis. If the effect changes as respondents grow older, the average effect in the pooled sample will differ from the effect that FRT

[8] We would ideally increase the sample size by adding additional districts or cohorts since that is the level we cluster on. This is not possible; we already include all districts and cohorts targeted by the IOC programs. The second best is to increase the number of observations within clusters, which increases precision as long as the within-cluster correlation is not perfect.

[9] The main differences are that the DHS surveys also conduct a more in-depth interview with a selected number of respondents in each household, mainly women. While the additional information available for these respondents is not vital to the analysis, it provides more accurate measurements of birth dates. The NPS 2008 is the first wave in a long-running panel data set under construction, with highly detailed data. The sample size is subsequently lower than in the cross-sectional surveys. The THBS 2000 is, instead, a large survey with less-detailed information.

estimate. The difference could be both negative (e.g., if untreated children attain the same level of education as treated children but at a slower rate) and positive (e.g., if untreated children tend to drop out of school and treated children tend to continue their education). The low average educational attainment and high drop-out rates at baseline suggest that the effect may grow with age.

In our new treatment specification, as detailed below, we use information from a number of goiter surveys that were conducted in Tanzania during the 1980s. We obtain these data from the original data sources. The purpose of these surveys was to estimate how widespread and severe iodine deficiency was in Tanzania. Due to the complexity and diversity of the disorders following from iodine deficiency, a common procedure is to use goiter prevalence, which is a highly noticeable iodine deficiency disorder, as a proxy for iodine deficiency in general. In particular, we use the district average *total goiter rate*—the ratio between individuals with an enlarged thyroid gland and the total population—measured by goiter prevalence amongst school children as a proxy for iodine deficiency.[10]

Results with additional data using FRT's specification

Table 3 presents the estimated effects of the IOC programs for the pooled data set and for the five surveys separately when using the closest possible specification to that used by FRT. It is not possible to use FRT's exact specification since the data sets contains different sets of variables. We attempt to impose as minimal changes as possible. If a variable used in FRT's specification is missing from a data set, we set the value to zero and add an indicator for missingness to the specification. In the pooled analysis, we include both cohort and age fixed effects (FRT only use age effects). Since we observe the same cohorts in different years in the pooled analysis, age and birth year are not collinear. For identification, only the cohort fixed effects are needed, but since educational attainment is strongly correlated with age, the age fixed effect should greatly improve precision. Note that none of the potential issues discussed in the narrow replication (nor those discussed below) are addressed in Table 3. In particular, the standard errors are estimated by clustering at the district-cohort level as in FRT's original analysis.[11]

Each cell of the table presents the results from a separate regression. Columns indicate different subsamples (girls, boys and both), and rows indicate different data sets. The coefficients should be interpreted as the estimated impact of iodine deficiency protection in utero on educational attainment. The third row (labeled "THBS 2000") exactly replicates the results in FRT; the specification and data set are identical. We see large and significant effects, especially for girls. The results in the other rows paint another picture. The estimates are occasionally large, but none is statistically significant at conventional levels. The sample sizes using the other data sources are, however, smaller, and we would expect more

[10] One of the treated districts, Kasulu in the Kigoma region, was not included in the goiter surveys. We impute the missing value with the average of the treated districts in the Kigoma region.

[11] In the online appendix, we also present the results when two of the specifications from the narrow replication are used with the five surveys. Table S5 shows the result when all available birth date information is used to derive treatment probabilities and corresponds to Column 4 in the first panel of Table 2. Table S6 presents the results when all issues discussed in the narrow replication are addressed and corresponds to the first column in the second panel of Table 2.

Table 3: Effect of iodine deficiency on educational attainment by data source using FRT's specification

	All	Females	Males
Pooled	0.089	0.141	0.024
	(0.131) [2356]	(0.192) [1125]	(0.179) [1231]
DHS 1999	0.210	0.265	0.273
	(0.242) [266]	(0.426) [140]	(0.346) [126]
THBS 2000	0.347**	0.594***	0.190
	(0.148) [1395]	(0.170) [678]	(0.160) [717]
DHS 2004	-0.200	0.464	-0.643
	(0.449) [461]	(0.538) [220]	(0.622) [241]
NPS 2008	2.482		1.311
	(1.571) [91]		(3.475) [58]
DHS 2010	-1.758*	-7.264	-1.084
	(0.900) [143]	(9.169) [54]	(1.033) [89]

Notes: The table presents the result when FRT's specification is used to estimate the effect in the pooled sample and in the five surveys separately. The first row contains the full, pooled sample. Each of the subsequent rows presents the effect of the IOC programs when estimated separately in each survey. Each cell reports point estimates, estimated standard errors clustered at the district-cohort level within parentheses and the number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

variation in the estimated effects for that reason. In fact, for the NPS 2008 data set, there are only 33 girls left after imposing FRT's sample restrictions, and the fixed effects consume all the variation in the dependent variable.

The pooled data set almost doubles the number of observations relative to FRT's main analysis. The estimated effects are considerably closer to zero and no longer statistically significant. The increased sample size is expected to improve the precision of the estimates. The estimated standard error does decrease for the overall sample including both girls and boys. However, the standard errors increase when the two subsamples are analyzed separately. This is puzzling. The increase might reflect underlying differences between the surveys (e.g., that the treatment effect becomes more heterogeneous as the cohorts grow older). The differences are, however, small, and a possible explanation is sampling variability in the standard error estimation, which may be of particular concern due to the clustering.

FRT use the DHS 2004 data set in their supplementary analysis (see Part 3 of Table 3 in FRT). However, they use different cohorts in their main and supplementary analyses. In their main analysis, the effect is estimated for children aged 10-13 in the years 2000 and 2001 (i.e., those born between 1987 and 1991). In the supplementary analysis, which uses data from the years 2004 and 2005, they focus on children aged 10-14 (i.e., those born between 1990 and 1995).[12] FRT do not report the estimated effects for the cohorts examined in their main analysis using the DHS 2004 data. The fourth row of Table 3 reports these estimates. The point estimates for the overall sample and for boys are considerably smaller than in the pooled analysis (even negative), but the point estimate for girls is closer to those found in FRT. The sample sizes are, however, smaller, and the standard errors are larger. Subsequently, none of

[12] As we note above, it is questionable whether the identification strategy is valid for cohorts born in the 1990s since salt iodization had started by that time. Based on the discussion in Assey et al. (2009), it is possible that the absence of an IOC program in the 1990s is associated with *greater* protection from iodine deficiency since the the IOC programs targeted districts without salt iodization during this period.

the effects are statistically significant. The increase in the standard error we observe for the older cohorts also occurs for the younger cohorts in FRT's supplementary analysis. However, in FRT's supplementary analysis, the point estimates also increase (especially for girls with an intent-to-treat effect of 1.2 years), and the estimate for girls therefore remains significant at the 5% level.

We conclude that FRT's specification applied to the pooled data set and separately to the four additional surveys does not provide evidence of an effect of the IOC programs on educational attainment. We do not know why this is. A possible explanation is that the effect vary with age, and a noticeable effect may only exist for the ages FRT investigate. Another explanation is that FRT's specification masks an effect in the other data sources due to measurement error and attenuation bias. Furthermore, FRT's sample restrictions reduce the number of observations that can be used from each of the surveys, which may hurt precision. In the subsequent sections, we seek to address these concerns.

Alternative specifications

The changes to FRT's specification beyond those discussed in the narrow replication are primarily intended to address the measurement error in the treatment variable. The remaining differences between the specifications are minor and intend to harmonize the variables across the five surveys (e.g., unlike FRT, we use only standard grades in the formal schooling system when we define our educational attainment variable). We discuss these remaining differences in detail in Appendix A. The most notable aspect is that we use a more parsimonious specification with respect to the control variables.

The parsimonious specification makes it easier to harmonize the data sets (since some variables do not exist in all data sets), but it also has an econometric motivation. Although the control variables may increase precision, they do not have an obvious motivation given the identification strategy. The main threats to identification are either that women can time their births in a way to exploit the timing of the IOC programs or that the programs targeted districts based on location- and time-specific shocks. The controls FRT use are unlikely to capture these aspects. The inclusion of control variables may, in fact, threaten identification. Since treatment occurred before birth, most of the variables in the data sets are potentially affected by treatment themselves, and they would then bias the estimates if included.[13]

Apart from the changes documented here and in Appendix A, the specification is identical to the narrow replication presented in Table 2. We adopt the same fixed-effects approach, in which treatment is considered to be as-if randomly assigned conditional on district and cohort effects. The standard errors are estimated by clustering at the district level to account for serial correlation, and the null distribution of the test statistic in our hypothesis tests is derived using wild cluster bootstrap. The overall conclusions are, however, not sensitive to this choice. Tables S7 and S8 in the online appendix follow FRT's approach to standard error estimation and testing, and the results differ only in that one of the point estimates

[13] Table S17 in the online appendix presents our main results in the wide replication using a set of control variables as close as possible to FRT (with indicators for missingness in the data sets that lacks the corresponding variables).

for the THBS 2000 subsample becomes statistically significant at the 5% level.

New treatment specification. While the available data do not allow us to address all of the numerous sources of uncertainty in the treatment variable, we will focus on two important aspects of the specification that we are able to improve. The first improvement is an updated model of iodine uptake, depletion and need during pregnancy. The model used by FRT is common in the medical literature, but we argue that it is not the best choice in the Tanzanian setting. Of particular concern is the model's inability to account for overlapping programs and depletion heterogeneity. Medical studies tend to focus on a single intervention in a fairly homogeneous sample of people, and the standard models are constructed for this setting. In the Tanzanian setting, however, nearly all districts have multiple, overlapping programs, and we observe large heterogeneity in initial iodine deficiency rates. Second, as discussed in the narrow replication, FRT's treatment specification does not exploit all information on the respondents' birth date. As treatment is assigned relative to when the respondent was in utero, this could introduce substantial imprecision. We update the specification to exploit all available birth date information, including reported birth year and month when available.

While we believe that our updated treatment specification is an improvement on FRT, we acknowledge that many uncertainties remain. To avoid being overly reliant on one specific treatment model, we also investigate a number of alternative specifications.

Iodine metabolism. We begin the description of our updated treatment model by discussing the biological properties of iodine metabolism. An exponential function is often used to model the depletion of micronutrients and the excretion of toxicants. Several studies have, however, found that iodine supplements do not deplete exponentially. Instead, the rate of depletion diminishes over time after administration (see, e.g., Untoro et al. 2006). Hyperbolic functions allow for diminishing rates, and they are often used to model iodine depletion. FRT's treatment model is based on a hyperbolic function.[14]

The main advantage of the hyperbolic function is that it accounts for depletion patterns fairly well without increasing the number of parameters that must be estimated. It is, however, recognized that the hyperbolic function should not be seen as a representation of the structural depletion process. As discussed in, for example, Furnée et al. (1995), the model is best understood as an approximation of an underlying multi-compartment process, where iodine is stored in several places in the body, and the stores deplete exponentially at different rates.

[14] There is, however, an important difference between FRT's model and established medical models. FRT use a hyperbolic function to model the stocks of iodine in the body, while, to our knowledge, the function has been exclusively used to model urine iodine concentration, which is a flow measure. As a possible consequence, their model may overstate the length of protection. Notably, FRT's model assigns full protection of the fetus for 24 months and partial protection for an additional 16 months. Most studies have found that administering 400 mg of oral iodine, in the form of fortified poppyseed oil (e.g., Lipiodol, which was used in the Tanzanian programs), offers protection from iodine deficiency for at most 24 months (Wolff, 2001), and some studies have estimated the period to be less than a year (Ingenbleek et al., 1997). Considering the increased requirements during pregnancy, this would indicate that the length, as anticipated, is overstated.

Given its prevalence in the medical literature, the hyperbolic depletion function would seem to be a natural choice to calculate treatment probabilities, but its use entails several disadvantages. First, the hyperbolic model has exclusively been used to study the depletion pattern of a single supplementation intervention. The extent to which the model is able to account for multiple interventions (as in the Tanzanian setting) is unclear. The depletion rate in the hyperbolic model is given purely by the time that has elapsed since administration. If the same person received IOC supplements from several interventions, the model would provide conflicting depletion rates. A possible solution is to let the iodine from each intervention deplete separately at a rate calculated from its respective administration date. This would, however, result in unreasonably low depletion rates for certain levels of stored iodine and, thus, overstate the length of protection. The route taken by FRT is exactly to calculate the probability of protection given by each intervention separately and then sum the probabilities in a second step. This yields a probability of protection higher than 100% in some instances, which they truncate to 100%.

Second, it is not obvious how to incorporate depletion heterogeneity into the hyperbolic model. While altering the depletion parameter is straightforward, it is unclear how a particular setting would translate to a particular parameter value. One explanation for the observed diminishing depletion rate is that there is continuous dietary iodine intake after supplementation. The natural availability of iodine varies regionally, which affects dietary iodine intake. If the baseline iodine intake level in a district is high (relative to other districts with endemic goiter), we would expect that the IOC supplement offers protection for a longer period. The hyperbolic model does not allow for the explicit inclusion of continuous iodine intake, and one must rely on parameter adjustments to account for such heterogeneity. In the current setting, we have access to a proxy for baseline iodine intake in the form of pre-intervention goiter rates. A model that directly accounts for intake could exploit such information and would endogenously assign higher depletion rates in districts where iodine intake is low.[15]

The specifics of the Tanzanian setting warrant an improved depletion model. We will base our model on the discussion in Furnée et al. (1995) and use a *multi-compartment* model. This specification adjusts for depletion heterogeneity by explicitly modeling intake. The model also derives the depletion rate as a function of stored iodine (rather than time since administration) and can, therefore, naturally account for possible interaction effects between overlapping IOC programs. In short, we presume there to be two compartments where iodine can be stored in the body, one with a low depletion rate (representing the thyroid gland) and the second with a high rate (representing all other storage mechanisms). Given baseline iodine intake and eventual IOC supplementation, we directly model the stores and flows of iodine (albeit in an approximate sense) and can thereby calculate the probability of in utero protection.

While this model, like any depletion model, encompasses some strong assumptions, and while we cannot confirm that it characterizes the true depletion process, we argue that it should be preferred

[15] In addition to dietary iodine intake, goiter rates capture other factors affecting iodine availability that vary regionally, for example exposure to goitrogens (i.e., substances that hinder iodine uptake and metabolism).

in light of its merits. Notably, the multi-compartment model predicts urine iodine concentration levels that follow the characteristic hyperbolic functional form when shocked with a one-off intake higher than baseline. Furthermore, the model predicts iodine deficiency protection ranging from one to two years depending on baseline intake when shocked with a 400 mg IOC supplement, in line with existing studies (Ingenbleek et al., 1997; Wolff, 2001).

We investigate two versions of the multi-compartment model. The two versions differ in the allocation of iodine between the mother and the fetus during pregnancy when iodine stores are insufficient for both. In the first model (labeled "multi-compartment 1"), the mother's need is assumed to be filled first. The second model (labeled "multi-compartment 2") assumes that the fetus and the mother share the available iodine.[16] We have not found evidence in the medical literature indicating which of these two models is closer to the truth, and thus, we opt to present the results from both. We note, however, that the first model is closer to FRT's specification, as they also assume that the mother's need is filled first.

The precise details of the depletion models are presented in Appendix B. Formally, the models result in a function $T(y, m)$ that gives the probability of iodine deficiency protection in utero of a child born in year y and month m given the history of IOC programs in the relevant district.

Exploiting full birth date information. The function $T(y, m)$ assigns a probability of iodine deficiency protection based on the respondent's birth month. This information is only available for a small fraction of the respondents. We need to estimate the birth dates for all other children. We do this by calculating an interval of possible birth months. The treatment variable is the average of $T(y, m)$ over the predicted interval of birth dates adjusted for birth seasonality at the regional level. As an example, if we know that a respondent was born in either January or February 1988, she would be assigned the average of $T(1988, 1)$ and $T(1988, 2)$.

Some respondents report both age and birth year (i.e., we only need to impute birth month). We can derive whether these children have had their birthday in the year surveyed. Let Y_s and M_s be the survey year and month, Y_b be the reported birth year, M_b be the unreported birth month and A the reported age. If $Y_s - A > Y_b$, we know that the respondent has not yet had her birthday in the year she was surveyed. This implies that the respondent was born in a month after the survey month: $M_b \geq M_s$. Subsequently, if $Y_s - A = Y_b$, the birthday must have been in a month preceding the survey month, i.e., $M_b \leq M_s$.[17] We assign the average probability of treatment over these possible birth months to the respondent. The month of interview is always possible but less likely relative to the other months and is thus included with a weight of one half.

A majority of the respondents only report their age, as in FRT's original sample. In this setting, both the birth year and month (Y_b and M_b) are unreported and must be imputed. Accounting only for

[16] We thank an anonymous referee for suggesting the second model.

[17] In the few cases with impossible values, for example if the reported and predicted ages differ by more than one year, we continue as if the birth year were missing.

the year these respondents were surveyed, we can derive an interval of 24 possible birth months: from January in the year $Y_s - A - 1$ to December in the year $Y_s - A$. The interval can, however, be narrowed by also exploiting the reported survey month. If a respondent has had his birthday in the year surveyed, we know that $M_b \leq M_s$ and $Y_b = Y_s - A$. Conversely, if a respondent has not had his birthday in the year surveyed, we know that $M_b \geq M_s$ and $Y_b = Y_s - A - 1$. This yields a span of 13 possible birth months: from month M_s in year $Y_s - A - 1$ to month M_s in year $Y_s - A$. We average over all these months to derive the final treatment probability.

Alternative treatment specifications. In addition to the two multi-complement models, we will explore three alternative definitions of treatment. These are chosen to alter the level of model dependency. All depletion models entail a certain level of guesswork. While a detailed model can increase precision if the specification is sound, this comes at a cost of sensitivity to misspecification. Thus, to complement the main analysis, we include two simpler models and a slight modification of the model used by FRT.

1. The first alternative specification is a hyperbolic depletion model with an initial half-life of three months as used by FRT. We attempt to remain as close as possible to FRT's original specification. In particular, while we are skeptical of their "correction factor"[18] and the cut-off levels of stored iodine for iodine deficiency protection,[19] we opt not to change them. They are needed to obtain reasonable depletion patterns, and a version without them would require larger structural changes (i.e., closer to the multi-compartment model). Instead, we make a set of more modest changes:

   (a) We assign treatment using the full set of birth date information available, as discussed above.

   (b) FRT write that women younger than 23 years received a lower IOC dose. We have found no records of this other than FRT, and one of the coauthors of this paper (Swartling Peterson), who worked on the implementation and original evaluation of the programs, has no recollection of such a practice. We disregard the mother's age in our specification.

   (c) FRT assume that the IOC dose was 380 mg, while Peterson et al. (1999) and others report that it was 400 mg. We depart from FRT and use the figure reported in the literature. However, due to the very high initial depletion rate in this model, this change is essentially inconsequential.

   (d) FRT's model implicitly assumes that all respondents were conceived on the first day of the month. We instead use the expected date, i.e., the middle of the month. In addition, FRT

[18] The correction factor captures the portion of the assigned treatment probability that relies on the assumptions of partial protection and is intended to counteract eventual misspecification in depletion. It seems to mainly be an *ad hoc* solution to the extensive protection FRT's depletion model yields. We have not been able to find any medical motivation for its inclusion.

[19] As detailed in their online appendix, FRT use a cut-off for full protection above 6.5 mg of stored iodine and partial protection down to 4.2 mg. These levels are based on the supposedly daily recommended iodine intake for pregnant women of 1.4 to 2.1 mg per day. However, they do not provide the source of these recommendations. Notably, this level is 10 times the intake of 200 μg per day that is recommended for pregnant women by the World Health Organization et al. (2001).

assign the maximum monthly protection probability in the first trimester, implicitly assuming that any month can fully compensate for low protection in the others. We find no support for full compensation in the medical literature and, therefore, average over all three months in the first trimester.

(e) The "correction factor" in FRT is based solely on whether at least one IOC program started 4 or 5 years prior to a respondent's birth date. The factor does not account for possibly overlapping programs. This leads to an overcorrection of the treatment probabilities. For example, if one program started four years ago and another started one year ago, FRT's model will assign a lower probability of protection than if only the second program occurred. We account for overlapping programs when defining the correction factor to avoid this artifact.

2. The second alternative specification is a cruder model that does not make any explicit assumptions regarding depletion rates. In particular, we assume that each IOC program provides protection from iodine deficiency during pregnancy for a fixed number of months after administration. We explore three versions of this model that vary the period of protection: 12, 18 and 24 months after administration. This model still accounts for the fact that we do not know the exact timing of the IOC programs, by averaging over all possible start dates.

3. The third alternative specification does not model iodine protection at all but assigns treatment as a binary indicator solely depending on the time past since the latest IOC program in the relevant district. The model is therefore closest to a pure intent-to-treat specification: it does not account for that programs started at different times during the year nor potential interactions between overlapping programs. We use two versions of this model. The first defines treatment such that all children born one year after an IOC program are considered treated (i.e., children who were in utero directly following a program). The second version regards children born one or two years after a program as treated.

Our crude and binary models differ substantively from FRT's binary model. FRT state that their binary treatment model regards only children born one to three years after an IOC program as treated. We were, however, unable to replicate this. Instead, whenever their hyperbolic depletion model predicts a treatment probability greater than two-thirds, their binary treatment indicator is one. In other words, their binary specification appears to be a discretization of their hyperbolic model. This approach is sensitive to misspecification of the underlying depletion model in a way that a simple binary specification would not be.

The calculated monthly probability resulting from each treatment specification when there is a single IOC program is presented in Figure 1. The multi-compartment models (for which the protection varies with baseline iodine intake) are plotted with the mean intake; the models predict shorter protection

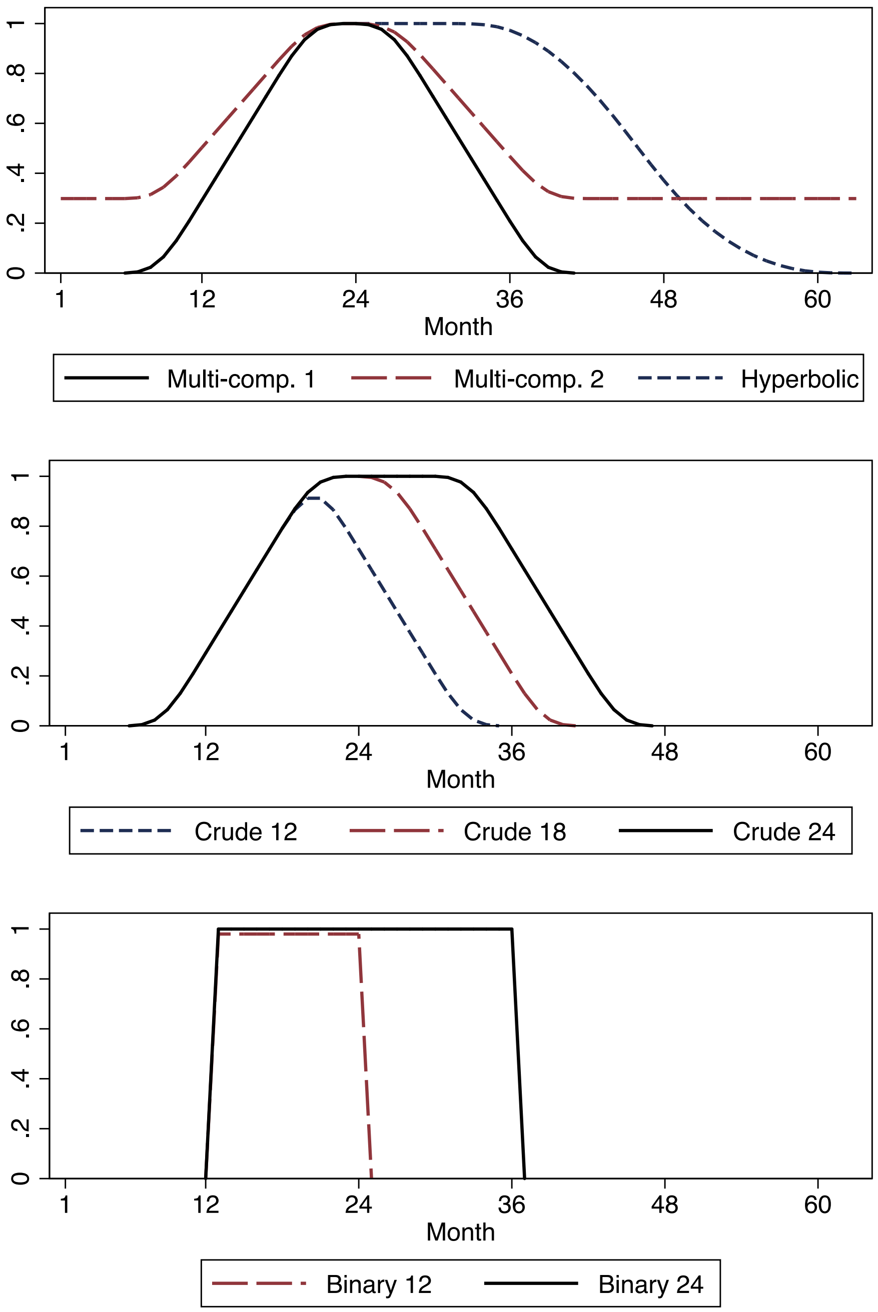

Figure 1: Treatment probability by birth month. The three panels present the calculated probability of being protected from iodine deficiency in utero by birth month relative to January in the year of the IOC program (labeled with 1). The first panel presents the three specifications that make explicit assumptions on the depletion pattern. The two multi-compartment models (labelled "Multi-comp. 1 & 2") are presented with the district mean baseline iodine intake levels. The second panel presents the cruder models that only implicitly make assumptions on the depletion pattern, representing 12, 18 or 24 months of protection. The third panel presents the binary specification that does not model depletion but assumes that the program grants full protection for 12 or 24 months starting with children born the year after the program.

in districts with low baseline intake and longer protection in districts with above-average intake. The first multi-compartment model resembles the cruder models (the main difference is its ability to account for heterogeneity in baseline iodine intake and overlapping programs). The second multi-compartment model predicts some protection from iodine deficiency even in the absence of an IOC program since the average baseline intake is not zero. The hyperbolic model predicts protection for a considerably longer time than the other models, but the "correction factor" could mitigate the problem.

Results with new specifications

The results from the wide replication are presented in Tables 4 and 5. Both tables follow the structure of Table 3. Each cell presents the result from a separate regression, columns indicate different subsamples (girls, boys and both), and rows indicate the separate specifications. The estimated coefficients should be interpreted as the estimated impact of iodine deficiency protection in utero on educational attainment.

Table 4 presents the effects of iodine deficiency protection across different treatment specifications using the pooled sample. Overall, the estimated effects are close to zero. The first multi-compartment specification ("Multi-comp. 1") suggests that in utero protection from iodine deficiency leads to a decrease in completed grades by 0.052 grades, but the estimate is not statistically significant. The effect is higher for girls than for boys, but there is no indication of a positive effect for either group. The second multi-compartment specification ("Multi-comp. 2") also suggests negative but statistically insignificant effects. The following row (labeled "Hyperbolic") presents the results when using our version of the hyperbolic depletion model. The estimates increase for girls and decrease for boys. Relative to FRT, they are still close to zero and not statistically significant at conventional levels.

The subsequent three rows (labeled "Crude") present the estimated effect when using the cruder model of depletion that assumes full protection for 12, 18 or 24 months after administration. Similar to the previous models, the estimates are small and not statistically significant. Note that as the length of assumed protection increases from 12 to 24 months, the estimated effects grow, and the effect particularly grows for girls. This roughly mirrors the change in the length of protection from the first multi-compartment model to the hyperbolic model, and could explain the difference between the two models. The last two rows (labeled "Binary") are the binary treatment specification that assumes that a program grants full protection for 12 or 24 months starting with children born the year after an IOC program. The estimates are close to zero and not statistically significant. In sum, the analysis provides no evidence for a positive effect.

As discussed in Section IV, the use of the pooled sample could conceal potential age heterogeneity. The remaining analyses aim to address this concern. Table 5 separates the estimation by data source. For these estimates, we use the first multi-compartment depletion model.[20] Positive effects are found using

[20] Tables S9 to S15 in the online appendix present the results when the other treatment specifications are used. The

Table 4: Effect of iodine deficiency on educational attainment using different treatment models

	All	Females	Males
Multi-comp. 1	-0.052	0.002	-0.121
	(0.134)	(0.201)	(0.209)
Multi-comp. 2	-0.099	-0.042	-0.182
	(0.174)	(0.277)	(0.251)
Hyperbolic	0.023	0.150	-0.152
	(0.133)	(0.187)	(0.170)
Crude 12m	-0.136	-0.049	-0.208
	(0.148)	(0.230)	(0.248)
Crude 18m	-0.052	0.004	-0.120
	(0.132)	(0.201)	(0.203)
Crude 24m	-0.001	0.080	-0.113
	(0.126)	(0.183)	(0.171)
Binary 12m	-0.121	-0.085	-0.114
	(0.086)	(0.154)	(0.107)
Binary 24m	0.002	0.010	-0.008
	(0.085)	(0.141)	(0.136)
Observations	5208	2658	2550

Notes: The table presents the results from the wide replication incorporating the changes discussed in Section IV and Appendix A. The sample pools observations of the 1986-1990 cohorts from the five surveys discussed in the text. The outcome is the number of completed grades in the formal educational system in Tanzania. We control for the respondents' gender, age and birth year; their relationship with the household head; an indicator of whether the household resides in an urban setting; indicators for the household head's and spouse's educational levels; and indicators of the surveys interacted with survey date. Each row reports the results from one of the treatment specifications discussed in the text. The first two rows are the two multi-compartment specifications. The next row (labeled "Hyperbolic") is our version of FRT's hyperbolic depletion model. The subsequent three rows (labeled "Crude") present the results from the cruder models that only implicitly make assumptions on the depletion pattern, representing 12, 18 or 24 months of protection. The final two rows are the binary specification that does not model depletion but assumes that a program grants full protection for 12 or 24 months starting with children born the year after a program. The columns present the results separately for each gender and pooled results for both genders. The last row presents the number of observations in the corresponding column. Each cell reports point estimates and estimated standard errors clustered at the district level within parentheses. Asterisks indicate statistical significance at the 10% (*), 5% (**) or 1% (***) level using wild cluster bootstrap to derive the null distribution of the test statistic.

the samples from the THBS 2000 and NPS 2008 surveys. The remaining three surveys show negative effects except for girls in the DHS 2010 survey and boys in DHS 1999. The estimates are, however, never statistically different from zero at conventional levels. We note that the pattern from Table 3 is repeated here, and the estimated standard error is lower for the THBS 2000 sample than in the pooled analysis despite its smaller sample size. The conclusion from the two analyses are, however, qualitatively the same.

The variability of the results in Table 5 is notable. While never statistically significant, some surveys and subsamples produce large and economically significant point estimates. In principle, this could be a reflection of some true underlying heterogeneity in the treatment effect over time. This seems, however, implausible in the current context. If iodine deficiency protection increased cognitive ability, we would not expect the sign of the effects to alternate between years in the way displayed. Instead, they should change monotonically as the cohorts grew older.

Another possible explanation is potential differences in sampling methods across surveys. The samples

estimated effects vary across specifications and are occasionally statistically significant at the 10% level and once significant at the 5% level. The overall conclusions are, however, unchanged.

Table 5: Effect of iodine deficiency on educational attainment by data source

	All	Females	Males
Pooled	-0.052	0.002	-0.121
	(0.134) [5208]	(0.201) [2658]	(0.209) [2550]
DHS 1999	-0.264	-0.399	0.001
	(0.224) [462]	(0.330) [230]	(0.247) [232]
THBS 2000	0.212	0.205	0.223
	(0.125) [3044]	(0.183) [1526]	(0.157) [1518]
DHS 2004	-0.414	-0.712	-0.144
	(0.291) [910]	(0.619) [457]	(0.238) [453]
NPS 2008	0.494	1.484	-2.250*
	(1.280) [224]	(1.358) [128]	(1.588) [96]
DHS 2010	-0.316	0.367	-1.236
	(0.444) [568]	(0.996) [317]	(0.880) [251]

Notes: The table presents the results from the wide replication, separating the estimates by survey. The first row contains the full, pooled sample as in Table 4. Each of the subsequent rows present the effect of the IOC programs when estimated separately in each survey. The specification is otherwise identical to the first row in Table 4 (see its notes for further details). Each cell reports point estimates, estimated standard errors clustered at the district level within parentheses and number of observations within square brackets. Asterisks indicate statistical significance at the 10% (*), 5% (**) or 1% (***) level using wild cluster bootstrap to derive the null distribution of the test statistic.

could, for example, be drawn from different (but partially overlapping) populations. However, since all surveys were intended to be representative of the Tanzanian population and were conducted by the same agency, there is no reason to expect such differences. Ultimately, the most plausible explanation appears to be ordinary sampling variability; we expect precision to decrease when separately investigating each survey because the samples are smaller. The fact that the smallest surveys show the most variation is indicative of this.

We present several extensions of the wide replication in the online appendix. First, Table S16 presents the estimates from the pooled analysis divided by age groups. Age is strongly correlated with the surveys, as we focus on the 1986-1990 cohorts, and it is difficult to separate possible age-specific effects from differences in the surveys. There is, however, sufficient overlap in age between surveys for a somewhat informative analysis. The only group for which we find some indication of an effect is girls aged 10-13 (i.e., FRT's age restriction), where the effect is statistically significant at the 10% level. The estimates are not statistically significant for girls aged 10-12 or girls aged 10-14. Second, Table S17 reports the results when we alter aspects of our the specification along the lines of the robustness checks that FRT conduct. The point estimates remain close to zero, and they are never statistically significant at conventional levels.

V Concluding remarks

We have revisited the Tanzanian experience with iodine supplementation in the late 1980s. We find no empirical support for a causal link between mild iodine deficiency in utero, preventable with IOC supplementation, and schooling later in life. We can, however, not conclude that no effect exists. The treatment variable may suffer from severe measurement error, which could mask economically important effects. In particular, the available data never reveal which respondents were protected from iodine deficiency in utero. One must resort to crude approximations of treatment assignment, and imprecision and attenuation bias is the consequence.

Our narrow replication highlights that the positive and statistically significant results reported by FRT are restricted to particular subsamples under particular specifications. Under alternative specifications, the estimated effects are considerably smaller and not significant at conventional levels. In our wide replication, we increased the sample size almost fourfold by pooling observations from five surveys conducted in Tanzania. In an effort to minimize attenuation bias and maximize power, we sought to improve the treatment model along several dimensions. Despite these efforts, the estimates remain small and do not support the conclusion of a positive effect.

Our conclusion is that the IOC programs in Tanzania are too poorly documented to be used in an impact study using quasi-experimental methods. The available data simply do not allow an answer to the substantive question.

References

Adhvaryu, A., Bednar, S., Nyshadham, A., Molina, T., and Nguyen, Q. (2018). When It Rains It Pours: The Long-run Economic Impacts of Salt Iodization in the United States. NBER Working Paper 24847.

Adhvaryu, A. and Nyshadham, A. (2016). Endowments at Birth and Parents' Investments in Children. *The Economic Journal*, 126(593), 781–820.

Almond, D. and Currie, J. (2011). Killing Me Softly: The Fetal Origins Hypothesis. *Journal of Economic Perspectives*, 25(3), 153–72.

Assey, V., Peterson, S., Kimboka, S., Ngemera, D., Mgoba, C., Ruhiye, D., Ndossi, G., Greiner, T., and Tylleskar, T. (2009). Tanzania National Survey on Iodine Deficiency: Impact After Twelve Years of Salt Iodation. *BMC Public Health*, 9(1), 319.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-differences Estimates? *The Quarterly Journal of Economics*, 119(1), 249–275.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, 90(3), 414–427.

Cao, X.-Y., Jiang, X.-M., Dou, Z.-H., Rakeman, M. A., Zhang, M.-L., O'Donnell, K., Ma, T., Amette, K., DeLong, N., and DeLong, G. R. (1994). Timing of Vulnerability of the Brain to Iodine Deficiency in Endemic Cretinism. *New England Journal of Medicine*, 331(26), 1739–1744.

Chan, S. Y., Andrews, M. H., Lingas, R., McCabe, C. J., Franklyn, J. A., Kilby, M. D., and Matthews, S. G. (2005). Maternal Nutrient Deprivation Induces Sex-specific Changes in Thyroid Hormone Receptor and Deiodinase Expression in the Fetal Guinea Pig Brain. *The Journal of Physiology*, 566(2), 467–480.

Deng, Z. and Lindeboom, M. (2018). A Bit of Salt, A Trace of Life: Long-run Impacts of Salt Iodization in China. Mimeo.

Dugbartey, A. T. (1998). Neurocognitive Aspects of Hypothyroidism. *Archives of Internal Medicine*, 158(13), 1413–1418.

Feyrer, J., Politi, D., and Weil, D. N. (2017). The Cognitive Effects of Micronutrient Deficiency: Evidence from Salt Iodization in the United States. *Journal of the European Economic Association*, 15(2), 355–387.

Field, E., Robles, O., and Torero, M. (2009). Iodine Deficiency and Schooling Attainment in Tanzania. *American Economic Journal: Applied Economics*, 1(4), 140–69.

Friedhoff, A. J., Miller, J. C., Armour, M., Schweitzer, J. W., and Mohan, S. (2000). Role of Maternal Biochemistry in Fetal Brain Development: Effect of Maternal Thyroidectomy on Behaviour and Biogenic Amine Metabolism in Rat Progeny. *The International Journal of Neuropsychopharmacology*, 3(2), 89–97.

Furnée, C. A., Pfann, G. A., West, C. E., van der Haar, F., van der Heide, D., and Hautvast, J. G. (1995). New Model for Describing Urinary Iodine Excretion: Its Use for Comparing Different Oral Preparations of Iodized Oil. *The American Journal of Clinical Nutrition*, 61(6), 1257–1262.

Haddow, J. E., Palomaki, G. E., Allan, W. C., Williams, J. R., Knight, G. J., Gagnon, J., O'Heir, C. E., Mitchell, M. L., Hermos, R. J., Waisbren, S. E., Faix, J. D., and Klein, R. Z. (1999). Maternal Thyroid Deficiency During Pregnancy and Subsequent Neuropsychological Development of the Child. *New England Journal of Medicine*, 341(8), 549–555.

Hassanien, M. H., Hussein, L. A., Robinson, E. N., and Mercer, L. P. (2003). Human Iodine Requirements Determined by the Saturation Kinetics Model. *The Journal of Nutritional Biochemistry*, 14(5), 280–287.

Ingenbleek, Y., Jung, L., Férard, G., Bordet, F., Goncalves, A. M., and Dechoux, L. (1997). Iodised Rapeseed Oil for Eradication of Severe Endemic Goitre. *The Lancet*, 350(9090), 1542–1545.

Lavado-Autric, R., Ausó, E., García-Velasco, J. V., del Carmen Arufe, M., Escobar del Rey, F., Berbel, P., and Morreale de Escobar, G. (2003). Early Maternal Hypothyroxinemia Alters Histogenesis and

Cerebral Cortex Cytoarchitecture of the Progeny. *The Journal of Clinical Investigation*, 111(7), 1073–1082.

Murcia, M., Rebagliato, M., Iñiguez, C., Lopez-Espinosa, M.-J., Estarlich, M., Plaza, B., Barona-Vilar, C., Espada, M., Vioque, J., and Ballester, F. (2011). Effect of Iodine Supplementation During Pregnancy on Infant Neurodevelopment at 1 Year of Age. *American Journal of Epidemiology*, 173(7), 804–812.

Peterson, S., Assey, V., Forsberg, B. C., Greiner, T., Kavishe, F. P., Mduma, B., Rosling, H., Sanga, A. B., and Gebre-Medhin, M. (1999). Coverage and Cost of Iodized Oil Capsule Distribution in Tanzania. *Health Policy and Planning*, 14(4), 390–399.

Pharoah, P., Buttfield, I., and Hetzel, B. (1971). Neurological Damage to the Fetus Resulting from Severe Iodine Deficiency During Pregnancy. *The Lancet*, 297(7694), 308–310.

Pharoah, P. and Connolly, K. (1987). A Controlled Trial of Iodinated Oil for the Prevention of Endemic Cretinism: A Long-term Follow-up. *International Journal of Epidemiology*, 16(1), 68–73.

Politi, D. (2010). The Impact of Iodine Deficiency Eradication on Schooling: Evidence from the Introduction of Iodized Salt in Switzerland. SIRE Discussion Paper 2010-02.

Pop, V. J., Brouwers, E. P., Vader, H. L., Vulsma, T., Van Baar, A. L., and De Vijlder, J. J. (2003). Maternal Hypothyroxinaemia During Early Pregnancy and Subsequent Child Development: A 3-year Follow-up Study. *Clinical Endocrinology*, 59(3), 282–288.

Pop, V. J., Kuijpens, J. L., van Baar, A. L., Verkerk, G., van Son, M. M., de Vijlder, J. J., Vulsma, T., Wiersinga, W. M., Drexhage, H. A., and Vader, H. L. (1999). Low Maternal Free Thyroxine Concentrations During Early Pregnancy are Associated with Impaired Psychomotor Development in Infancy. *Clinical Endocrinology*, 50(2), 149–155.

StataCorp (2007). Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

Untoro, J., Schultink, W., West, C. E., Gross, R., and Hautvast, J. G. (2006). Efficacy of Oral Iodized Peanut Oil is Greater than that of Iodized Poppy Seed Oil Among Indonesian Schoolchildren. *The American Journal of Clinical Nutrition*, 84(5), 1208–1214.

Wolff, J. (2001). Physiology and Pharmacology of Iodized Oil in Goiter Prophylaxis. *Medicine*, 80(1), 20–36.

World Health Organization, ICCIDD, and UNICEF (2001). *Assessment of the Iodine Deficiency Disorders and Monitoring Their Elimination.* Technical report, WHO publication, WHO/NHD/01.1.

Zimmermann, M. B. (2009). Iodine Deficiency. *Endocrine Reviews*, 30(4), 376–408.

Zoeller, R. T. and Rovet, J. (2004). Timing of Thyroid Hormone Action in the Developing Brain: Clinical Observations and Experimental Findings. *Journal of Neuroendocrinology*, 16(10), 809–818.

A Additional details on the wide replication

In addition to the differences documented in Sections III and IV, our wide replication includes minor changes of FRT's specification that we document here. The motivation for these changes is primarily to harmonize the five data sets discussed in Section IV (i.e., making coding conventions identical for all observations in the pooled sample). The specification differs from FRT in three aspects:

- We use a less restricted sample than FRT. As discussed in the narrow replication, FRT restrict their analysis only to children of the household head. Instead, we include all non-adopted children that are permanent residents of the household and are related to the household head. Furthermore, our cohort restriction (i.e., those born between 1986 and 1990) is based on the estimated birth year, while FRT restrict the sample based on reported age.
- We change the definitions of some of the included variables:
  - As in FRT, the outcome variable is educational attainment as measured by the number of completed grades. Non-standard grades are reported differently across surveys, and we restrict our attention to grades that correspond to the Tanzanian formal education levels (namely the seven grades in primary school, the four grades in lower secondary and the two grades in upper secondary school). FRT use the variable as reported in the THBS 2000 data set, which includes vocational education and other informal schooling.
  - As in FRT, we include the educational level of the household head and spouse as covariates. However, in an effort to harmonize the data sets and avoid measurement error, we use indicators for major educational achievements rather than linearly include the number of completed grades. This has the added benefit of being a non-parametric adjustment.[21]
  - FRT use the month of the interview as a covariate. However, they do not interact this variable with the year of the interview. Subsequently, two respondents interviewed the same month but in different years are assigned the same value on this covariate. We use indicators for the full year, month and survey interactions.

[21] The educational levels of the household head and spouse are defined in six categories, corresponding to no education, some primary, completed primary, some secondary, completed secondary, and higher education. For households without a spouse (2,326 respondents or 23.3% of the sample), the educational level of the spouse is unreported. To avoid excluding these observations from the analysis, we form a separate category for them. For an additional 81 respondents (0.8% of the sample), the educational level is unreported for either the head or the spouse. Similarly, we include a separate category for them. A third group comprises respondents who themselves are either a household head or spouse. Including the educational level as a control variable would, for these respondents, be to include the outcome variable as an independent variable. We form a separate category for these respondents. This specification, thus, assumes that the IOC program does not affect the probability that a respondent becomes household head.

- We use a different set of control variables. In general, we use a more parsimonious specification. As the treatment effect is identified through the fixed-effect approach, these changes will only affect the precision of the estimates. The primary motivation for this change is that some of these variables are not reported in all data sets we use in the pooled analysis. An important secondary reason is that some of these variables are potentially endogenous, as these variables are observed long after the IOC programs. In detail, the changes are as follows:
  - We drop the indicator of whether the respondent's mother was younger than 23 at the time of birth. As discussed in the main text, we fail to find the medical or institutional motivation for its inclusion.
  - We drop variables on family composition (birth order and number of children in the household).
  - We drop variables on whether the respondent's family are home owners, whether there is a grass roof, whether they report food problems, distance to closest secondary school and primary care clinic.
  - We include both age and birth year indicators for the respondents. In our pooled analysis, we have information on the cohorts from several years. We can, therefore, differentiate between age and cohort effects.
  - We add an indicator for the respondents' relationship to the household head. As we include all children who are permanent residents in the household in our analysis, this variable could be an important predictor of educational attainment and thus increase precision.

Table A1 presents a step-wise specification change from our wide replication to our narrow replication. This demonstrates the effect of the specification changes discussed in this section. In the first column, the specification is identical to the wide replication for the THBS 2000 subsample (i.e., it is exactly the third row in Table 5). The second column changes the treatment specification to our version of the hyperbolic depletion model, and in column 3, we use FRT's version of the hyperbolic depletion model. Our multi-compartment and hyperbolic treatment models produce roughly the same results. In the fourth column, we impose the sample restriction as in the narrow replication. That is, we restrict the analysis to children aged 10-14 (rather than restricting the sample based on birth year) and children or grandchildren of the household head or spouse. This restriction reduces the point estimates slightly. In the remaining two columns, we first change the definition of the covariates and then add and remove covariates as documented above. With these changes, the last column in Table A1 is the same specification as when all issues are addressed in our narrow replication (i.e., the left-most column in the second panel in Table 2).

Table A1: Step-wise specification change from wide to narrow replication

	(1)	(2)	(3)	(4)	(5)	(6)
All	0.212	0.220*	0.202*	0.150	0.125	0.082
	(0.125)	(0.114)	(0.099)	(0.110)	(0.110)	(0.135)
	[3044]	[3044]	[3048]	[2739]	[2721]	[2610]
Females	0.205	0.230	0.215	0.204	0.161	0.074
	(0.183)	(0.145)	(0.137)	(0.126)	(0.128)	(0.149)
	[1526]	[1526]	[1525]	[1348]	[1338]	[1286]
Males	0.223	0.221*	0.183	0.115	0.096	0.079
	(0.157)	(0.119)	(0.108)	(0.129)	(0.128)	(0.162)
	[1518]	[1518]	[1523]	[1391]	[1383]	[1324]

Notes: This table presents the consequences of the specification changes discussed in this appendix. Similar to the third panel in Table 2, we impose the changes step-wise in a cumulative fashion to show the consequences of each change: (1) presents the results from the wide replication for the THBS 2000 subsample (i.e., the third row in Table 5); (2) changes the treatment specification to our version of the hyperbolic depletion model; (3) changes the treatment specification to FRT's version of the hyperbolic depletion model; (4) imposes the sample restrictions discussed in this appendix; (5) changes the definition of the control variables as discussed in this appendix; and (6) adds and removes control variables to the set of variables used in the narrow replication. The last column is the results from the narrow replication (i.e., the first column in the second panel in Table 2). Each cell reports point estimates, estimated standard errors clustered at the district level within parentheses and number of observations within square brackets. Asterisks indicate statistical significance at the 10% (*), 5% (**) or 1% (***) level using wild cluster bootstrap to derive the null distribution of the test statistic.

B The multi-compartment depletion model

We here detail the multi-compartment depletion model used in our preferred treatment specification. Let I_t denote the total amount of stored iodine in the body at month t. This amount can be stored in either of two compartments (or types of compartments).[22] The first compartment represents the pool of stored iodine in follicular cells in the thyroid. This compartment has a good ability to retain iodine over a long period of time, which is modeled with a low depletion rate. However, the thyroid can only store a limited amount of iodine, which we set to 15 mg based on its estimated maximum storage capacity (Hassanien et al., 2003). The second compartment represents all other storage mechanisms in the body, which are less suited to the purpose and thus modeled with a higher depletion rate. We assume that iodine can be transported between the compartments freely and without cost,[23] but because the thyroid is the preferred location of storage, it will be filled first. The second compartment is assumed to have unlimited capacity, which is reasonable in the relevant interval considering the high depletion rate.

In addition to depletion, iodine stores are affected by consumption and intake. The thyroid can maintain an euthyroid state when at least 50 μg of iodine can be utilized daily to synthesize thyroid hormones (Zimmermann, 2009). In the current context, a reasonable approximation is that *at most* 50 μg per day (or 1.5 mg per month) will be consumed by the thyroid if the stores last. This implies that monthly consumption is given by $C_t = \min(I_t, 1.5)$. Intake is dietary iodine from food and potential IOC

[22] While we model the stored iodine with two compartments, the flows would be described by a semi-exponential, three-compartment model due to how we model consumption.

[23] Under normal circumstances, the thyroid is limited in how much iodine it can trap per day, which would motivate us to also limit monthly transport ability. However, iodine deficiency induces the pituitary gland to produce thyroid stimulating hormones that increases uptake. Thus, given the upper limit of 15 mg and the availability in the second compartment, modeling the transport, on a monthly basis, as unlimited is reasonable.

supplements. Their sum in milligram on a monthly basis is denoted N_t. Approximately 30% of orally administered iodine is subject to instantaneous fecal excretion (Hassanien et al., 2003), and thus never enters the blood stream, implying that $0.7N_t$ mg are added to the current stores in t.

We set the first compartment to deplete exponentially with a half-life of 12 months, corresponding to a monthly depletion factor of 0.056.[24] The second compartment depletes at a very high rate; some studies indicate a half-life of less than one month (Wolff, 2001). We will set the depletion rate to exactly one month (i.e., a monthly depletion factor of 0.5). Thus, the iodine retained in the first compartment from t to the following month is $0.944 \min(I_t - C_t, 15)$, while $0.5 \max(I_t - C_t - 15, 0)$ is retained in the second. This yields the following process of iodine storage and consumption:

$$\begin{aligned}
I_t^P &= \min(I_{t-1} - C_{t-1}, 15), && \text{(A1)}\\
I_t^S &= I_{t-1} - C_{t-1} - I_t^P, && \text{(A2)}\\
I_t &= 0.944 I_t^P + 0.5 I_t^S + 0.7 N_t, && \text{(A3)}\\
C_t &= \min(I_t, 1.5), && \text{(A4)}
\end{aligned}$$

where I_t^P is the amount of iodine stored in the first compartment between month $t-1$ and t, and I_t^S is the amount stored in the second compartment.

Note that if uptake is less than the required 1.5 mg, the stored iodine will diminish over time until the stores are completely depleted, at which point consumption will only consist of the current uptake. Thus, in steady state (where $I_t = I_{t-1}$), consumption and uptake must be equal if consumption is less than 1.5 mg. We will use this fact to derive the baseline iodine intake level through the goiter surveys. In addition to assuming that the population is in steady state at the time of observation, we will assume that the goiter rate is proportional to the average iodine uptake. Thus, if there is virtually no goiter (induced by iodine deficiency) in a given population, uptake must be at least 1.5 mg per month (i.e., an intake of at least 2.14 mg); if the entire population suffers from goiter, we assume that average uptake is so low that it can be approximated by zero. Consequently, average monthly intake in district d would be $2.14(1 - g_d)$, where g_d is the goiter rate in the district.[25] The assumptions underlying these calculations are strong and unlikely to hold exactly. Goiter rates are, however, the only available proxy for iodine intake for the relevant time period, and they may provide a useful approximation.

Iodine consumption is increased during pregnancy. One could therefore argue that the limit of 1.5 mg is not an adequate level to offer complete protection for the fetus. The necessary level of *stored* iodine has, to our knowledge, not been studied. Required *intake*, however, has been studied extensively. Based

[24] A higher depletion rate is often discussed for this compartment (see, e.g., Wolff, 2001). However, this includes consumption, which we model separately.

[25] If goiter rates are very low, we only know that intake is at least 2.14 mg; naturally, it could be higher. This is not relevant in the current context, as the lowest goiter rate among the treated districts is 28%. Thus, no district falls outside the interval that can be predicted by the formula.

on the ratio between the recommended intake for pregnant women and the recommend intake levels for adults of 4/3 (World Health Organization et al., 2001), we consider women with a calculated level of stored iodine higher than 2 mg to offer full fetal protection. The two multi-compartment models differ in how they assign partial protection. The first model predicts that women with stores between 1.5 and 2 mg offer partial protection proportionally within that interval. The second model predicts that women with stores between 0 and 2 mg offer partial protection also proportionally within the interval. We denote the level of fetal protection offered in month t by $S(t, \mathcal{H})$, where $\mathcal{H}$ describes the history of IOC administration in the relevant district. The two models thereby predict:

$$S(t,\mathcal{H}) = \begin{cases} 1 & \text{if } I_t \geq 2, \\ (I_t - 1.5)/0.5 & \text{if } 2 > I_t \geq 1.5, \\ 0 & \text{else,} \end{cases} \qquad \text{and} \qquad S(t,\mathcal{H}) = \begin{cases} 1 & \text{if } I_t \geq 2, \\ I_t/2 & \text{if } I_t < 2, \end{cases} \tag{A5}$$

where the stored iodine I_t is given by the complete iodine intake process prior to t implied by $\mathcal{H}$ (i.e., $N_{t'}$ for $t' \leq t$) and the relevant g_d.

Since we are interested in the probability of protection during the complete prenatal period, not a specific month, the measure above must be aggregated to an overall probability of in utero protection of a child born on a certain date. Although clear evidence exists that the development of the fetus is highly sensitive to deviations from optimal iodine levels, as previously discussed, the relative importance during the pregnancy is not entirely settled. There are indications that the first trimester is especially important. However, the late prenatal stage is not completely insensitive to deviations (especially for the development of higher cognitive ability; see, for example, Zoeller and Rovet 2004). Nevertheless, we will follow FRT and assume that *only* the first trimester is sensitive to iodine deficiency. A child born in month t would, on average, have been conceived in the middle of month $t - 9$. The probability of fetal protection for that child, denoted $P(t, \mathcal{H})$, is thus derived by averaging over the first three months of the pregnancy:

$$P(t,\mathcal{H}) = \frac{0.5S(t-9,\mathcal{H}) + S(t-8,\mathcal{H}) + S(t-7,\mathcal{H}) + 0.5S(t-6,\mathcal{H})}{3}. \tag{A6}$$

This function requires that we know the exact month when the mothers were given the IOC supplement. However, as discussed in the paper, we do not know the exact starting date of the programs or their length. As in FRT, we assume that they began uniformly during the specified year with a length of three months. This implies that a resident in a targeted district could have received the IOC at any time between January of the program year and February of the following year. As the interaction effect of overlapping programs could be important, we will consider all possible combinations of distribution dates. Let $\mathcal{H}(m_1, \cdots, m_K)$ denote the history of the IOC programs, where IOC program $i \in \{1, \cdots, K\}$ reached the respondent in month m_i relative to January of the program year. The resulting probability of protection for an individual born in year y and month m is given by:

$$T(y,m) = \frac{1}{36^K} \sum_{a_1=0}^{11} \sum_{b_1=0}^{2} \cdots \sum_{a_K=0}^{11} \sum_{b_K=0}^{2} P(12y+m, \mathcal{H}(a_1+b_1, \cdots, a_K+b_K)). \tag{A7}$$

In words, $T(y,m)$ takes the average of $P(t,\mathcal{H})$ over all possible combinations of starting months of the programs in a district. The hyperbolic and cruder treatment specifications discussed in the main text follow the same structure as the multi-compartment models but change $S(t,\mathcal{H})$ to their respective depletion model.