Fetal Iodine Deficiency and Schooling*

A Replication of Field, Robles and Torero (2009)

Niklas Bengtsson† Fredrik Sävje‡ Stefan Swartling Peterson§

July 3, 2017

Abstract

Scholars have theorized that congenital health endowment is a critical determinant of economic outcomes later in a person's life. In an important contribution, Field, Robles and Torero [*American Economic Journal: Applied Economics*, **1**, 4 (2009)] use iodine supplementation programs in Tanzania to estimate the impact of fetal iodine deficiency on educational attainment. The study is one of the first validations of the fetal origins hypothesis. Based on their large estimated effects, the authors conclude that geographic variation in iodine deficiency plausibly accounts for a substantial share of the variation in educational attainment in the developing world. We revisit the Tanzanian iodine supplementation programs through a narrow and wide replication of Field, Robles and Torero's study. We are able to exactly replicate the original results, but we find that they rest on a set of undocumented and unmotivated specification choices and sample restrictions. With a better motivated specification, we cannot establish an effect of fetal iodine protection on educational attainment. The result is unchanged after we increase the sample size fourfold and improve the precision of the treatment variable by incorporating new institutional and medical insights. We conclude that the available data do not provide sufficient power to detect an eventual effect since treatment cannot be measured with sufficient precision.

Keywords: *Iodine deficiency, Education, Prenatal exposure, Replication*
JEL: I12, I21, J16, O15

*We extend our gratitude to Douglas Almond, Sebastian Axbard, Per Engström, Rita Ginja, Linna Martén, Eva Mörk and seminar and conference 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. Please direct correspondence to Fredrik Sävje.

†Department of Economics, Uppsala University and Uppsala Center for Labor Studies.
‡Department of Political Science, Yale University.
§Department of Women's and Children's Health, Uppsala University and UNICEF.

1 Introduction

How does improved health *in utero* affect educational outcomes later in life? Field, Robles and Torero (2009, hereafter FRT) aim to shine light on this question by exploiting iodized oil capsule (IOC) distribution programs launched 1986 in Tanzania. 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 and estimate that being protected from iodine deficiency (ID) in utero increased educational attainment by 0.35 years on average. The estimated effect is particularly large for girls. Using the programs' coverage rates to derive the hypothetical case of achieving full coverage, the authors calculate that, internationally, "the expected increase in grade attainment for a child protected from fetal ID 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. FRT suggest that their work provides 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 in a field setting.[1] Since the publication of the original article, the empirical strategy and data have been used to address other aspects child health and household behavior (Adhvaryu and Nyshadham, 2016).

We revisit the Tanzanian experience in this paper by replicating FRT's study. We first try to exactly reproduce FRT's results using original data. The exercise is successful, but it reveals a series of undocumented or poorly motivated sample restrictions and other specification choices. We investigate 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 large estimates are primarily driven by three details of FRT's specification. First, in an effort to increase precision, FRT use the education level of the spouse of the household head as a control variable. In households where the head does not have a spouse, this variable is missing and, as a result, all single parent households are inadvertently dropped from the analysis. The dropped observations constitute about 21% of the sample. Second, FRT exclude all 14-year olds from their main analysis. As this cohort was targeted by the IOC programs and FRT include 14-year olds in their supplementary analyses, this appears to be an oversight. Including the dropped 14-year olds increases the sample size by 22%. Third, treatment is defined as whether the respondent's mother was protected from ID 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, however, disregards that a quarter of the sample was surveyed in 2001 rather than in 2000. Combined, the three specification choices inflate the estimated effect of ID protection by 300% for the overall sample and by 450% for girls. When we address the three issues, the estimated effects are close to zero and not significant at convention levels despite a substantial decrease in the standard errors.

Prior studies have documented health effects of fetal ID, and it is plausible that some of these effect carry on to schooling outcomes (see, e.g., Feyrer et al., 2013). A relevant critique against a null result is that with better data, it would be possible to detect positive effects. In particular, treatment assignment is never actually observed in the original study, and FRT's independent variable is the probability of treatment

[1] See Almond and Currie (2011) for a survey of this literature.

as approximated by a model. Treatment is defined as being protected from ID during the first trimester, and the IOC programs provide variation in such protection. The IOC 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 contain accurate information of when and where the respondents were born. We, subsequently, do not know when the respondents were in utero. Finally, the biological details of iodine intake and depletion are not known. These uncertainties introduce measurement error which could lead to imprecision and attenuation bias.

We try to mitigate these concerns in a wide replication of the original study. We maintain the baseline 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 so that the sample size almost quadruples. Both changes improve our ability to detect non-zero effects. The estimates are, however, still close to zero and not statistically significant at conventional levels. We conclude that the existing data do not provide evidence of the fetal origins hypothesis using FRT's method.

The remainder of this article is structured as follows. In the next section, we provide a short background to iodine deficiency disorders and the IOC programs. 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 report the result from the wide replication. Section 5 concludes.

2 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 are referred to as *iodine deficiency disorders* (IDD). Although the consequences of such deficiencies might be reinforced or counteracted by social mechanisms during childhood, FRT focus on the reduced-form relationship between iodine deficiency (ID) exposure and educational attainment.

The negative health effects of severe ID have long been known. Recent studies have found that milder forms of ID during pregnancy could be associated with hindered cognitive development as well (Lavado-Autric et al., 2003; Pop et al., 1999; Haddow et al., 1999). While there is an ongoing discussion concerning the exact period of greatest sensitivity to ID, and how persistent the effects of mild ID are, there is evidence that the early prenatal period is particularly sensitive. Notably, during the first trimester, the fetus cannot itself synthesize thyroid hormones and is completely reliant on the mother's hormone production. In addition, experimental studies have demonstrated that severe IDDs (cretinism) can be prevented in an iodine deficient population only if supplementation is given before conception (Pharoah et al., 1971), and that iodine supplements fail to prevent hindered cognitive development to a measurable level if given in the third trimester (Cao et al., 1994).

FRT's identification strategy exploits the possibility that the early pregnancy is particularly sensitive to ID. In particular, they use the delayed roll-out of IDD prevention programs in Tanzania to get variation in in utero protection from ID. Following the increased awareness of the benefits of IDD prevention, several large-scale iodized oil capsule (IOC) supplementation programs were introduced in the late 1980s in Tanzania. The aim of the programs was to target the populations most affected with IDD until universal salt iodization (USI) began in the early 1990s. In total, 27 districts with

severe ID (as measured by high goiter prevalence) were selected for inclusion. The intended structure of the programs was to distribute iodized oil capsules every second year starting in 1986. At each distribution round, each person aged from 2 to 45 years was 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; 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 still reached a substantial number of individuals. A conservative estimate is that the programs provided 12 million person-years of protection from ID (Peterson et al., 1999).

In the early 1990s, the USI had started.[2] By that time the focus of the IOC programs shifted from districts with high IDD prevalence to districts not yet reached by the USI program, namely districts where less than 75% of households had access to iodized salt. Therefore, the absence of an IOC program does not necessary indicate that the population is unprotected from ID after USI had started, even in districts with previously very high levels of ID.

2.1 FRT's empirical strategy

FRT's data consist of children surveyed at school-going age that 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 ID in utero, effectively comparing treated and untreated children in the targeted cohorts. In particular, 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 at birth date t. The treatment variable, T_{dt}, is the calculated probability of in utero protection from ID for an individual born in district d at date t, thus β_1 is the coefficient of interest. μ_d and λ_t are district and birth date fixed effects, and $\mathbf{X}_{idt}$ is a vector of control variables measured at either the individual or household level. 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. They do not account for correlation in the error term within districts or cohorts.

The treatment variable T_{dt} is the probability of being protected from ID in utero. Ideally, this variable would be an indicator taking on the value one if a respondent was protected from ID during the relevant part of the pregnancy, and zero otherwise. Such 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 and depletion of iodine during pregnancy. None of these factors are known. Instead, FRT's treatment variable is an approximation of the probability of ID protection based on the limited data available.

FRT's specification includes several control variables in $\mathbf{X}_{idt}$.[3] The purpose of the control variables are not entirely clear. Although FRT state that "identification [...]

[2] Data on the exact dates of the USI programs are, to our knowledge, not available.

[3] In particular, they 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,

requires that the error term be uncorrelated with treatment" conditional on the fixed-effects and covariates, their discussions seem to indicate that the fixed-effects alone suffice. The main threats to identification are either that women can time their births in a way to exploit the timing of the IOC programs (e.g., that more healthy women delay their pregnancies to after receiving the IOC) or that the programs targeted districts based on location- and time-specific shocks. It is unlikely that the covariates FRT include would solve such problems if they exist. Furthermore, as treatment was assigned on the district level, the most relevant level to control for confounding is at the district or, ideally, at the district-birth cohort level. All covariates used by FRT are on the respondent or household level which account for confounding only insofar that they are correlated with district level counterparts.

Importantly, given the hypothesis tested, the inclusion of control variables at the individual and household levels comes at the risk of biased estimates. Since treatment occurred before birth, the majority of the potential control variables are themselves outcomes. In particular, fertility decisions (birth order and number of children), access to health and schooling facilities (distance to secondary education and health clinic) and the different poverty measures (food security, home ownership, and house quality) could potentially be affected by treatment under the working hypothesis that ID affects educational attainment. Our reading is that FRT assume that the control variables are not affected by treatment and elect to include them to increase precision.

2.2 FRT's results

FRT find that the IOC programs 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' covarage rate, FRT conclude that ID protection in utero increases educational attainment by 0.73 years. However, this latter figure is likely an underestimated effect as well. 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 from ID during the first trimester did not benefit from the IOC programs at all. While the first trimester likely is most sensitive to ID, there is evidence that several skills (including cognitive ability) are sensitive to ID during other periods of development (Zoeller and Rovet 2004 provide an overview). Treatment is, thus, likely to spill-over into the control group and further attenuate the estimates.

Even though FRT's results are qualitatively in line with the medical literature, there are some unexpected details. First, it is hard to reconcile FRT's large education effects with the lack of health effects. The existing medical literature has been able to establish clear health effect of ID, but FRT are, to our knowledge, the first study that establishes effects of mild ID on long-term, real-life outcomes. It has been well-documented since the early 20th century that severe ID deficiency during pregnancy lead to *cretinism*. This is a permanent disorder that entails both severe and salient health problems as well as cognitive disabilities. If the IOC programs were able to prevent cases of this disease, the large and persistent effects FRT estimate would be easily understood (in line with, e.g., Pharoah and Connolly 1987). However, since FRT do not find any health effects, the cognitive effects must run primarily through milder forms of ID. The medical literature is not consistent on the cognitive effects of mild ID—especially at the magnitude that FRT estimate (Pop et al., 2003; Dugbartey, 1998).

The discrepancy between girls and boys observed in FRT is a second unexpected

whether the main dwelling has a grassroof and whether the household experienced food problems; and distance to nearest secondary school and nearest health clinic.

result. As discussed by FRT, there are two laboratory studies on rodents that provide some indications of gender differentials. In particular, Chan et al. (2005) report different responses between male and female guinea pigs in the regulation of thyroid hormone receptors. However, Chan et al. note that this may be the result of different compensating mechanisms rather than a manifestation of increased sensitivity. Further, Friedhoff et al. (2000) observe behavioral differences between male and female rats when their mothers' thyroid glands were completely removed, thereby inducing severe ID during the prenatal period. It is, however, questionable how well the rodent model with severe ID generalize to humans with mild ID. Evidence of a gender difference from studies on human subjects is scant. Most of the medical studies on human subjects do not show a gender differences in ID sensitivity. To our knowledge, the only investigation on humans that find gender differences besides FRT is an observational study of one-year olds in Spain (Murcia et al., 2011). That study documents a correlation between a diet low in iodine (proxied by self-reported fish consumption and mineral supplement intake) among pregnant mothers and infant neurodevelopment. However, the result does not account for possible omitted variables. A recent study of iodine deficiency on educational attainment in humans in Switzerland does not find support for a stronger effect for females (Politi, 2014).

3 Narrow replication

In this section, we aim to replicate the findings of FRT using original data sources. In a first part, we try to reproduce the exact results of FRT without any changes to their specification. In a second part of the narrow replication, we try to improve the analysis by addressing issues discovered in 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 was cleaned and how variables were defined is not publicly available. The lack of a detailed description of the treatment variable is particularly problematic as the results are sensitive changes in the specification and, in many cases, there are several alternatives that are equally sound. We managed to replicate undocumented parts of FRT's analysis by trial and error.[4] Our recreated data set was compared observation by observation to the original in order to detect any differences.

We were able to replicate the FRT data set using original data sources net of these variables:

1. We did not try to replicate the outcome variables in the supplementary analyses. That is, we did not replicate the sickness variables presented in FRT's Table 6 nor the alternative schooling measures presented in FRT's online appendix.

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

3. We tried but did not succeed to fully replicate the program coverage rate variable used in FRT's main analysis. We could not find the rule they used to resolve overlapping programs with respect to coverage of the IOC programs. We found a rule that replicated the coverage rate variable for children aged 10-14 (i.e., FRT's main sample). However, this rule does not replicate the variable for the extended sample not used in this replication.

[4]Complete documentation and code to replicate the analyses in this study from original data sources are available on request from the authors.

Table 1: Replication of FRT's main results

	(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: This table replicates the analysis presented in part 1 and 2 in Table 3 in FRT. The outcome in all columns are the number of completed grades as reported in the THBS 2000 survey (including non-standard grades). The first column is the results from FRT's main specification where time and district fixed-effects are included, and the treatment variable is the estimated probability of being targeted by the IOC programs while in utero, as discussed in Section 2.1. The second column changes the treatment variable to a binary indicator derived from FRT's treatment probability model (see the discussion in 4.2.3). The third column interacts the treatment variable from the first specification with the coverage rate of the IOC programs. This inflates the estimates to adjust for that not everyone in the targeted districts was reached by the programs. 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 on the district-cohort level within parentheses and number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

4. We could not find how FRT derived the birth seasonality information used to calculate their treatment probabilities. We therefore use the partially assembled data set provided by FRT to construct the birth seasonality variable. This is the only variable in the replication data set that is not derived from original data sources.

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 full replication of FRT's main analyses. Table 1 presents the replicated results, which corresponds to the first two parts of Table 3 in FRT. The results are replicated exactly.

3.1 Issues discovered during replication

During the replication effort, we noted several issues in FRT's specification. Some of these issues only become apparent in light of new institutional and medical insights, and were presumably not available to the original authors. Other issues are judgement calls, entailing, for example, a bias-variance trade-off. Issues of those two kinds are discussed in the wide replication in Section 4. In this narrow replication, we will focus on issues that were undocumented or poorly motivated in the original study. Of particular concern is a set of sample restrictions that lack medical and institutional motivation. Such restrictions not only decreases the precision of the estimates, but also make them less representative of the target population.

The first set of issues—labeled 1-4 below—are not discussed by FRT and unjustified

from the viewpoint of identifying the causal effect of the IOC programs. The two subsequent sample restrictions—labelled 5 and 6—are documented by FRT, but they are not clearly motivated and their implications are understated. The final issue is an econometric issue with estimation of standard errors. In detail, the issues are:

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 gives the correct birth order.

2. FRT add the educational level of the spouse of the household head as a covariate. Spouses' education is unreported when there is no spouse in the household. As the statistical software used by FRT silently drops observation with missing values, FRT's specification implicitly excludes all single parent households. Approximately 21% of the children in the sample live in such households. As this type of household tend to be poorer, and is thus more likely to benefit from the IOC programs, this sample restriction is hard to rationalize and is likely to bias the results towards zero.

   We add the excluded children and include an indicator for missing spouses as a control variable.

3. FRT include children aged between 10 and 13 years in their main analysis. The explanation is partly that these cohorts were potentially affected by the IOC program in utero, and partly that the age span represents "the modal age of enrollment." Although FRT investigate how the estimates are affected by including younger cohorts in their robustness checks, they never include children aged 14 when investigating the THBS 2000 data. Given the stated motivation, this appears to be an oversight.

   Since the THBS survey was conducted in 2000 and 2001, children with a reported age of 14 are born either in 1986 or 1987. These children would, thus, have benefited from ID protection in the counterfactual setting where all programs started in time. In practice, IOC programs had started only in few districts when this cohort was in utero, so the estimated treatment probabilities are, in general, small for these children. This does, however, not constitute a threat to identification; given that these children were targeted by the programs and could potentially been treated, they are best consider to be part of the natural experiment.

   The low estimated treatment probabilities could, in fact, be beneficial to the analysis. In some districts, there are so many overlapping programs that all children aged 10-13 have treatment probabilities close to one. As a result, there are no effective control group in these districts, and the fixed effects are estimated by extrapolation from children will high, but not full, protection. Similar to a classical difference-in-difference approach, if children aged 14 have low treatment probabilities in such districts, their inclusion will lead to more credible identification by making the fixed effects more precise and less model dependent.

   FRT's second remark—the modal age of enrolment—would make one worried that there is not enough variation in the outcome variable for children aged 14 to warrant 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.

Notably, FRT include children aged 10 to 14 years in their auxiliary investigation (part 3 of their Table 3). This is not documented in their article.

We include children age 14 in the analysis. As FRT calculate their 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. This procedure ignores the information provided by the date of the interview. Most notably, it disregards that 25.5% of respondents are interviewed in 2001.

   It is not obvious how one would correct this issue. Respondents' probable birth years depend on when during the year they are interviewed (the THBS data contain both year and month of the interview). Conditional on reported age in years, birth year is positively correlated with the date of the interview. For example, if a respondent is interviewed in January 2000, he is unlikely to have had his birthday yet: his birth year is more likely to be 1999 minus his age rather than 2000 minus his age. While such information can be exploited to better predict when the respondents were born, FRT's treatment specification assumes that no such information exists. Using the interview date yields an interval of possible birth dates that typically spans parts of two years (see Section 4.2.2 for more details). However, FRT's main specification assigns treatment on a yearly basis adjusting for birth seasonality during a whole year. A prediction that spans parts of two years would, therefore, lead to additional measurement error if used with FRT's treatment variable. We try to remain as true as possible to FRT and derive birth year as the year of the interview minus the age of the respondent. However, in the wide replication, we fully exploit the information provided by the interview date and adjust the treatment variable accordingly.[5]

5. FRT drop all children that cannot be linked to a *unique* mother in the household. The stated 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 the need for matching to mothers. As FRT's matching procedure is based solely on potential mothers' ages, it would, in any case, not be able to differentiate between adopted and biological children in the household. The only remaining rationale to link the child to a unique mother is to calculate the mothers' age at birth, which FRT use as a covariate. However, as we discuss in the wide replication, this covariate is not necessary for identification and lack medical and institutional motivation (see Section 4.2.3, point 2). The main, implicit 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" covariates are missing for these children, thus we include an indicator for missingness as a control variable.

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

[5]Table 6 in the appendix presents the narrow replication exploiting the full information contained in the interview date. The estimates are slightly higher in this setting, but the qualitative conclusions remain. Note, however, that we still use FRT's specification for the treatment variable in this table; it is, thus, misspecified.

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 being the most notable group. For all these children the needed outcome variables are readily available. We include children that are grandchildren of the household head.

7. Finally, we note that FRT cluster their standard error estimators on the district-cohort level. As the treatment variable is perfectly correlated for respondents in the same cohort and district, such clustering is motivated. However, FRT treatment model estimates that protection from ID continued for several years after an IOC program ended. In other words, the treatment variable is also correlated between 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 of the IOC programs 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 a downward bias in the estimated standard errors.

   It is beyond the scope of this replication study to construct an estimator that fully account for the dependence structure in the data. However, as an indication of the consequence of the issue, we re-estimate FRT specification with the standard error estimator clustered on the district level rather than the district-cohort level. This accounts for any serial correlation within districts, but still disregards spatial dependencies between districts.

The impacts of our changes are reported in Table 2. The first column, labeled (0), is the exact specification used in FRT's main analysis (i.e., it is exactly the first column in Table 1 in this paper, or the first panel in Table 3 in FRT). The subsequent columns each address one of the issues listed above. The changes are cumulative (e.g., column 3 shows the total results when addressing issues 1, 2 and 3). Table 7 in the appendix presents the influence of the seven issues in isolation rather than their accumulated impact.

In the second column of Table 2, we see that the first issue—the erroneous use of the `sort` command—is largely inconsequential to the results, and the estimates are close to those in original specification. The next two issues, however, lead to large 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 the size of those reported by FRT and not statistically different from zero at conventional significance levels. With the next three corrections (column 4-6), the point estimates are further reduced and effectively become zero. The last correction—accounting for serial correlation between cohorts—increases the estimated standard errors, as expected. However, as the point estimates are close to zero at this point, the increased standard errors do not affect the qualitative conclusions.

We conclude that FRT's findings are mainly driven by a set of poorly motivated sample restrictions. FRT do not use all cohorts that were affected by the IOC programs. 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 find no institutional, medical or econometric reasons to restrict the study to this subsample. When we relax these restrictions, the large and significant effects estimated by FRT disappear.

A possible explanation to the large changes in the estimated effects is treatment effect heterogeneity; respondents in the subsample that FRT focus on could have particularly benefited by the intervention. We find this explanation unlikely. In fact, one could argue that the effect is expected to be lower in FRT's subsample. Single parent households

Table 2: Addressing issues found in the replication of FRT

	FRT's specification						All issues corrected	
	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All	0.347**	0.340**	0.250**	0.155	0.087	0.070	0.026	0.026
	(0.148)	(0.145)	(0.118)	(0.112)	(0.090)	(0.090)	(0.089)	(0.124)
	[1395]	[1395]	[1691]	[2056]	[2056]	[2333]	[2610]	[2610]
Females	0.594***	0.627***	0.542***	0.258*	0.108	0.055	-0.044	-0.044
	(0.170)	(0.171)	(0.156)	(0.147)	(0.130)	(0.132)	(0.131)	(0.147)
	[678]	[678]	[819]	[1009]	[1009]	[1143]	[1286]	[1286]
Males	0.190	0.170	0.088	0.086	0.049	0.057	0.035	0.035
	(0.160)	(0.156)	(0.144)	(0.150)	(0.118)	(0.115)	(0.119)	(0.158)
	[717]	[717]	[872]	[1047]	[1047]	[1190]	[1324]	[1324]

Notes: This table presents the results of the main specification of FRT when the errors discovered in the narrow replication is corrected. In the first column, labeled (0), the replicated estimates from FRT are presented (i.e., the first column in Table 1 in this paper or the first panel of Table 3 in FRT). Cumulatively, each subsequent column addresses one issue discussed in the text: (1) uses `gsort` rather than `sort` to derive birth order; (2) includes children in single parent households; (3) adds children that are 14 years old; (4) derives birth years based on reported survey year; (5) adds children that cannot be linked to a unique mother; (6) adds grandchildren till household heads; (7) clusters standard errors at the district level rather than district-cohort level to account for serial correlation. Apart from these changes, the specifications are identical to the main specification in FRT as described above. Each cell reports point estimates, estimated standard errors within parentheses and number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

are, in general, poorer and, thus, more likely to suffer from ID. We would, therefore, expect the effect to be larger among the excluded respondents.

While FRT's study does not grant support to the conclusion that the IOC programs had a positive effect on educational attainment, neither can we conclude the opposite; it is possible that the programs had a fairly large effect that we are unable to detect. The main issue is that we do not observe which respondents were treated; at best, we can derive a rough estimate of treatment. This introduces a measurement error that biases 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 with 60% to account for this. However, this is just one among a host of uncertainties that introduce errors in the treatment variable. There are primarily two ways to combat the power loss that follows from measurement error: obtain better measures and increase the sample size. In the wide replication, we employ both of these approaches to maximize our ability to detect an effect of the IOC programs.

4 Wide replication

Based on institutional, medical and econometric considerations, we argue that FRT's analysis can be improved in several aspects. We believe these improvements reduce the risk of biases and increase our power to detect an eventual positive effect. Apart from the issues already discussed in the narrow replication, there are primarily two differences between our specification and that of FRT. 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. Second, in an effort to reduce the measurement error, we improve the definition of the treatment variable based on

new institutional and medical insights.

The remaining differences between the specifications are minor and discussed in detail in Appendix A. In sum, changes are made to harmonize the variable definitions between the included surveys (e.g., unlike FRT, we use only standard grades in the formal schooling system when we define our educational attainment variable). The only substantive difference is that we use a more parsimonious specification with respect to the control variables included in $\mathbf{X}_{idt}$.

Besides the documented changes, the specification is identical to the one presented in the narrow replication in Section 3.1. In particular, we adopt the same fixed effects approach where treatment is considered to be as-if randomly assigned conditional on district and birth date effects. While there is serial correlation in the treatment variable as discussed in the narrow replication, we will disregard this fact and cluster the standard error estimator on the district-cohort level as in FRT. This will bias the estimated standard errors downwards and, thus, understate the amount of uncertainty in the analysis. However, as we do not find significant effects with this less conservative strategy, an estimator that correctly account for the dependence structure in the data would not change our conclusions (i.e., we would still not find significant effects).

4.1 Additional data sources

We investigate the effects of protection from ID using data of the cohorts born between 1986-1990 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 all of them are representative of the whole of Tanzania and contain the relevant data needed for the analysis.[6] All surveys were conducted by the National Bureau of Statistics in Tanzania with the involvement of international organizations.

We will investigate each survey separately, as well as all surveys collected in a pooled analysis. The pooled analysis has the benefit of maximizing sample size, and thus the power to detect an eventual effect. However, as the surveys are conducted at different years, the respondents will be of different ages. If the effects of the IOC programs are changing over time (e.g., if untreated children attain the same level of education as treated children but at a slower rate), the effect in the pooled sample could mask effects that differ between age groups. We do not deem this to be a sufficiently acute concern to warrant an exclusion of the pooled analysis. For the cohorts born between 1986-1990, the main reason for the low educational attainment was high drop-out rates. In 1998 the probability that a student prematurely left education, on a yearly basis averaged over all primary grades, was 6.6%. One large contributing factor is the fourth grade exam, after which 8.8% of the students drop out. The exam at the end of primary education, the *Primary School Leaving Examination*, which confers eligibility for secondary school, had a pass rate of 27.1% in 2002 (Ministry of Education and Culture, 2002). As a child's educational level is frozen when she exits the educational system, we expect the effect

[6]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.

of the IOC programs to be persistent. In fact, given that the average attainment is low at baseline, we would expect that the treatment effect would grow over time if the IOC programs had a positive effect; children that stayed in school due to the program could potentially continue to secondary school at a higher rate and, thus, making the difference between the treatment groups more salient over time.

In our preferred treatment specification, as detailed below, we use information from a number of goiter surveys that were conducted in Tanzania during the 1980s. We obtain this data from the original data sources. The purpose of these surveys was to estimate how widespread and severe IDD was in Tanzania. Due to the complexity and diversity of the disorders following from iodine deficiency, a common procedure is to use goiter prevalence as a proxy for all IDDs. 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 ID.[7]

4.2 New treatment specification

A concern in the study of the IOC programs—affecting both us and FRT—is that treatment assignment is unobserved. The treatment variable in the analyses is approximated based on incomplete data. Such measurement errors are problematic as it will attenuate the estimates and, thus, reduce our power to detect an eventual effect. Making the treatment variable more accurate would lead to less bias and improved power. While the available data do not allow us to address all of the numerous sources of uncertainty in the treatment variable, we will focus two important aspects of the specification that we are able to improve.

The first improvement is an updated model for 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 well-suited for the Tanzanian setting. Of particular concern is the model's inability to account for overlapping programs and depletion heterogeneity. The typical medical study focuses on a single intervention in a fairly homogeneous sample of people, in which case the standard model works well. However, in the current study, nearly all districts have overlapping interventions and we observe large heterogeneity in initial IDD 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 so to exploit all available information.

While we believe our updated treatment specification is an improvement over FRT, we acknowledge that many uncertainties remain. In order to not be too reliant on one specific treatment model, we also investigate a number of alternative specifications—one of which is very close to FRT's specification.

4.2.1 Iodine metabolism

We begin the description of our updated treatment specification by discussing the biological properties of the iodine metabolism. An oft-used model for the depletion of micronutrients or the excretion of toxicants is the exponential function. 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). The hyperbolic function is, for this reason, often used to model iodine depletion as it allows for a diminishing depletion rate. FRT's treatment model is based on the

[7] 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.

hyperbolic function.[8]

The main advantage of the hyperbolic function is that it accounts for depletion patterns fairly well without increasing the number of free parameters that must be estimated. It is, however, acknowledged 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 instead best understood as an approximation of an underlying multi-compartment process, where iodine is stored in several places in the body and each store depletes exponentially at different rates.

Given its prevalence in the medical literature, the hyperbolic depletion function would seem to be the natural choice to base the calculation of treatment assignment on, but its use entails certain disadvantages. First, the hyperbolic model has exclusively been used to study the depletion pattern of a single supplementation invention. The extent to which the model is able to account for overlapping interventions (as in the Tanzanian setting) is unclear. As the depletion rate in the hyperbolic model is given purely by the time that has elapsed since administration, overlapping inventions would provide conflicting depletion rates. A possible solution would be to let the iodine from each intervention deplete separately with the 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 to calculate the probability of protection given by each intervention separately, as if there were no overlap, and then sum the probabilities in a second step. This yields a probability of protection higher than 100% in some instances, which they truncated to 100%.

Second, it is not obvious how to incorporate depletion heterogeneity in the hyperbolic model. While altering the depletion parameter is straightforward, it is unclear how a particular setting would translate to a particular parameter value. An important 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 was 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 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 are low.[9]

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

[8]There is, however, an important difference between FRT's model and established medical models. FRT use the hyperbolic model for the stocks of iodine in the body, although, to our knowledge, this model 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 ID 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.

[9]In addition to dietary iodine intake, goiter rates capture other factors affecting iodine availability that varies regionally, for example exposure to goitrogens (i.e., substances that hinder iodine uptake and metabolism). An analysis using this proxy would therefore benefit from the possibility of accounting for a wide range of other sources of heterogeneity in depletion.

than time since administration) and can, therefore, naturally account for overlapping IOC programs, including possible interaction effects between 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 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 an one-off intake higher than baseline. Furthermore, the model predicts goiter protection ranging from one to two years depending on baseline intake when shocked with a 400 mg IOC supplement, in line with the existing studies (Ingenbleek et al., 1997; Wolff, 2001).

Formally, the model results in a function $T(y, m)$ which gives the probability of ID protection in utero of a child born in year y and month m given the history of IOC programs in the relevant district. The precise details of the depletion model are explained in Appendix B.

4.2.2 Exploiting full birth date information

The function $T(y, m)$ assigns probability of ID protection based on the respondent's birth month. This information is only available for a fraction of the respondents. We need to estimate the birth dates for all remaining 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 adjusting for birth seasonality at the regional level. As an example, if we know that a respondent is born in either January or February 1988, she would be assigned the average of $T(1988, 1)$ and $T(1988, 2)$. Of course, if the exact birth month is reported, no averaging is done.

For some respondents, we know their 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., $M_b \leq M_s$.[10] 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.

The only available birth date information for most respondents is their age, which, together with the survey year, gives an interval of 24 possible birth months. We can make the interval narrower by also exploiting the survey month. Again, let Y_s and M_s be the survey year and month, and A the reported age. Furthermore, let Y_b and M_b be the unreported birth year and 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

[10] 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 was missing.

probability.

4.2.3 Alternative treatment specifications

In addition to our preferred treatment specification, we will explore three alternative definitions of treatment. These are chosen to alter the level of model dependency. As we discuss above, 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 also use two simpler models that include fewer parameters. In addition to the simpler models, we investigate an updated version 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 try to remain close to FRT's original specification. In particular, while we are skeptical to their "correction factor"[11] and the cut-off levels of stored iodine for ID protection[12], we opt not to change them. They are needed to get reasonable depletion patterns, and a version without them would require larger structural changes. Instead, we make a set of more modest changes:

   (a) We assign treatment using the full set of birth date information available, as documented in section 4.2.2.

   (b) FRT claim that women younger than 23 years received a lower IOC dose. We have found no records of this beside FRT, and Swartling Peterson, who aided in the original implementation of the programs, has no recollection of such 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. Due to the very high initial depletion rate in this model, this change makes only a very slight difference.

   (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 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. As a result, the factor does not account for possible overlapping programs. This leads to an overcorrection of the treatment probabilities. We account for overlapping programs when defining the correction factor.

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

[11] 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 results in. We have not been able to find any medical motivation for its inclusion.

[12] As detailed in their online appendix, FRT use a cut-off with 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).

IOC program provides protection from ID 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 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 assign 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 version of this model. The first defines treatment so that all children born one year after an IOC program are consider treated (i.e., children that was in utero directly following a program). The second version considers 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 considers only children born one to three years after an IOC program as treated. We were, however, not able to replicate this. Instead, it seems that 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 is a discretization of their hyperbolic model. Unlike a truly binary treatment specification, this specification is sensitive to misspecification of the underlying depletion model. Particularly worrying is that it is sensitive to assumptions regarding overlapping programs 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. Notably, protection in the multi-compartment model varies with baseline iodine intake, while the hyperbolic model predicts protection for considerably longer time than the other models. The inclusion of the "correction factor" (also presented in the figure) in the hyperbolic specification could mitigate this problem, however.

4.3 Results from the wide replication

The main results from the wide replication are presented in Tables 3 and 4. Both tables follow the same structure; each cell presents a separate regression. Columns indicate different subsamples (females, males and both), and rows indicate the separate specifications. The estimated coefficients should be interpreted as the impact of ID protection in utero on educational attainment.

Table 3 presents the effects of ID protection across different treatment specifications. Overall, the estimated effects are close to zero. Our preferred multi-compartment specification (labeled "Main") suggests that in utero protection from ID leads to a decrease in completed grades by 0.068 grades, which is not statistically different from zero. Although still negative, the effects are slightly higher for females than for males, but the difference is neither statistically nor economically significant. The following row (labeled "Hyperbolic") presents the results when using our version of the hyperbolic depletion model. The estimates increase for females and decreases for males. Relative to FRT, they are still close to zero and not significant at conventional levels.

The subsequent three rows (labeled "Crude") present the estimated effect when using the cruder model that assumes full protection from in utero ID for 12, 18 or 24 months after administration. Similar to the previous two models, the estimates are close to zero and statistically insignificant. Note that as the length of protection increases from

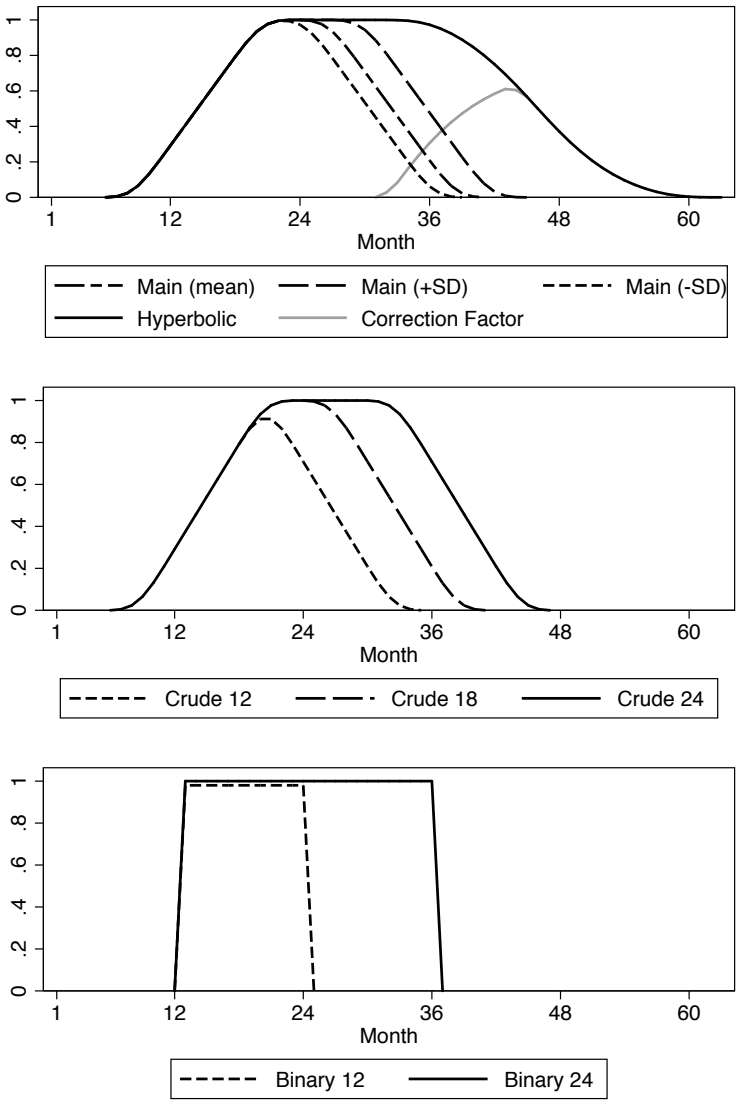

Figure 1: Treatment probability by birth month. The three panels present the calculated probability of being protected from ID in utero by birth month relative to January the year of the IOC program (labeled with 1). In the first panel, the two specifications that make explicit assumptions on the depletion pattern are presented. The multi-compartment model (labelled "Main") is presented with three different baseline iodine intake levels, corresponding to the district mean and with one standard deviation from the mean. This illustrates how this model can account for heterogeneity in depletion rates in a way that the other models cannot. For the hyperbolic model, the base probability is presented as well as the "correction factor" discussed in the text. The second panel presents the cruder models that only implicitly make assumptions on the depletion pattern, representing 12, 18 and 24 months of protection. The third panel present 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.

Table 3: Effect of ID on educational attainment using different treatment models

	All	Females	Males
Main	-0.068	-0.016	-0.143
	(0.111)	(0.170)	(0.150)
Hyperbolic	0.005	0.139	-0.188
	(0.112)	(0.166)	(0.144)
Crude 12m	-0.150	-0.066	-0.225
	(0.126)	(0.203)	(0.173)
Crude 18m	-0.066	-0.014	-0.137
	(0.109)	(0.168)	(0.146)
Crude 24m	-0.018	0.062	-0.137
	(0.105)	(0.158)	(0.137)
Binary 12m	-0.117	-0.112	-0.097
	(0.086)	(0.132)	(0.110)
Binary 24m	-0.003	-0.006	-0.019
	(0.076)	(0.120)	(0.100)
Observations	5204	2654	2550

Notes: This table presents the result from the wide replication incorporating the improvements and changes discussed in Section 4 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; indicator of whether the household resides in an urban setting; indicators for the household head's and spouse's educational level; and indicators of the surveys interacted with survey date. Each row reports the results from one of the treatment specifications discussed in Section 4.2. The first tow (labeled "Main") is our preferred multi-compartment specification. The next row (labeled "Hyperbolic") is our version of FRT hyperbolic depletion model as discussed in the text. 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 and 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 separated for each gender and pooled results for all genders. The last row presents the number of observations in the corresponding column. Each cell reports point estimates and estimated standard errors clustered on the district-cohort level within parentheses. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

12 to 24 months, the estimated effect grows—particularly for females. This roughly mirrors the change in the length of protection from the multi-compartment model to the hyperbolic model, and could explain the difference between the multi-compartment and hyperbolic 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. Both versions of the binary specification produce negative estimates. The estimates are, however, close to zero and not significant. In sum, there appears to be little evidence for a positive effect of the IOC programs on educational attainment.

We use the pooled sample for the analyses presented in Table 3. As discussed in Section 4.1, this could hide potential age heterogeneity. The remaining analysis tries to address this concern. Table 4 separates the estimation by data source. For these estimates, we use our preferred multi-compartment depletion model. Positive effects are found using the samples from the THBS 2000 and NPS 2008 surveys. For the THBS 2000 sample, the effect is almost significant at conventional levels when both genders are included in the analysis. The remaining three surveys exhibit negative effects with the exception for females in the DHS 2010 survey and males in DHS 1999. The estimates are, however, never significantly different from zero at conventional levels.

The variability of the results in Table 4 is notable. While never statistically significant, some surveys and subsample produces large and economically important 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 ID protection increased cognitive ability, we would not expect the sign of the effects to alternate between years in the way displayed, but rather change monotonically as the cohorts became older.

Another possible explanation is potential differences in sampling methods across surveys, so that the samples were 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 *a priori* reason to expect such differences. In the end, the most plausible explanation appears to be ordinary sampling variability; we expect precision to decrease when investigating each survey separately as the samples are smaller.[13] The fact that the smallest surveys show the most variation is indicative of this. If, indeed, sampling variation is the culprit, the pooled approach is likely to yield the most reliable results.

Next, we estimate the pooled analysis broken up in age groups. Age is strongly correlated with surveys as we focus on the 1986-1990 cohorts. There is, however, sufficient overlap in age between survey for an informative analysis. The estimates are presented in Table 8 in the appendix. Notably, the only age group where we find some indication of an effect for those aged 10-13 (i.e., FRT's sample restriction). However, for those aged 10-12 and those aged 10-14 there is no significant effect. As discussed in the narrow replication, we have no reason to exclude those aged 14, and it is hard to rationalize the observed pattern other than by sampling error.

In this wide replication, we have mainly investigated whether different definitions of the treatment variable affect the estimates. The results are, however, qualitatively robust to numerous other specification changes. In particular, in Table 9 in the appendix, we report the results when we alter other aspects of our preferred specification. These changes are based on the robustness checks FRT conduct. The point estimates remain close to zero, and they are never statistically significant at conventional levels.

[13]The estimated standard errors presented here does not account for eventual correlation over time and between districts. The point estimates have, therefore, likely more variation than the reported standard errors indicate.

Table 4: Effect of ID on educational attainment by data source

	All	Females	Males
Pooled	-0.068	-0.016	-0.143
	(0.111) [5204]	(0.170) [2654]	(0.150) [2550]
DHS 1999	-0.264	-0.399	0.001
	(0.176) [462]	(0.310) [230]	(0.257) [232]
THBS 2000	0.195*	0.183	0.190
	(0.104) [3040]	(0.160) [1522]	(0.137) [1518]
DHS 2004	-0.414	-0.712	-0.144
	(0.257) [910]	(0.480) [457]	(0.305) [453]
NPS 2008	0.494	1.484	-2.250
	(0.889) [224]	(1.091) [128]	(1.432) [96]
DHS 2010	-0.316	0.367	-1.236*
	(0.420) [568]	(0.816) [317]	(0.694) [251]

Notes: This table presents the result from the wide replication separating the estimates by survey. The first row contains the full, pooled sample as in Table 3. Each of the subsequent row presents the effect of the IOC programs when estimated in each survey separately. The specification is otherwise identical to the first row in Table 3 (see its note for more details). Each cell reports point estimates, estimated standard errors clustered on the district-cohort level within parentheses and number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

5 Concluding remarks

We have revisited the Tanzanian experience with iodine supplementation in the late 1980s. We find that there is no empirical support for a causal link between mild ID in utero, preventable with supplementation, and schooling later in life. The main issue is the severe measurement error in the treatment variable. The available data never reveal which respondents were protected from ID in utero; like FRT, we must resort to crude approximations of treatment assignment. This leads to imprecision and possible attentution bias.

Our narrow replication highlights that the positive and statistically significant results reported in FRT are restricted to particular subsamples under particular specifications. We argue that this restricted setting is poorly motivated. When we address these issues, the estimated effects are close to zero and not significant at conventional levels. In our wide replication, we increase the sample size almost fourfold by pooling observations from five surveys conducted in Tanzania. We also improve FRT's treatment specification. This is in an effort to minimize attenuation bias and maximize power against the null hypothesis of no effect. Despite our improvements, the estimates remain close to zero and do not support a positive effect.

Our conclusion is that the IOC program in Tanzania is too poorly documented to be used in an impact study using quasi-experimental methods. This also means that we can not conclude that the programs did *not* have an effect. The available data simply do not allow an answer to the substantive question.

The investigation of the Tanzanian IOC experience provides an illustrative example of the challenges associated with causal investigations of *in utero* stress on long-term outcomes, not only in this particular setting but in all cases where data is collected after the intervention (as is typically the case in natural experiments). There are two aspects that make this type of research particularly prone to econometric pitfalls. The first is that there is usually no way of investigating a "first stage"—that is, observing whether an infant was actually protected from health hazards *in utero*. The relevant

treatment definition may depend on complex medical details and small differences in the mothers' environment when pregnant. Without an intermediate outcome known to be affected by the factor under study, misspecification of the treatment variable is an important concern. The second problem is that it is not possible to use covariates to improve identification and precision. Neither can we use covariate balance tests to validate the identification strategy. This is because all observed individual variables are potentially affected by treatment themselves as they are determined after birth. Given these two shortcomings, we believe replication studies are a particularly important tool in this field.

References

Adhvaryu, Achyuta and Anant Nyshadham (2016) "Endowments at Birth and Parents' Investments in Children," *The Economic Journal*, Vol. 126, No. 593, pp. 781–820.

Almond, Douglas and Janet Currie (2011) "Killing me softly: The fetal origins hypothesis," *The Journal of Economic Perspectives*, Vol. 25, No. 3, pp. 153–172.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) "How much should we trust differences-in-differences estimates?" *The Quarterly Journal of Economics*, Vol. 119, No. 1, pp. 249–275.

Cao, Xue-Yi, Xin-Min Jiang, Zhi-Hong Dou, Murdon Abdul Rakeman, Ming-Li Zhang, Karen O'Donnell, Tai Ma, Kareem Amette, Nancy DeLong, and G. Robert DeLong (1994) "Timing of vulnerability of the brain to iodine deficiency in endemic cretinism," *New England Journal of Medicine*, Vol. 331, No. 26, pp. 1739–1744.

Chan, Shiao Y., Marcus H. Andrews, Rania Lingas, Chris J. McCabe, Jayne A. Franklyn, Mark D. Kilby, and Stephen G. Matthews (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*, Vol. 566, No. 2, pp. 467–480.

Dugbartey, Anthony T. (1998) "Neurocognitive aspects of hypothyroidism," *Archives of Internal Medicine*, Vol. 158, No. 13, pp. 1413–1418.

Feyrer, James, Dimitra Politi, and David N Weil (2013) "The cognitive effects of micronutrient deficiency: Evidence from salt iodization in the United States,"Technical report, National Bureau of Economic Research.

Field, Erica, Omar Robles, and Maximo Torero (2009) "Iodine deficiency and schooling attainment in Tanzania," *American Economic Journal: Applied Economics*, Vol. 1, No. 4, pp. 140–69.

Friedhoff, Arnold J., Jeannette C. Miller, Mary Armour, Jack W. Schweitzer, and Sandhya Mohan (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*, Vol. 3, No. 2, pp. 89–97.

Furnée, Carina A., Gerard A. Pfann, Clive E. West, Frits van der Haar, Daan van der Heide, and Joseph G.A.J. Hautvast (1995) "New model for describing urinary iodine excretion: its use for comparing different oral preparations of iodized oil.," *The American Journal of Clinical Nutrition*, Vol. 61, No. 6, pp. 1257–1262.

Haddow, James E., Glenn E. Palomaki, Walter C. Allan, Josephine R. Williams, George J. Knight, June Gagnon, Cheryl E. O'Heir, Marvin L. Mitchell, Rosalie J. Hermos, Susan E. Waisbren, James D. Faix, and Robert Z. Klein (1999) "Maternal thyroid deficiency during pregnancy and subsequent neuropsychological development of the child," *New England Journal of Medicine*, Vol. 341, No. 8, pp. 549–555.

Hassanien, Mohammed H., Laila A. Hussein, Erica N. Robinson, and L. Preston Mercer (2003) "Human iodine requirements determined by the saturation kinetics model," *The Journal of Nutritional Biochemistry*, Vol. 14, No. 5, pp. 280–287.

Ingenbleek, Y., L. Jung, G. Férard, F. Bordet, A. M. Goncalves, and L. Dechoux (1997) "Iodised rapeseed oil for eradication of severe endemic goitre," *The Lancet*, Vol. 350, No. 9090, pp. 1542–1545.

Lavado-Autric, Rosalía, Eva Ausó, José Victor García-Velasco, María del Carmen Arufe, Francisco Escobar del Rey, Pere Berbel, and Gabriella Morreale de Escobar (2003) "Early maternal hypothyroxinemia alters histogenesis and cerebral cortex cytoarchitecture of the progeny," *The Journal of Clinical Investigation*, Vol. 111, No. 7, pp. 1073–1082.

Ministry of Education and Culture, The (2002) *Basic education statistics in Tanzania 1998-2002*, The United Republic of Tanzania, Dar es Salaam.

Murcia, Mario, Marisa Rebagliato, Carmen Iñiguez, Maria-Jose Lopez-Espinosa, Marisa Estarlich, Belén Plaza, Carmen Barona-Vilar, Mercedes Espada, Jesús Vioque, and Ferran Ballester (2011) "Effect of iodine supplementation during pregnancy on infant neurodevelopment at 1 year of age," *American Journal of Epidemiology*, Vol. 173, No. 7, pp. 804–812.

Peterson, Stefan, Vincent Assey, Birger Carl Forsberg, Ted Greiner, Festo P. Kavishe, Benedicta Mduma, Hans Rosling, Alfred B. Sanga, and Mehari Gebre-Medhin (1999) "Coverage and cost of iodized oil capsule distribution in Tanzania," *Health Policy and Planning*, Vol. 14, No. 4, pp. 390–399.

Pharoah, P.O.D., I.H. Buttfield, and B.S. Hetzel (1971) "Neurological damage to the fetus resulting from severe iodine deficiency during pregnancy," *The Lancet*, Vol. 297, No. 7694, pp. 308–310.

Pharoah, P.O.D. and K.J. Connolly (1987) "A controlled trial of iodinated oil for the prevention of endemic cretinism: A long-term follow-Up," *International Journal of Epidemiology*, Vol. 16, No. 1, pp. 68–73.

Politi, Dimitra (2014) "The impact of iodine deficiency eradication on schooling: evidence from the introduction of iodized salt in Switzerland."

Pop, Victor J., Evelien P. Brouwers, Huib L. Vader, Thomas Vulsma, Anneloes L. Van Baar, and Jan J. De Vijlder (2003) "Maternal hypothyroxinaemia during early pregnancy and subsequent child development: a 3-year follow-up study," *Clinical Endocrinology*, Vol. 59, No. 3, pp. 282–288.

Pop, Victor J., Johannes L. Kuijpens, Anneloes L. van Baar, Gerda Verkerk, Maarten M. van Son, Jan J. de Vijlder, Thomas Vulsma, Wilmar M. Wiersinga, Hemmo A. Drexhage, and Huib L. Vader (1999) "Low maternal free thyroxine concentrations during early pregnancy are associated with impaired psychomotor development in infancy," *Clinical Endocrinology*, Vol. 50, No. 2, pp. 149–155.

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

Untoro, Juliawati, Werner Schultink, Clive E. West, Rainer Gross, and Joseph G.A.J. Hautvast (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*, Vol. 84, No. 5, pp. 1208–1214.

Wolff, J. (2001) "Physiology and pharmacology of iodized oil in goiter prophylaxis," *Medicine*, Vol. 80, No. 1, pp. 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, Michael B. (2009) "Iodine deficiency," *Endocrine reviews*, Vol. 30, No. 4, pp. 376–408.

Zoeller, R. T. and J. Rovet (2004) "Timing of thyroid hormone action in the developing brain: Clinical observations and experimental findings," *Journal of Neuroendocrinology*, Vol. 16, No. 10, pp. 809–818.

Appendices

A Additional details on the wide replication

In addition to the specification differences documented in Section 3.1 and Section 4, our wide replication includes minor changes which we document here. The motivation behind these changes are mainly to harmonize the five data sets discussed in Section 4.1 (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 as described in Section 4.2.2, while FRT restrict the sample based on reported age.
- We change the definition of some of the included variables:
    1. As in FRT, the outcome variable is educational attainment as measured by the number of completed grades. However, as non-standard grades are reported differently across surveys, we restrict our attention to established grades in formal education (namely the seven grades in primary school, the four grades in lower secondary and the two in upper secondary). FRT use the variable as reported in the THBS 2000 data set which includes vocational education and other informal schooling.
    2. 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 achievement rather than include the number of completed grades linearly. This has the added benefit of being a non-parametric adjustment.[14]
    3. FRT use the month of the interview as a covariate. They do, however, 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.
- 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 behind this change is that some of these variables are not reported in all data sets we use in the pooled analysis. However, an important secondary reason is that some of these variables are potentially endogenous as all these variables are set long after the IOC programs. For example, fertility decisions could

[14] The educational level 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. In order not to exclude 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 are 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.

be affected by the "quality" of children and, thereby, by treatment. Furthermore, if the IOC programs affected the health of the adults in the household (which was the primary goal), there could be a direct effect on household wealth. In detail, the changes are:

1. We drop variables on family composition (birth order and number of children in household).
2. We drop variables on household wealth (whether the respondent's family is home owners, whether there is a grass roof, if they report food problems, distance to closest secondary school and primary care clinic).
3. 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, this variable lacks institutional motivation.
4. 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.
5. We add an indicator for the respondents' relationship to the household head. As we include all children that are permanent residents in the household in our analysis, this variable could be an important predictor of educational attainment and thus increase precision.

Table 5 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 4). The second column changes the treatment specification to our version of the hyperbolic depletion model (i.e., as discussed in Section 4.2.3), and in column 3, we use FRT's version of the hyperbolic depletion model. As apparent from the table, our two treatment specifications produce roughly the same results, while the point estimate for females with FRT's model is lower and flips sign. 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 and children or grand children of the household head or spouse. This restriction reduces the point estimates further. The remaining two columns, where we first change the definition of the covariates and then add and remove covariates as documented above, only entails small changes in the point estimates. With these changes, the last column in Table 5 is the same specification as in our narrow replication. However, the standard error estimator does not account for serial correlation (i.e., the estimates corresponds to column 6 in Table 2).

Table 5: Step-wise specification change from wide to narrow replication

	(1)	(2)	(3)	(4)	(5)	(6)
All	0.195*	0.199**	0.075	0.008	-0.007	0.026
	(0.104)	(0.098)	(0.104)	(0.084)	(0.085)	(0.089)
	[3040]	[3040]	[2952]	[2739]	[2721]	[2610]
Females	0.183	0.209	-0.043	-0.029	-0.070	-0.044
	(0.160)	(0.138)	(0.177)	(0.135)	(0.135)	(0.131)
	[1522]	[1522]	[1493]	[1348]	[1338]	[1286]
Males	0.190	0.172	0.192*	0.032	0.032	0.035
	(0.137)	(0.128)	(0.112)	(0.113)	(0.115)	(0.119)
	[1518]	[1518]	[1459]	[1391]	[1383]	[1324]

Notes: This table presents the consequences of the specification changes discussed in this appendix. Similar to Table 2, we impose the changes step-wise in a cumulative fashion to show the consequence of each change: (1) presents the results from the wide replication for the THBS 2000 subsample (i.e., the third row in Table 4); (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 some of the variables as discussed in this appendix; and (6) adds and removes covariates to the set of variables used in the narrow replication. The last column is exactly the results in the narrow replication net of the corrected standard error estimates (i.e., column 6 in Table 2). Each cell reports point estimates, estimated standard errors clustered on the district-cohort level within parentheses and number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

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).[15] 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 for the purpose and thus modeled with a higher depletion rate. We assume that iodine can be transported between the compartments freely and without cost,[16] 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 used 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 consumption and any eventual IOC supplements. Their sum in mg 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.[17] 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, or 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 the iodine storage and consumption:

$$\begin{aligned} I_t &= 0.944 \min(I_{t-1} - C_{t-1}, 15) + 0.5 \max(I_{t-1} - C_{t-1} - 15, 0) + 0.7N_t \\ C_t &= \min(I_t, 1.5) \end{aligned}$$

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, in which case 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

[15] 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.

[16] 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.

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

from goiter, we assume 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.[18]

During pregnancy, iodine consumption is increased. 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 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 will consider women with a calculated level of stored iodine higher than 2 mg to offer full fetal protection, while women with stores between 1.5 and 2 mg offer only partial protection (proportionally within that interval). We denote the level of fetal protection offered in month t^* with $S(t^*, \mathcal{H})$, where $\mathcal{H}$ describe the history of IOC administration in the district. We thereby have:

$$S(t^*, \mathcal{H}) = \begin{cases} 1 & \text{if } I_{t^*} \geq 2 \\ (I_{t^*} - 1.5)/0.5 & \text{if } 2 > I_{t^*} \geq 1.5 \\ 0 & \text{else} \end{cases} \tag{2}$$

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 in $t \leq t^*$) 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 at 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 are 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 of importance. A child born in month t would, on average, 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 pregnancy:

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

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 nor 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 the program year and February the following year. As the interaction effect of overlapping programs could be of importance, 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 the program year. The resulting probability of protection for an individual born in year y and month m is given by:

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

[18] 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.

The hyperbolic and cruder treatment specifications discussed in Section 4.2.3 follows a similar pattern, but changes $S(t^*, \mathcal{H})$ to their respective depletion model.

Additional tables

Table 6: The cumulative impact of the issues discussed in the narrow replication, exploiting full birth date information

	FRT's specification						All issues corrected	
	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All	0.347**	0.340**	0.250**	0.155	0.158	0.143	0.123	0.123
	(0.148)	(0.145)	(0.118)	(0.112)	(0.121)	(0.116)	(0.107)	(0.123)
	[1395]	[1395]	[1691]	[2056]	[2056]	[2333]	[2610]	[2610]
Females	0.594***	0.627***	0.542***	0.258*	0.260*	0.245*	0.156	0.156
	(0.170)	(0.171)	(0.156)	(0.147)	(0.147)	(0.149)	(0.152)	(0.153)
	[678]	[678]	[819]	[1009]	[1009]	[1143]	[1286]	[1286]
Males	0.190	0.170	0.088	0.086	0.093	0.072	0.098	0.098
	(0.160)	(0.156)	(0.144)	(0.150)	(0.166)	(0.143)	(0.134)	(0.152)
	[717]	[717]	[872]	[1047]	[1047]	[1190]	[1324]	[1324]

Notes: This table presents the results of the main specification of FRT when the errors discovered in the narrow replication is corrected. Unlike Table 2, we here allow further misspecification in FRT treatment model and use the all available information to estimate the respondents' birth year (as discussed in Section 3). In all other aspects, the analysis is identical to the one behind Table 2 (see its note for further details). Each cell reports point estimates, estimated standard errors within parentheses and number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

Table 7: Each issue discussed in the narrow replication investigated separately

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All	0.347**	0.340**	0.262**	0.244*	0.130	0.329**	0.434***	0.347**
	(0.148)	(0.145)	(0.119)	(0.127)	(0.124)	(0.139)	(0.135)	(0.176)
	[1395]	[1395]	[1691]	[1705]	[1395]	[1558]	[1500]	[1395]
Females	0.594***	0.627***	0.521***	0.328**	0.248	0.573***	0.687***	0.594***
	(0.170)	(0.171)	(0.157)	(0.160)	(0.184)	(0.183)	(0.172)	(0.185)
	[678]	[678]	[819]	[839]	[678]	[760]	[730]	[678]
Males	0.190	0.170	0.113	0.195	0.046	0.156	0.222	0.190
	(0.160)	(0.156)	(0.147)	(0.163)	(0.139)	(0.151)	(0.149)	(0.182)
	[717]	[717]	[872]	[866]	[717]	[798]	[770]	[717]

Notes: This table presents the results of the main specification of FRT when the errors discovered in the narrow replication is corrected separately. That is, in each column we correct *only* the corresponding issue and leave all other aspects as in FRT's original specification. In the first column, labeled (0), the replicated estimates from FRT are presented (i.e., without any issues corrected). Each subsequent column corrects one issue discussed in the text: (1) use `gsort` rather than `sort` to derive birth order; (2) include children in single parent households; (3) add children that are 14 years old; (4) derive birth years based on reported survey year; (5) add children that cannot be linked to a unique mother; (6) add grandchildren till household heads; (7) cluster estimates of the standard error at the district level rather than district-cohort level to account for serial correlation. Apart from these changes, the specifications are identical to the main specification in FRT as described above. Each cell reports point estimates, estimated standard errors within parentheses and number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

Table 8: Effect of ID on educational attainment by age

	All	Females	Males
Age 10-12	0.157	0.211	0.090
	(0.102)	(0.153)	(0.155)
	[2191]	[1088]	[1103]
Age 13-15	-0.475	-0.661	-0.203
	(0.318)	(0.492)	(0.295)
	[1502]	[759]	[743]
Age 16-18	-0.087	-0.770	0.060
	(0.359)	(0.897)	(0.503)
	[537]	[270]	[267]
Age 19-21	0.013	0.868	-0.812
	(0.400)	(0.814)	(0.640)
	[502]	[263]	[239]
Age 10-13	0.210**	0.297**	0.137
	(0.102)	(0.140)	(0.150)
	[2833]	[1403]	[1430]
Age 10-14	0.168*	0.170	0.162
	(0.097)	(0.159)	(0.124)
	[3509]	[1749]	[1760]

Notes: This table presents the result from the wide replication separated by age group. The sample in all rows is the full, pooled data set from all five surveys discussed in Section 4.1. However, since we restrict the study to the 1986-1990 cohort, the age restrictions will implicitly restrict the analyses to a subset of the surveys. For example, when estimating the effect for the younger ages (10-12 and 10-13), only the DHS 1999 and THBS 2000 surveys are used. The first four rows present the effect for 3 year wide age groups from age 10 to 21. The two final rows present the effect for ages 10-13 (as in FRT) and ages 10-14 (as in our narrow replication). The specification is otherwise identical to the first row in Table 3 (see its note for more details). Each cell reports point estimates, estimated standard errors clustered on the district-cohort level within parentheses and number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.

Table 9: Robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
All	-0.068	-0.063	-0.023	0.052	-0.005	0.045	-0.113	-0.060	-0.081	-0.128	-0.068
	(0.111)	(0.112)	(0.102)	(0.062)	(0.101)	(0.155)	(0.118)	(0.106)	(0.124)	(0.109)	(0.132)
	[5204]	[5705]	[4189]	[9972]	[5204]	[2850]	[5204]	[5204]	[3913]	[5204]	[5204]
Females	-0.016	-0.016	0.074	0.063	-0.041	0.158	-0.084	-0.001	0.018	-0.158	-0.016
	(0.170)	(0.168)	(0.157)	(0.086)	(0.189)	(0.222)	(0.169)	(0.171)	(0.164)	(0.169)	(0.205)
	[2654]	[2908]	[2123]	[5016]	[2654]	[835]	[2654]	[2654]	[1879]	[2654]	[2654]
Males	-0.143	-0.135	-0.095	0.037	-0.004	0.037	-0.139	-0.139	-0.157	-0.097	-0.143
	(0.150)	(0.151)	(0.151)	(0.083)	(0.149)	(0.249)	(0.157)	(0.143)	(0.179)	(0.157)	(0.200)
	[2550]	[2797]	[2066]	[4956]	[2550]	[1459]	[2550]	[2550]	[2034]	[2550]	[2550]

Notes: This table investigates how numerous specification changes affect the results presented in the wide replication. The first column (1) is the main specification unchanged (i.e., correspond exactly to the first row of Table 3). The subsequent columns changes a single aspect of the analysis: (2) We include two districts (Mbinga and Bukoba Rural) that was targeted by the IOC programs very late, presumably on other grounds that the other districts. FRT also exclude these districts. (3) We exclude the 1986 cohort. The specification is, thus, closer to FRT with respect to included cohorts. (4) We include the 1991-1994 cohorts. These children were in utero when universal salt iodization had started in Tanzania. There is, therefore, likely severe contamination of the control group for these cohorts. (5) We use ward fixed-effects rather than effects on the district level. As treatment was assigned on the district level, this will not aid identification but could improve precision if inter-ward migration is not too common. (6) We use household fixed-effects. Again, this would only increase precision. However, as this exclude all household with only one child in the included cohorts, the sample size is severally reduced with a decrease in precision as a consequence. (7) We use a specification will less controls. We here only include indicators for age, cohort, gender and survey date. (8) We use a specification that include more controls—more closely mirroring FRT original specification. In addition to the control in our wide replication, we here include the number of children in the household; birth order of the respondent within the included cohorts; whether the household owns their dwelling, have grass roofs, have experience food shortages, and its distances to the nearest secondary school and health clinic. There are, however, slight differences between the surveys in the coding of these variables. Notably, some of the household wealth indicators are missing from the DHS surveys. We include missingness indicators in those cases. (9) We exclude all children that are not sons or daughters of the household head or spouse as in FRT. This mainly excludes grandchildren to the household head. (10) We include district specific cohort trends in addition to fixed-effects. (11) We cluster the estimation of the standard errors on the district level rather than on the district-cohort level. Each cell reports point estimates, estimated standard errors clustered on the district-cohort level within parentheses (unless otherwise noted) and number of observations within square brackets. Asterisks indicate significant estimates at the 10% (*), 5% (**) or 1% (***) level.