Introduction

What is restricted early life growth and why is it potentially important?

I use the term *restricted early life growth* as a general term to refer to low weight and/or height in the first 1000 days of life, relative to genetic growth potential, from conception to approximately age two. This term encompasses several subcategories: stunting, low birth weight, and small for gestational age. Generally, *stunting* is defined as being markedly small for age postnatally. More specifically, the World Health Organization defines stunting as a height shorter than two standard deviations below its Child Growth Standards median (1,2). Two related concepts that apply to prenatal development are *low birth weight* and *small for gestational age*. Low birth weight can result from slower weight accrual during gestation and/or premature birth, and is typically defined as a birth weight less than 2,500 grams (5 pounds, 8 ounces) (3). Small for gestational age refers only to slower weight accrual during gestation and is often defined as a birth weight in the bottom ten percent of a standard age-specific weight distribution (4).

Prenatal and postnatal development over the first 1,000 days of life is often considered a key period with disproportionate impact on later life outcomes. Advocates in the research, public health, and nonprofit communities suggest that this period is particularly important for the development of adult physical stature, cognitive function, health, and other determinants of well-being and adult economic status.[1] The primary purpose of this report is to evaluate a portion of this argument.

[1] From the 1,000 Days website (5): “Good nutrition is critical to support the rapid growth and development of babies and young children during their first 1,000 days. Without good nutrition however, a young child can suffer serious and often permanent damage to his developing brain and body. We can’t really see this damage but we can measure it by looking at how well a child is or isn’t growing. A child who doesn’t grow well and is too short for their age suffers from a condition known as stunting… The effects of stunting last a lifetime: impaired brain development, lower IQ, weakened immune systems, and greater risk of serious diseases like diabetes and cancer later in life.”

From Victora and colleagues 2008, page 340 (6):

- “Poor fetal growth or stunting in the first 2 years of life leads to irreversible damage, including shorter adult height, lower attained schooling, reduced adult income, and decreased offspring birthweight.
- Children who are undernourished in the first 2 years of life and who put on weight rapidly later in childhood and in adolescence are at high risk of chronic diseases related to nutrition…
- The prevention of maternal and child undernutrition is a long-term investment that will benefit the present generation and their children.”

From Galasso and colleagues 2017, page 5 (7): “Stunting in childhood matters because it is associated with adverse outcomes throughout the life cycle. The undernourishment and disease that cause stunting impair brain development, leading to lower cognitive and socioemotional skills, lower levels of educational attainment, and hence lower incomes. Health problems in terms of non-communicable diseases are more likely in later life, leading to increased health care costs. Stunting in childhood also leads to reduced stature in adulthood, which, due to the persistence of shortness over the lifetime, and the negative (and independent) effect of height on income, further reduces income in adulthood.”

This hypothesis influences the public health approach of the World Health Organization, governments, and nonprofit organizations, and money spent on these efforts is viewed as an investment in the future that will be repaid to society many-fold.[2]

What causes restricted early growth?

Stunting, low birth weight, and small for gestational age are quantitative descriptions that do not imply a specific mechanism. They are thought to reflect multiple mechanisms alone or in combination, including undernutrition of the mother during gestation, undernutrition of infants and children after birth, twinning, cigarette smoking during gestation, infectious disease, and genetics (9–11).[3] Undernutrition is a condition that results from a deficiency of nutrients, often calories and/or protein.

In low-income countries, undernutrition, short stature and/or underweight, and infectious disease of mothers and children appear to be predominant causes of restricted early growth, and stunting is much

[2] From page 1 of the World Health Organization document “Essential Nutrition Actions” 2013 (8): “This document provides a compact summary of WHO guidance on nutrition interventions targeting the first 1000 days of life. Focusing on this package of essential nutrition actions, policy-makers could reduce infant and child mortality, improve physical and mental growth and development, and improve productivity.”

From the 1,000 Days website page on stunting (5): “Beyond the individual impacts of this problem, stunting is an enormous drain on economic productivity and growth. Economists estimate that stunting can reduce a country’s GDP by as much as 12%.”

From Victora and colleagues 2008, page 340 (6): “The prevention of maternal and child undernutrition is a long-term investment that will benefit the present generation and their children.”

From Galasso and colleagues 2017, pages 5 and 6 (7): “Stunting among children today reduces a country’s future income per capita… Our estimates suggest that implementing the Bhutta et al. program, and factoring in the annual trend decline of 1.5% p.a., will leave the stunting rate in 2025 at 36% below to its 2010 value – 4 percentage points shy of the 40% target reduction adopted by the 65th World Health Assembly. We estimate a rate-of-return for the 34 countries as a whole of 17%, with a benefit-cost ratio of 15:1.”

[3] From Kramer 1987, page 663 (9): “Factors with well-established direct causal impacts on intrauterine growth include infant sex, racial/ethnic origin, maternal height, pre-pregnancy weight, paternal weight and height, maternal birth weight, parity, history of prior low-birth-weight infants, gestational weight gain and caloric intake, general morbidity and episodic illness, malaria, cigarette smoking, alcohol consumption, and tobacco chewing. In developing countries, the major determinants of [intrauterine growth restriction] are Black or Indian racial origin, poor gestational nutrition, low pre-pregnancy weight, short maternal stature, and malaria.”

Magnus 1984 suggests that genetics exert an important influence on birth weight (12). From the abstract: “The results indicate that more than 50% of the total variation in birth weight is caused by variation in fetal genes, and that less than 20% is caused by variation in maternal genes. The remaining variance (20-30%) could be explained by random environmental effects.”

From Behrman and Rosenzweig 2004, page 591 (10): “For example, twins are significantly smaller at birth than are nontwins. Table 1 reports descriptive statistics for the two samples of female twins that we use. As seen in table 1, the average birthweight of twins is 5 lb. 10 oz. The average birthweight in the general population is 7 lb. 6 oz. Moreover, by the standard definition of low birthweight—below 5 lb. 8 oz. (2.5 kg)—almost half of the twins were low birthweight.”

more common in these countries than in affluent nations with adequate food supplies and effective sanitation (9).[4]

The fact that early growth integrates multiple potential causes has advantages and disadvantages. The advantage is that early growth is a marker of the overall quality of the developmental environment. A notable disadvantage is that small body size can result from different causes in different environments. This may increase the variability of data linking early growth to later life outcomes, make it more challenging to identify ultimate causes and solutions, and limit the external validity of the evidence.

Our outcome measure of economic status also integrates a variety of influences, including physical and cognitive abilities and health. This has similar advantages and disadvantages to our measure of early growth.

The goal of this report

The primary goal of this report is to evaluate the hypothesis that being born small for gestational age and/or stunted in the first two years of life meaningfully reduces individual adult economic outcomes. We are particularly interested in nongenetic influences on these measures because they are potentially modifiable. Secondary goals include:

1. Determine the magnitude of the effect.
2. Determine if growth during the first 1,000 days impacts adult economic productivity more than other periods of development.
3. Determine the relative importance of the prenatal vs. postnatal portions of the first 1,000 days.

My prior beliefs

Before writing this report, I did not have strong prior beliefs about the impact of early life growth restriction on economic outcomes. Based on general plausibility and the fact that this connection is commonly asserted in high-profile review papers and taken seriously by major organizations such as the World Health Organization and the World Bank, I had a weak prior belief that restricted early life growth negatively impacts adult economic outcomes, but I had no beliefs about the size of the effect.

Methods

My initial strategy was to read a wide variety of review papers, identify useful categories of evidence, and identify particularly important meta-analyses and primary studies. I focused in particular on the Lancet 2008 and 2013 series on maternal and child nutrition, materials from the website of the

[4] From Kramer 1987, page 663 (9): “In developing countries, the major determinants of [intrauterine growth restriction] are Black or Indian racial origin, poor gestational nutrition, low pre-pregnancy weight, short maternal stature, and malaria.”

de Onis and colleagues list the 2010 prevalence of stunted preschool children at 6.0% in “developed countries” and 29.2% in “developing countries” (13).

nonprofit organization 1,000 Days, review papers from the World Bank, and primary research papers from the National Bureau of Economic Research (7,14–21).

To collect evidence on biological plausibility and animal research, I relied on review papers in the animal literature that I identified using Google Scholar, as well as my own general knowledge of biology. Search terms included “early development”, “growth”, “stunting”, “malnutrition”, “rat”, “primate”, and “review”. I did not conduct a systematic search of the literature in this area, and I focused primarily on review papers, none of which were systematic.

To identify observational evidence, I focused on review papers, and particularly meta-analyses when available. I performed literature searches for meta-analyses and systematic review papers using the initial search strategy described previously, as well as Google Scholar, the economics literature database RePEc via the website EconPapers, and the reference list of review papers I had identified. Search terms included “early development”, “height for age”, “length for age”, “stunting”, “malnutrition”, “birth weight”, “economic”, “earnings”, “meta-analysis”, “systematic review”, and “review”. I did not perform a systematic literature search or conduct a detailed analysis of the primary literature in this area.

I performed a systematic search of the primary research literature to identify all relevant natural experiment studies. I located studies using review papers and literature searches via Google Scholar and EconPapers. Search terms included “twins”, “natural experiment”, “early development”, “height for age”, “length for age”, “stunting”, “nutrition”, “malnutrition”, “birth weight”, “economic”, and “earnings”. I then performed backward and forward citation searches using Google Scholar, EconPapers, and the reference lists of the papers I had identified. Due to the high value of this category of evidence, I focused on the primary literature and read each paper in full.

I also performed a systematic search of the primary research literature to identify all relevant randomized controlled trials. I located studies using review papers and literature searches via Google Scholar and EconPapers. Search terms included “randomized”, “early development”, “height for age”, “length for age”, “stunting”, “malnutrition”, “birth weight”, “economic”, and “earnings”. I then performed backward and forward citation searches using Google Scholar, EconPapers, and the reference lists of the papers I had identified. I initially focused on trials that reported direct measures of individual economic status such as income, employment characteristics, and material possessions. Due to the high value of this category of evidence, I read each paper in full. After finding that the research literature relevant to my primary outcome is very limited, I expanded my search to three indirect measures of future economic status: height, years of schooling, and cognitive abilities. I evaluated the literature on these indirect measures in less detail, relying on a 2015 systematic review paper to identify studies, and supplementing with additional relevant studies that I encountered in my previous searches (17).

Throughout this report, I interpret a p-value of less than 0.05 as statistically significant and everything above 0.05 as nonsignificant. Many of the included econometric papers applied a significance threshold of $p < 0.10$ but also reported when an outcome met $p < 0.05$ and < 0.01 thresholds. Due to this discrepancy in significance thresholds, my interpretation of results occasionally differs from that of the authors.

Sources of evidence and their strengths and weaknesses

Several independent approaches are available to evaluate our hypothesis, and I believe that incorporating all of them increases my likelihood of obtaining reliable conclusions (22,23). Yet not all approaches are equally informative.

Basic plausibility and animal research

The first source of evidence I will consider is basic plausibility and animal research. Plausibility cannot itself provide strong supporting evidence, but an implausible hypothesis requires stronger experimental evidence to support.

Animal research has several advantages over other categories of evidence. Since it involves true experiments, it can clearly identify cause-and-effect relationships. It also allows tightly controlled experiments with full compliance with experimental procedures, which is difficult or impossible in human studies. Finally, it allows detailed studies of the mechanisms underlying the effects.[5]

A key disadvantage of animal research is that animals differ from humans, so findings aren't necessarily relevant to our species. All nonhuman species have a different developmental timeline than humans, and differences in adult brain size and structure demonstrate that developmental processes are not identical even after correcting for the duration of development.[6]

This may be particularly true of brain development, due to the uniqueness of the human brain. In general, species that are more closely related to humans are considered more accurate models of human physiology, and by analogy this seems likely to be true of development as well (26). Another disadvantage of animal research is that laboratory conditions such as housing and feed may have little in common with those that humans experience, potentially reducing external validity further.

[5] From Levitsky and Strupp 1995 (24): "The study of animals is essential for estimating the potentially harmful effects of malnutrition to humans because only through experiments with animals can causal relationships between early malnutrition, alterations in brain structures and resulting behavioral and cognitive consequences be established." I disagree with the above quote that animal experiments are the only way to establish causality in this domain. Human randomized controlled trials can theoretically do the same, but in practice they are difficult to implement and those performed to date have limited evidence value (discussed in later sections).

[6] From Rice and Barone 2000, page 514 (25): "A significant difference between the rodent and human (and other primate) brains can be observed in the size and contour of the neocortex. Rodents have a smooth (lissencephalic) cerebral cortex with a relatively small neocortex. Humans, on the other hand, have brains with a highly convoluted surface (gyrencephalic brain) resulting from the enormous phylogenetic expansion of the neocortex. The ontogeny of specific regions of the brain incorporate the timely progression and completion of these developmental processes. This sequence for a specific region may generalize readily between rodents and humans; however, at the level of integration and connectivity among structures, with the exception of sensory systems, it is often only speculative to extrapolate results from animals to humans because of the limited number of comparative studies."

This paper was published in the year 2000. We now have more detailed information on brain homology between humans and rodents, but the general concern remains.

Furthermore and most obviously, nonhuman animals aren't economically productive in the ways that are most relevant to our hypothesis, so we have to rely on proxy measures like adult body size, health, and cognitive abilities.[7] Due to these limitations, I will not attempt to systematically review the animal literature on this topic, but rather conduct a shallow investigation via review papers with the goal of forming reasonable priors about our hypothesis that I can update using evidence from human studies.

Observational studies

Observational studies measure and correlate variables in free-living people without altering a variable. For example, researchers may measure the birth weight of one thousand people and see how naturally-occurring differences in birth weight correlate with naturally-occurring differences in adult earnings 25 years later. Observational studies are conducted on people living their typical lives and so the results can be highly relevant to real-life conditions. They are also relatively inexpensive and easy to implement so they can include sample sizes up to tens of thousands of people (27).

The primary limitation of observational studies is that they are typically not as effective at identifying cause-and-effect relationships as controlled studies. It is often unclear whether associations reflect a causal relationship or whether they result from other confounding factors that either were not measured or were not corrected for appropriately. There is compelling evidence that confounding factors can substantially distort the results of observational studies relevant to our hypothesis.[8]

In our current context, genetics is a particularly challenging confounding factor. Small size is often assumed to be a sign of suboptimal development, yet it is not difficult to imagine that genetic variants might exist that favor smaller babies and smaller adults even under optimal conditions. It also seems possible that unlucky genetic variants could reduce early growth and also lead to poor health, and/or impaired brain function, reducing adult economic status. If this is true, then correcting low birth weight in these cases would not completely correct detrimental adult outcomes. To the extent that genetic effects cannot be corrected by improving living conditions, they may lead observational studies to misestimate the potential benefits of doing so.[9]

[7] From Levitsky and Strupp 1995 (24): "Unfortunately, the major disadvantage of using animals is a greater dependence on interpretation and extrapolation to infer events in humans than if humans could be studied directly."

[8] From Behrman and Rosenzweig 2004, page 599 (10): "The effect of increasing birthweight on schooling, moreover, is underestimated by 50% if there is no control for genetic and family background endowments as in cross-sectional estimates, and the use of standard family background variables—parental schooling and father's earnings—to reduce endowment heterogeneity does not reduce these biases significantly…our estimates also suggest that the cross-country correlation between incomes and birthweight substantially overstates the reduction in world earnings inequality that would arise from reducing cross-country disparities in birthweight."

[9] From Behrman and Rosenzweig 2004, page 586 (10): "However, the literatures concerned with the consequences of weight gain at birth generally do not distinguish between the policy-relevant effects of increasing the nutrients received by a fetus and possible genetic and family background influences on fetal development. It is possible that infants with genetically determined low weights at birth also have genetic endowments that make them less healthy as adults, and that increasing their weight at birth would have little lasting effect on adult life achievements. It also is possible that the genetic endowments of such infants are correlated with those of their parents and thereby associated with the household resources available to support child development. What is of policy interest is the effect of increasing the birthweight of a child with given genetic endowments and family

Another common limitation of observational studies is that the practical challenges of working with large numbers of people impose constraints on data collection methods that can result in low data quality (28). Fortunately, for our purposes birth weight and the length of gestation can often be measured accurately even decades in the past by examining birth records (10).

Natural experiments

Perhaps the most important class of study for evaluating our hypothesis is the *natural experiment*. This is technically a type of observational study, but it exploits natural conditions that approximate a randomized controlled trial. For example, in the Chinese Great Famine of 1959-1961, some counties were hit harder than others. Researchers subsequently compared the adult outcomes of people who experienced different levels of famine severity (20).

The Great Famine studies also illustrate a limitation of most natural experiments; counties that experienced famine differed in habitual grain production levels from counties that did not, suggesting that the comparison groups were not well "randomized".[10]

Baseline differences between groups of comparison can lead to systematic biases in results, and these biases can be difficult to detect or correct. However, I believe one type of natural experiment can be as well randomized and controlled as the best randomized, controlled trials: monozygotic (identical) twin studies. These studies compare siblings that are genetically identical, developed in the same uterus, had the same gestation length, were born at the same time, were weighed at the same time on the same scale, and were raised in the same household. Yet due to the random chance involved in placental placement and function, one sibling often received more nutrients than the other and was born significantly larger (10). This offers an exceptionally well-controlled measurement of the impact of *in utero* growth on later outcomes. Twin studies that do not distinguish between monozygotic and dizygotic (fraternal) twins are relatively well controlled for most of the reasons listed above, however they are not fully controlled for genetic differences due to the fact that dizygotic siblings only share half their genes.

background. Conventional cross-sectional and longitudinal data sets cannot answer that question unambiguously."

From Black and colleagues 2005, page 410 (18): "Until recently, analysis of birth weight effects has relied primarily on cross-sectional variation and has established a relationship between low birth weight and poor health, cognitive deficits, and behavioral problems among young children. It has also provided evidence that this relationship persists for longer term outcomes such as health status, educational attainment, employment, and earnings [for example, Barker 1995, Currie and Hyson 1999, Case et al. 2004]. However, it is possible that there are no underlying causal relationships, as low birth weight may be correlated with many difficult-to-measure socio-economic background and genetic variables"

[10] From Meng and Qian 2009, page 10 (20): "For the purposes of this paper, the most important finding of MQY (2009) is the strong positive correlation between grain production in 1959 and famine intensity, which was a reversal from normal years when production was negatively correlated with mortality."

One limitation of twin studies is that they tend to be conducted in affluent countries that keep accurate records of birth and family information, such as the US and Norway (10,18). Low birth weight of twins in affluent countries could result from different mechanisms than low birth weight in the general population of low-income countries. Because of this, external validity to the general population of lower-income countries may be limited. This is significant because this demographic is the primary target of interventions to address restricted early growth. However, this problem may be mitigated by the fact that twins tend to have low birth weights, increasing the relevance of twin studies to situations in which low birth weight is common.[11]

Despite these limitations, the ability of twin studies to provide credible estimates of causal relationships makes them particularly informative.

Randomized controlled trials

In *randomized controlled trials*, people are randomly allocated to separate groups, an exposure variable is altered in at least one group, and an outcome measure is compared between groups. They are generally considered the gold standard for investigating cause-and-effect relationships in humans. However, the practical realities of these experiments mean that they can be difficult to apply to large populations and/or long follow-up periods. In our case, the outcome variable follows the exposure variable by 16 or more years. Such a long follow-up period tends to reduce sample size, obscure outcome data and increase its variance. In addition, it can be difficult to obtain good adherence of the participants to the research protocol, and sometimes it is difficult to know whether or not good adherence has been achieved.

The latter limitation applies to the research on this topic in general. The fact that our outcome variable often lags our exposure variable by decades imposes challenges for any study design. For this reason, at times in my examination of the randomized controlled trial literature I will rely on proxy measures of adult economic productivity that can be measured before adulthood, such as physical stature, cognitive ability, and years of schooling. Since these are proxy measures with uncertain causal impacts on our outcome of interest, they must be interpreted cautiously.

Category-specific conclusions

Basic plausibility and animal research

Basic plausibility

[11] From Behrman and Rosenzweig 2004, page 591 (10): "For example, twins are significantly smaller at birth than are nontwins. Table 1 reports descriptive statistics for the two samples of female twins that we use. As seen in table 1, the average birthweight of twins is 5 lb. 10 oz. The average birthweight in the general population is 7 lb. 6 oz. Moreover, by the standard definition of low birthweight—below 5 lb. 8 oz. (2.5 kg)—almost half of the twins were low birthweight."

Development requires energy and nutrients, and it is biologically plausible that an insufficient supply of either could impair tissue development, including the tissues that determine stature, musculature, immune function, brain function, and other determinants of adult economic status. Likewise, any substantial physical stressor such as infectious disease could plausibly impair tissue development, with adverse effects on adult economic productivity. Human brain volume increases most rapidly in the first 1,000 days of life (29). In addition, different periods of human development involve qualitatively different developmental processes (30). It therefore seems plausible that impaired development in the first 1,000 days could lead to irreversible deficits.

Animal research

Animal research presents a way to form reasonable priors about the hypothesis that restricted early growth can impact characteristics that are relevant to adult economic status in humans. Experiments beginning in the 1960s demonstrated that calorie and/or protein restriction of pregnant females and/or in early postnatal life leads to a permanent reduction of adult offspring brain size in mice, rats, guinea pigs, and pigs (24). This is consistent with human observational studies suggesting that malnutrition is associated with reduced brain size in children (31). Furthermore, malnutrition in animals causes substantial alterations in brain microstructure.[12]

These alterations in brain structure correspond with deficits of behavioral flexibility, emotional reactivity, learning, and memory (32). These attributes are analogous to human skills that contribute to economic status. Importantly, in animal models of malnutrition, microstructural and behavioral deficits can be partially or entirely corrected by adequate nutrition and/or environmental enrichment later in development (24,33).[13]

[12] From Levitsky and Strupp 1995 (24): “Despite the sparing of the total number of cortical neurons from the ravages of malnutrition, more sophisticated analyses of cortical structures continued to reinforce the idea that malnutrition caused permanent structural damage to the brain. Studies using Golgi staining techniques show that malnutrition causes a significant disruption in pyramidal cells of the cerebral cortex (Angulo-Colmenares et al. 1979, Cordero et al. 1985, Leuba end Robinowicz 1979b, Noback and Eisenman 1981, Salas et al. 1974, Schonheit 1981, Schonheit and Haensel 1989, West and Kemper 1976), reduction in the density of cortical dendritic spines (Angulo-Colmenares et al. 1979, Leuba and Rabinowicz 1979b, Noback and Eisenman 1981, Salas et al. 1974, Sarkar et al. 1990, Schonheit 1982, Schonheit and Haensel 1989, West and Kemper 1976), a decrease in the width of cortical cells (Angulo-Colmenares et al. 1979, Leuba and Rabinowicz 1979b, Salas et al. 1974) and the complexity of the dendritic branching of the cortex (Leuba and Rabinowicz 1979a, Leuba and Rabinowicz 1979b, Schonheit 1982, Yoshida 1985). In addition, the total number of cortical glial cells is significantly reduced by early malnutrition (Leuba and Robinowicz 1979a). Although the density of cortical synapses appeared to be unaffected by malnutrition Cragg 1972, Gambetti et al. 1974, Warren and Bedi 1984), the total number of synapses in visual cortex is clearly reduced by malnutrition (Warren and Bedi 1984). The length and the width of synaptic reactive zones are also reduced, and the number of cisterns embedded within the spinous apparatus is also significantly altered by malnutrition (Medvedev and Babichenko 1988).”

[13] From Levitsky and Strupp 1995 (24): “Given these rather profound anatomical effects of early malnutrition observed during or immediately after the period when brain was growing at its maximum or near maximum rate, it is not surprising that neuroscientists predicted that enduring alterations in cognitive function would persist. More recent evidence, however, suggests that the term irreversibly may have been premature and that many of the earlier animal studies may not have allowed sufficient time for recovery of various structures to occur. Studies of the recovery of the brain after the period of malnutrition revealed that the period of mitotic activity of the cortex of the rat is prolonged after early malnutrition (Gopinath et al. 1976), allowing the period of maximal brain protein

Early deficiency of calories and/or protein can substantially retard linear growth, which is relevant because adult stature is thought to contribute to economic status, particularly in contexts where manual labor is most prevalent (34,35). However, as with the behavioral deficits of malnutrition, animals appear to be flexible in their developmental timelines and (at least in some cases) they can largely make up for low birth weight and/or early stunting by faster later growth if they gain access to adequate nutrition. This appears to depend on the severity, duration, and timing of the malnutrition, among other factors, with more severe, longer, and earlier exposures believed to lead to more pronounced deficits.[14] In humans, the ability to recover partially or completely from early growth restriction is known as "catch-up growth".[15]

synthesis to continue (Hamberger and Sourander 1978)… Remarkable recovery of other brain parameters from early malnutrition were also demonstrated… Historically, the predictions that children will be intellectually damaged by early malnutrition were based on anatomical perturbations. As indicated above, the results of long-term studies do not support the assertion that the anatomical changes observed in the cortical regions during or immediately after malnutrition are irreversible."

[14] From Boersma and Wit 1997, page 649 (36):

"After mild nutritional restriction, mammals and birds generally achieve their normal body size and conformation at a later stage of growth (30). Periods of prolonged or excessive restriction may, however, cause permanent stunting. Wilson et al. (30) have identified six factors that appear to be important for the extent of compensatory growth:

1. The nature of undernutrition; very severe protein restriction may have a more harmful effect than very severe energy restriction.
2. The severity of undernutrition; the more severe the restriction, the greater is the initial rate of gain immediately after realimentation, but the smaller is the ultimate weight.
3. The duration of the period of undernutrition; excessively long periods of restriction may result in permanent stunting.
4. The stage of development of the body at the start of undernutrition; it was suggested that undernutrition in the earlier stages of growth has a more harmful effect for an animal than restriction at a later stage, and that the ability to recover and to reach normal mature size is consequently reduced.
5. The relative rate at which the species matures; slower maturing realimentated animals make a more rapid recovery from undernutrition than faster maturing ones.
6. The pattern of realimentation; the higher the plane of nutrition upon realimentation, the more rapid and the greater the recovery in weight of cattle."

From the abstract of Elias and Samonds 1977 (34): "The growth and development of 32 cebus monkeys were studied during a period of insult in nutritional or rearing conditions and after rehabilitation… The period of insult from 2 to 6 months of age was followed by 6 months of rehabilitation in both diet and rearing conditions… All groups caught up in physical growth during rehabilitation but the protein-calorie restricted groups failed to recuperate completely in exploratory behavior."

From Lizárraga-Mollinedo and colleagues 2015, page 19,361 (35): "After switching to ad libitum diets, the undernourished rats' [lean body mass] increased more markedly, which explains why their body lengths reached similar (with chow) or identical (with cafeteria formula) values to that of the age-matched controls, despite the severity and duration of food restriction."

[15] From Boersma and Wit 1997, page 646 (36): "As early as the beginning of this century, it was reported that animals with retarded growth due to undernutrition can achieve a growth rate higher than normal for chronological age after removal of the food restriction (1–3). A supranormal height velocity may also be observed in children recovering from starvation or illness. In medical literature, little or no attention was paid to this

Although brain function and linear growth can catch up to nearly normal levels following restricted early growth in animal models, research suggests that it may leave a lasting susceptibility to obesity and metabolic disease that is not corrected by nutritional rehabilitation (35,37,38).[16] Body fatness and metabolic health are important in their own right, but they are also mechanisms that could underlie a causal relationship between early life growth and later economic outcomes.

Tentative conclusions

Animal research suggests that restricted early growth due to calorie and/or protein restriction can lead to alterations in the development of attributes that are analogous to human attributes that impact economic productivity, including smaller brain size, alterations of brain microstructure, cognitive and emotional deficits, reduced linear growth, and increased susceptibility to obesity and metabolic disease. However, the trajectory of nonhuman animal development is flexible and these alterations appear to be partially or entirely reversible with adequate nutrition later in development, apparently depending on the severity, duration, and timing of the restriction. More severe, longer, and earlier restriction appear to produce more pronounced permanent deficits in animal models.

This animal research suggests that although early-life growth is an important determinant of adult attributes, substantial and permanent deficits require that adverse conditions persist later into development, preventing catch-up growth. Because of this, it seems possible that observational studies might overestimate the importance of the first thousand-day period if they do not accurately control for the fact that poor conditions in early life tend to be followed by poor conditions later.

Smaller brain size and susceptibility to obesity and metabolic disease appear to be less sensitive to adequate nutrition later in development than other traits discussed in this section. Although the behavioral deficits associated with smaller brain size are often reversible in animals, it seems possible (although speculative) that this wouldn't be as true of humans due to our exceptional dependence on higher brain functions.

phenomenon of accelerated linear growth until the second half of the 20th century. One of the first reports on this subject was published in 1963 (4). In that article, Prader et al. (4) introduced the term catch-up growth to describe the phase of rapid linear growth that allowed the child to accelerate toward and, in favorable circumstances, resume his/her pre-illness growth curve."

[16] From Cottrell and Ozanne 2007, page 18 (38): "In rats and mice, impaired early growth due to maternal protein restriction, the well-defined model used in our laboratory, results in offspring of a lowered birth weight, and who exhibit asymmetric alteration in tissue growth, leading to a preservation of brain growth at the expense of other organs such as the pancreas and kidneys [24]. These animals then go on to develop impairments in glucose tolerance with age, contributed to by a reduction in pancreatic beta cell mass, reduced insulin secretion and peripheral tissue insulin resistance, in low-protein exposed offspring [25–29]. These effects are worsened by the presence of obesity, in an additive manner[30], and in addition the presence of accelerated postnatal growth, induced by suckling in small litters (4 pups) by a control-fed dam, both accentuates the effects of fetal growth restriction and reduces longevity in rats and mice [24,31,32]. In a rat model of total caloric restriction during pregnancy, offspring are hyperphagic, hyperinsulinemic, develop obesity and hypertension, and exhibit reduced activity levels [33,34]. Other models of early growth restriction have produced similar findings, and again there is an amplification of metabolic disturbances in the presence of a highly-palatable or high-fat diet in later life [31,33]."

Since these studies were not conducted in humans and didn't directly measure economic outcomes, we must be very cautious in interpreting them. My primary source of uncertainty in applying these findings to our hypothesis is in the degree to which they are relevant to humans. However, the fact that similar effects are observed in multiple species including nonhuman primates leads me to believe that they are relevant enough to form a reasonable set of priors.

Observational studies

Due to my judgment that observational studies offer limited evidence value relative to other study designs (with the exception of natural experiment studies, which strictly speaking are observational), I limited my investigation to a shallow non-systematic literature search, and I did not spend significant time critically evaluating study methods. The primary reason for reviewing this evidence is that it is commonly cited as a rationale for early life interventions intended to improve adult outcomes, and it therefore seems useful to have a sense of.[17] I chose to focus on one relatively recent, highly cited systematic review and meta-analysis of the observational literature.

In 2008, Victora and colleagues published a meta-analysis as part of a *Lancet* series on maternal and child undernutrition (6). It includes data from five cohorts in Brazil, Guatemala, India, the Philippines, and South Africa, but only the Brazilian, Guatemalan, and Indian cohorts report direct measures of economic status. Presumably due to the small number of studies that met their inclusion criteria, the authors did not pool results or report a single quantitative estimate for economic outcomes. Outcomes were inconsistent between studies but overall suggested an association between indicators of early undernutrition and lower adult income.[18]

[17] From pages 340 and 353 of Victora and colleagues (6): "In this paper we review the associations between maternal and child undernutrition with human capital and risk of adult diseases in low-income and middle-income countries. We analysed data from five long-standing prospective cohort studies from Brazil, Guatemala, India, the Philippines, and South Africa and noted that indices of maternal and child undernutrition (maternal height, birthweight, intrauterine growth restriction, and weight, height, and body-mass index at 2 years according to the new WHO growth standards) were related to adult outcomes (height, schooling, income or assets, off spring birthweight, body-mass index, glucose concentrations, blood pressure)." "Our results strongly suggest that undernutrition leads to long-term impairment. This evidence, combined with the well-known short-term effects of undernutrition, is sufficient for giving the prevention of undernutrition high priority in national health, education, and economic agendas in low-income and middle-income countries."

From pages 6 and 7 of the 2013 paper "The World Health Organization's global target for reducing childhood stunting by 2025: rationale and proposed actions" (39): "Evidence demonstrates that stunting in early life is associated with adverse functional consequences including poor cognition and educational performance, low adult wages, lost productivity and, when accompanied by excessive weight gain later in childhood, increased risk of nutrition-related chronic diseases (Victora et al. 2008)." [Victora et al. 2008 is the paper referenced previously in this footnote] "This paper summarises the rationale for the global target on stunting, describes stunting trends from 1990 until 2025, presents a methodology to adapt the global target at the national level, and reviews what can be done to reduce stunting."

[18] From page 346 of Victora and colleagues 2008 (6): "Most indicators of undernutrition were associated with lower income in Brazil and fewer assets in India, but in Guatemala few associations were significant. The most consistent results for men were for height-for-age: 1 Z score was associated with an 8% increase in income in Brazil

The authors also report indirect measures that could contribute to economic outcomes. In their analysis, one kilogram of additional birth weight is associated with 3.3 centimeters of additional adult height and 0.3 years of additional schooling, and one standard deviation of additional height at age two is associated with 3.2 centimeters of additional adult height and 0.5 years of additional schooling.

Tentative conclusions

Observational studies tend to suggest that birth weight and early postnatal growth are associated with individual adult economic outcomes. They also suggest that birth weight and early postnatal growth are associated with attributes that may contribute to economic outcomes, such as height and schooling. However, this evidence will have little impact on my overall conclusions because more informative study designs are available.

Natural experiments

I began my investigation by performing a systematic literature search for natural experiment studies relevant to the hypothesis, as described in the methods section. My inclusion criteria were:

- Studies had to either report a measure of growth in the first thousand days of life, or exposure to a situation that would very plausibly impact growth in the first thousand days of life (e.g., famine during gestation).
- Studies had to report a measure of individual adult economic status.
- The early life exposure had to be at least somewhat randomly distributed, in other words not obviously strongly correlated with variables that could confound the relationship between restricted early life growth and economic outcomes. Note that this is simply the definition of a natural experiment.
- Manuscripts had to be complete, rather than incomplete drafts.
- When I encountered duplicate (or very similar) manuscripts published in different places, I only included one version.

Types of studies identified, and their relevance

Each type of natural experiment study approaches the ideal of randomization in its own way. In twin studies, researchers compare adult outcomes within twin pairs that differ in birth weight.[19]

(p<0·0001) and Guatemala (p=0·07), as well as with an increase of 0·27 household assets in India (p<0·0001). Associations with weight were less consistent."

[19] From Behrman and Rosenzweig 2004, page 590 (10): "In our sample, the average absolute value of the difference in birthweights within pairs of [monozygotic] twins is 10.5 ounces, with substantial variation across pairs (figure 2). The corresponding figure for same-sex nonidentical twin pairs is 11.2 ounces. Thus there is ample and real within-twin variation to obtain precise estimates of birthweight input effects using twin difference methods."

When a study compares adult outcomes within twin pairs, it is automatically controlled for many thorny confounding factors such as fetal environment, gestation length, time of birth, and home environment. When the comparison is only between monozygotic twins, it is additionally controlled for genetic differences. Because differences in birth weight between monozygotic twins are due to seemingly random differences in how effectively each fetus competes for nutrients, the overall quality of "randomization" seems likely to be superior to other types of natural experiments.[20]

Dizygotic twins are more genetically similar to one another than two unrelated people, but they share only half their genes so comparisons between them are only partially controlled for genetic differences.

Famine studies take advantage of random or semi-random natural or human-made disasters that plausibly influenced nutrient availability during early growth. Researchers compare between cohorts that were exposed vs. spared, or experienced famine with different degrees of intensity. Comparisons between cohorts can be across time (birth year) and/or space (birth location). Famine studies could plausibly be confounded by a variety of variables that don't act via direct impacts on early growth. For example, a disaster may damage schools resulting in lower educational attainment, or famine severity may correlate with baseline local characteristics that themselves impact adult economic prospects.[21]

Ramadan fasting studies take advantage of the fact that observant Muslims do not eat during daylight hours for approximately one month of each year, and that pregnant women are not generally exempt.[22]

[20] From Behrman and Rosenzweig 2004, page 588 (10): "The key identifying assumption is that although the average of the nutrient intakes for any twin pair may be correlated with their common endowment, which may reflect their genetic heritage and resource allocation decisions of their parents, the birthweight difference within [a monozygotic] twin pair reflects purely random differences in nutrient intakes that may arise, for example, from differences in womb position. Twin-specific intakes that are expressed in birthweight differences are thus orthogonal to their identical endowments."

[21] The following passage from page 10 of Meng and Qian 2009 explains that during the 1959-61 Great Famine in China, areas with higher typical grain production levels were affected most strongly by the famine (20). Since typical grain production level is probably correlated with many other features of society, this suggests that the comparison between counties that differed in famine intensity may not have been "apples to apples" and could have suffered from confounding that is difficult to identify or correct. "For the purposes of this paper, the most important finding of MQY (2009) is the strong positive correlation between grain production in 1959 and famine intensity, which was a reversal from normal years when production was negatively correlated with mortality."

[22] From Almond and Mazumder 2008, appendix A.1.2 (40): "Although pregnant women may request an exemption from fasting, they are expected to 'make up' the fasting days missed during pregnancy after delivery and this requirement may discourage pregnant women from seeking the exemption since they may be the only member of the household fasting [Hoskins, 1992, Mirghani et al., 2004]. Mirghani et al. [2004] noted: 'Most opt to fast with their families rather than doing this later'… As far as we are aware, comprehensive data on Ramadan fasting during pregnancy do not exist. Various surveys of Muslim women suggest that fasting is the norm. For example, of the 4,343 women delivering in hospitals in Hamadan, Iran in 1999, 71% reported fasting at least 1 day, 'highlighting the great desire of Muslim women to keep fasting in Ramadan, the holy month'[Arab and Nasrollahi, 2001]. In a study in Singapore, 87% of the 181 muslim women surveyed fasted at least 1 day during pregnancy, and 74% reported completing at least 20 days of fasting [Joosoph and Yu, 2004]. In a study conducted in Sana'a City, Yemen, more than 90 percent fasted over 20 days [Makki, 2002]. At the Sorrento Maternity Hospital in Birmingham, England, three quarters of mothers fasted during Ramadan [Eaton and Wharton, 1982]. In a study conducted in Gambia, 90 percent of pregnant women fasted throughout Ramadan [Prentice et al., 1983]. In the US, a study of 32 Muslim women in Michigan found that 28 had fasted in at least one pregnancy and reported that 60-90 percent of women from their communities fast during pregnancy [Robinson and Raisler, 2005]."

Ramadan fasting can lead to a decrease in calorie intake and body weight, including in pregnant women, although based on the supporting data cited in the papers I reviewed it's not currently clear to me how common this is.[23] Research suggests that pregnant women are particularly vulnerable to adverse effects of fasting, a phenomenon called "accelerated starvation".[24] Although the observant are allowed to eat as much as desired after sunset, this temporary energy restriction may restrict fetal development.[25]

[23] From Almond and Mazumder 2008, appendix A.1.1 (40): "Ramadan fasting in the adult population (i.e. not conditioning on pregnancy) has been associated with modest but statistically significant declines in the weight of fasters of around 1 to 3 kg (Husain et al. [1987]; Ramadan et al. [1999]; Adlouni et al. [1998]; Mansi [2007]; Takruri [1989]) Reductions in weight are sometimes (but not always) accompanied by declines in caloric intake and likely depend on dietary customs in specific countries. Two studies are of particular relevance. First, in a study of 185 pregnant women, Arab [2003] found that over a 24 hour period encompassing the Ramadan fast, over 90 percent of the women had a deficiency of over 500 calories relative to the required energy intake and 68 percent had a deficiency of over 1000 calories. Second, in the only large scale population-based study we are aware of, Cole [1993] found striking evidence of sharp weight changes during Ramadan for women in Gambia."

[24] From Almond and Mazumder 2008, pages 4 and 5 (40): "Writing in The Lancet, Metzger et al. [1982] documented a set of divergent biochemical measures among pregnant women who skipped breakfast in the second half of pregnancy. Relative to twenty-seven non-pregnant women with similar characteristics, "circulating fuels and glucoregulatory hormones" changed profoundly in twenty-one pregnant women the "overnight fast" was extended to noon on the following day (relative to post- prandial baseline). Further, plasma glucose and alanine was lower in the pregnant women than in the non-pregnant women after 12 hours of fasting while levels of free fatty acids and beta-hydroxybutyrate, a ketone, were significantly higher. This set of biochemical changes, also known as "accelerated starvation", occurred after only "minor dietary deprivation" for both lean and obese women. Metzger et al. [1982] concluded that meal skipping "should be avoided during normal pregnancy." Meis and Swain [1984] found that daytime fasts during pregnancy caused significantly lower glucose concentrations than nighttime fasts." "Following the study of breakfast skipping by Metzger et al. [1982], Ramadan fasting was likewise found to cause accelerated starvation among pregnant women in Gambia [Prentice et al., 1983] and in England [Malhotra et al., 1989]. Mirghani et al. [2004] found that maternal glucose levels were lower in the fasting state compared to the postprandial baseline, a difference accentuated by the number days fasted: "the effect on maternal glucose levels during Ramadan fasting is cumulative." Several studies of maternal fasting during Ramadan have found adverse effects carried over to measures of fetal health: fetal breathing movements and fetal heart rate accelerations [Mirghani et al., 2004, 2005]."

[25] Evidence of this comes from (modest) differences of birth weight that occur in pregnancies exposed to Ramadan. From Almond and Mazumder 2008, page 17 (40): "Overall, in utero exposure to Ramadan is associated with lower birth weight among Michigan's Arab mothers. A full month of exposure to Ramadan during the peak period of daylight hours could lead to a reduction in birth weight of about 40 grams especially when Ramadan falls in the first month of gestation. Nevertheless, the size of this effect is relatively small: 40 grams is only about 1.2 percent of the mean birth weight for Arabs."

Evidence that Ramadan fasting may have adverse effects on fetal development also comes from differences in the sex ratio at birth. A lower male:female sex ratio suggests in utero stress because male fetuses are more likely to be lost when stressed. From Almond and Mazumder 2008, page 18 (40): "Mathews et al. [2008] found that poor maternal nutrition (possibly due to breakfast skipping), around the time of conception skews the sex ratio in favor of girls, most likely through the selective attrition of male conceptuses. Similarly, Almond et al. [2009] found that severe morning sickness in early pregnancy is associated with female births, but also a 50% fetal death rate due to severe nausea and vomiting.28 More generally, maternal nutrition among mammals close to conception is positively associated with the likelihood of male offspring Cameron [2004]. We consider Ramadan's effect on sex at birth and the number of lives births in Table 3. Using our full sample of Arab mothers (column 1) we find a large effect of -3.7 percentage points (p-value = 0.06) on the likelihood of a male birth from exposure to Ramadan during the longest diurnal fast in month 1 of pregnancy. In column (2) when we restrict the zipcodes to those with fewer Chaldeans relative to Arabs, this point estimate rises substantially to -6.6 percentage points and is significant at the 1 percent level."

Usefully, the month of Ramadan shifts by approximately 11 days each year, meaning that its dates cover an entire year over the course of 33 years. Given a sufficient number of years of data, this allows researchers to separate the effects of Ramadan *per se* from the time of year during which it occurs, avoiding confounding due to prevailing conditions in one season vs. another. However, there are still potential limitations to this type of study, for example it is difficult to disentangle the effects of fasting *per se* from other aspects of the environment that differ during Ramadan.[26]

Ramadan fasting studies don't actually measure who fasted and who didn't; they measure "Ramadan exposure" in Muslim populations that probably fast. The disadvantage of this approach is that effects may be diluted by the fact that some people don't fast. The advantage is that it avoids dividing cohorts into "Muslims who fast" vs. "Muslims who don't fast", and the potential confounding that could arise from all of the other personal characteristics that correlate with being more vs. less observant of Muslim traditions. Instead, studies compare adult outcomes of people who were in an early growth window vs. outside that window during Ramadan, yielding a measurement that is more suited for causal inference at the expense of diluting the effect size.

The adoption study approaches randomization because adoptees were "quasi-randomly" assigned to families. Because adoptee characteristics such as birth weight were semi-independent of adoption family characteristics such as income, this type of study design may be less confounded by family characteristics.[27] However, this study design does not address confounding due to genetic differences between adoptees, and the authors' term "quasi-random" implies that assignment was not totally random. If adoptee characteristics were somewhat matched to adoption family characteristics, then confounding due to family characteristics may persist to some degree.

Summary of findings

Majid 2015 found evidence consistent with this as well, however statistical significance was not reported (41): "The overall sample is representative of males and females with a 1:1 sex ratio. And among those exposed, there are fewer men. Given that men are known to be more responsive to nutritional deficits in utero, this is consistent with the findings in Almond and Mazumder (2011) and Van Ewijk (2011), who find that males are more likely to die from Ramadan exposure."

[26] Majid 2015 raises this concern, listing several possible influences besides fasting per se that could account for observations (41). No references are provided to support these statements and it is unclear to me how important and widespread these specific behaviors are among Muslims. Still, the general point is valid that "Ramadan exposure" likely entails more than just fasting, and although fasting seems like the most plausible explanation for the observed outcomes, we can't be certain of this. "Although this paper identifies the Ramadan effect, it is not clear whether religious fasting by mothers is driving these results. A change in eating behavior after sunset (iftaar), which involves eating greasy, oily and generally unhealthy foods may be causing the real harm rather than calorie restriction during fasting. Changes in sleep patterns may also occur. People may also work less during Ramadan due to fatigue. All these factors may confound the Ramadan effect from the fasting effect."

[27] From Beach and Saavedra 2015, page 2 (42): "To study how these measures interact with LBW, we use data from Sacerdote (2007) in which an adoption agency quasi-randomly assigned Korean orphans to American adoptive families. Because we have random assignment, childhood environment is not confounded by genetics, prenatal health, or neonatal healthcare. Consistent with the findings of Currie and Hyson (1999) and Cheadle and Goosby (2010), the interaction of LBW with parent's education, family income, or family size is not statistically significant. The interaction between LBW and median neighborhood income, however, is positive and significant."

Studies are summarized in Table 1. Of the 23 natural experiment studies, 11 report predominantly statistically significant associations between early life development and adult economic outcomes that are in the direction predicted by the hypothesis being evaluated (10,18,40–48). Eight offer mixed results that are somewhat supportive of the hypothesis (20,49–55). For example, a statistically significant association is observed in women but not in men and not in both sexes combined. Four studies are predominantly unsupportive of the hypothesis (56–59).[28] No studies support the *opposite* hypothesis, i.e. that early life growth restriction is associated with superior economic outcomes.

Among the eight twin studies, four predominantly support the hypothesis (10,18,46,47), two are somewhat supportive (50,55), and two are predominantly unsupportive (56,57). Among the eleven famine studies, four are predominantly supportive (43–45,48), six are somewhat supportive (20,49,51–54), and one is predominantly unsupportive (59). Among the three Ramadan fasting studies, two are predominantly supportive (40,41) and one is predominantly unsupportive (58). The adoption study is predominantly supportive (42).

Superficially, this appears to be a fairly consistent body of evidence suggesting that early life growth restriction impairs individual economic outcomes in adulthood. Although not all studies are supportive, this predominantly supportive pattern is what I expect from a body of literature that uses imperfect methods to study a real effect. The fact that similar results have been obtained from very different study designs is reassuring. However, it is worth noting that the "vote counting" method I have just applied (counting supportive vs. unsupportive studies) is not a well-regarded method for assessing a body of research.[29]

Rather than relying on a vote counting approach, I will evaluate each study in detail and weight them according to their informativeness in my conclusions. First, I will explain a few general reasons to be skeptical of this literature.

[28] Note that classifying these four studies as unsupportive is my interpretation, not necessarily that of the study's authors. My interpretation may differ from the authors due to 1) my more stringent significance threshold of p less than 0.05, rather than the threshold of p less than 0.10 commonly used in this literature; 2) the fact that some authors chose to focus on a minority of statistically significant findings even if the findings were predominantly null; and 3) the fact that my outcome of interest (adult economic status) was not always the primary focus of the paper.

[29] From Borenstein and colleagues, page 251 (60): "One question we often ask of the data is whether or not it allows us to reject the null hypothesis of no effect. Researchers who address this question using a narrative review need to synthesize the p-values reported by the separate studies. Since these are discrete pieces of information and the narrative review provides no statistical mechanism for synthesizing these values, narrative reviewers often resort to a process called vote counting. Under this process the reviewer counts the number of statistically significant studies and compares this with the number of statistically nonsignificant studies… One might think that summarizing p-values through a vote-counting procedure would yield more accurate decision than any one of the single significance tests being summarized. This is not generally the case, however. In fact, Hedges and Olkin (1980) showed that the power of vote-counting considered as a statistical decision procedure can not only be lower than that of the studies on which it is based, the power of vote counting can tend toward zero as the number of studies increases. In other words, vote counting is not only misleading, it tends to be more misleading as the amount of evidence (the number of studies) increases!"

Reference	Year	Category	Location	Findings	Notes
Behrman and Rosenzweig (10)	2004	Twins (monozygotic)	United States	1 lb (0.45 kg) higher birth weight-for-gestation-length associated with 7% higher adult earnings	Studied women only. Cohort spans 1936-55 and may have less relevance to modern births.
Black et al. (18)	2005	Twins (mixed, with one monozygotic sample)	Norway	1 lb (0.45 kg) higher birth weight associated with 1.6% higher earnings and 1.7% higher full-time earnings in mixed-twin sample. Smaller (~0.9%) and nonsignificant in monozygotic sample.	Effect size for earnings was smaller and nonsignificant in monozygotic subsample. Effect observed in total sample, and among men but not women when considered separately.
Miller et al. (55)	2005	Twins (monozygotic)	Australia	Birth weight was not significantly associated with adult earnings	Although nonsignificant, the effect size is very similar to Behrman and Rosenzweig 2004 (6.4% higher adult earnings per lb) and almost statistically significant. Birth weight is self-reported and subject to recall error. Smaller sample size.
Oreopoulos et al. (56)	2006	Twins (mixed; ~25% monozygotic)	Canada	Little or no association between birth weight and social assistance use in adulthood	Genetic confounding is possible due to inclusion of dizygotic twins
Almond et al. (43)	2007	Famine	China	Higher famine severity during gestation associated with unemployment of adult men, smaller house size of adult women	“Randomization” may not have been ideal, e.g. counties with worse famines had higher typical grain production levels
Chen and Zhou (44)	2007	Famine	China	Famine exposure during gestation and postnatal years 1 and 2 associated with 28% fewer work hours, lower home garden income, smaller houses in adults	“Randomization” may not have been ideal, e.g. counties with worse famines had higher typical grain production levels. Small sample size.
Almond and Mazumder (40)	2008	Ramadan fasting	United States, Uganda, Iraq	Adult Ugandan men exposed to Ramadan in 1st month of gestation 2.6% less likely to own a home. Adult Iraqis exposed to Ramadan in 1st month of gestation less likely to have multiple wives or own a home. Adult US subjects exposed to Ramadan in first trimester have 5.5-6% lower earnings.	Did not measure whether or not subjects fasted, although text suggests that most pregnant Muslim women fast during Ramadan. Likely religious affiliation of US sample was inferred from ancestry. Dates of Ramadan shift each year, avoiding time of year confounding.
Maccini and Yang (49)	2008	Famine	Indonesia	Adult women but not men who experience higher rainfall in the first	Assumes that higher rainfall in rural areas is correlated with greater

				postnatal year have a higher asset index. No significant association between higher rainfall *in utero* and asset index.	income and food availability. Lower rainfall in the first postnatal year could theoretically impact food security for the first two years of life. Asset index is "log total value of household assets and indicators for ownership of television, refrigerator, private toilet, and stove".
Mu and Zhang (45)	2008	Famine	China	Greater famine severity *in utero* through age 2 associated with higher nonworking rate in adults	"Randomization" may not have been ideal, e.g. counties with worse famines had higher typical grain production levels.
Meng and Qian (20)	2009	Famine	China	Exposure to famine at ages 1-6 associated with 13.9% fewer hours worked in adulthood. No significant association between *in utero* exposure and hours worked.	Focuses on the top decile of outcomes to mitigate survivorship bias. Attempts to control for inadequate "randomization" due to different levels of typical grain production between counties.
Royer (50)	2009	Twins (mixed)	United States	Below 2.5 kg only, one SD higher birth weight (~0.54 kg) associated with $1,300 greater mean income of county of residence in adulthood	Genetic confounding is possible due to inclusion of dizygotic twins. Studied women only.
Neelsen and Stratmann (51)	2011	Famine	Greece	Famine exposure *in utero* or during year 1 or 2 of life associated with lower-prestige adult employment. Similar employment effect for exposure *in utero*, in year 1, or in year 2. Result became nonsignificant after controlling for rural vs. urban birth.	"Randomization" may not have been ideal, e.g. cohorts born in different years may have experienced environmental differences besides famine exposure. Cohorts span 1936-46 (famine lasted 6-8 months in 1941-42) so outcomes may be less relevant to modern births. Urban areas were hit harder than rural areas, potentially biasing results upward. Short duration of famine allows separate examination of effects *in utero*, in year 1, and in year 2.
Shi (52)	2011	Famine	China	Greater famine intensity in year 1 and 2 after birth associated with lower household wealth of adult women but not men. No significant association	"Randomization" may not have been ideal, e.g. counties with worse famines had higher typical grain production levels. To address this, they

				with labor supply of men or women.	control for provincial time trends. Also controlled for fertility selection. Speculate that gender difference is because parents preferentially fed males during famine.
Rosenzweig and Zhang (46)	2012	Twins (mixed, same-sex)	China	One SD higher birth weight (0.6 kg) associated with 12% higher wage in adult men but not women	Same-sex twins are enriched for monozygotic twins. Genetic confounding is possible due to inclusion of dizygotic twins. Authors speculate that gender difference is because the mechanism is via body size/strength rather than intelligence.
Jurges (53)	2013	Famine	Germany	Men *in utero* during the worst part of the famine 0.9% more likely to be unemployed and 3% more likely to have a blue collar job. No significant association for women.	Worst famine period was Feb-July 1945 but low food availability persisted until at least 1948. Hard to distinguish between effects of famine and other post-war hardships.
Nakamuro et al. (57)	2013	Twins (monozygotic)	Japan	No association between birth weight and adult earnings when comparing between twins	Data on both twins (including birth weight) self-reported by one twin, introducing error. Quality of data is unclear.
Schultz-Nielsen et al. (58)	2014	Ramadan fasting	Denmark	Males in 7th month of gestation during Ramadan were 2.6% less likely to be employed as adults. No significant associations with salary, wage rate, or annual hours worked. No significant associations for months 0-6 or 8-9	Finding is likely attributable to chance (see text). Studied men only. Did not measure whether or not subjects fasted, although text suggests that most pregnant Muslim women fast during Ramadan. Likely religious affiliation was inferred from country of origin. Dates of Ramadan shift each year, avoiding time of year confounding.
Beach and Saavedra (42)	2015	Adoption	United States	Birth weight less than 2.5 kg is associated with 20% lower adult earnings	Sample is 75% women. "Quasi-random" assignment of adoptees reduces family background confounding but not genetic confounding. Birth weight data are not high quality.
Bharadwaj et al. (47)	2015	Twins (mixed)	Sweden	10% higher (+250 g) birth weight associated with 1% higher permanent adult income	Genetic confounding is possible due to inclusion of dizygotic twins. Primary cohort spans 1926-1958 and may have less

					relevance to modern births, although results are consistent with a more recent cohort (1973+).
Fan and Qian (59)	2015	Famine	China	No association between *in utero* or early postnatal famine exposure and adult income	“Randomization” may not have been ideal, e.g. counties with worse famines had higher typical grain production levels. Small sample size.
Majid (41)	2015	Ramadan fasting	Indonesia	Ramadan exposure *in utero* associated with lower birth weight, 10% fewer hours worked per week, and a 7.8% higher likelihood of being self-employed in adulthood (indicating low-skill work). Strongest association in highly religious families; no effect in non-Muslims.	Indonesia is the most populous Muslim country. Sample size was very large and representative. Sample allowed comparison between Ramadan-exposed and non-exposed sibs. Date of birth was self-reported. Religious affiliation was measured. Dates of Ramadan shift each year, avoiding time of year confounding.
Scholte et al. (54)	2015	Famine	Netherlands	1st trimester famine exposure associated with ~3% lower adult employment likelihood but no significant association with income. No significant labor associations in 2nd or 3rd trimesters.	Studied men only due to low participation of women in the labor market. Studied urban areas only. Compared famine intensity across both time and location to control for location and time confounding. Famine is “sharply defined in time and space” (Dec-May 1944-1945 in West Netherlands). Effect sizes could be underestimated due to survival selection.
Mussa (48)	2017	Famine	Malawi	10% lower relative corn yield while a future farmer is *in utero* associated with 4.2% lower relative adult corn production. No association with postnatal years 1 or 2.	Corn is the primary staple crop and “the best direct indicator of [rural] incomes” in Malawi. Consumption of corn is 133 kg/yr which equals more than 50% of calorie needs. Small sample size.

Table 1. Natural experiment studies on the relationship between restricted early life growth and individual adult economic outcomes.

General reasons for skepticism

My background is in biomedical science, and the studies in this section use econometric methods that I am not very familiar with. Given my limited familiarity with what lies “under the hood” of these methods, it is difficult for me to fully assess the studies' quality. However, I noticed design features that would be problematic in any scientific context, which I will explain.

I have not seen evidence that the authors adjusted their p-value thresholds for the fact that they performed multiple hypothesis tests on the same dataset. This is called the "multiple comparisons problem" and it can greatly increase the likelihood of a false positive finding.[30] To explain the problem, the most commonly used statistical significance threshold allows a five percent false positive rate due to random chance, so if we perform ten significance tests, there is a 40 percent chance that at least one of them will return a false positive.[31]

This problem can be addressed by making the significance threshold more stringent in proportion to the number of tests performed (e.g., Bonferroni correction). Due to the substantial number of significance tests in most of these studies, many results would not survive correction for multiple hypothesis testing, which reduces my confidence in the findings. I can address this problem crudely by applying Bonferroni correction myself, and I will do so for a few particularly informative studies, with the understanding that this correction can be viewed as conservative (i.e., can sometimes produce false negatives).[32]

A related problem is that none of the study methods appear to have been preregistered. Preregistration of methods allows the reader to be confident that authors have not adjusted data analysis methods and/or outcome reporting (whether intentionally or unintentionally) to favor a particular outcome. Without preregistration, authors are left with an ability to tinker with methods that can bias results. Judging by the papers I encountered, preregistration does not appear to be common in this field, although I did find evidence that it sometimes occurs in the broader field of econometrics.[33] Preregistration has become an important component of rigorous study design in other domains such as biomedical randomized controlled trials (64).

Publication bias is another concern I have about this body of literature, particularly since preregistration appears to be rare. Publication bias occurs where statistically nonsignificant or inconvenient results are less likely to be published, biasing the literature toward studies that support a hypothesis. This is also

[30] From the abstract of Streiner 2015 (61): "Testing many null hypotheses in a single study results in an increased probability of detecting a significant finding just by chance."

[31] The calculation is 1 - ((1-0.05)^10).

"When 10 statistically independent tests are performed, the chance of at least one test being significant is no longer 0.05, but 0.40." (62)

[32] I will use Behrman and Rosenzweig 2004 as an example here (10). Table 2 reports the paper's primary findings and includes 12 hypothesis tests on the associations between fetal growth and later schooling, body mass index, height, and adult wages using three different models. If we apply Bonferroni correction to this table, a finding would have to cross a threshold of p = 0.004 to be statistically significant. If we are less stringent and only consider the number of hypotheses tested by the most stringent model (4), which is the one the paper focuses on, the Bonferroni-corrected significance threshold is p less than 0.0125. Using this new threshold, the key finding of the paper that birth weight is associated with adult income becomes nonsignificant by a slim margin (p = 0.0137; calculated using t-statistic, sample size, and two-tailed t-test).

[33] From the "instructions for authors" page of the website of the journal Econometrics (63): "Preregistration: Where authors have preregistered studies or analysis plans, links to the preregistration must be provided in the manuscript."

called the “file drawer effect” (65). It would be difficult for me to perform meaningful tests of publication bias in this setting because the studies in question lack features that are required for such tests.[34] So the concern about publication bias remains, and increases my uncertainty about the degree of support provided by this literature.

Measurement error is another general concern that applies more to some study designs than others. Two monozygotic twin studies illustrate this problem. Behrman and Rosenzweig used hospital records containing objectively measured birth weights of monozygotic twins to quantify the association between birth weight and adult economic outcomes (10). Nakamuro and colleagues used self-reported birth weights of monozygotic twins for the same purpose, and in fact they asked one person of each twin pair to report the birth weight of both twins via an online survey (57). It is not hard to imagine that this might result in substantially lower data quality in the latter study, weakening associations. Does this explain the fact that Behrman and Rosenzweig reported a statistically significant association, while Nakamuro and colleagues did not? It is difficult to know, and there are other potential explanations, however it is possible. Other potential sources of measurement error include limitations of the questionnaires and government databases used to measure adult economic outcomes.

Observational studies are prone to confounding, meaning that associations may not accurately reflect a cause-and-effect relationship due to the impact of additional (often unmeasured) variables. A classic example of this is the finding that men who shave every day have a lower overall mortality rate than men who shave less often (69). The naïve interpretation is that shaving reduces the risk of death. A more sophisticated interpretation acknowledges that the two groups of men differed in many ways, including height, employment, and smoking rates, that shaving frequency was probably just a marker of lifestyle factors that impact death risk, and concludes, as the authors did, that “the association between infrequent shaving and… mortality is probably due to confounding”.

[34] The most commonly used test of publication bias is the funnel plot, which plots the effect size vs. sample size of included studies. For a discussion of publication bias and how to identify it, see Borenstein and colleagues, chapter 30 (60). The underlying assumption is that the results of large studies are more likely to be published than those of small studies because more resources were invested in them and more people are expecting results, yielding a greater incentive to publish regardless of outcome. If small studies tend to report different effect sizes than large studies, particularly in a direction that favors a preferred hypothesis, this suggests the possibility of publication bias in the overall literature.

Unfortunately, this tool is not applicable to the current body of literature because the underlying assumption that large sample sizes correlate with a larger resource investment by the researchers (and greater incentive to publish regardless of outcome) is not valid. In the current context, researchers often rely on databases that were collected by other institutions such as hospitals and governments, not specifically created for the current work (one exception is Berman and Rosenzweig 2004, which used a mailed survey for some of its data (10)). Furthermore, it is not obvious that econometric methods are substantially more demanding for large vs. small datasets, and nearly all the papers under consideration had 1-4 authors. Compare this with large biomedical randomized controlled trials, which can involve more than 100 authors (see appendices in the following papers) (66,67).

Another tool for detecting publication bias is the p-curve (68). This method has the advantage of not requiring that some studies have higher publication pressure than others. However, it does require exact p-values, or the means to calculate them, which are not present in many of the natural experiment papers. Often, p-values are reported as ranges rather than exact numbers (e.g., $p < 0.05$ or $p < 0.01$).

Natural experiment studies are observational studies that are trying to be randomized controlled trials, and they succeed to varying degrees. The better the "randomization"—that is, the less the cause of growth restriction (e.g., famine severity) is correlated with baseline characteristics of the affected people (e.g., household earnings)—the lower the potential for confounding. As argued earlier, monozygotic twin studies appear to have the lowest potential for confounding of all study designs under consideration. Ramadan fasting studies may also approximate randomization fairly well because exposed vs. unexposed individuals can be compared within the same communities and families and at many times of year. It is not difficult to imagine how most famine studies might suffer from confounding. Typically, these studies compare either a) populations in different locations that were affected with different severity at the same time, or b) populations that were born at different times in the same location, thereby being exposed with different severity. People who are born and raised in different places or at different times have baseline individual differences that cannot be completely attributed to famine exposure, and these differences could plausibly confound our relationship of interest.[35]

For these reasons, I am somewhat skeptical of this body of evidence as a whole, although my degree of skepticism varies greatly between studies.

A closer look at seven studies

Among the 23 studies I identified, I believe seven stand out as the most informative so I will discuss them each in detail. Due to their designs, these seven studies appear to have the lowest risk of confounding and therefore, in my opinion, support the strongest causal inference. I include all twin studies that examine a monozygotic sample, except Nakamuro and colleagues 2013 due to its less compelling method for measuring birth weight (discussed previously). I also include all three Ramadan fasting studies due to their compelling study designs that appear fairly resistant to confounding. Lastly, I include Mussa 2017, a famine study that exploits regional variation over many years of data on grain harvest, potentially avoiding time- and place-specific confounding factors that bedevil other famine studies.

The earliest natural experiment study I identified is also one of the most informative: Behrman and Rosenzweig 2004 (10). The authors collected data on birth weight, gestation length, and wages at the most recent job (among other variables) from 804 female monozygotic twins from the Minnesota Twin Registry. Birth weight and gestation length data are high-quality because they were collected from

[35] On page 10, Meng and Qian 2009 provide an example of a potential confounding factor (20): "For the purposes of this paper, the most important finding of MQY (2009) is the strong positive correlation between grain production in 1959 and famine intensity, which was a reversal from normal years when production was negatively correlated with mortality." Note that typical grain production levels probably correlate with many other unmeasured variables.

hospital records.[36] The authors collected wage data by mailing out surveys.[37] Although wages were self-reported, it seems likely that respondents were able to provide a reasonably accurate account of their wages.

I believe Behrman and Rosenzweig 2004 is among the two most convincing natural experiment studies for three reasons. First, the monozygotic twin design is extremely well controlled, as previously discussed. Second, Behrman and Rosenzweig is one of only two studies that included a sample of monozygotic twins *and* used objectively measured birth weight from hospital birth records, resulting in an accurately measured exposure variable. Third, as discussed below, the study used a measure of fetal growth restriction rather than birth weight itself, making it more directly relevant to our hypothesis.

Rather than using birth weight itself to estimate correlations, the authors calculate a measure of fetal growth rate, which is simply birth weight divided by gestational age at birth. Another way of saying this is "weight for gestational age". This is more indicative of fetal growth restriction than simple birth weight because a prematurely born child can have a low birth weight despite being on a normal growth trajectory. Therefore, this study tests the impact of fetal growth restriction specifically, rather than birth weight *per se*. However, it is worth noting that this will be effectively true of any within-twin comparison, since both twins are born at the same time and therefore gestation length is automatically factored out of the comparison.

Comparing within twin pairs, the authors found that a one pound (0.45 kg) difference of birth weight for gestational age was associated with 7 percent higher wages as an adult. This association was highly statistically significant, although as discussed previously it narrowly fails to survive adjustment for multiple hypothesis testing (using Bonferroni correction). However, I am not overly concerned about this problem.[38]

[36] From page 590 of Behrman and Rosenzweig 2004 (10): "…the birthweight information is based on measures from birth certificates, and thus the birthweights for the twins are not subject to recall error. It is well known that estimates based on within-twin or sibling-pair differences (Bishop, 1976; Griliches, 1979) are particularly prone to bias from measurement error, so that accurate measures of birthweight are critical. Moreover, because the birthweights of twins are assessed at the same time and by the same measurer, most of the measurement error will be common to the twins and thus will be eliminated using within-twin estimators."

[37] From page 589 of Behrman and Rosenzweig 2004 (10): "Our survey instrument was mailed out in May 1994 to the 5,862 members of same-sex twin pairs who had filled out the BQ and for whom the MTR had current addresses. An additional 776 members of same-sex pairs for whom updated addresses had been located between May and September 1994 were sent questionnaires in November 1994. Altogether 3,682 twins returned completed questionnaires, for a response rate of surviving twins of over 60%. Information obtained included the height and weight of the twins at the time of the survey, their schooling attainment, their work experience and wages on the last job, their parents' characteristics, and the birthweights for their first four children.7 The estimates in this paper are based on the returned questionnaires from (i) all women (N 5 1418), (ii) women in MZ twin pairs (N 5 804) , (iii) twin mothers (N 5 1207), and (iv) MZ-twin–mother pairs (N 5 608) for whom the key birthweight, gestation, and self-reported height, weight, and earnings variables are available."

[38] Here is why. First, of the four major outcomes reported in Table 2, three are statistically significant. This is not what would typically be observed of false positives in a null data set, where one would expect only one in twenty findings to cross a p-value threshold of 0.05. Second, from my perspective as the reader, there is only one "primary outcome" of this study: the association between birth weight for gestational age and adult wages. Either that finding is statistically significant, in which case I judge the paper to be supportive of the hypothesis; or it is not, in which case I judge the paper to be unsupportive. Although it would have been preferable if the authors had

A one pound (450 g) higher birth weight was also associated with one third year longer schooling and 0.6 inches of additional adult height, but no meaningful difference of adult body mass index (a measure of body fatness). Since the mean difference in birth weight within twin pairs was 10.5 ounces, the average twin pair would have a wage difference of 4.6 percent as a result of birth weight differences, which seems fairly significant. It is worthwhile to note that when the authors compared all individuals to one another rather than comparing solely within twin pairs, there was no significant association between birth weight for gestational age and adult wages, suggesting that genetics and/or family characteristics can substantially confound associations when comparing between unrelated individuals (which is the design used in the studies briefly discussed in the observational studies section of the report (6)).

To address the fact that twins tend to have a substantially lower birth weight than the average newborn, the authors then reweighted the distribution to match the general weight distribution of newborns and rederived the association. This attenuated the association between birth weight and wages by 25 percent and caused it to lose statistical significance, suggesting that differences in birth weight matter less when births are not on the lower end of the weight distribution.

Applying their twin-derived model to global birth weight statistics, the authors estimate that "the world inequality in birthweight could account for less than 1% of world earnings inequality". In other words, low birth weight in low-income countries is not a major explanation for economic differences between nations. However, according to their estimates the impact of birth weight can be substantial in specific countries. The authors report that closing the US-Malaysia and US-India birthweight gap over the bottom half of the birth weight distribution would increase earnings in Malaysia and India by 4.8 percent and 9.2 percent, respectively. It is worth noting that at the time the authors wrote their paper, India had the highest incidence of low birth weight of all countries for which national data were available.

Behrman and Rosenzweig 2004 has other notable limitations in addition to the multiple comparisons problem discussed previously. Due to the disproportionate weight I place on this study in forming my conclusions, I will review these in detail. The most obvious is that it did not include men, limiting its external validity. The paper provides the following rationale for this: First, that the effect size of the impact of birth weight on wages may be larger in women than in men, and second, that one of their hypotheses of interest is the impact of a woman's birth weight on the birth weight of her children, which is irrelevant to men.[39]

declared a primary outcome themselves and adjusted secondary outcomes appropriately for multiple hypothesis testing, from my perspective as the reader of this particular paper there is only one key statistical test being performed, so adjusting for multiple hypothesis testing is less important. If they had examined the association between birth weight for gestational age and wages in multiple ways, some results were statistically significant and others not, and it was unclear which result was the most appropriate for judging the hypothesis, this would be more of a problem.

[39] From page 587 of Behrman and Rosenzweig (10): "We focus on women in this paper because of the suggestion in most previous studies that the earnings of women are more sensitive to anthropometrics and the claim that lack of sufficient weight for women at birth has important effects across generations (Averett & Korenman, 1996; Conley & Bennett, 2000; Gortmaker et al., 1993; Haskins & Ransford, 1999; Sarlio-Lahteenkorva & Lahelma,

Josh Rosenberg and I had a conversation with Jere Behrman, one of the authors of the study, to ask for more details about the reason for not including men. He explained that the survey response rate was lower for men than for women, and they ended up with a male cohort that was not large enough to test the hypotheses of interest. He stated that the associations were similar in men and women, but with less precision in men. It would be useful to reanalyze the data with men included.

Insofar as the results of Behrman and Rosenzweig may be used to inform interventions in low-income communities, the setting of the study is a possible limitation. It was conducted in a cohort of twins born in Minnesota between 1936 and 1955, and the United States is an affluent setting by global standards (mitigated by the fact that the cohort includes several years of the Great Depression, which lasted until 1941). The quantity and quality of nutrition for this cohort may be higher, and the infectious disease burden may be lower, than is typical of low-income settings in many other parts of the world, not to mention numerous other differences between the US and other parts of the world that could potentially influence the relationship between birth weight and adult economic outcomes.

Birth weight differences within monozygotic twin pairs are thought to be caused by "random" differences of nutrient delivery that result from placental and umbilical cord placement and function in the uterus.[40] To consider the results of Behrman and Rosenzweig externally valid to low-income settings, we must rely on the assumption that low birth weight caused by this mechanism has the same adult impacts as low birth weight caused by the mechanisms that are most prevalent in low-income settings, such as undernutrition and infectious disease. This seems plausible in the case of undernutrition since it also causes relatively straightforward fetal nutrient restriction, but perhaps less so for infectious disease where the mechanism may not be as similar. The fact that many studies in diverse settings yield similar results (Table 1) provides some reassurance that the finding has good external validity.

We attempted to think of additional reasons for skepticism but did not identify any that were compelling. One possibility was mentioned by Black and colleagues, who suggested that the incomplete response rate to the survey administered in Behrman and Rosenzweig 2004 may have led to bias in the association between birth weight and adult outcomes.[41] Yet the key analyses of Behrman and Rosenzweig 2004 compared outcomes *within* twin pairs, both of whom had responded to the survey by

1999)… We also estimate the extent to which the intergenerational birthweight relation is due to the heritability of body mass and preferences."

[40] From Black and colleagues 2005, page 10 (18): "Given that gestation is the same among twins, evidence suggests that much of the difference is birthweight is due to differences in nutritional intake.7 In the case where there are two placentas (called dichorionic, including all fraternal twins and about 30% of identical twins), nutritional differences can arise because one twin is better positioned in the womb. Among single-placenta (monochorionic) twins, nutritional differences are related to the location of the attachment of the two umbilical cords to the placenta and the placement of the fetus within the placenta. (Bryan 1992, Phillips 1993). Hence, since there are no genetic differences, birth weight differences within monozygotic twin pairs appear to come primarily from differences in nutritional intake."

[41] From a passage of Black and colleagues 2005 referring to Behrman and Rosenzweig 2004, on page 7 (18): "…there is substantial attrition and item non-response that may not be random."
On page 589, Behrman and Rosenzweig 2004 state that the survey response rate was "over 60%", which in a non-twin study design would raise questions about selection bias (10).

design. It is difficult to understand how an incomplete response rate could bias associations in this situation.

A second possibility arises due to the fact that monozygotic twins are not perfectly identical. Occasional mutations randomly occur during cell division and these accumulate as development proceeds, leading to slight genetic differences that could impact both fetal growth and adult traits (70). Therefore, even monozygotic twin studies are not perfectly controlled for genetic differences. However, this seems unlikely to result in genetic differences between twins that are significant enough to bias outcomes.

Overall, my assessment is that Behrman and Rosenzweig 2004 offers quite convincing evidence that restricted early growth *in utero* impacts adult earning potential, particularly on the lower end of the birth weight distribution, and that the effect is large enough to be meaningful. The results are also consistent with the possibility that the mechanism responsible for this association includes reduced educational attainment and smaller body size, although the study did not directly test this.

The second twin study I will discuss is Black and colleagues 2005, which does not focus primarily on monozygotic twins but reports outcomes for a monozygotic subsample (18). The researchers obtained data on 33,346 twin births in Norway from the Medical Birth Registry of Norway, which includes objectively measured birth date, birth weight, and gestation length, among other data. This sample does not distinguish between monozygotic and dizygotic twins. They drew matched adult data from the Norwegian Registry Data database, including educational attainment, labor market status, earnings, and demographic variables like age and sex. They also drew matched data from military records, which represents every "able" Norwegian man and includes data on height and IQ, among other things.

The smaller monozygotic sample is a subset of the larger mixed-twin sample, and as such it is also linked to administrative data on birth weight and earnings. To distinguish monozygotic from dizygotic twins, the authors rely on a questionnaire that is intended to determine zygosity. At face value, this method appears to be quite reliable, although I have not investigated it.[42] I was unable to locate the total number of twins represented in this sample, however Table 5 suggests that it represents at least 1,606 monozygotic twin pairs (3,212 individuals).[43]

Comparing within mixed (monozygotic and dizygotic) twin pairs, the authors report that ten percent higher birth weight is statistically significantly associated with 0.9 percent higher adult earnings, and 1 percent adult full-time earnings (Table 4 (18)). To put this into terms comparable to Behrman and Rosenzweig 2004, by my calculation a one-pound (450 g) increase in birth weight would result in a 1.6 percent increase in earnings and a 1.7 percent increase in full-time earnings.[44] This effect size is 78 percent smaller than the estimate of Behrman and Rosenzweig 2004. Using the average within-twin

[42] From Black and colleagues 2005, page 13 (18): "Zygosity assignment is based on questionnaire items about co-twin similarity during childhood. These classification techniques are considered to have very high rate of correct classification (greater than 96%). See Harris, Magnus, and Tambs (2002) for more details."

[43] From Black and colleagues 2005, page 13 (18): "Our final dataset is a survey of twins born from 1967 through 1979 that contains information on zygosity and can be matched to the administrative data. The survey includes information on twin pairs that were intact at age 3 and was collected in two waves, one in 1992 and one in 1997. This is the only survey we use that is based on voluntarily self-reported information."

[44] Based on the mean twin birth weight of 2,607 g provided in table 2 (18).

difference of birth weights (320 g; 11.3 oz), this suggests that the average twin pair would have an earnings difference of 1.1 percent as a result of birth weight differences. This seems rather small.

The authors go on to estimate relationships separately for all twins, for same-sex twins (mixed but enriched for monozygotic), and for monozygotic twins (Table 5 (18)). These three analyses are progressively less likely to be confounded by genetic factors. The effect of birth weight on adult earnings becomes progressively smaller moving from mixed twins, to same-sex twins, to monozygotic twins, suggesting that the mixed-twin results previously discussed may be inflated by genetic confounding. The effect size in the monozygotic sample is approximately 0.9% higher earnings per additional pound of birth weight, and it is not statistically significant. For full-time earnings, the effect size is similar to the mixed sample but not statistically significant. It is worth noting that the sample size of monozygotic twins is smaller than the mixed sample, increasing variance and reducing the ability to detect significant effects. However, Table 5 does raise the possibility that the (already small) previously discussed earnings result is inflated due to genetic confounding.

When the authors divided the (mixed twin) results by gender, birth weight showed a statistically significant impact on total and full-time earnings in men but not women. The authors suggest that the lack of effect among women may be due to selection issues that result from lower employment rates (and therefore non-response), but it is unclear to me why this would bias within-twin pair comparisons.[45] The authors also report that in the mixed-twin military sample representing only men, higher birth weight is associated with greater height, body mass index, and IQ. Among men and women, higher birth weight is associated with a greater likelihood of high school completion (Table 4 (18)). These findings provide plausible mechanisms by which higher birth weight might increase adult earnings.

Black and colleagues 2005 has three key advantages over Behrman and Rosenzweig 2004. First, it includes both men and women, increasing its external validity. Second, both birth weight and earnings data come from administrative records, meaning that none of the key exposure or outcome variables are self-reported. Third, the sample size (3,212 individuals for the earnings measure) was approximately four times larger than Behrman and Rosenzweig 2004. For these reasons, I find Black and Colleagues 2005 to be more informative. However, the external validity of Black and colleagues 2005 to low-income situations is still debatable, as discussed for Behrman and Rosenzweig 2004.

From the perspective of multiple comparisons, Black and colleagues 2005 is not ideal. The paper tests many hypotheses and does not declare primary outcomes, although the text does focus predominantly on the mixed-twin sample, implying that it is the primary sample. This problem does limit my confidence in the results, but not greatly.[46]

[45] From pages 23 and 24 of Black and colleagues 2005 (18): “Of course, the earnings of women in our sample are particularly prone to selection problems due to nonparticipation and this may be partly responsible for this result.”

[46] Table 4 lists ten hypothesis tests using within-twin comparisons in the mixed-twin sample, nine of which are statistically significant using a p-value threshold of 0.05. Since the authors only state whether a p-value is below 0.10 or 0.05, it is not informative to apply Bonferroni correction, however it seems likely that at least some of the findings would lose significance. At the same time, one would not expect nine out of ten significant findings due to chance, so failure to correct for multiple comparisons seems less concerning than if only one or two outcomes

I consider both Behrman and Rosenzweig 2004 and Black and colleagues 2005 to be high-quality studies, yet they report quite different effect sizes in their monozygotic cohorts: 7 percent higher earnings per pound of body weight vs. 0.9 percent, respectively, and the latter finding is nonsignificant. I do not have a satisfying explanation for this discrepancy. It is possible that different genetic and/or environmental contexts powerfully modify the relationship between fetal growth and adult outcomes. It is also possible that one or both of these studies have serious limitations that I have not identified.

The third and final twin study I will discuss is Miller and colleagues 2005 (55). The authors collected data from male and female identical twin volunteers listed in the Australian Twin Register born between 1964 and 1971. The sample size, at 278 twin pairs (556 individuals), is smaller than the previous two studies. All exposure and outcome data used in the study come from a questionnaire administered to twins in the Register, including birth weight and adult outcomes.[47]

The authors report that each ounce of additional birth weight is associated with 0.4 percent higher adult earnings. Translating this into terms comparable to Behrman and Rosenzweig 2004 and Black and colleagues 2005, one pound (450 g) higher birth weight was associated with 6.4 percent higher earnings. The authors do not discuss the statistical significance of the finding, but they do provide the means to calculate it, and it narrowly fails to achieve statistical significance according to my predesignated significance threshold.[48]

Regarding multiple comparisons, the paper is somewhat problematic because it tests multiple hypotheses without obvious correction. I will not dwell on this because the relevant result is already nonsignificant.

In my opinion, there are three primary aspects of the study that limit its informativeness. The first is the small sample size, which is 31 percent smaller than the next smallest study among the seven considered in this section (Behrman and Rosenzweig 2004). The second is the fact that birth weight is self-reported, and the paper discloses evidence of significant inaccuracy in this measure. The third is the multiple comparisons problem. For these reasons, I do not find this study as informative as Behrman and Rosenzweig 2004 or Black and colleagues 2005, despite the use of a monozygotic twin design. In addition, it suffers from the same external validity limitation previously described.

were significant. In the monozygotic sample, our primary outcomes of interest are nonsignificant so the problem of multiple comparisons is irrelevant.

[47] The following statement from page 389 of Miller and colleagues alludes to the error that self-report of birth weight can introduce (55): “There is agreement among twins on the weight at birth (that is, one twin’s self-reported birth weight matches the co-twin’s report on the first twin’s birth weight) in approximately 70% of cases.”

[48] I calculated a p-value using the t-statistic of 1.79 and the sample size of 556 (278 twin pairs) in Table 2, assuming a two-tailed test (which is more stringent but leaves open the possibility that birth weight would be inversely associated with earnings) (55). The resulting p-value is 0.074, which narrowly fails to clear my threshold of 0.05. If we assume that birth weight could not be inversely associated with earnings and apply a one-tailed t-test, the resulting p-value would be 0.037 and the finding would be statistically significant. It is not totally clear to me which test is more appropriate, however I prefer the two-tailed test because although an inverse association seems unlikely, it is not impossible.

Despite narrowly failing to achieve statistical significance, the earnings result in Miller and colleagues 2005 actually appears quite consistent in magnitude with Behrman and Rosenzweig 2004. The lack of statistical significance may simply be due to the smaller sample size of the former paper, giving it insufficient statistical power to detect what may have been a real effect. It may also be attributable to the error inherent in self-reported birth weight, which would be expected to dilute any association with later-life outcomes.

The first Ramadan fasting study I will discuss in detail is Majid 2015, the Ramadan fasting study I find the most compelling because 1) it is based on a large, nationally representative sample of people with high levels of Muslim religiosity, and 2) it used within-household and within-sibling analyses to further control possible confounding (41). The author collected data on 13,251 Indonesian men and women from the Indonesian Family Life Survey, which the author suggests offers high quality data.[49]

Indonesia is the most populous Muslim country, approximately 88 percent of the sample population is Muslim, and mean religiosity of the sample is high, offering a compelling setting for testing the impacts of Ramadan fasting.[50] Furthermore, the sample offers the possibility of comparing exposed vs. non-exposed individuals within the same household, increasing the degree of control over potential confounding factors. The sample includes individuals born between 1942 and 1993, offering 51 birth years of exposures. Since the dates of the one-month Ramadan period shift by approximately 11 days each year, this sample represents Ramadan exposure at all times of year.

The author used self-reported birth dates to determine the degree of overlap between gestation and Ramadan. The accuracy of the self-reported birth date data is unclear to me because I don't know if dates of birth are systematically recorded or culturally valued in Indonesia, particularly in the older portion of the cohort. A quick Google search using the search terms "birthdays Indonesia" suggests that at least some people in modern Indonesia celebrate birthdays.

In the overall sample, Muslim subjects exposed to Ramadan *in utero* worked 4.5 percent fewer hours per week than non-exposed subjects, and the former were 3.2 percent more likely to be self-employed than the latter. These associations were statistically significant. The author suggests that self-employment rather than wage labor tends to indicate less desirable lower-skill, lower-income work in

[49] From Majid 2015 (41): "IFLS collected a great amount of information at the individual, household and community level on a large collection of economic, health and social indicators. Sampling took place at the household level. Great care was taken to assure representativeness of the sample for the reference population. IFLS covers 13 of the (then) 26 provinces of Indonesia, which, in total, represent 83 percent of the Indonesian population."

[50] From Majid 2015 (41): "Another feature of the data set, which is conducive to my study, is that around 88 percent of the sample population is Muslim, which gives me a large enough sample size to compare siblings and household members in Muslim families."
"Average religiosity among families is rather high in Indonesia with a mean of 2.796 (on a scale of one to four, where four is the highest value possible) and a standard deviation of 0.463."

Indonesia, and therefore that a higher self-employment rate should be interpreted as a negative impact of Ramadan exposure during gestation.[51]

Because of its importance for the interpretation of the paper's findings, I evaluated this claim further by following references cited in Majid 2015 and considering additional relevant outcomes reported in the paper. I have relegated this evaluation to a footnote in the interest of reading efficiency.[52] Overall, the

[51] From Majid 2015 (41): "Banerjee and Duo (2008) and Fields (2011) suggest that self-employment is a low skill and low income sector in many developing countries. Though there may be significant heterogeneity in this sector, I am interested in the average. In the IFLS, I find that self-employed workers have fewer years of completed schooling and have worse general health."
"Smith et al. (2002) document the economic impact of the East Asian financial crises in Indonesia on adults. They find very modest changes in total employment rates, but find that male employment declined by 3.7 percent in the wage sector, with a 1.74 percent increase in the self-employment sector as a result. My estimate of the effects of mother's fasting during pregnancy on children's self-employment probabilities (3.2 percent) is broadly similar, and if anything larger, compared to the labor market response during the financial crises."

[52] Banerjee and Duflo 2008 states, on the basis of a large global data set, that low-income people are more likely to be self-employed in rural agricultural settings. From pages 6 and 7 (71): "One difference is that in rural areas, the middle class seem less directly connected to agriculture than those with low incomes. Strikingly, the rural middle class are actually less likely to own land than the rural poor in all but three of our countries. Correspondingly, the middle class are also less likely to be self-employed in agriculture. For example, in Nicaragua, the fraction of households self-employed in agriculture goes from 56 percent among the extremely poor (daily per capita expenditures below $1 per day) to 36 percent for those with daily per capita expenditures between $2 and $4; in Panama, the comparable figures are 65 and 32 percent (and it drops further to 18 percent among those with daily per capita expenditure between $6 and $10)."

However, they also state on page 7 that this is not the case in urban settings: "In urban areas, the broad occupation patterns are remarkably similar between the poor and the
middle class."

Fields 2011 includes relevant statements about self-employment in low-income settings, however it does not provide supporting data or references for these (72). It was written by a prominent professor of economics at Cornell University so presumably it is based on evidence of some sort, however it is difficult to evaluate since the evidence is not provided or referenced in this manuscript. The relevant statements are as follows. Page S17: "Typically, the better jobs are in wage employment, not self-employment. But within wage employment, the regular wage jobs are better than casual wage jobs. For these reasons, "everybody" in developing countries wants a regular wage job."
"The problem the poor face is that not enough regular wage employment is available for all who would like wage jobs and are capable of performing them. Would-be wage employees could respond to the lack of wage jobs by remaining unemployed and continuing to search. However, few do, for the simple reason that they cannot afford to. They find it better to create their own self-employment opportunities."

Smith and Colleagues 2002 did find that a major financial crisis that struck Indonesia in the late 1990s caused a shift from wage labor to self-employment, consistent with the claim that self-employed occupations tends to be less desirable in Indonesia (73).

Majid 2015 includes two measures of cognitive performance in children 8 to 15 years: General intelligence as measured by Raven's matrices, and math skills as measured by a standardized set of math problems (41). For both measures, Ramadan exposure is significantly associated with poorer performance. Since people with poorer cognitive function may have fewer employment options and would be expected to sort into less desired occupations, this is consistent with the suggestion by Majid that self-employed occupations tend to be less evidence presented by Majid appears to support his suggestion that self-employed occupations tend to be less desirable than wage labor in Indonesia, and therefore that a higher self-employment rate should be interpreted as a negative impact of Ramadan exposure during gestation.

Associations between gestation during Ramadan and adverse employment outcomes were not observed among non-Muslims in the sample.

When comparing between Muslim individuals born in the same household (expected to reduce confounding), the magnitude of the association increased; Ramadan exposure was associated with 10 percent fewer hours worked and a 7.8 percent higher likelihood of being self-employed. These associations were statistically significant and driven primarily by third-trimester exposure, although none of the associations by individual trimester were statistically significant. These associations were almost completely absent in the sample of non-Muslims analyzed in the same way. In a smaller sample comparing only between siblings, similar statistically significant associations between Ramadan exposure and poorer adult labor status among Muslims were observed. Lastly, when the sample was divided into self-reported "highly religious Muslims" and "less religious Muslims", the association between Ramadan exposure and poorer labor market outcomes was approximately twice as strong among highly religious Muslims. Collectively, these findings are what one would expect if Ramadan exposure itself is driving the association.

The author also reports that Ramadan exposure is associated with lower birth weight by about 270 grams (0.59 lbs) and a sex ratio skewed toward women, both consistent with the hypothesis that Ramadan fasting is a significant nutritional stressor for fetuses. Ramadan exposure is also associated with lower scores on tests of cognitive ability, providing a potential mechanism for the labor outcomes observed. Effects on height are not reported.

From the perspective of multiple comparisons, this paper is also a bit problematic, both from the authors' perspective and from my perspective as a reader interested in economic outcomes. However, I think at least some of the economic outcomes can reasonably be considered statistically significant.[53]

desirable and potentially lower-skill and lower-wage in Indonesia. In my opinion, this is the most compelling piece of evidence supporting Majid's argument.

[53] If we focus on the within-household comparisons among Muslims (Table 3), which seem to be the most logical candidates for primary outcomes, there are two hypothesis tests of economic outcomes: one for hours worked per week, and one for the likelihood of being self-employed. Bonferroni correction would suggest that these outcomes need to meet a p-value threshold of 0.025 to be considered statistically significant. The result for hours worked is reported as having a p-value of less than 0.01, so it meets that criterion. The result for likelihood of being self-employed is reported as having a p-value of between 0.05 and 0.01, so it may or may not be statistically significant after this adjustment. We cannot know without further information because the exact p-value is not provided, nor is the means to calculate it.

If we take a more stringent view of multiple comparisons correction, table 3 actually contains a total of 24 hypothesis tests, and table 2 contains an additional 4 on economic outcomes, for a total of 26. Bonferroni correction in this case suggests a p-value threshold of 0.0019, and it is unclear if any of the current results would survive. To my knowledge, determining the number of relevant hypothesis tests for Bonferroni correction is somewhat subjective. I prefer the former adjustment because from my perspective as a reader interested in

My greatest source of skepticism in interpreting these findings is that it seems remarkable that an ostensibly mild form of nutrient restriction for one month out of a 9-month gestation period would have such pronounced impacts on adult outcomes. During Ramadan, Muslims are allowed to eat without limit before dawn and after sunset, so it seems surprising that there would be a major deficit of calories or specific nutrients. Superficially, Ramadan exposure appears to be a much milder nutritional stressor than serious famines such as the Chinese Great Famine and the Dutch Hunger Winter Famine, which lasted for months or years and killed many adults. There is some evidence that Ramadan fasters eat fewer calories and lose a modest amount of weight during Ramadan (see earlier discussion of this). There is also evidence that pregnant women are especially sensitive to the adverse effects of fasting. And lastly, evidence of lower birth weights and skewed sex ratios among those exposed *in utero* seems to provide fairly strong support for the hypothesis that Ramadan is a substantial stressor to the fetus. However, it is difficult to know the exact cause of the associations with adult economic outcomes. It is possible that fetal stress resulting from a variable nutrient supply (or something else associated with Ramadan), rather than growth restriction *per se*, is responsible for the effect.

Overall, I believe Majid 2015 offers fairly strong evidence that Ramadan exposure during fetal development negatively impacts adult economic outcomes. The most straightforward explanation is that Ramadan imposes nutrient restriction on the fetus, however other explanations are possible. Due in part to my inability to think of compelling alternative explanations, I believe it's likely that at least some of the effect is caused by nutrient restriction *per se*.

The second Ramadan fasting study I will discuss is Almond and Mazumder 2008 (40). Although the data sets used in this study are of variable quality, the fact that the authors observe similar findings in three independent samples that report adult outcomes makes this study fairly convincing. The first sample is birth records for 2.5 million Michigan residents from Michigan's Division for Vital Records and Health Statistics. Michigan contains a large population of Arabs, which have approximately 50,000 live births per year (page 8 (40)). To enrich their sample with Muslims, the authors used information from the previously mentioned database to identify the ancestry of the mother and retain only those who are of Arab descent. Since approximately one quarter of Arabs in Michigan are Christian, in some analyses the authors excluded counties that are known to contain a high proportion of Christian Arabs.[54]

economic outcomes using the most rigorous model available, there are really only two hypothesis tests that are directly relevant.

[54] From page 8 of Almond and Mazumder 2008: "Although, there is no information on religion, ancestry of the mother is reported (ancestry information is not recorded in the national vital statistics data produced by NCHS). This feature of Michigan's natality data allows us to construct a proxy for whether the mother is Muslim based on reported "Arab" ancestry (or reported ancestries from predominantly Muslim countries).9 Compared to other US states, Michigan has a relatively large Arab population.10 There are a total of about 50,000 births to mothers of Arab ancestry (about 2.2 percent of MI births) over this period. While there is a large population of Arabs around Detroit, the Arabs are reasonably dispersed throughout the State (see Appendix Figure A2, Panel A).

Since a large fraction of Arabs in Michigan are actually Chaldeans – a sect of Christianity – our proxy may misclassify many mothers and thereby attenuate our estimated effects.11 We use the 2000 US Census SF3 (1 in 6 sample) data to identify Michigan zipcodes with heavy concentrations of Chaldeans – who presumably do not observe the fast – relative to Arabs (see Appendix Figure A2, Panel B). In some specifications, we will drop observations from these zipcodes to compare ITT estimates."

This sample contains between 42,000 and 47,000 male and female births, depending on the measure (Table A2 (40)), and the birth data including birth date are from hospital records. It is unclear to me what proportion of this sample is actually Muslim, which reduces my confidence in the results of analyses based on this sample. The authors only report early-life outcomes for this sample because the offspring were not adults at the time of the study.

The second sample consists of male and female data from the 2002 Uganda census, which includes birth date, disability measures, home ownership, employment, and religious affiliation. The authors state that 11 percent of Ugandans are Muslim. The sample includes approximately 80,000 Muslim men and women.[55]

The third sample is the 1997 Iraq census, including approximately 250,000 men and women with birth month data. The data set also includes information on home ownership and polygyny (more than one wife), a proxy for wealth and status. The sample does not measure religious affiliation but the authors suggest that 97 percent of Iraqis are Muslim (page 9). Of note, the census year was six years after the end of the Gulf War and six years before the beginning of the Iraq War, so it did not occur during wartime.

The fourth sample is composed of male and female immigrants to the US from predominantly Muslim countries drawn from the 1980 US census. It is linked to records from the American Community Surveys that include information on earnings. The sample size for earnings ranges from approximately 6,500 to 22,000 people, depending on the analysis. The data only include the quarter of birth rather than the exact birth date, which "dulls the empirical comparisons that we can make". The data on birth year are also calculated indirectly and are imprecise.[56] In addition, not all of these immigrants are expected to be

[55] From page 9 of Almond and Mazumder (40): "Our sample of Muslim adults includes approximately 80,000 men and women between the ages of 20 and 80 in 2002. Muslims constitute about 11% of Uganda's population and have more schooling and lower rates of disability than non-Muslims (Appendix Table A3). Both Muslims and non-Muslims share a strong seasonality in the number of births. Muslims tend to live in the southeastern portion of the country.

Unlike other national censuses, the Uganda Census asks a battery of questions about specific disabilities, including: blindness or vision impairments, deafness or hearing impairments, being mute, disabilities affecting lower extremities, disabilities affecting upper extremities, mental/learning disabilities, and psychological disabilities (lasting six months or longer). As only about 5% of adults report a disability compared to over 10% in the US Census, disabilities recorded in the Uganda Census may be more severe. Further, Uganda reports information on the origin of disabilities: congenital, disease, accident, aging, war injury, other or multiple causes. In the absence of direct measures of economic status we use home ownership. We also consider several other socioeconomic outcomes such as literacy, schooling, and employment."

[56] From page 10 of Almond and Mazumder 2008 (40): "Our third Census sample is composed of immigrants to the US who were born in predominantly Muslim countries.14 We use a 6% sample from the 1980 Census along with a pooled sample of the American Community Surveys (ACS) for the years 2005 through 2007 (3% sample). The ACS is modelled on the long form of the decennial Census. We use these years because they provide the quarter of birth.15 Not observing birth month dulls the empirical comparisons that we can make. On the positive side, we obtain a large, national sample of US immigrants from Muslim countries in which to implement our ITT analysis. Data quality is high, and includes additional outcomes beyond disability, such as earnings."

Muslim, and data from Canada suggest that approximately one third may not be.[57] These limitations reduce my enthusiasm for this fourth data set.

Using the first (Michigan) data set, the authors report that birth weight is approximately 40 grams lower among infants who may have been exposed to Ramadan early in gestation. This represents 1.2 percent of mean birth weight in this sample. The authors apply several different models and it is unclear which result to focus on, however most of the models returned statistically significant findings similar to the one I described.

The authors also examine associations between Ramadan exposure in the first month of gestation and the offspring sex ratio, a marker of fetal stress. Ramadan exposure is associated with a 3.7 percent lower likelihood of male birth that is not quite statistically significant (p = 0.06). This becomes a highly significant 6.6 percent decline when counties with a high proportion of Christian Arabs are excluded from the analysis.

Using the second (Uganda) data set, they again find that Ramadan exposure in the first month of gestation is significantly related to a lower likelihood of being male. Restricting the analysis to men, they find that Ramadan exposure in the first month of gestation, but not other months, is associated with a statistically significant 2.6 percent lower likelihood of owning a home.[58] They also report that Ramadan exposure in the fourth month of gestation is associated with a statistically significant 1.9 percent lower likelihood of being employed at the time of the census. Associations with other months were not significant, and no trend was observed for exposure in the first month.

Using the third (Iraq) data set, Ramadan exposure in the first month of gestation is associated with a lower likelihood of having multiple wives or of owning a home, and a higher likelihood of being employed. The authors suggest that employment may actually be a sign of low socioeconomic status in Iraq, since homeowners were less likely to report employment than non-owners.[59] They report

"In addition to not knowing birth month, we do not know the exact birth year in the ACS since it is not asked and age is not reported as of a specific enumeration date as it is in the decennial Census. Given that Ramadan exposure shifts by only 11 days from year to year, using survey year - age provides a good approximation of birth year for the purposes of constructing Ramadan exposure measures at the quarterly level. The correlation between Ramadan exposure using survey year minus age, and survey year minus age minus 1, is about 0.93."

[57] From page 10 of Almond and Mazumder 2008 (40): "Specifically, we use countries with at least an 80 percent Muslim population. To asses the magnitude of misclassification attributable to this birth-county proxy, the same proxy variable was created in the 2001 Canada Census, which includes self-reported religion. 67% of Canadian immigrants from these 80% Muslim countries reported being Muslim."

[58] The authors restrict the analysis to men because they say that "men are the vast majority of property owners in Uganda" (page 24). They offer the following footnote on page 24 to support that statement, but neither the statement nor the footnote cite a reference (40): "Uganda is a patriarchal society where land is passed down through sons. Although women are not prevented from owning land, by one estimate, 93 percent of Ugandan land is owned by men."

[59] From page 27 of Almond and Mazumder 2008 (40): "We note that among males, home owners are less likely to be employed (73%) than non-home owners (82%) suggesting that employment may be a poor proxy for economic status in Iraq and may actually signal lower status."

significant associations with other months of gestation, particularly for home ownership, but the months are not consistent across outcomes. They find no association between Ramadan exposure and the sex ratio.

Using the fourth (US census) data set, Ramadan exposure in the first quarter of gestation is associated with lower adult earnings. They report a nonsignificant trend for Ramadan exposure in the first quarter of gestation to be associated with a sex ratio skewed toward women.

To summarize the economic outcomes, all three samples that reported economic outcomes showed some evidence of a negative impact of Ramadan on adult economic status. All three also reported that this effect occurred in the first month (or first three months) of gestation.

In all three samples that reported adult outcomes, the authors found strong associations between Ramadan exposure and adult disability, and particularly mental disability including psychiatric disorders. These associations were concentrated around the first month of gestation in the Uganda and Iraq samples and in the first quarter of gestation in the US census sample. This suggests the possibility that negative impacts on early brain development could be a mechanism whereby Ramadan exposure constrains adult economic outcomes, although this hypothesis was not directly tested.

It is important to note that the effects reported in this paper are likely diluted by the indirect measure of Ramadan exposure employed (gestational month) and several sources of measurement error, suggesting that the true effects of Ramadan exposure in these samples could be greater than observed.[60] At the same time, indirect and error-prone measures introduce a higher risk of bias. For this reason, Almond and Mazumder 2008 does not influence my beliefs about the overall hypothesis as much as Majid 2015.

An additional limitation of this study is the multiple comparisons problem. For most adult outcomes, the authors applied nine hypothesis tests corresponding to the nine months of gestation, and while individual months were often statistically significant, the overall p-value for all of gestation was often not significant. Applying Bonferroni correction to the findings suggests that most if not all of the economic outcomes are nonsignificant. However, the fact that similar findings were reported in three independent samples mitigates this problem substantially. Although these samples individually are not very compelling after Bonferroni correction, it is extremely unlikely that all three would have identified the same effect by chance. So I feel that the overall argument that Ramadan exposure is associated with less favorable adult economic outcomes is probably correct.

In sum, I think Almond and Mazumder 2008 provides fairly convincing evidence that Ramadan exposure in the first month of gestation adversely impacts adult economic outcomes, possibly by impacting early

"If we control for home ownership and multiple wives the instances of positive effects of Ramadan exposure on male employment are eliminated."

[60] Sources of measurement error: Month of gestation was inferred from birth date in some samples, which requires assuming a fixed gestation length for all fetuses and will misclassify some. In the US census sample, only birth quarter was observed, meaning that the data were an imprecise measure of Ramadan exposure. Some adult outcomes were self-reported, which introduces error.

brain development. The most obvious mechanism for this would be a restricted and/or variable nutrient supply *in utero*, although this is uncertain.

The third Ramadan fasting study I will discuss in detail is Schultz-Nielsen and colleagues 2014 (58). The authors used data from multiple Danish registers maintained by Statistics Denmark. The sample consists of 38,637 men who are first-generation immigrants from countries that are at least 90 percent Muslim, or whose parents are from such countries.[61] Ninety-two percent of the sample are first-generation immigrants and eight percent were born in Denmark (Table 1 (58)). Data on economic status are based on tax data reported by employers, likely making them more accurate than questionnaire data used in some other studies, but also missing non-employment income.

Birth date for first-generation immigrants in the sample is self-reported, meaning that birth dates in the overall sample are nearly all self-reported. The fact that this introduces error is clearly demonstrated by the observation that 1,704 people in the sample report a birth date of January 1 and 1,049 report a birth date of July 1, when only 106 births would be expected on each date.[62] This suggests that many people in the sample may not know their birth date, and some of them may have been assigned a date administratively. Although the authors excluded individuals with birth dates on January 1 and July 1 from their analysis, the broader concern about inaccuracy of self-reported birth date in this sample remains.

The sample does not include women. The authors' reasoning is that 1) Muslim women living in Denmark are less likely to be employed than Muslim men, and less likely to be stably employed; and 2) For cultural reasons, employment among Muslim women in Denmark may actually reflect low socioeconomic status, so the direction of employment effects may actually be reversed.[63] I do not find this type of rationale for excluding a gender very compelling. I would prefer if the authors analyzed the data in both genders, presented it, offered their interpretation of it, and allowed readers to decide if they agree.

[61] From page 14 of Schultz-Nielsen and colleagues 2014 (58): "We categorized an individual as Muslim if he/she or his/her parents migrated from a country with a population comprised of at least 90 percent Muslims.21 According to this criterion, 29.6 percent of Muslim immigrants in Denmark are of Turkish origin, followed by Iraqis (13.4 percent), Iranians (11.9 percent), Palestinians (11.1 percent), and Pakistanis (10.4 percent)."

[62] From page 14 of Schultz-Nielsen and colleagues 2014 (58): "In addition, we observed spikes in the proportion of individuals reporting birthdays as January 1st and July 1st. Note that misreporting is only possible among first generation Muslim immigrants because we have the exact birthday information for the descendants. To eliminate measurement error due to heaping, we excluded first generation Muslim immigrants with a birthday on January 1st (1,704 persons) and July 1st (1,049 persons) from our analysis."

[63] From page 5 of Schultz-Nielsen and colleagues 2014 (58): "Our analysis considers only men because among Muslims living in Denmark, women have a particularly low and unstable labor market attachment with frequent interruptions."

"The roles expected from or assigned to Muslim women in their communities may also be different from those of native Danish women. Accordingly, being employed need not necessarily be interpreted as a positive outcome, but rather be an indication of a poor financial situation. For example, on one hand Muslim women who have poorer health as a result of being exposed to Ramadan in utero may find it difficult to work. On the other hand, these women may also have poorer marriage prospects in terms of spousal income, an effect that may then exert pressure on them to seek jobs, including even low-paying ones."

The authors measure four different economic outcomes, each broken down by gestational month of Ramadan exposure: Likelihood of employment, annual salary, hourly wage rate, and annual hours worked. They report that Ramadan exposure in the seventh month of gestation is associated with a 2.6 lower likelihood of employment. No other associations by gestational month are statistically significant at the 5 percent level, although there are nearly significant associations between Ramadan exposure in the seventh month and lower salary income and fewer annual hours worked.

The multiple comparisons problem is devastating to this study. Given that Table 4 tests 36 hypotheses about gestational impacts on adult economic outcomes (nine months by four outcomes), there is an 84 percent chance of obtaining at least one statistically significant finding as a result of chance.[64] Since one significant finding is what we observe, my interpretation is that it is likely attributable to chance and the overall finding does not support the hypothesis that Ramadan exposure during gestation impacts adult economic outcomes. If other Ramadan fasting studies had also found clear associations in the seventh month of gestation, that would increase the probability that the finding is real, but that is not the case.

This problem, coupled with birth date self-report error and the fact that we don't know what percentage of the sample was Muslim, limits the informativeness of this study relative to other Ramadan fasting studies.

The only famine study I will discuss in detail is Mussa 2017 (48). In fact, this study examines the impact of year-to-year variations of staple food harvest rather than famine *per se*. The author obtained data on annual field corn (maize) harvest from the Malawi government Agriculture Production Estimates Survey, which includes metric tons of corn per hectare for each district and each year.[65] The author also obtained data on adult farmer year of birth, district of birth, and corn production.[66]

This resulted in a sample of 1,275 male and female rural corn farmers. The author then determined the association between relative regional corn yields near the time of birth and relative adult corn yields. The author used relative yield (yield relative to a district's typical yield) to avoid confounding that could be caused by persistent differences between regions, for example if one region was consistently more productive, wealthier, and had better early-life and adult outcomes than another.[67]

[64] Calculation: 1 - ((1-0.05)^36)

[65] From page 3 of Mussa 2017 (48): "The data is collected annually and in every district through the Agriculture Production Estimates Survey (APES) in which extension workers act as data collectors. The APES collects data on area cultivated, yield, and production of crops. It also collects data on livestock and fisheries. Of interest in this paper is the maize yield which is measured in metric tonnes per hectare. For each district and year, the maize yield is calculated as a total of local maize, hybrid maize, and composite maize."

[66] From page 3 of Mussa 2017 (48): "Using year of birth and district of birth, this data is then linked to adult life production and farmer characteristics data taken from the Third Integrated Household Survey (IHS3). The IHS3 is statistically designed to be representative at national, district, urban and rural levels. The survey was conducted by the National Statistical O¢ ce from March 2010 to March 2011. The survey collected information from a sample of 12271 households; 2233 (representing 18.2%) are urban households, and 10038 (representing 81.8%) are rural households. The survey collected socio-economic data at the household level and on individuals within the households. It also collected data on farming activities including crop output, land, labour and other inputs."

[67] From page 4 of Mussa 2017 (48): "Maize yield in a farmer's early life may depend on observed and unobserved local characteristics thus making them potentially endogenous. To ensure that the early life maize yields are not confounded by these omitted local characteristics, I follow Burke et al. (2014) and Flatø et al. (2016) and transform

I find this study design somewhat more compelling than others in the famine category for two reasons. First, it is not based on one famine that occurred in one specific place at one specific time, which creates difficulty in separating the effects of famine from confounding time-and location-specific factors that associate with famine exposure. The current study incorporates eleven years of data from 24 districts, permitting a study design that I believe is inherently less susceptible to confounding by time- and location-specific factors, similar to how the moving dates of Ramadan average out the confounding effect of seasonality. Second, although Maccini and Yang 2008 use a similar design to investigate the impact of early life rainfall on adult economic status in Indonesia, rainfall is a less direct measure of nutritional and economic status than the yield of a primary staple food (49).

Corn supplies more than half of total calorie intake in Malawi, and corn cultivation is a major source of rural income. According to the author, corn is “the best direct indicator of incomes especially rural incomes.”[68] For this reason, corn yield seems like a plausible marker of nutritional status in Malawi (calorie availability and diet diversity due to direct consumption and sale of corn), although this is somewhat speculative because the paper does not provide direct data on calorie intake or other aspects of nutritional status. The uncertain link between corn yield and nutritional status, and nutritional status and early life growth, is a disadvantage of this study vs. other famine studies, which often have more informative data on famine severity and its impacts on growth such as calorie availability, mortality rate, birth weight, and/or adult stature.

The paper reports that a ten percent lower relative corn yield while a farmer is *in utero* is associated with 4.2 percent lower relative yield when that farmer is an adult. No statistically significant associations are reported for the first and second postnatal years.

I believe there are reasons to be skeptical of this finding. First, if we reflect on what this result means, the effect size is almost implausibly large. The result suggests that a poor corn harvest during pregnancy will cause the offspring to produce harvests that are 42 percent as poor as the one that occurred during his/her gestation, for the person’s entire adult life. This implies that the effects of corn yield on fetal development in Malawi are extremely profound. This result would be more convincing if the author had

the actual maize yields into relative maize yields by using a cumulative gamma distribution. This transformation ensures that in each year, each district receives a value which reflects the probability of having a maize yield at that level or below in that particular district. This in turn means that the level of relative maize yield in a given year is orthogonal to local characteristics.”

[68] From page 2 of Mussa 2017 (48): “Similar to many African countries, maize is a primary staple crop in Malawi, and is the best direct indicator of incomes especially rural incomes (Burke et al., 2014). Maize ac- counts for more than two-thirds of caloric availability (Ecker and Qaim, 2011). Compared to neighbouring countries, food consumption is less diversified in Malawi. For instance, Malawi’s per capita maize consumption of 133.1 kg/per person per year is 2.5 times that of Mozambique, and 2.3 times that of Tanzania. Only Zimbabwe (110.4 kg/per person) and Zambia (110.2 kg/per person) are the closest to Malawi (Mussa, 2015). As a result of this low food diversification, national food security continues to be defined in terms of access to maize. It is not just food consumption which is skewed towards maize, crop production by smallholder agriculture is dominated by maize production. For instance, NSO (2012) found that 85% of households in Malawi cultivated maize (69% in urban areas, and 88% in rural areas). According to Smale (1995) given its importance "maize is life" in Malawi. As a result of this, maize availability takes a special place in political, social, and economic discourse.”

been able to provide evidence suggesting that corn yield during gestation also impacts predictors of adult economic success such as height and intelligence. In my opinion, the very large effect size raises the concern that confounding could be inflating the results, although I do not have a clear idea of what its source might be. I do find it reassuring that the author reported no associations with the first and second postnatal years, suggesting that the association with gestation is probably not driven by parents' higher income leading to later-life benefits to children as a result of their parents' greater wealth.

A second source of skepticism, already mentioned, is that corn yield is an indirect measure of nutritional status, which itself is an indirect measure of early growth. The connection between corn yield and early life growth is not entirely clear, and the paper does not provide evidence that might help clarify it.

The author does not attempt to adjust for multiple comparisons, however the primary result survives Bonferroni correction so this isn't concerning.[69]

Overall, I believe Mussa 2017 adds to the evidence that early-life growth impacts adult economic status, but its evidence value is substantially limited for the reasons listed above.

After carefully reviewing these seven studies, I quantified the degree to which each influences my beliefs about whether early life growth influences individual adult economic outcomes. This was not based on a formal quantitative process but is simply an attempt to clearly communicate my beliefs about the informativeness of each study, given its design. Out of a total of 100 percent for the natural experiment literature, I assign a weight of 35 percent to Black and colleagues 2005, 23 percent to Behrman and Rosenzweig 2004, 11 percent to Majid 2015, 9 percent to Almond and Mazumder 2008, 7 percent to Miller and colleagues 2005, 5 percent to Schultz-Nielsen and colleagues 2014, 5 percent to Mussa 2017, and 5 percent to the other 16 studies in Table 1.

Tentative conclusions

Despite its limitations, in my opinion the natural experiment literature tends to support the hypothesis that early life growth restriction impairs individual adult economic outcomes. This conclusion is based disproportionately on the seven studies I focused on in the previous section, which support the hypothesis overall, although it is notable that the highest-quality study identified only a very small and nonsignificant effect (in the monozygotic twin sample, which I view as the most informative). The hypothesis is independently supported by several types of natural experiment studies, including twin studies, famine studies, Ramadan fasting studies, and an adoption study. Not all studies are supportive, yet this is what one would expect from a large body of literature using imperfect methods to study a real effect.

[69] The association between in utero and adult corn yield is statistically significant with a p-value of less than 0.01. The paper tests three main hypotheses related to the relative corn yield during gestation, postnatal year one, and postnatal year two. Applying Bonferroni correction to a p-value threshold of 0.05, the result would have to have a p-value smaller than 0.016 to be considered significant.

The hypothesis is best supported for gestational growth restriction, and less well supported for postnatal growth restriction due to a smaller and more conflicting evidence base. I will consider effect size in the last section of the report.

My remaining uncertainty about this conclusion results from two primary sources. First, methodological problems I identified in the natural experiment literature, including the fact that none of the authors preregistered their research plans, none appear to have adjusted for multiple hypothesis tests, and some of the study designs are inherently more susceptible to bias. Second, the two (in my opinion) highest-quality studies report conflicting results. My conclusions could be altered by more high-quality monozygotic twin studies using objectively measured birth weights, preregistered study design, and adjustment for multiple hypothesis testing where applicable. It would also be informative to see more high-quality studies on the relative importance of different developmental windows as a way of understanding which periods are the strongest leverage points for intervention.

Randomized controlled trials

In addition to the search strategy described in the methods section, I applied the following inclusion criteria:

- Studies had to be randomized and controlled trials of a nutrition intervention.
- Studies had to report a measure of growth in the first thousand days of life.
- Intervention must have increased a measure of growth in the first thousand days relative to a control group.
- Studies had to report a measure of individual adult economic status.

I identified two randomized controlled trials that modified early life growth and measured its impact on individual adult economic outcomes (see Table 2). The first trial was conducted in Guatemala, and the second in Jamaica. I will discuss each in detail.

Reference	Year	Location	Intervention	Findings	Notes
Martorell (74) Schroeder et al. (75) Hoddinott et al. (76)	1992 1995 2008	Guatemala	Four villages received unlimited access to either a drink high in micronutrients, calories, and protein, or a drink high in micronutrients, low in calories, and containing no protein	Increased growth rate between ages zero and 3 years in the high-calorie group. No effect of supplementation on adult income in unadjusted comparison. In adjusted comparison, high-calorie supplement increased wage rate but not annual salary of men and impacted neither in women.	Large sample size. Too few units of randomization (see text). High attrition rate. Economic study used *post hoc* methods to address original trial design shortcomings.

Walker et al. (77) Gertler et al. (78,79)	1991 2013 2014	Jamaica	9- to 24-month-old stunted children received formula weekly providing 66% of calorie and protein needs for two years	Supplementation likely increased length, weight, and head circumference after one year (see text). No effect of supplementation on earnings at age 22, with no trend toward positive impact.	Trial was likely underpowered. Follow-up studies used *post hoc* methods to address original trial design shortcomings. Impact of supplementation on growth is somewhat uncertain.

Table 2. Randomized controlled trials testing the impact of modifying early life growth on individual adult economic outcomes.

Early life growth and individual adult economic outcomes

In the Guatemalan trial, researchers selected four rural villages for the study, two smaller in size and two larger. One larger village and one smaller village were randomly selected for an intervention in which all residents received unlimited access to a drink high in micronutrients, calories, and protein, while residents of the other two villages received unlimited access to a drink containing micronutrients, few calories, and no protein.[70] The study focused on 2,392 children aged zero to seven. The intervention began in 1969 and lasted eight years (76).

This study uses the method of "cluster randomization" in which groups of people, rather than individuals, are randomized. While the cluster randomization method is valid, similar to individual randomization it requires a certain number of units of randomization to be effective. If a study uses too few units of randomization (i.e., too few villages), the randomization process cannot be expected to even out all the baseline differences between intervention and control groups that can confound results. The current trial had four units of randomization, which is well below the number required for effective randomization.[71]

[70] From pages 411 and 412 of Hoddinott and colleagues 2008 (76): "Two villages, one from each pair matched on population size, were randomly assigned in March 1969 to receive a nutritious supplement called atole. Atole is a gruel-like drink made from Incaparina (a vegetable protein mixture), dry skimmed milk, and sugar that provided 6·4 g protein and 380 kJ (91 kcal) energy per 100 mL. In the other two villages, residents were given fresco, a drink that contains no protein, and 138 kJ (33 kcal) per 100 mL from sugar. From October, 1971, both supplements were fortified with micronutrients in equal concentrations by volume, sharpening the contrast in protein content. The supplements were available to all villagers twice daily throughout the study at a central location in each village."

[71] Tanner and colleagues discuss this problem on pages 117 and 118 of a systematic review conducted for the World Bank (17): "Despite these promising results, much of the potential impact of this program remains unknown because of challenges in the initial evaluation design that were not corrected for in subsequent studies. From a pair of large villages and a pair of small villages, the design randomly selected one village from each pair to receive the intervention, resulting in one large and one small village being in the treatment group with the other large and small village being assigned to the control.2 Although the sample size is large (more than 1,000 children), there are too few units of randomization (the four villages in this case) to rely on the Law of Large Numbers to generate treatment and comparison groups that are statistically equivalent. The randomization process allowed only four possible permutations of groupings. As a result, the validity of the counterfactual relies on the strength of the pre-randomization matching exercise, which used only size and geography to match the large and small villages. There again, though, the small number of matching variables is not sufficiently large for matching to be credible over all

Furthermore, a true randomized controlled trial compares outcomes between units that were randomized. For example, if we randomized four people to medical treatment or placebo such that two people are in each group, we would compare mean outcomes between individuals of the two groups. If we instead took blood samples on ten different days from each individual, and treated each blood sample as an independent measurement in our model (rather than averaging them together into a mean value for each individual), this would not be a true randomized trial and the result could be misleading. Yet this is essentially what the Guatemalan study did: It treated each individual as an independent measurement despite the fact that individuals were not randomized, as opposed to treating village averages as the independent measurements. This was presumably done because a sample size of four would obviously not provide enough statistical power to test the hypotheses of interest, and it is much more practical to implement a nutrition intervention in a whole village than to randomize individuals to different interventions within villages. However, this limitation of study design renders it more difficult to causally attribute outcomes to the supplementation intervention.

Based on diet surveys, children who were offered the high-calorie supplement consumed 94 more total Calories and 8.7 more total grams of protein per day than children offered the low-calorie supplement. This represents a 10 percent and 29 percent difference of calorie and protein intake, respectively (76). Children between ages zero and three grew more rapidly in length/height when offered the high-calorie supplement when compared with children offered the low-calorie supplement, while children older than this showed no significant increase in growth rate.[72]

relevant observed and unobserved characteristics. Early authors indicate systematic baseline differences between the treatment and control group over other important characteristics (for example, Pollitt and others 1993)."

Martorell 1992 also discusses this issue in a review of ten papers related to this trial (74): "Preceded by substantial preparatory work, the study began in 1968 with the selection of the villages and collection of baseline information. Many characteristics, including language and culture, population size and structure, and nutrition and disease patterns, were taken into account in selecting communities that would be as homogeneous as possible [2]. Nonetheless, in-depth analyses carried out later, some based on new information, revealed important differences among the villages. For example, one of the villages (Espritu Santo, the small fresco village) differed from the rest in terms of geography, ecology, and economy [3]; and subjects from the atole villages were less educated than those from the fresco villages [4]. The atole villages also had higher mortality rates than the fresco villages before the beginning of the longitudinal study [5]. Nonetheless, children as well as adults in the various villages were similar in their anthropometric characteristics before 1969 [6]."

[72] Table 1 of Schroeder and colleagues 1995 reports the growth rate of each group during specific intervals between three months and seven years of age (75). Between ages three and twelve months, the high-calorie group grew 15.6 cm in length, vs. 14.7 cm in the low-calorie group ($p < 0.001$). This represents a six percent increase in growth rate. Between ages twelve and 24 months, the high-calorie group grew 9.2 cm in length, vs. 8.2 cm in the low-calorie group ($p < 0.001$), representing a twelve percent increase in growth rate. Between ages 24 and 36 months, the high-calorie group grew 8.5 cm in length, vs. 8.1 cm in the low-calorie group ($p < 0.01$), representing a five percent increase in growth rate. Growth rate between ages 3 and 7 was approximately equal, with no statistically significant differences between groups.

Summing the growth figures, the high-calorie group gained 58.3 cm between ages three months and seven years, while the low-calorie group gained 55.7 cm, a difference of 2.6 cm and a 4.7% growth rate advantage for the high-calorie group.

The evidence supporting increased early growth of the high-calorie supplement group is more convincing than evidence on later-life outcomes because the former data are not adjusted in any way; they are a straightforward comparison between groups. Since the study involved too few units of randomization, we can't be certain that these growth differences are attributable to supplementation *per se*, rather than intrinsic differences in the early life environment provided by the four villages.

Studies also report increased birth weights in the villages offered the high-calorie supplement, as well as large declines in infant mortality rates and large declines in the prevalence of stunting at age three.[73] The overall picture is consistent with what one would expect from increasing calorie and protein intake in a setting where infant mortality and restricted early growth are due in part to insufficient calorie and protein intake.

Between 2002 and 2004, researchers attempted to contact participants in the original trial and administer surveys measuring adult economic outcomes (76). They successfully administered surveys to 60 percent of the original cohort of children, who at the time of survey were 25 to 42 years old. While an attrition rate of 40 percent is relatively low given the length of follow-up, high mortality rate, and low-income setting, it is nevertheless high enough to potentially introduce bias into outcome measures. This problem is further exacerbated by the exclusion of survey respondents, predominantly women, who were "not engaged in economic activities", leaving only 49 percent of the original male sample and 43 percent of the female sample.[74]

[73] From Martorell 1992, a review of ten papers on this intervention (74): "The nutrition intervention benefited the children in many ways. Supplementation during pregnancy improved birth weights: the risk of delivering a low-birth-weight baby was half as great for women who ingested more than 20,000 kcal from the supplements during pregnancy as for those who ingested less than that [11]. Infant mortality rates were markedly reduced: in 1969-1977 they declined by 66% in the atole villages from the 1949-1968 rates, compared with 24% in the fresco villages and 19% in non-intervened (i.e., control) villages [5]. Whereas the number of days children were ill with diarrhoea was not reduced by the intervention [12], atole did protect against the negative effects of diarrhoea on growth [13]. Atole also promoted speedy recovery from wasting [14] Although the effects of atole on growth in children were important [9], they were confined to the first three years of life [15]. Specifically, atole intake was not related to growth from three to seven years of age. Another improvement associated with the intervention was enhanced motor development [16]. Finally, atole had only a minor effect on mental development, certainly much less than anticipated [17]. The results of the impact of the nutrition intervention on physical growth in three-year-olds may be seen in figures 3 (see FIG. 3. Changes over time in percentages of three-year-olds with severe growth failure (length 3 SD or more below the reference median), by supplement type, sexes combined) and 4 (See FIG. 4. Changes over time in percentages of three-year-olds with severe growth failure, by supplement type and sex), which show, by supplement type and calendar time, the percentages of children who were severely stunted, defined as being shorter than three standard deviations (SD) below the NCHS mean.' When the study began in 1969, the prevalence of severe stunting was extremely high, around 45%, but was similar in the atole and fresco villages. Over the course of the study, the figure was reduced by half in the atole villages but stayed about the same in the fresco villages."

[74] From Hoddinott and colleagues 2008, page 413 (76): "We excluded from the analyses respondents (12 men and 238 women) who were not engaged in economic activities (ie, not participating in the labour market) since an hourly wage rate could not be calculated. We also excluded 41 men and 26 women who reported an extreme number of hours worked (ie, more than 12 h per day for all 365 days of the year) because of the concern that these implausibly high values would bias estimates of wage rates towards a lower value. The final samples analysed had 602 men and 505 women. For men, this sample represents 48·9% of the 1230 original male participants enrolled in the study (49·2% and 48·7% in atole and fresco villages, respectively). For women, the corresponding proportion is 43·2% (41·6% and 45·6 % in atole and fresco villages, respectively)."

The authors estimate the impact of supplementation exposure between ages zero and 24 months on adult economic outcomes, focusing on this age window because "this is the priority target age for nutrition programmes" (page 413 (76)). They also considered two other age windows: zero to 36 months and 36 to 72 months.

Straightforward comparison of mean income between adult subjects exposed to high- vs low-calorie supplements during ages zero to 24 months showed no statistically significant differences in either men or women (see Table 1 and Supplementary Table 3 (76)). While there was a trend toward higher income in male subjects exposed to the high-calorie supplement during this age, the same trend was observed in subjects residing in the same villages but not exposed during the zero to 24 month age window, suggesting that the trend may not be attributable to supplement exposure during that age window *per se*.[75]

The authors then applied methods including multiple regression to control for confounding variables that could potentially mask the effects of supplementation on adult outcomes.[76] The problem with this

[75] See Table 1 of Hoddinott and colleagues (76). Also, from page 414: "The mean annual earned income and wage rates of men exposed to either supplement during 0–24 months were $3334 and $1·45, respectively in atole villages, compared with $3117 and $1·27, respectively, for the fresco villages, but these differences were not significant. Annual hours worked for those exposed to supplementation from 0–24 months were lower in atole villages than in fresco villages, but these differences also were not significant. Those not exposed to any supplementation at 0–24 months of age had higher incomes and wage rates than those exposed to supplementation at 0–24 months, perhaps because a large proportion of them were older and thus had higher earnings from longer work experience. We show results for effects of exposure to atole in women in the appendix, since they were non-significant."

[76] From Hoddinott and colleagues 2008, pages 412, 413, and 414 (76):"We used a linear regression model to estimate the impact of exposure to atole on incomes; the analyses also controlled for potentially confounding variables."

"To improve the validity, as well as the precision, of our estimates,20 we controlled for individual, family, and community characteristics. The villages also underwent substantial socioeconomic changes during the 25 years that elapsed between the original intervention and the survey in 2002. For this reason, we also controlled for variables that capture changes and events in the community that might have affected income-generating activities and income-earning potential. The individual characteristics included sex and age, and family characteristics included the logarithm of the mother's age when the child was born, mother's and father's completed grades of schooling, the logarithm of mother's height, and an index of household wealth just before the intervention, all from the original 1969–77 study.

"Village fixed effects were represented by dummy variables for three of the four villages, capturing all fixed characteristics of these localities that might affect wages and incomes. Using census and village histories, we documented key demographic, social, and economic changes.21 From these community developmental histories, we constructed and then incorporated several "historical" variables into our analyses to control for community change, relating them to each individual's age. Proxy measures for schooling availability and quality included were a binary indicator of the availability of a permanent, cement-block structure for the primary school and student-teacher ratios in primary schools, in both cases when the individual was 7 years old. Several variables were created with age 18 years as the reference point, when most youths were entering the labour market. These included binary indicators for whether electricity was available, which could affect returns to labour in the village; the occurrence of events that might have reduced incomes such as earthquakes or other natural disasters sufficiently large to result in food aid being supplied to the village; and whether the village had increased demand for specific agricultural products (vegetables and yuquilla, a starchy tuber). Also included were the producer price for maize, approach is that it strays from randomized controlled trial territory and into observational study territory, weakening our ability to make causal inferences. Such methods can also be applied *post hoc* to dredge for significant effects when straightforward comparisons of means between groups are found to be nonsignificant. The paper does not indicate whether these methods were selected in advance or applied in response to a lack of statistically significant results, leaving open both possibilities.[77]

As discussed previously, another related problem is that the authors do not use the original randomized groups (i.e. villages) as the basis for comparison between the two forms of nutrition supplementation; their analysis uses individuals instead, which were not randomized in the original trial.[78] While it seems clear that these methods were necessary to extract statistically significant results from the data, it means that the analysis does not fully leverage the advantages of the randomized controlled trial design, again placing it somewhere between a randomized controlled trial and an observational study in its ability to identify cause-and-effect relationships.

In the fully adjusted model, exposure to high-calorie vs. low-calorie supplementation during the zero-to-24 month age window resulted in a significantly higher wage rate (by US $0.67 per hour) in men, but no significant increase in annual income or annual hours worked. There was a trend toward higher annual income (an additional US $870 per year) in men but this was not statistically significant (p = 0.12). The only other statistically significant differences were an increase in wage rate, and a decrease in annual hours worked, in the zero to 36 month age window in men only (see Table 2 (76)). No significant effects were identified in women (see Supplementary Table 5 (76)).

the main agricultural commodity in the region, and the level of wages in the construction sector, at age 18 years, to assess the demand for labour."

[77] In an ideal randomized controlled trial, researchers nudge an input variable in one group and directly compare outcome measures with a similar control group that has not been nudged. In an observational study, researchers simply measure input and output variables in people who are dissimilar in many ways, then typically attempt to isolate variables of interest by mathematically adjusting for other ways in which subjects differ. When groups being compared are not intrinsically comparable aside from the variable of interest—as they should be in a randomized controlled trial that is well designed and powered—this requires additional math and assumptions, increases the probability of confounding, and ultimately weakens causal inference. This is why randomized controlled trials are generally considered to provide more reliable evidence than observational studies.

[78] From Hoddinott and colleagues 2008, page 414 (76): "The small number of villages randomised did not provide enough statistical power to estimate the effect of exposure to atole at the village level. Thus our models use individuals and not villages as the unit of analysis, since the duration and timing of exposure to the intervention for particular children depended on village of residence and year of birth."

This problem is discussed in Martorell 1992 (74): "Technically speaking, the longitudinal study design calls for the village to be the unit of analysis. A drawback of this approach is that it can be followed only for certain outcomes for which baseline data exist (e.g., infant mortality, physical growth). The low power of the village-level design, with only two villages per treatment, makes it unfeasible as the approach to be used routinely. Instead most analyses have used the "subject" as the unit of analysis but adjusted for relevant differences between and within villages, such as in socio-economic status and education. However, such subject-level analysis cannot be said to address causality in the sense that village-level analysis does. Rather, it can be viewed as enhancing plausibility, or the internal consistency of results and hence their persuasiveness [2]. For example, subject-level analyses allow one to explore whether a dose-response relationship exists between the supplement and outcomes of interest. The persuasiveness of the findings is enhanced to the degree that this and other expectations are met."

As with other papers discussed in this document, I was not able to find evidence that the authors took steps to address the multiple comparisons problem. This is a major concern because Table 2 lists nine economic outcomes. When I adjust for multiple comparisons by applying Bonferroni correction, all results become nonsignificant, although only narrowly so for the wage rate increases in men exposed during the zero to 36 month age window.[79] As a reminder, Bonferroni correction is conservative, so it would not be unreasonable to view the former result as statistically significant.

My overall assessment is that this study provides weak evidence of an impact of early life growth on individual adult economic outcomes. Here is a summary of the study limitations that lead me to discount the study's evidence value:

1. The original trial design was flawed due to the small number of units of randomization.
2. The results reflect fewer than half of the original study participants, potentially introducing bias.
3. Straightforward (non-adjusted) comparison between mean income of adults who received high-calorie vs. low-calorie supplements between ages zero and 24 revealed no significant differences and no obvious trends.
4. The methods applied by Hoddinott and colleagues fall somewhere between those of a randomized controlled trial and an observational study, weakening our ability to draw conclusions about cause and effect.
5. In the adjusted model, statistically significant results were observed in men but not women.
6. The authors identified statistically significant results for wage rate but not annual income.
7. The results do not appear to survive adjustment for multiple comparisons, although wage rate in men comes close.

I do not think this study produced statistically significant results by the best-practice standards of biomedical research. Yet we only have two randomized controlled trials to draw from, and as I will explain shortly, both have substantial limitations. The question is: Despite not meeting best-practice standards, should this study impact my beliefs about the relationship between early life growth and economic outcomes? I believe the calorie-protein supplement probably increased early life growth in our age window of interest. Despite substantial limitations, I think the later-life outcomes add slightly to the case that increasing early life growth in a low-income setting can improve economic outcomes in adulthood, at least in men.

In the Jamaican trial, researchers recruited 129 stunted children ages 9-24 months from "poor disadvantaged neighborhoods" of the Kingston area of Jamaica, stratified them by age and sex, and randomized the strata to four groups: control (no intervention), nutrition supplementation, psychosocial stimulation, or nutrition supplementation plus psychosocial stimulation (77,78). The nutrition supplementation group received 1 kg per week of infant formula intended to supply two-thirds of total calorie, protein, and micronutrient needs for the infant. Each family member received 1 kg of skim milk powder and 1 kg of cornmeal weekly to reduce the temptation to consume the infant formula. This continued for two years. Families reported via questionnaires that total calorie intake increased by 106

[79] Since there are nine statistical tests in Table 2, Bonferroni correction suggests that any single test would need to cross a p-value threshold of 0.0056 (0.05 / 9) to achieve statistical significance. The smallest reported p-value in Table 2 is 0.007.

Calories per day for infants in the stunted group.[80] Thirty-two nonstunted children were also recruited for comparison but were not part of the randomized trial.

The researchers recorded length, weight, head circumference, arm circumference, and subscapular and triceps skinfold thickness (a measure of the amount of fat under the skin) at 6 and 12 months after starting the intervention. This is not explicitly stated, but there appears to have been no significant differences in any of these measures at the 6 or 12 month timepoints when comparing between the original four randomized groups without further mathematical adjustment (see Table 3 (77)).[81]

The authors applied multiple regression to attempt to isolate the effect of supplementation and stimulation on growth measures from the "noise" generated by different starting ages, genders, initial nutrient intakes, and unintended between-group differences caused by the randomization process.[82] As

[80] From Walker and colleagues 1991, page 643 (77): "One kilogram of a milk-based formula containing 2195 kJ (525 kcal) and 14 g protein/100 g was given each week. This provided 3135 kJ (750 kcal) and 20 g protein/d, which is approximately two-thirds of the energy requirement and all of the protein requirement (28). Because of logistical problems the brand of supplement used was changed twice during the study but the energy and protein contents remained constant. In addition, 1 kg each of skimmed-milk powder and cornmeal were provided for other household members to reduce sharing of the target child's milk supplement. All the supplements were delivered to the home each week by a community health worker who reinforced the need for the child to be given the milk supplement. The home diets of the children were measured by two 24-h recalls before the interventions (24) and were repeated 6 mo later. On enrollment, the mean (±SD) daily intakes of the stunted and nonstunted children were similar [stunted: 3984 ± 1873 Id (953 ± 448 kcal), 26.6 ± 17.8 g protein; nonstunted: 4067 ± 1496 kJ (973 ± 358 kcal), 29.0 ± 14.1 g protein). This unexpected finding was discussed previously (24). The mean intake of supplement per day at 6 mo was 1442 ± 1037 kJ (345 ± 248 kcal). However, the energy intake from the home diet was significantly less in the supplemented than in the nonsupplemented children (P < 0.001). The net increase in daily energy intake in the supplemented children was 443 kJ (106 kcal) after adjusting for initial intakes and child's age and weight."

[81] Walker and colleagues 1991 does state on page 645 that there was no significant difference between the control vs. stimulation groups, or the supplementation vs. supplementation and stimulation groups (77): "There were no significant differences between the control group and the stimulated group or between the supplemented children and those who received both interventions."

I am fairly confident that there were no significant differences between any groups as originally randomized despite the fact that this was not stated explicitly, because (1) typically, significant results in a table would be highlighted and discussed in the text, yet Table 4 shows no evidence of statistically significant comparisons and none are reported in the text; (2) such results would be the most straightforward and convincing evidence of efficacy and therefore if they were statistically significant we would expect the authors to emphasize them; and (3) the authors go on to lump groups together to increase statistical power, a post hoc method that would not be expected if initial results were statistically significant.

[82] From page 645 of Walker and colleagues (77): "To avoid bias in analyzing change (3 1), the interventions were evaluated by using a multiple-regression model. The basic approach was to estimate final status, including as independent variables initial status and age and supplementation and stimulation status. For evaluating effects from 6 to 12 mo, status at 6 mo was included."

"In each of the regression analyses a set of covariates was offered: sex, birth weight (> or < 2.5 kg), and a quadratic term for age to account for nonlinear relationships with age, maternal stature, guardian's age, employment and occupation, quality of housing, and initial energy and protein intakes from the home diet. The interaction terms offered were supplementation X stimulation, age X supplementation, initial length-for-age X supplementation, and initial weight-for-length X supplementation. Intervention effects were considered over three periods for each dependent variable (0-6 mo, 6-12 mo, and 0-12 mo). If a covariate entered only one or two of the regressions it was forced into the remaining ones."

explained previously, these methods increase the likelihood of extracting statistically significant findings from trial data and may be useful in some cases, but they straddle the line between randomized controlled trial and observational methods, reducing our ability to make causal inferences. They are sometimes used to dredge for significant effects when straightforward comparisons of means between groups are nonsignificant.

After multivariate adjustment, the authors report that supplementation was associated with significant increases in length, weight, head circumference, arm circumference, and triceps skinfold thickness in the first six-month period but not the second six-month period (see Table 5 (77)). Supplemented children remained shorter in length than the comparison group of children who were not stunted at baseline, suggesting that supplementation did not fully correct their growth deficit. Psychosocial stimulation had no effect on growth measures even after multivariate adjustment, which is somewhat reassuring.

These growth results provide the foundation for interpreting the results of long-term follow-up of this randomized controlled trial. If we believe that supplementation increased early life growth, then this randomized controlled trial is a valid test of the hypothesis that early life growth impacts later-life outcomes. I do not find the growth results totally convincing due to the fact that the most straightforward interpretation of the data is null and multivariate methods may have been applied after this observation.

However, we have very little randomized controlled trial evidence to choose from, so even though I believe this result does not qualify as statistically significant by the best-practice standards of biomedical science, it's worth considering my degree of belief that the intervention had the claimed effect despite not meeting this standard. In this case, I think it probably did. My reasoning is that the result has the hallmarks of a real effect that was studied using an underpowered design. Accelerated growth is what one would expect from nutritional supplementation of a cohort that is stunted in part due to undernutrition, so we should expect a positive result. Sample size was relatively small (32 children per group), I was unable to locate a power calculation by the authors to select appropriate sample size, and the wide (9 – 24 month) age window of enrollment would likely have created substantial variability between individuals (77). Consistent with this argument, a follow-up study describes the initial sample size as "small" and views this as sufficiently limiting that the authors lump treatment groups together to increase statistical power in the main analysis (page 9 (78)).

When examining the multivariate-adjusted results in table 5, nearly all of the growth outcomes were statistically significant in the expected direction in the first six months in the supplementation group, while no measures of growth were impacted at any time point in the psychosocial stimulation group. This is what one would expect from a higher food availability and higher apparent calorie intake, while there is no clear reason to expect that psychosocial stimulation would increase physical growth. For these reasons, putting conventional statistical significance tests aside for a moment, my best interpretation is that the supplementation intervention was probably effective at increasing growth but the study design was simply underpowered to detect this effect in the most convincing manner. If it is useful for me to quantify my degree of confidence, based on the limited information available to me, I believe the intervention increased growth with approximately 75 percent confidence. We should expect any follow-up study on the same cohort to be underpowered to an equal or greater degree.

Gertler and colleagues published a 20-year follow-up study in two similar manuscripts in 2013 and 2014 that measured the impact of the interventions on earnings at approximately age 22 (78,79). Only the 2014 publication analyzed the supplemented arm separately (79). They were able to locate and study 105 of the original group of 129 subjects, further reducing the original statistical power of the study. To address the problem of insufficient statistical power, they applied several techniques including lumping groups together and adjusting for potential confounding variables using multivariate regression.[83] As explained previously, this puts the study somewhere between a randomized controlled trial and an observational study, with a corresponding reduction in our ability to interpret findings as causal. Applying *post hoc* methods to a compromised trial design does not fully repair its ability to produce informative results.

The authors found no impact of nutrition supplementation on average earnings, earnings at a person's first job, or earnings at the current job (supplementary tables 10, 11, and 12 (79)). The p-values are large (0.37 – 0.96) and nearly all of the associations are actually inverse, meaning that supplementation is nonsignificantly associated with *lower* adult earnings. Although it is possible that nutrition supplementation actually increased adult earnings and the reported result arose due to chance in an underpowered design, this seems unlikely. The more likely explanation is that nutrition supplementation did not increase adult earnings in the context of the current study.

It is worthwhile to note that the same study identified a large positive impact of the psychosocial stimulation intervention on adult earnings, demonstrating that the study design was capable of detecting significant effects if they were sufficiently large.[84]

Overall, the Jamaican randomized trial does not provide evidence that increasing early life growth among stunted infants improves individual adult economic outcomes. This is despite the fact that the supplementation probably did increase early life growth, although not sufficiently to catch up to a nonstunted comparison group.

[83] From Gertler and colleagues 2013, pages 6, 7, and 11 (78): "While the stimulation arms had strong and lasting effects, the nutrition-only arm had no long-term effect on any outcome (Walker et al., 2005, 2000). Hence, we combine the two psychosocial stimulation arms into a single treatment group (N=64) and combine the nutritional supplementation only group with the pure control group into a single control group (N=65). Henceforward we use the term stimulation effects of stunted participants to designate the analysis that compares groups 1 and 3 against groups 2 and 4."
"While randomization guarantees that any baseline variable Z is independent of the vector of treatment status D conditional on variables X used in the randomization protocol, the realization of baseline variables can turn out to be imbalanced across treatment groups. In the case of the Jamaica Study, three potentially important characteristics were not balanced at baseline. In order to control for potential bias, we estimate treatment effects by linear regression controlling for these variables when relevant for explaining outcomes."

[84] From the abstract of Gertler and colleagues 2014 (79): "The authors reinterviewed 105 out of 129 study participants 20 years later and found that the intervention increased earnings by 25%, enough for them to catch up to the earnings of a nonstunted comparison group identified at baseline (65 out of 84 participants)."
Although this is a large effect size, due to the low statistical power of the study, the confidence interval is likely to be wide. This means that the true effect could have been smaller than the study's estimate.

Does the Jamaican study suggest that early life growth interventions are not an effective means of increasing adult economic outcomes? Due to the substantial limitations of this study, I think it only provides weak evidence to support this conclusion. If postnatal early life growth has an impact on adult economic outcomes that is similar in magnitude to the effect of birth weight on adult economic outcomes (estimated at 7 percent increase of wages for each additional pound of birth weight), one would need a larger study to detect the effect, particularly if, as in the current study, the nutrition intervention was only partially effective and there was substantial variability between subjects.

Overall, randomized controlled trials unfortunately do not provide strong evidence on the impact of early-life growth on individual adult economic outcomes. Only two trials are directly relevant, both have substantial limitations, and their results are somewhat conflicting. They only provide evidence on postnatal early life growth, not on growth during gestation. My assessment is that the randomized controlled trial literature as a whole does not provide much evidence value that is directly relevant to our hypothesis of interest.

Indirect outcomes

Since the randomized controlled trial literature includes so few studies on the direct link between early life growth and individual adult economic outcomes, I expanded my search to include three outcomes with indirect relevance to adult economic status: Adult stature, years of schooling completed, and general intelligence. For time efficiency, rather than conducting a systematic literature search I relied on a recent systematic literature review by the World Bank to identify studies (17). I include an additional relevant study that I encountered incidentally in my previous literature search, which was not published in time to be included in the World Bank review. I also include the Guatemalan study discussed previously, which for certain outcomes did not meet the inclusion criteria of the World Bank review. I only include interventions that demonstrated increases in early life growth.

Below, I include a table summarizing the findings.

Reference	**Year**	**Location**	**Intervention**	**Findings**	**Notes**
Walker (80) Walker (81)	1996 2000	Jamaica	9- to 24-month-old stunted children received formula weekly providing 66% of calorie and protein needs for two years	Supplementation likely increased length, weight, and head circumference after 1 yr (see text). No effect on height at 7.7 or 11.5 years. No effect on schooling or general intelligence at age 22.	Trial was likely underpowered. Impact of supplementation on growth is somewhat uncertain.
Hawkesworth (82) Alderman (83)	2008 2014	The Gambia	Pregnant or post-partum women were given biscuits high in protein, calories, and micronutrients	Supplementation during pregnancy increased birth weight by 136 g. No height, schooling, or general intelligence difference at 19.6 years.	
Devakumar (84)	2014	Nepal	Pregnant women received a multiple micronutrient	Birth weight 77 g higher in the multiple micronutrient group.	

			supplement or iron and folic acid only	No difference in height at 8.5 years.	
Rivera (85)	1995	Guatemala	Four villages received unlimited access to either a drink high in micronutrients, calories, and protein, or a drink high in micronutrients, low in calories, and containing no protein	Increased growth rate between 0 - 3 years in the high-calorie group. At age 16.4, girls who received high-calorie supplement were 2.1 cm taller. No significant effect in boys. At age 32.3, girls who received high-calorie supplement had completed 1.2 more years of schooling. At 32.3, men and women in the high-calorie group had greater general intelligence.	Large sample size. Too few units of randomization (see text). High attrition rate.
Santos (86)	2015	Brazil	Mothers of children age zero to 17.9 mo months received counseling on feeding practices, vs. no counseling	Children 12 - 17.9 mo in intervention group were 0.34 kg heavier than controls. No difference in younger age groups. At age 15, boys in the intervention group were 3.4 cm taller. No significant effect in girls. At ages 15-16, general intelligence was lower in the intervention group, although this disappeared after adjusting for confounding.	Cluster-randomized

Table 3. Randomized controlled trials testing the impact of modifying early life growth on outcomes indirectly related to individual adult economic status.

Adult stature

The World Bank Review includes the Jamaican study. As discussed previously, the calorie/protein supplementation intervention probably increased early life growth, but I do not have full confidence in this due to the study's limitations. A follow-up study when the subjects were a mean age of 7.7 years showed no statistically significant effect on stature, and another follow-up at age 11.5 also reported no statistically significant effect and virtually no trend toward an effect (80,81). Although the World Bank review does not include data on stature in adulthood, it seems unlikely that a height difference would re-emerge in adulthood.

In the Gambia, researchers provided high calorie/protein/micronutrient biscuits to women who were either pregnant or had already given birth. Among 2,047 live births, supplementation during pregnancy increased birth weight by 136 grams overall, and 201 grams during the "hungry season" of lower food availability (82). At a mean offspring age of 19.6 years, there was no difference in height between groups, and no indication of a nonsignificant trend toward larger stature in the *in utero* supplemented group (83).

In Nepal, researchers randomized 1,200 pregnant women to receive either a multiple micronutrient supplement or a control supplement containing only iron and folic acid. Birth weight was 77 grams higher in the intervention group than in the control group (84). At a mean offspring follow-up age of 8.5 years, there was no difference in stature between groups (84). Adult stature has not been reported but it seems unlikely that a difference will emerge later in life.

The World Bank review does not include the Guatemalan study for this outcome, but I will review it here. As discussed previously, the high-calorie/protein/micronutrient supplement probably increased growth between ages zero and three years. At a mean follow-up age of 16.4, in unadjusted comparisons there was no statistically significant height difference between groups for boys, while high-calorie-supplemented girls were significantly taller by 1.7 cm. There was a nonsignificant trend for high-calorie-supplemented boys to be shorter by 1.3 cm. In comparisons adjusted for potential confounding factors, high-calorie-supplemented boys and girls were taller by 1.2 and 2.1 cm, respectively, but this difference was only statistically significant for girls (85). It is likely that a portion of the sample had not yet achieved adult height.

I also include a study published after the World Bank review, which was conducted in Brazil (86). This was a cluster-randomized trial in which 28 communities were randomized to an intervention or a control group. In total, the study included 424 children. In the intervention group, mothers of children age zero to 17.9 months were counseled on feeding practices. Recommendations were as follows:

> Increase the frequency of breastfeeds and complementary feeds; provide animal protein and micronutrient-rich foods (e.g., egg, chicken liver, and shredded chicken and beef); increase the energy by adding 1 teaspoon (volume capacity of ~5 mL) of oil, butter, or margarine to the child's plate; and increase nutrient and energy density by providing mashed beans instead of broth and by giving thick mashes instead of soup.

Mothers in the control group did not receive this advice. 180 days later, children between 12 months and 17.9 months of age in the intervention group were 0.34 kg heavier than the control group, while length did not differ between groups, and neither differed in younger age groups.

A follow-up study focused on the children in the age category that had experienced increased early-life growth as a result of the intervention, including 363 subjects aged 15 years (86). Boys in the intervention group were 3.4 cm taller than those in the control group. Although girls in the intervention group were 1.4 cm taller, this difference did not approach statistical significance.

The randomized controlled trial literature reviewed here provides modest and inconsistent support to the hypothesis that nutrition interventions that increase early life growth result in larger adult stature.

Only the Guatemalan and Brazilian studies support the hypothesis, and they are only partially supportive.

Years of schooling

For this outcome, the World Bank review includes the Jamaican study, the Guatemalan study, and the Gambian study discussed above. Having already reviewed the nature of the interventions, I will not do so again here, focusing instead on the results of follow-up studies.

In the Jamaican trial, one study provides follow-up data on years of school completed at 22 years of age (87). It reports no significant effect of nutrition supplementation on years of schooling completed, and no nonsignificant trends in the direction predicted by the hypothesis. In contrast, the group that received psychosocial stimulation completed significantly more years of schooling than controls, suggesting that the study design is capable of detecting effects that are sufficiently large.

In the Guatemalan trial, one study provides follow-up data on years of schooling completed at a mean age of 32.3 years (88). The authors report that exposure to the high-calorie supplement between ages zero and three significantly increased total years of schooling by 1.2 grades among women. Among men, high-calorie supplement exposure nonsignificantly decreased total years of schooling by 0.4 years.

In the Gambian trial, one study provides follow-up data on years of schooling completed at a mean age of 19.6 years (83). The authors report that maternal nutrition supplementation in pregnancy led to a nonsignificant decline of 0.49 total years of schooling in offspring relative to the control group in the sample of both men and women.

Altogether, the randomized controlled trial literature does not support the hypothesis that increasing early life growth increases years of schooling, although one study reported a positive effect.

General intelligence

For this outcome, the World Bank review includes the Jamaican, Guatemalan, and Gambian studies. In addition, I will consider the Brazilian trial that was published after the World Bank review, as mentioned previously.

In the Jamaican trial, one study reported follow-up data on general intelligence at a mean age of 22 years (87). Walker and colleagues report that supplementation did not impact general intelligence as measured by the Wechsler Adult Intelligence Scale.

In the Guatemalan trial, one study reported follow-up data on general intelligence at a mean age of 32.3 years (88). Supplementation increased general intelligence in men and women together, as measured by Raven's matrices. It also increased reading comprehension.

In the Gambian trial, one study reported follow-up data on general intelligence at a mean age of 19.6 years (83). The study reports no difference in general intelligence between groups, as measured by

Raven's matrices. It also reports no differences in verbal intelligence or working memory between groups.

In the Brazilian trial, one study reported follow-up data on general intelligence between the ages of 15 and 16 (89). Munhoz and colleagues report significantly lower general intelligence in the intervention group as measured by the Wechsler Adult Intelligence Scale, although this became nonsignificant after adjusting for potential confounding variables.

Overall, the randomized controlled trial literature does not support the hypothesis that increasing early life growth increases adult general intelligence, although individual studies suggest both positive and negative effects.

Indirect outcomes conclusions

Including indirect outcomes did expand the number of studies available for analysis from two to five, which was the primary goal. Yet this larger number of studies does not offer more support to the hypothesis that early life growth impacts adult economic outcomes. These trials provide some evidence that increasing early life growth increases adult stature, but this evidence is not very strong due to the fact that an effect has not been consistently observed across studies.   Taken as a whole, randomized controlled trials do not offer evidence that increasing early life growth increases years of schooling or measures of general intelligence.

These predominantly null results could be due to the fact that predicted effect sizes are small enough that they may be difficult to detect without very large numbers of subjects. The monozygotic twin study Behrman and Rosenzweig 2004 provides an estimate of the impact of birth weight on adult stature: a one pound increase of birth weight increases adult stature by 1.5 cm (0.6 inches) (10). An intervention that increased birth weight by one pound would be rather impressive, and the single prenatal intervention reviewed above only achieved 14 percent of this (136 grams). According to the estimate of Behrman and Rosenzweig, this prenatal intervention would produce a predicted increase of only 0.2 cm of adult height, which would require substantial statistical power to measure in a field study of this nature. Therefore, one obvious potential explanation for the inconsistent results of these studies is that the effect sizes are small and the studies to date have not had enough statistical power to detect it reliably. I believe this is the most likely explanation for the findings, based on context from the natural experiment and animal literature. However, I cannot rule out the possibility that the interventions are simply not effective.

We also must face the concerning possibility that in some contexts, nutrition supplementation that increases early-life growth may actually impair later-life outcomes. In the Brazilian trial, nutrition counseling increased early life growth and adult stature, but reduced a measure of general intelligence. There are several possible explanations for this finding. The most straightforward is that the intervention actually reduced general intelligence. However, based on the animal and natural experiment literature, this seems unlikely.

A second possibility is that the result is due to residual confounding, despite adequate randomization. This can occur when bad luck during random assignment creates baseline differences between groups that impact outcome measures. In support of this explanation, when the authors adjusted for baseline differences between groups, the finding became nonsignificant.[85]

A third possibility is that the result arises from the multiple comparisons problem encountered repeatedly in this report. The more hypotheses that are tested using data from the same experiment, the higher the likelihood of a false positive outcome. False positives can occur in either direction, i.e. they are just as likely to suggest a reduction of general intelligence as an increase. The unadjusted p-value for the negative effect on general intelligence is 0.047—scarcely crossing the significance threshold. Table 2 contains eleven hypothesis tests, and applying Bonferroni correction places this p-value well short of statistical significance.

My view is that the second and third explanations are more likely than the first, and the Brazilian study does not provide sufficient reason to be concerned about negative impacts of early-life nutrition interventions on later-life outcomes.

Randomized controlled trials tentative conclusions

Randomized controlled trials do not provide compelling evidence that increasing early-life growth via nutrition interventions improve individual adult economic outcomes, however they also do not provide compelling evidence that there is no effect. Of the two studies that directly tested the hypothesis, only one supported it and the result was inconsistent and not very convincing overall (e.g., effects only in men, only for one economic outcome, and only after multivariate adjustment). Expanding the search to indirect measures of adult economic potential identified three additional studies, but these did not provide further support to the hypothesis overall.

Among the five studies considered, the Guatemelan trial had the most positive influence on indirect outcomes. The reason for this is unclear, however there are two obvious possibilities. The first is that the Guatemalan trial may have had the clearest beneficial effects later in life because it had the largest effects early in life. It appears to have increased linear growth, reduced stunting prevalence, and reduced early-life mortality. The second possibility is that the result is an artifact caused by ineffective randomization. As discussed previously, the trial did not include sufficient units of randomization to be fully randomized. This allows the possibility that later-life observations are due to differences that were intrinsic to individual villages rather than caused by the intervention. The adjustments applied by the authors using multiple regression do not fully address this problem.

The animal and natural experiment literature suggest that partially relieving early life growth restriction should enhance traits that contribute to adult economic status. Yet this is not convincingly observed in

[85] The legend of Table 2 describes the baseline variables that the authors adjusted, resulting in a nonsignificant outcome (89): “Adjusted for the baseline variables: maternal schooling in years and child weight-for-age Z-score, height-for-age Z-score and weight-for-height Z-score”
Interestingly, this could be viewed as a possible instance of p-hacking in order to seek a nonsignificant result where a significant result is undesirable.

the randomized controlled trial literature. One possibility, discussed previously, is that the effects are too small to be reliably observed without larger sample sizes. This seems most likely to me, given the effect sizes predicted by the natural experiment literature. Yet it is also possible that nutrition interventions that partially relieve early life growth restriction are simply not effective at enhancing traits that contribute to economic status, at least in the contexts where they have been studied. Whether or not there is a real effect that is simply difficult to measure, my view is that these largely null results should inspire skepticism about the real-world effectiveness of nutrition interventions to improve outcomes related to economic status in low-income settings.

Overall tentative conclusions

Summary

I began the research for this report with a weak prior belief that early life growth probably impacts adult economic outcomes. This belief originated from the general plausibility of the idea, the fact that it is commonly asserted in high-profile review papers on the subject, and the fact that it is taken seriously by major organizations such as the World Bank and the World Health Organization.[86] However, at the outset my knowledge of the evidence was very limited and my beliefs about it were correspondingly weak.

My prior is strengthened by the animal literature, which supports certain commonly-asserted arguments about the hypothesis, in particular:

- Early growth restriction imposed by insufficient calorie and/or protein intake alters development in ways that could plausibly limit adult economic status if they occurred in humans (e.g., reduced linear growth, reduced adult brain size, alterations in brain structure, behavioral deficits).
- More severe, more prolonged, and/or earlier nutrient restriction leads to more severe developmental deficits.

[86] From page 1 of the World Health Organization document “Essential Nutrition Actions” 2013 (8): “This document provides a compact summary of WHO guidance on nutrition interventions targeting the first 1000 days of life. Focusing on this package of essential nutrition actions, policy-makers could reduce infant and child mortality, improve physical and mental growth and development, and improve productivity.”

From the 1,000 Days website page on stunting (5): “Beyond the individual impacts of this problem, stunting is an enormous drain on economic productivity and growth. Economists estimate that stunting can reduce a country’s GDP by as much as 12%.”

From Victora and colleagues 2008, page 340 (6): “The prevention of maternal and child undernutrition is a long-term investment that will benefit the present generation and their children.”

From Galasso and colleagues 2017, pages 5 and 6 (7): “Stunting among children today reduces a country’s future income per capita… Our estimates suggest that implementing the Bhutta et al. program, and factoring in the annual trend decline of 1.5% p.a., will leave the stunting rate in 2025 at 36% below to its 2010 value – 4 percentage points shy of the 40% target reduction adopted by the 65th World Health Assembly. We estimate a rate-of-return for the 34 countries as a whole of 17%, with a benefit-cost ratio of 15:1.”

The animal literature also suggests that deficits in early life development can often be mostly or completely reversed if the animal has access to adequate nutrition later in development, suggesting that developmental deficits acquired in the first thousand days may not be totally irreversible.

Due to interspecies differences, we cannot assume that animal research translates well to humans, but I think it forms a reasonable set of priors that can be updated using evidence from research in humans. This is particularly true since researchers have reported similar results in several nonhuman species including primates.

Excluding natural experiment studies, my opinion is that the observational literature is not very informative about the link between early life growth and individual adult economic outcomes. This is because more informative study designs are available and the natural experiment literature suggests that typical observational methods may be heavily confounded.

In my opinion, the natural experiment literature is the most informative class of evidence in this report because it is less susceptible to confounding than other observational studies and has fewer design and size constraints than the randomized controlled trial literature. This literature supports the hypothesis overall, from several independent angles, yet there are two key caveats. First, the highest-quality study did not report a significant effect (in the monozygotic twin sample, which I believe is the most informative) and its results conflict with those of a similar high-quality study, raising questions about the source of the discrepancy. Second, the natural experiment literature is primarily informative about the impacts of gestational growth, and not very informative about the impacts of postnatal early growth.

Randomized controlled trials, a mainstay of causal inference in many fields, are only weakly informative in this case. It appears to be very difficult to conduct trials that test this hypothesis effectively. Low-income settings create challenges for study implementation, small effect sizes are hard to detect, and very long follow-up periods reduce sample sizes, introduce attrition bias, and increase variation between individuals that can drown out effects. Researchers have dealt with these problems by using data analysis methods that ostensibly increase the signal-to-noise ratio, but these methods also undermine causal inference.

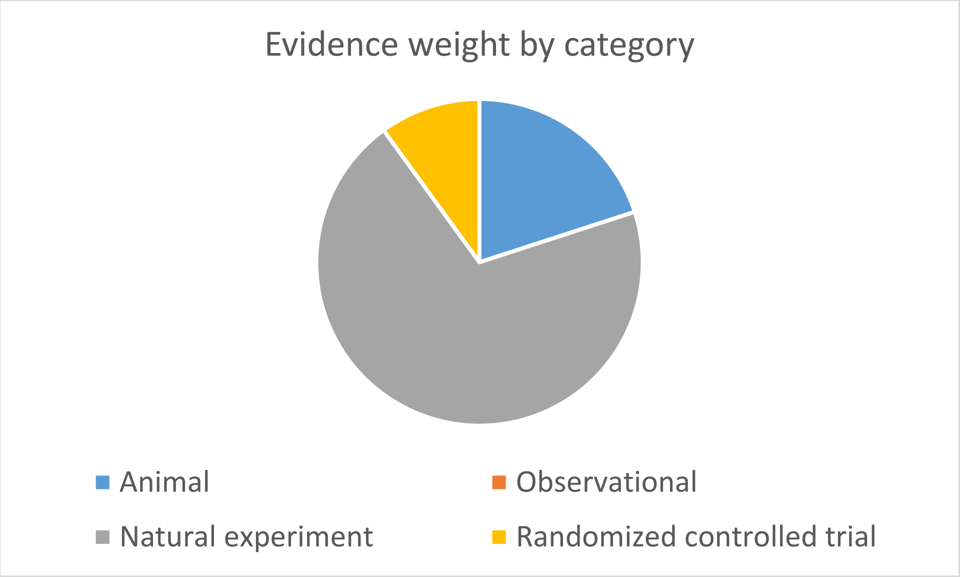
The randomized controlled trial literature adds slightly to the case that alleviating early postnatal growth restriction increases economic prospects in adulthood. This conclusion rests almost exclusively on the results of the Guatemalan trial, since the Jamaican trial was unsupportive and indirect indicators of economic potential were not clearly supportive. The randomized controlled trial literature in this area highlights the difficulty of implementing early life growth interventions and the substantial challenge of evaluating their efficacy, particularly in later life.

Tentative conclusions

Based on my review of the evidence, I believe with approximately 78 percent confidence that growth restriction in the first thousand days of development limits individual adult economic outcomes. Since this is not a meta-analysis my reasoning process is somewhat subjective, but I will attempt to make it as explicit as possible.

My strategy was to identify all areas of evidence that are potentially informative, with the idea that my conclusions will be more robust if they are based on multiple independent lines of evidence. Then I evaluated each area of evidence and formed an opinion about its informativeness. I weighted each category of evidence according to its informativeness and used the weighted category-specific conclusions to form my overall conclusions.

When evaluating the primary hypothesis that early growth restriction limits adult economic outcomes, I assigned an approximate weight of 20 percent to general plausibility and the animal literature, zero percent to the observational literature, 70 percent to the natural experiment literature, and 10 percent to the randomized controlled trial literature. I have represented this graphically below.



Within the natural experiment category, approximately 95 percent of my evidence weight comes from the seven studies I discussed in detail (67 percent of total). Fifty-eight percent comes from the two studies I find most informative of all, both of which are twin studies: Behrman and Rosenzweig 2004 and Black and colleagues 2005 (41 percent of total) (10,18).

All major lines of evidence I evaluated offer some degree of support to the hypothesis, which increases my confidence in it. Natural experiment studies, the most informative line of evidence, are fairly supportive of the hypothesis, further increasing my confidence. Yet the fact that the most informative study of all (Black and Colleagues 2005) is not very supportive limits my confidence significantly. Together, this leads me to an overall confidence level of 78 percent.[87] Concerns about publication bias, methodological problems, and conflicting results prevent this from rising higher.

[87] I did not arrive at this number via an objective quantitative process; it is simply an attempt to clearly communicate my degree of belief in the hypothesis after having reviewed the evidence.

External validity

The primary populations that would be targeted by early life growth interventions are those experiencing high rates of early life growth restriction. These are typically low-income settings with a high prevalence of undernutrition and/or infectious disease.[88] Although the evidence supports the general conclusion that early-life growth impacts adult economic outcomes, its usefulness for justifying early-life growth interventions depends on the degree to which we believe it applies to low-income settings.

The studies with the greatest influence on my conclusions are natural experiment studies performed in monozygotic twins (10,18,55). These were conducted in affluent, industrialized countries with a low burden of undernutrition and infectious disease by global standards. Superficially, they appear to have low external validity to our primary populations of interest. However, as discussed in the natural experiment section, twin pregnancies inherently result in fetal growth restriction, generally more pronounced for one fetus than the other, as a result of competition for nutrients. This appears similar to fetal growth restriction caused by maternal undernutrition in low-income settings, although perhaps less similar to fetal growth restriction caused by infectious disease. It seems reasonable to believe that these studies probably do have some degree of external validity when fetal growth restriction is caused by simple nutrient restriction.

Expanding our field of view to other study designs in the natural experiment literature, a number of studies tested the hypothesis in low-income settings with a higher prevalence of undernutrition and infectious disease. In my opinion, after the monozygotic twin studies the most informative natural experiment studies are the three Ramadan fasting studies and the one famine study previously discussed in detail (40,41,48,58). These studies reported supportive findings in the United States, Uganda, Iraq, Indonesia, and Malawi, while findings in Denmark were not supportive. This suggests that the impact of early life growth restriction on adult economic outcomes probably occurs in a broad range of contexts and the findings of monozygotic twin studies are probably externally valid to low-income settings.

Since several studies have supported the hypothesis in low-income settings with high rates of undernutrition and/or infectious disease, and results are fairly similar across populations, I believe the conclusions of this report should have good external validity for early-life growth interventions.

Also worth considering is the possibility that early-life growth restriction may have different long-term effects in affluent vs. low-income settings. In a setting of resource abundance, parents and society can increase investment in growth-restricted children and attempt to catch them up to a normal growth curve. Returning to Behrman and Rosenzweig 2004, the results reveal that a one z-score lower birth

[88] From Kramer 1987, page 663 (9): "In developing countries, the major determinants of [intrauterine growth restriction] are Black or Indian racial origin, poor gestational nutrition, low pre-pregnancy weight, short maternal stature, and malaria."

de Onis and colleagues list the 2010 prevalence of stunted preschool children at 6.0% in "developed countries" and 29.2% in "developing countries" (13).

weight leads to a 0.26 z-score shorter adult height and no difference of adult body mass index (10). This offers some evidence that in an affluent setting, catch-up growth occurs, narrowing the gap between normal and low-birth-weight infants.

I do not have monozygotic twin data from a low-income setting to compare with, so it is unclear whether catch-up growth would be less in this setting. However, it seems possible that catch-up growth would be constrained a resource-limited setting since nutrient restriction and/or infectious disease pressure may continue after birth. For this reason, it is not difficult to imagine that effect sizes may actually be larger in low-income settings than studies like Behrman and Rosenzweig 2004 predict.

Effect size

To quantify the magnitude of the association between early life growth restriction on economic outcomes, I focus on monozygotic twin studies from the natural experiment literature, for three reasons. First, they directly quantify growth restriction. Second, they provide what I believe are the most reliable estimates of the causal impact of *in utero* growth restriction on adult economic outcomes. Third, the exposure variable of birth weight is directly relevant to the early life growth interventions whose cost-effectiveness we would like to evaluate.

I base my estimate on the three highest-quality monozygotic twin studies: Behrman and Rosenzweig 2004, Black and colleagues 2005, and Miller and colleagues 2005 (10,18,55). These reported that one pound of additional birth weight results in 7, 0.9, and 6.4 percent higher adult earnings. Applying the study weights described in the natural experiment study section, I estimate that one pound of additional birth weight in a low-birth-weight context yields 3.7 percent higher adult earnings.[89]

Since the natural experiment literature has not allowed me to estimate the relative importance of different periods of development, I will not attempt to determine the magnitude of the effect of other periods of early life growth restriction on adult economic outcomes. On the basis of the animal literature, it seems reasonable to suppose that a longer duration of growth restriction may cause larger effects in adulthood, however the natural experiment literature does not allow me to update this prior effectively. If true, it suggests that growth restriction that continues after birth might diminish adult earnings by more than 3.7 percent.

I have not considered whether this extra income would be realized in a zero-sum or positive-sum manner, in other words whether boosting someone's birth weight would increase their income at the expense of others in their community, yielding no increase in average income. This important question is beyond the scope of my investigation.

I will put this effect size into context in two ways. The first is to compare it to established thresholds for concern about birth weight. One is "small for gestational age", which is defined as the bottom ten percent of a gestational-age-specific weight distribution (4). If we assume a normal distribution of

[89] I assigned a weight of 23, 35, and 7 percent to each of the three studies, respectively. Therefore, each study receives 35, 54, and 11 percent of the total weight. Multiplying these weights by the effect sizes reported in each study and summing yields an estimate of 3.7 percent.

weights, this translates to a z-score of -1.28 or smaller (1.28 standard deviations below the mean). Using the standard deviations reported in Behrman and Rosenzweig 2004 and Black and colleagues 2005, as well as the effect size estimated above, we can estimate that a z-score of -1.28 would be expected to reduce adult earnings by 5.8 percent, a small but meaningful impact.[90]

The second way to put this effect size into context is to compare it to the much-studied relationship between education and adult earnings. I have not spent significant time searching the literature or evaluating the quality of the evidence underlying estimates of this relationship, so I do not claim the following estimate can be taken as an accurate measure of the causal impact of education on earnings. I only provide it as a means of comparing our current data with a relationship that is commonly considered to be quantitatively important. Montenegro and Patrinos 2013 compiled data from 131 economies and 545 household surveys for the World Bank, finding that one year of additional schooling is associated with 10.4 percent higher adult earnings (90). This suggests that the returns to one pound of additional birth weight of an infant on the lower end of the weight distribution are approximately as large as one third of a year of additional schooling. If growth restriction continues after birth due to poor postnatal conditions, its impact could conceivably be larger than one-third of a year of schooling, although this is speculative.

One pound may seem like a large amount, yet low birth weight is defined as a weight below 5.5 pounds (2,500 grams). According to Behrman and Rosenzweig 2004, the average birth weight in the general US population in 1988 was 7.4 pounds (p. 591, (10)). This means that low birth weight infants are 1.9 pounds or more below average US weight, leaving much room for increase. In 2004, the World Health Organization and UNICEF estimated the incidence of low birth weight among 154 countries (3). They report that globally, 15.5 percent of all infants exhibit low birth weight. South-central Asia has the highest regional incidence of 27 percent, driven primarily by India at 30 percent. Western Asia and Western Africa are tied for the second highest regional incidence at 15.4 percent. US incidence is reported as 8 percent. According to Behrman and Rosenzweig 2004, mean birth weight in India in 1998/99 was 6.1 pounds, which is 1.3 pounds below the US mean (Table 4, (10)). Given these figures, it seems likely that low birth weight meaningfully constrains adult economic outcomes, particularly in low-income regions where low birth weight is common. On the basis of the natural experiment literature, it is difficult to determine whether the same is true for early postnatal growth.

Since low birth weight is tied to modifiable risk factors such as undernutrition, infectious disease, and cigarette smoking, it is conceivable that small but meaningful gains of adult earnings could be realized by interventions intended to increase birth weight.[91] This is in addition to gains that may be realized for

[90] A simple mean of standard deviations of birth weight reported in the twin samples of Behrman and Rosenzweig 2004 and Black and colleagues 2005, expressed in ounces, is 19.75 oz. This is equal to 1 z-score unit. Our effect size estimate is 3.7 percent higher earnings per 16 ounces of birth weight. Multiplying 3.7 by (19.75/16) equals 5.8.

[91] From Kramer 1987, page 663 (9): “Factors with well-established direct causal impacts on intrauterine growth include infant sex, racial/ethnic origin, maternal height, pre-pregnancy weight, paternal weight and height, maternal birth weight, parity, history of prior low-birth-weight infants, gestational weight gain and caloric intake, general morbidity and episodic illness, malaria, cigarette smoking, alcohol consumption, and tobacco chewing. In developing countries, the major determinants of [intrauterine growth restriction] are Black or Indian racial origin, poor gestational nutrition, low pre-pregnancy weight, short maternal stature, and malaria.”

other adult outcomes relevant to human well-being. However, it is unclear how much of these regional differences in birth weight are due to environmental factors and how much are due to genetics, leaving unanswered the question of how much they can be modified.

What is the critical age window?

Public health advocacy organizations frequently state that the first 1,000 days of development—from conception to age two—is a critical age window with disproportionate impacts on later life outcomes, and that optimizing it is a key leverage point for improving health and economic status.[92] The animal literature is consistent with this position, suggesting that restricting early development causes more severe and lasting deficits than restricting later development.

The natural experiment studies I identified are primarily concerned with *in utero* effects and provide very little basis for comparing the importance of *in utero*, early postnatal, and later periods of development. The ideal resource for this would be studies that provide independent estimates for all three developmental periods using the same methods. I have identified one study that does so, although it is not very informative for reasons I will explain. Several studies in Table 1 provide data on both *in utero* and early postnatal exposures, yet as detailed below, they are not very helpful for resolving the current question.

Chen and Zhou 2007 studied the long-term economic outcomes of survivors of the Chinese Great Famine that occurred in 1959 through 1961 (44). The Chinese Great Famine was a long and severe famine that caused an estimated 15-30 million excess deaths.[93] Famine exposure varied considerably by

[92] From the 1,000 Days website (5): "Good nutrition is critical to support the rapid growth and development of babies and young children during their first 1,000 days. Without good nutrition however, a young child can suffer serious and often permanent damage to his developing brain and body. We can't really see this damage but we can measure it by looking at how well a child is or isn't growing. A child who doesn't grow well and is too short for their age suffers from a condition known as stunting… The effects of stunting last a lifetime: impaired brain development, lower IQ, weakened immune systems, and greater risk of serious diseases like diabetes and cancer later in life."

From Victora and colleagues 2008, page 340 (6): "Poor fetal growth or stunting in the first 2 years of life leads to irreversible damage, including shorter adult height, lower attained schooling, reduced adult income, and decreased offspring birthweight. Children who are undernourished in the first 2 years of life and who put on weight rapidly later in childhood and in adolescence are at high risk of chronic diseases related to nutrition… The prevention of maternal and child undernutrition is a long-term investment that will benefit the present generation and their children."

From Galasso and colleagues 2017, page 5 (7): "Stunting in childhood matters because it is associated with adverse outcomes throughout the life cycle. The undernourishment and disease that cause stunting impair brain development, leading to lower cognitive and socioemotional skills, lower levels of educational attainment, and hence lower incomes. Health problems in terms of non-communicable diseases are more likely in later life, leading to increased health care costs. Stunting in childhood also leads to reduced stature in adulthood, which, due to the persistence of shortness over the lifetime, and the negative (and independent) effect of height on income, further reduces income in adulthood."

[93] From Chen and Zhou 2007, page 659 (44):

county. Most studies of this event use regional differences of famine severity to estimate the effects of famine exposure on adult outcomes, and compare cohorts born at different times relative to the famine. By analyzing the cohorts by birth year, the authors are able to observe outcomes between individuals who differ in the timing and length of early life exposure.

In this study, the largest associations with economic outcomes were observed in people born in 1959 or 1960, with no significant associations for those born in 1961. Taking the results at face value, the former finding suggests that the postnatal early growth period is particularly important, and/or perhaps simply that longer exposure to famine yields more severe outcomes. The latter finding suggests that the developmental period encompassing gestation and very early postnatal life may not be critical for adult outcomes, although it is notable that this contradicts the majority of the rest of the literature. This study does not appear to be especially informative for the current question.

Maccini and Yang 2008 studied the long-term economic impacts of variation of rainfall in Indonesia, which cause variation of the food supply and economic status of rural households that are the study's focus (49).[94] From the information presented in the paper, it is unclear how tight of a relationship exists between rainfall and early life growth, but the relationship is clearly indirect because one would have to propose that rainfall impacts crop yields, which in turn impact food availability and income, which in turn impact early life growth.

The authors report no association between rainfall during gestation and asset index in later life. Only rainfall in the first postnatal year was associated with asset index, however applying my p-value threshold of 0.05, that finding is also not statistically significant, although it is larger in magnitude than other exposure periods. No association was observed for rainfall in the second, third, or fourth years. This study provides weak evidence that growth in the first postnatal year is more important for adult economic outcomes than gestation and later growth, including growth after the first 1,000 days.

Meng and Qian 2009 studied the long-term economic outcomes of survivors of the Chinese Great Famine (20). They focus on the top of the distribution of outcomes (i.e., the 90th percentile of height, educational attainment, hours worked, etc.) to avoid selection bias whereby famine may have

"China's 1959–1961 famine stands out as the worst in human history: there were about 15–30 million excess deaths and about 30 million lost or postponed births."

[94] From Maccini and Yang, pages 6 and 7 (49): "Rainfall is the most important dimension of weather variation in Indonesia. Because of its equatorial location, temperature shows very little variation in Indonesia, either within years or across them. In any particular year, the length of the wet season and the intensity of drought during the dry season vary markedly across Indonesia's 5,100-kilometer east-west span. The specific trajectories of the monsoons vary from one year to the next, and lead to wide variation in precipitation across the archipelago both within year and across years (Library of Congress 2003)."

"Levine and Yang (2006) find that deviations of rainfall from the district-level mean are positively associated with deviations of rice output from the district-level mean in Indonesian nonurban districts in the 1990s (years in which rainfall is unusually high have unusually high rice output, and years in which rainfall is unusually low have unusually low rice output)."

"Because of the importance of the seasonal cycle of rain, food security also tends to vary seasonally. Food insecurity tends to the highest at the end of the dry season and beginning of the following wet season when stocks of food from the previous wet season are low and physical demands are high with the initiation of planting (Herdt 1989). Accordingly, dry season droughts amplify food scarcity."

preferentially killed weaker individuals, leaving a stronger cohort less likely to show adverse long-term effects of famine exposure. The results suggest that postnatal exposure to famine at ages one to six is associated with fewer hours worked in adulthood, but *in utero* exposure is not. Similar to Chen and Zhou 2007, this comparison suffers from the limitation that individuals experiencing famine solely *in utero* were exposed for a shorter period of time than individuals experiencing famine after birth. For this reason, the result appears difficult to interpret for our current purpose.

Neelsen and Stratmann 2011 studied the long-term economic outcomes of survivors of the 1941-42 Greek famine, an event caused by a combination of military occupation and naval blockade during World War II (51). The Greek famine lasted six to eight months and therefore, like the Dutch famine, is suitable for disentangling the effects of famine on different periods of early development.[95] It was severe, causing a 3- to 10-fold increase in mortality rates.[96]

The authors' results suggest that exposure to famine *in utero*, in postnatal year one, or in postnatal year two are associated with lower-prestige adult employment, with little difference between the three age groups. This result became nonsignificant after controlling for urban vs. rural location of exposure, suggesting that it may result partially or entirely from confounding. This study offers weak evidence that growth restriction *in utero*, in postnatal year one, and postnatal year two results in similar adult economic outcomes.

As previously discussed, Mussa 2017 studied the long-term economic outcomes of individuals exposed to variations of corn yields in early life in Malawi (48). Corn is the primary staple food of Malawi, and according to the authors, corn yields are "the best direct indicator of [rural] incomes" (page 2 (48)).

The author reports that 10% lower relative corn yield while a future farmer is *in utero* is associated with 4.2% lower relative adult corn production. No statistically significant association was observed for postnatal years 1 or 2, and the magnitude of the associations was much smaller. This study provides some evidence that *in utero* effects are more important than early postnatal effects, although as previously discussed there are reasons to be skeptical of this finding.

Overall, the natural experiment literature provides a rather weak basis for comparing the impacts of growth restriction *in utero* vs. years 1-2 vs. later development. I do not find the Chinese Great Famine studies to be very well suited to this question for reasons previously discussed, which leaves us with three studies on rainfall variability in Indonesia, the Greek famine, and corn yield variability in Malawi (48,49,51). None of these studies directly examine early life growth and they require the assumption

[95] From the abstract of Neelsen and Strantmann 2011 (51): "Given the short duration of the famine, we can separately identify the famine effects for cohorts exposed in utero, during infancy and at one year of age."

[96] From Neelsen and Strantmann 2011, page 7 (51): "The nutritional situation became critical in the summer of 1941 and in the fall turned into a full-blown famine. In the Greater Athens area, the calorific value of rations and food provided by public or charity soup kitchens deteriorated from 600 calories per day and person in July of 1941 to 320 in November of 1941. In many places, civil registration records were discontinued during the occupation (Valaoras 1960). Where they were not, the data suggest mortality increases between 300 and 1000 percent compared to pre-war years. Estimates of a country-wide death toll of the famine vary between 100,000 and 200,000 (Hionidou 2006) or 1.4 to 2.8 percent of the population, the large majority of which occurred between October 1941 and March 1942 (Helger 1949)"

that prevailing conditions during early life impacted growth. The study of Indonesia weakly suggests that the first postnatal year may be more important than gestation and the second, third, and fourth postnatal years. The Greek famine study suggests that gestation, year one, and year two are equally important. The study of Malawi suggests that gestation is more important than postnatal years 1 and 2.

The human evidence I reviewed, including natural experiment studies and randomized controlled trials, offers very little basis for comparing the impact of different periods of development. It is possible that I could gain more information about the relative importance of different developmental periods by considering other forms of evidence, but I have not done this yet.

Gestation is the only period of development that the evidence strongly links to adult economic outcomes. This is primarily because gestation is easier to target using natural experiment studies, particularly the most compelling study designs that exploit monozygotic twin pregnancies and exposure to Ramadan fasting *in utero* (children are not required to fast so the impact of Ramadan fasting on early postnatal development cannot be studied by this method).[97]

Certain natural experiment studies and one randomized controlled trial hint that early postnatal development may also be important, but the evidence for this is much less convincing and does not offer compelling direct comparisons between different periods of development. However, an absence of evidence is not evidence of absence, so I cannot resolve the relative importance of early postnatal development vs. gestation with the evidence I have reviewed.

What is the mechanism?

Presumably, if early life growth restriction impairs individual adult economic outcomes, it does so by impacting developmental processes that produce traits that influence later-life economic status. The animal literature provides several possible mechanisms by which this could occur: Reduced linear growth leading to smaller adult stature and lower physical work capacity, alterations in organ development leading to greater susceptibility to noncommunicable disease, and reduced brain size and/or altered brain microstructure leading to deficits in behaviors that contribute to economic status.

The human evidence I reviewed in this report does not directly investigate the mechanisms by which restricted early growth may impair adult economic status. However, some studies do report traits that are plausible mechanisms. Restricted early-life growth, or exposure to conditions expected to restrict growth, has been associated with smaller adult stature, a higher prevalence of mental illness, lower general intelligence, and higher susceptibility to metabolic disease, particularly in natural experiment studies (10,18,40,91). The evidence I have reviewed does not allow me to determine the degree to which these mechanisms contribute (or do not contribute) to poorer economic outcomes, but they are certainly plausible contributors, in addition to being important in their own right.

[97] From page 1 of Almond and Mazumder 2008 (40): “Muslims generally fast each day during the lunar month of Ramadan. Fasting includes abstaining from eating and drinking during daylight hours. Certain persons are automatically exempted from fasting: “children, those who are ill or too elderly, those who are traveling, and women who are menstruating, have just given birth, or are breast feeding” [Esposito, 2003].”

I did not cast a wider net and search for additional evidence that might link these mechanisms to economic outcomes, but this is something I could do in future work on this report.

Cost-effectiveness

I have performed a preliminary cost-effectiveness estimate of interventions intended to increase birth weight or early postnatal growth. It can be found here.

Other potential impacts of early-life growth interventions

This report focuses specifically on adult economic outcomes and does not attempt to evaluate other potential positive or negative effects of increasing early-life growth. Although detailed consideration of these topics is currently beyond the scope of this report, I will briefly discuss other potentially significant impacts here.

Some researchers have argued that restricted early-life growth, particularly during gestation, increases adult risk of noncommunicable disease. For example, the Dutch famine study reported that gestational famine exposure is associated with a higher risk of cardiovascular disease and obesity in adulthood (91). On the other hand, some evidence suggests that faster postnatal growth following gestational growth restriction is associated with higher adult obesity risk (92,93).

Related to this, one Ramadan fasting study discussed in this report found evidence that gestational Ramadan exposure increases the adult risk of mental disability, and another found less compelling "suggestive evidence" (40,58). A portion of this effect is probably captured in economic outcomes, but mental disability seems intrinsically undesirable beyond its economic benefits.

Some natural experiment studies, including one twin study, report that higher birth weight leads to higher adult general intelligence (18,41). A portion of this is probably captured in the economic outcomes, but again a higher general intelligence seems intrinsically desirable beyond its economic benefits.

Ways to modify early-life growth

This report could be relevant to several types of early-life growth interventions. Although it is beyond the current scope of this report to review these in detail, I will discuss some of them briefly.

Providing food rich in calories, protein, and micronutrients to mothers and/or young children in food-limited situations is an obvious way to address restricted early-life growth. I have not examined this evidence in detail but meta-analyses suggest that such interventions can increase body weight in early life by about one quarter of a standard deviation (94,95).

Micronutrient supplementation is a related intervention that costs less than food provision but is also probably less effective for increasing early-life growth. There is some evidence supporting the effectiveness of multiple micronutrient supplementation, and zinc specifically, but I have not examined it in detail (96–98).

Sometimes, nutritional problems are not caused by low food availability but by suboptimal dietary practices such as weaning infants onto low-protein, low-calorie, and/or low-micronutrient foods. Nutrition education is an intervention that addresses this problem and may cost less than food provision. A quick literature search suggests that complementary feeding education is effective for increasing early postnatal growth, but pregnancy nutrition education is not effective for increasing birth weight (94,95).

Nutrition-sensitive interventions are those that indirectly impact nutrition. For example, interventions that increase or stabilize agricultural production of calories and/or micronutrients, or social safety nets that increase food security (99). I have not investigated the efficacy of nutrition-sensitive interventions for increasing early-life growth.

Infectious disease control is another intervention that may impact early-life growth. For example, antimalarial drugs may decrease the risk of low birth weight (100). I speculate that the conclusions of this report may have lower external validity for such interventions because the mechanism is different than the studies I considered.

How I could be wrong

I believe it is possible that my conclusions are substantially inaccurate. One way to place approximate bounds on this uncertainty is to consider the range of outcomes among the most informative studies. Among the three most informative monozygotic twin samples—a design that also allows us to quantitatively estimate the relationship between birth weight differences and adult earnings—one pound of additional birth weight yields between 0.9 and 7 percent higher adult income.

The lower end of this bound suggests that the impact of birth weight on adult earnings may not be very meaningful, perhaps even small enough to leave the hypothesis effectively unsupported, particularly if we consider that the result was not statistically significant. The upper end of this bound suggests that the impact could be fairly meaningful, nearly twice as large as my estimate. Of course, these bounds are based on estimates that are themselves uncertain, so they should be interpreted with caution.

I have argued that the natural experiment literature provides the most informative test of this hypothesis, above even randomized controlled trials. Yet in many domains, randomized controlled trials are considered the gold standard for causal inference. Another way in which my conclusions could be inaccurate is if natural experiment studies such as monozygotic twin studies are not as effective for causal inference as I believe they are. Given the sparse, weak, and inconsistent evidence from the randomized controlled trial literature and the absence of convincing evidence that early life nutrition supplementation interventions increase adult intelligence, if randomized controlled trials were the most informative source of evidence then I would conclude that the hypothesis is poorly supported.

Publication bias is another possible pitfall. I rely heavily on natural experiment studies in this report, and they are not very resource-intensive to conduct relative to randomized controlled trials. Many are based on pre-existing databases that require few resources to obtain and analyze, and the number of authors per study is small. This is the type of situation in which one would expect a high risk of publication bias because there is little penalty for discarding a disappointing result. Add to this the fact that none of the studies were preregistered and therefore the authors had considerable latitude in conducting, presenting, and interpreting their analyses, and the potential for a biased literature is substantial. I adjusted for this possibility in my cost-effectiveness analysis (using a replicability adjustment) but did not factor it in to my effect size estimate in this report.

A related issue is that when we (GiveWell and Open Philanthropy Project) replicate econometric studies similar to the natural experiment studies included in this report, we often become more skeptical of the results originally reported by the authors. As an example, in his incarceration report, Open Philanthropy Project senior advisor David Roodman concludes the following: “Of the eight studies selected for this review whose data sets I could obtain or reconstruct, reanalysis revealed minor problems in one (Green and Winik) and significant issues of methodology or interpretation in seven (Helland and Tabarrok, Abrams, Levitt, Lofstrom and Raphael, Green and Winik, Kuziemko, Ganong), which led to major reinterpretations of four (Helland and Tabarrok, Abrams, Kuziemko, Ganong). There is no reason to believe that the studies for which data were unavailable are more reliable” (101). This possible source of inaccuracy is also reflected in the replicability adjustment of my cost-effectiveness analysis. I suspect this is less likely to be a problem with monozygotic twin studies due to their small number of assumptions and simplicity of design, but we have not tested this.

Given the complexity of this body of literature, there may be other sources of inaccuracy that I have not identified.

Future work

Future work on this report could include:

- Replicating key natural experiment studies to increase my confidence in the sources of evidence that have the greatest impact on my beliefs. In particular, it would be informative to replicate Behrman and Rosenzweig 2004 while including data from men.
- Expanding the animal research section to be more systematic and more focused on primary literature in order to evaluate/strengthen my conclusions in this area.
- Considering additional forms of evidence that may provide more information about the relative importance of different periods of development to adult outcomes.
- Developing and applying a systematic, objective process for evaluating studies, weighting them, and integrating them into overall conclusions. My current process is somewhat subjective, as with all narrative reviews.
- Refining my discussion of external validity by examining the specific populations that are currently being targeted by early-life growth interventions, identifying the likely causes of growth restriction in those settings, and determining how well they correspond with the settings of the studies in this report.

- Performing power calculations to estimate what size randomized controlled trial would be required to detect the impact of early life growth on adult economic outcomes, given the effect size predicted by natural experiment studies. This would help determine whether the randomized controlled trials had sufficient statistical power to detect an effect.
- Investigating whether income gains would occur in a positive-sum or zero-sum fashion.
- Considering impacts of early-life growth besides economic status, such as health.
- Having the report reviewed by experts and incorporating their input.

References

1. WHO child growth standards. World Health Organization; 2006.

2. WHA Global Nutrition Targets 2025: Stunting Policy Brief. World Health Organization; 2015.

3. Low birthweight. Country, regional, and global estimates. World Health Organization; 2004.

4. Physical status: The use and interpretation of anthropometry. World Health Organ Tech Rep Ser. 1995;854.

5. Stunting [Internet]. 1,000 Days. [cited 2017 May 31]. Available from: http://thousanddays.org/the-issue/stunting/

6. Victora CG, Adair L, Fall C, Hallal PC, Martorell R, Richter L, et al. Maternal and child undernutrition: consequences for adult health and human capital. Lancet Lond Engl. 2008 Jan 26;371(9609):340–57.

7. Galasso E, Wagstaff A, Naudeau S, Shekar M. The economic costs of stunting and how to reduce them. World Bank; 2017.

8. Essential Nutrition Actions: improving maternal, newborn, infant and young child health and nutrition. World Health Organization; 2013.

9. Kramer MS. Determinants of low birth weight: methodological assessment and meta-analysis. Bull World Health Organ. 1987;65(5):663–737.

10. Behrman JR, Rosenzweig MR. Returns to Birthweight. Rev Econ Stat. 2004 May 1;86(2):586–601.

11. Black RE, Victora CG, Walker SP, Bhutta ZA, Christian P, de Onis M, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. Lancet Lond Engl. 2013 Aug 3;382(9890):427–51.

12. Magnus P. Causes of variation in birth weight: a study of offspring of twins. Clin Genet. 1984 Jan;25(1):15–24.

13. de Onis M, Blössner M, Borghi E. Prevalence and trends of stunting among pre-school children, 1990-2020. Public Health Nutr. 2012 Jan;15(1):142–8.

14. Maternal and child undernutrition. Lancet. 2008;371.

15. Maternal and child nutrition. Lancet. 2013;382.

16. 1,000 Days [Internet]. 1,000 Days. [cited 2017 Jun 2]. Available from: https://thousanddays.org/

17. Tanner JC, Candland T, Odden WS. Later impacts of early childhood interventions: a systematic review. World Bank; 2015.

18. Black SE, Devereux PJ, Salvanes K. From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes [Internet]. National Bureau of Economic Research; 2005 Nov [cited 2017 May 23]. Report No.: 11796. Available from: http://www.nber.org/papers/w11796

19. Almond D, Chay KY, Lee DS. The Costs of Low Birth Weight. Q J Econ. 2005 Aug 1;120(3):1031–83.

20. Meng X, Qian N. The Long Term Consequences of Famine on Survivors: Evidence from a Unique Natural Experiment using China's Great Famine [Internet]. National Bureau of Economic Research; 2009 Apr [cited 2017 May 24]. Report No.: 14917. Available from: http://www.nber.org/papers/w14917

21. Hoynes H, Schanzenbach DW, Almond D. Long-Run Impacts of Childhood Access to the Safety Net. Am Econ Rev. 2016 Apr;106(4):903–34.

22. Tetlock PE, Gardner D. Superforecasting: The Art and Science of Prediction. Broadway Books; 2016. 352 p.

23. Karnofsky H. Sequence thinking vs. cluster thinking [Internet]. The GiveWell Blog. 2014 [cited 2017 May 24]. Available from: http://blog.givewell.org/2014/06/10/sequence-thinking-vs-cluster-thinking/

24. Levitsky DA, Strupp BJ. Malnutrition and the brain: changing concepts, changing concerns. J Nutr. 1995 Aug;125(8 Suppl):2212S–2220S.

25. Rice D, Barone S. Critical periods of vulnerability for the developing nervous system: evidence from humans and animal models. Environ Health Perspect. 2000 Jun;108(Suppl 3):511–33.

26. Getz GS, Reardon CA. Animal models of atherosclerosis. Arterioscler Thromb Vasc Biol. 2012 May;32(5):1104–15.

27. Liu S, Manson JE, Lee IM, Cole SR, Hennekens CH, Willett WC, et al. Fruit and vegetable intake and risk of cardiovascular disease: the Women's Health Study. Am J Clin Nutr. 2000 Oct;72(4):922–8.

28. Hu FB, Rimm E, Smith-Warner SA, Feskanich D, Stampfer MJ, Ascherio A, et al. Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. Am J Clin Nutr. 1999 Feb;69(2):243–9.

29. Dobbing J, Sands J. Quantitative growth and development of human brain. Arch Dis Child. 1973 Oct;48(10):757–67.

30. Grantham-McGregor S, Cheung YB, Cueto S, Glewwe P, Richter L, Strupp B. Developmental potential in the first 5 years for children in developing countries. Lancet. 2007 Jan 6;369(9555):60–70.

31. Brown RE. Organ weight in malnutrition with special reference to brain weight. Dev Med Child Neurol. 1966 Oct;8(5):512–22.

32. Laus MF, Vales LDMF, Costa TMB, Almeida SS. Early Postnatal Protein-Calorie Malnutrition and Cognition: A Review of Human and Animal Studies. Int J Environ Res Public Health. 2011 Feb;8(2):590–612.

33. Levitsky DA, Barnes RH. Nutritional and Environmental Interactions in the Behavioral Development of the Rat: Long-Term Effects. Science. 1972 Apr 7;176(4030):68–71.

34. Elias MF, Samonds KW. Protein and calorie malnutrition in infant cebus monkeys: growth and behavioral development during deprivation and rehabilitation. Am J Clin Nutr. 1977 Mar;30(3):355–66.

35. Lizárraga-Mollinedo E, Fernández-Millán E, García-San Frutos M, de Toro-Martín J, Fernández-Agulló T, Ros M, et al. Early and Long-term Undernutrition in Female Rats Exacerbates the Metabolic Risk Associated with Nutritional Rehabilitation. J Biol Chem. 2015 Jul 31;290(31):19353–66.

36. Boersma B, Wit JM. Catch-up Growth. Endocr Rev. 1997 Oct 1;18(5):646–61.

37. Gluckman PD, Hanson MA, Cooper C, Thornburg KL. Effect of In Utero and Early-Life Conditions on Adult Health and Disease. N Engl J Med. 2008 Jul 3;359(1):61–73.

38. Cottrell EC, Ozanne SE. Early life programming of obesity and metabolic disease. Physiol Behav. 2008 Apr 22;94(1):17–28.

39. de Onis M, Dewey KG, Borghi E, Onyango AW, Blössner M, Daelmans B, et al. The World Health Organization’s global target for reducing childhood stunting by 2025: rationale and proposed actions. Matern Child Nutr. 2013 Sep 1;9:6–26.

40. Almond D, Mazumder B. Health Capital and the Prenatal Environment: The Effect of Maternal Fasting During Pregnancy [Internet]. National Bureau of Economic Research; 2008 Oct. Report No.: 14428. Available from: http://www.nber.org/papers/w14428

41. Majid MF. The persistent effects of in utero nutrition shocks over the life cycle: Evidence from Ramadan fasting. J Dev Econ. 2015 Nov 1;117:48–57.

42. Beach B, Saavedra M. Mitigating the Effects of Low Birth Weight: Evidence from Randomly Assigned Adoptees. Am J Health Econ. 2015 Jul 1;1(3):275–96.

43. Almond D, Edlund L, Li H, Zhang J. Long-Term Effects Of The 1959-1961 China Famine: Mainland China and Hong Kong [Internet]. National Bureau of Economic Research; 2007 Sep. Report No.: 13384. Available from: http://www.nber.org/papers/w13384

44. Chen Y, Zhou L-A. The long-term health and economic consequences of the 1959–1961 famine in China. J Health Econ. 2007 Jul 1;26(4):659–81.

45. Mu R, Zhang X. Gender Difference in the Long-Term Impact of Famine. International Food Policy Research Institute; 2008.

46. Rosenzweig M, Zhang J. Economic Growth, Comparative Advantage, and Gender Differences in Schooling Outcomes: Evidence from the Birthweight Differences of Chinese Twins. Yale University Department of Economics; 2012.

47. Bharadwaj P, Lundborg P, Rooth D-O. Birth Weight in the Long Run [Internet]. National Bureau of Economic Research; 2015 Jul. Report No.: 21354. Available from: http://www.nber.org/papers/w21354

48. Mussa R. Long-term Effects of Early Life Maize Yield on Maize Productivity and Efficiency in Rural Malawi. Munich Personal RePEc Archive; 2017.

49. Maccini SL, Yang D. Under the Weather: Health, Schooling, and Economic Consequences of Early-Life Rainfall [Internet]. National Bureau of Economic Research; 2008 May. Report No.: 14031. Available from: http://www.nber.org/papers/w14031

50. Royer H. Separated at Girth: US Twin Estimates of the Effects of Birth Weight. Am Econ J Appl Econ. 2009;1(1):49–85.

51. Neelsen S, Stratmann T. Effects of prenatal and early life malnutrition: Evidence from the Greek famine. J Health Econ. 2011 May 1;30(3):479–88.

52. SHI X. Famine, fertility, and fortune in china. China Econ Rev. 2011 Jun 1;22(2):244–59.

53. Jürges H. Collateral damage: The German food crisis, educational attainment and labor market outcomes of German post-war cohorts. J Health Econ. 2013 Jan 1;32(1):286–303.

54. Scholte RS, van den Berg GJ, Lindeboom M. Long-run effects of gestation during the Dutch Hunger Winter famine on labor market and hospitalization outcomes. J Health Econ. 2015 Jan;39:17–30.

55. Miller P, Mulvey C, Martin N. Birth weight and schooling and earnings: estimates from a sample of twins. Econ Lett. 2005;86(3):387–92.

56. Oreopoulos P, Stabile M, Walld R, Roos L. Short, Medium, and Long Term Consequences of Poor Infant Health: An Analysis using Siblings and Twins [Internet]. National Bureau of Economic Research; 2006 Feb. Report No.: 11998. Available from: http://www.nber.org/papers/w11998

57. Nakamuro M, Uzuki Y, Inui T. The effects of birth weight: Does fetal origin really matter for long-run outcomes? Econ Lett. 2013 Oct 1;121(1):53–8.

58. Schultz-Nielsen ML, Tekin E, Greve J. Labor Market Effects of Intrauterine Exposure to Nutritional Deficiency: Evidence from Administrative Data on Muslim Immigrants in Denmark [Internet]. National Bureau of Economic Research; 2014 Dec. Report No.: 20723. Available from: http://www.nber.org/papers/w20723

59. Fan W, Qian Y. Long-term health and socioeconomic consequences of early-life exposure to the 1959-1961 Chinese Famine. Soc Sci Res. 2015 Jan;49:53–69.

60. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Wiley; 2009.

61. Streiner DL. Best (but oft-forgotten) practices: the multiple problems of multiplicity-whether and how to correct for many statistical tests. Am J Clin Nutr. 2015 Oct;102(4):721–8.

62. Feise RJ. Do multiple outcome measures require p-value adjustment? BMC Med Res Methodol. 2002 Jun 17;2:8.

63. Econometrics [Internet]. [cited 2017 Jun 27]. Available from: http://www.mdpi.com/journal/econometrics/instructions

64. Woolston C. Registered clinical trials make positive findings vanish. Nat News. 2015 Aug 20;524(7565):269.

65. Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. Proc Natl Acad Sci U S A. 2017 Apr 4;114(14):3714–9.

66. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002 Feb 7;346(6):393–403.

67. Ridker PM, Danielson E, Fonseca FAH, Genest J, Gotto AMJ, Kastelein JJP, et al. Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein. N Engl J Med. 2008 Nov 20;359(21):2195–207.

68. Simonsohn U, Nelson LD, Simmons JP. P-curve: a key to the file-drawer. J Exp Psychol Gen. 2014 Apr;143(2):534–47.

69. Ebrahim S, Smith GD, May M, Yarnell J. Shaving, Coronary Heart Disease, and StrokeThe Caerphilly Study. Am J Epidemiol. 2003 Feb 1;157(3):234–8.

70. Frank SA. Somatic evolutionary genomics: Mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. Proc Natl Acad Sci. 2010 Jan 26;107(suppl 1):1725–30.

71. Banerjee AV, Duflo E. What is Middle Class about the Middle Classes around the World? J Econ Perspect J Am Econ Assoc. 2008;22(2):3–28.

72. Fields GS. Labor market analysis for developing countries. Labour Econ. 2011 Dec 1;18:S16–22.

73. Smith JP, Thomas D, Frankenberg E, Beegle K, Teruel G. Wages, employment and economic shocks: Evidence from Indonesia. J Popul Econ. 2002 Jan 1;15(1):161–93.

74. Martorell R. Overview of long-term nutrition intervention studies in Guatemala, 1968-1989. Food Nutr Bull UNU. 1992;(14):270–7.

75. Schroeder DG, Martorell R, Rivera JA, Ruel MT, Habicht JP. Age differences in the impact of nutritional supplementation on growth. J Nutr. 1995 Apr;125(4 Suppl):1051S–1059S.

76. Hoddinott J, Maluccio JA, Behrman JR, Flores R, Martorell R. Effect of a nutrition intervention during early childhood on economic productivity in Guatemalan adults. Lancet Lond Engl. 2008 Feb 2;371(9610):411–6.

77. Walker SP, Powell CA, Grantham-McGregor SM, Himes JH, Chang SM. Nutritional supplementation, psychosocial stimulation, and growth of stunted children: the Jamaican study. Am J Clin Nutr. 1991 Oct;54(4):642–8.

78. Gertler P, Heckman J, Pinto R, Zanolini A, Vermeersch C, Walker S, et al. Labor Market Returns to Early Childhood Stimulation: a 20-year Followup to an Experimental Intervention in Jamaica [Internet]. National Bureau of Economic Research; 2013 Jun. Report No.: 19185. Available from: http://www.nber.org/papers/w19185

79. Gertler P, Heckman J, Pinto R, Zanolini A, Vermeersch C, Walker S, et al. Labor market returns to an early childhood stimulation intervention in Jamaica. Science. 2014 May 30;344(6187):998–1001.

80. Walker SP, Grantham-McGregor SM, Himes JH, Powell CA, Chang SM. Early childhood supplementation does not benefit the long-term growth of stunted children in Jamaica. J Nutr. 1996 Dec;126(12):3017–24.

81. Walker SP, Grantham-Mcgregor SM, Powell CA, Chang SM. Effects of growth restriction in early childhood on growth, IQ, and cognition at age 11 to 12 years and the benefits of nutritional supplementation and psychosocial stimulation. J Pediatr. 2000 Jul;137(1):36–41.

82. Hawkesworth S, Prentice AM, Fulford AJC, Moore SE. Dietary Supplementation of Rural Gambian Women during Pregnancy Does Not Affect Body Composition in Offspring at 11–17 Years of Age. J Nutr. 2008 Dec;138(12):2468–73.

83. Alderman H, Hawkesworth S, Lundberg M, Tasneem A, Mark H, Moore SE. Supplemental feeding during pregnancy compared with maternal supplementation during lactation does not affect schooling and cognitive development through late adolescence123. Am J Clin Nutr. 2014 Jan;99(1):122–9.

84. Devakumar D, Chaube SS, Wells JCK, Saville NM, Ayres JG, Manandhar DS, et al. Effect of antenatal multiple micronutrient supplementation on anthropometry and blood pressure in mid-childhood in Nepal: follow-up of a double-blind randomised controlled trial. Lancet Glob Health. 2014 Nov;2(11):e654–63.

85. Rivera JA, Martorell R, Ruel MT, Habicht JP, Haas JD. Nutritional supplementation during the preschool years influences body size and composition of Guatemalan adolescents. J Nutr. 1995 Apr;125(4 Suppl):1068S–1077S.

86. Santos IS, Matijasevich A, Assunção MCF, Valle NC, Horta BL, Gonçalves HD, et al. Promotion of Weight Gain in Early Childhood Does Not Increase Metabolic Risk in Adolescents: A 15-Year Follow-Up of a Cluster-Randomized Controlled Trial. J Nutr. 2015 Dec;145(12):2749–55.

87. Walker SP, Chang SM, Vera-Hernández M, Grantham-McGregor S. Early childhood stimulation benefits adult competence and reduces violent behavior. Pediatrics. 2011 May;127(5):849–57.

88. Maluccio JA, Hoddinott J, Behrman JR, Martorell R, Quisumbing AR, Stein AD. The Impact of Improving Nutrition During Early Childhood on Education among Guatemalan Adults*. Econ J. 2009 Apr 1;119(537):734–63.

89. Munhoz TN, Santos IS, Karam S de M, Martines J, Pelto G, Barcelos R, et al. Effect of childhood nutrition counselling on intelligence in adolescence: a 15-year follow-up of a cluster-randomised trial. Public Health Nutr. 2017 May 23;1–8.

90. Montenegro CE, Patrinos HA. Returns to Schooling around the World. World Bank; 2013.

91. Roseboom T, de Rooij S, Painter R. The Dutch famine and its long-term consequences for adult health. Early Hum Dev. 2006 Aug;82(8):485–91.

92. Ong KK, Ahmed ML, Emmett PM, Preece MA, Dunger DB. Association between postnatal catch-up growth and obesity in childhood: prospective cohort study. BMJ. 2000 Apr 8;320(7240):967–71.

93. Vickers MH, Breier BH, Cutfield WS, Hofman PL, Gluckman PD. Fetal origins of hyperphagia, obesity, and hypertension and postnatal amplification by hypercaloric nutrition. Am J Physiol Endocrinol Metab. 2000 Jul;279(1):E83-87.

94. Lassi ZS, Das JK, Zahid G, Imdad A, Bhutta ZA. Impact of education and provision of complementary feeding on growth and morbidity in children less than 2 years of age in developing countries: a systematic review. BMC Public Health. 2013;13 Suppl 3:S13.

95. Gresham E, Byles JE, Bisquera A, Hure AJ. Effects of dietary interventions on neonatal and infant outcomes: a systematic review and meta-analysis. Am J Clin Nutr. 2014 Nov;100(5):1298–321.

96. Mayo-Wilson E, Junior JA, Imdad A, Dean S, Chan XHS, Chan ES, et al. Zinc supplementation for preventing mortality, morbidity, and growth failure in children aged 6 months to 12 years of age. Cochrane Database Syst Rev. 2014 May 15;(5):CD009384.

97. Imdad A, Bhutta ZA. Effect of preventive zinc supplementation on linear growth in children under 5 years of age in developing countries: a meta-analysis of studies for input to the lives saved tool. BMC Public Health. 2011 Apr 13;11(Suppl 3):S22.

98. Ramakrishnan U, Nguyen P, Martorell R. Effects of micronutrients on growth of children under 5 y of age: meta-analyses of single and multiple nutrient interventions. Am J Clin Nutr. 2009 Jan;89(1):191–203.

99. Ruel MT, Alderman H, Maternal and Child Nutrition Study Group. Nutrition-sensitive interventions and programmes: how can they help to accelerate progress in improving maternal and child nutrition? Lancet Lond Engl. 2013 Aug 10;382(9891):536–51.

100. Muanda FT, Chaabane S, Boukhris T, Santos F, Sheehy O, Perreault S, et al. Antimalarial drugs for preventing malaria during pregnancy and the risk of low birth weight: a systematic review and meta-analysis of randomized and quasi-randomized trials. BMC Med. 2015 Aug 14;13:193.

101. Roodman D. The impacts of incarceration on crime [Internet]. Open Philanthropy Project; 2017. Available from: http://files.openphilanthropy.org/files/Focus_Areas/Criminal_Justice_Reform/The_impacts_of_incarceration_on_crime_10.pdf