NBER WORKING PAPER SERIES

SAMPLE-SELECTION BIASES AND THE "INDUSTRIALIZATION PUZZLE"

Howard Bodenhorn
Timothy W. Guinnane
Thomas A. Mroz

Working Paper 21249
http://www.nber.org/papers/w21249

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2015

For comments and suggestions we thank Shameel Ahmad, Cihan Artunç, Gerard van den Berg, Claire Brennecke, Raymond Cohn, Thomas Cvrcek, Jeremy Edwards, James Fenske, Amanda Gregg, Farly Grubb, Sukjin Han, Brian A'Hearn, Philip Hoffmann, Sriya Iyer, John Komlos, John Murray, Sheilagh Ogilvie, Jonathan Pritchett, Paul Rhode, Mark Rosenzweig, Gabrielle Santangelo, Richard Steckel, Jochen Streb, William Sundstrom, Werner Troesken, James Trussell, Christopher Udry, Marianne Wannamaker, John Warner, David Weir, anonymous referees, and participants in seminars at the University of Michigan, the University of Nuremberg, Queen's University (Ontario), the Rhein-Westfälisches Wirtschaftsintitut, Tulane University, and the 2012 Cliometrics meetings. We thank Emilia Arcaleni, John Murray and Richard Steckel for sharing data. We acknowledge financial support from the Yale University Economic Growth Center. Meng Liu, Yiming Ma, and Adèle Rossouw provided excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2015 by Howard Bodenhorn, Timothy W. Guinnane, and Thomas A. Mroz. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Sample-selection biases and the "industrialization puzzle"
Howard Bodenhorn, Timothy W. Guinnane, and Thomas A. Mroz
NBER Working Paper No. 21249
June 2015
JEL No. J11,N11,N31

ABSTRACT

Understanding long-term changes in human well-being is central to understanding the consequences of economic development. An extensive anthropometric literature purports to show that heights in the United States declined between the 1830s and the 1890s, which is when the US economy industrialized and urbanized. Most research argues that declining heights reflects the impact of the industrialization process. This interpretation, however, relies on sources subject to selection bias. Changes in that selection mechanism may account for the declining heights. We show that the evidentiary basis of the puzzle is not as robust as previously believed. Our meta-analysis of more than 150 studies shows that declining-heights finding emerges primarily in selected samples. Finally, we offer a parsimonious diagnostic test for revealing (but not necessarily correcting for) selection bias. The diagnostic applied to four samples that underlay the industrialization puzzle shows compelling evidence of selection.

Howard Bodenhorn
John E. Walker Department of Economics
College of Business and Behavioral Science
201-B Sirrine Hall
Clemson University
Clemson, SC 29634
and NBER
bodenhorn@gmail.com

Timothy W. Guinnane
Department of Economics
Yale University
PO Box 208269
New Haven CT 06520
timothy.guinnane@yale.edu

Thomas A. Mroz
Andrew Young School of Policy Studies
Georgia State University
33 Gilmer St SE
Atlanta, GA 30302
tmroz@clemson.edu

An online appendix is available at:
http://www.nber.org/data-appendix/w21249

1. Introduction

Introduced into economic history just over three decades ago, anthropometric history is now one component of the cliometrician's standard toolkit. Roderick Floud et al (2011, 374 hereafter FFHH) note that anthropometric history originated with the limited objective of establishing the heights and health of the inhabitants of North America and Western Europe in the eighteenth century. Since then, the study of heights has "mushroomed into a study of the long-term development of human society," drawing on the tools and insights of economics, statistics, and medicine, among others. It is fair to say that no other branch of cliometric history has had as many resources devoted to it in the past two decades as historical anthropometrics (Voth and Leunig 1996, 541).

Anthropometricians have developed a systematic approach to the study of the connection between changes in the human body and economic development labeled the "techno-physical" approach, which is built on four premises (FFHH 2011). First, the typical body size and shape of a population reflect that generation's longevity and work capacity. Second, a generation's work capacity and the technology available to it determine its outputs. Third, a generation's outputs and, therefore, its physical wellbeing, is both an inheritance from previous generations and a legacy to generations to follow. Thus, fourth, a society can embark on a path of long-run economic growth only if it witnesses improvements in each generation's physical wellbeing, which it then passes to subsequent generations.[1]

Physiological improvement – greater mean stature, increases in mean body mass index toward modern norms – takes a central place in the list of factors that influence

[1] Deaton (2013) offers a related argument.

economic development in the long run. But, as FFHH (2011) note in their summary of the British and American experiences, nutritional wellbeing in these now-rich economies did not grow monotonically. Most of the available evidence from these two countries reveals that heights stagnated or declined in the early phase of modern economic growth. In the United States, in particular, mean height apparently declined for cohorts born between (approximately) the 1830s and the 1890s. This apparent anomaly, now widely labeled the "antebellum puzzle," is one of the most-studied issues in the field of economic history (FFHH 2010, 298). Similar substantial, long-term declines have also been identified for Great Britain, Sweden, and the Habsburg monarchy, although at different times (Floud, Wachter and Gregory 1990, Sandberg and Steckel 1987, Komlos 1989). Komlos (1994a, 493) calls the decline in heights in the early phase of economic modernization in these countries "the most amazing discovery" of anthropometric history. Studies have identified a number of potential causes, though most focus on the failure of food supplies (measured by quantity, quality or both) and public health, broadly conceived, to keep up with population growth and urbanization.[2] "Economic growth in the nineteenth century," write FFHH (2011, 348), "was very costly because economic booms caused rapid population growth, internal and external migration, urbanization, sanitation problems, and rampant diseases; all these reduced people's productivity and their ability to accumulate human capital."

The apparent decline in heights in the United States, Great Britain, Sweden and Habsburg-era central Europe is indeed interesting. But we doubt the evidence adduced

[2] Komlos (2012 working paper) also provides a summary of the literature and identifies at least a dozen different explanations for the puzzle.

for this apparent decline. These countries had very different economies at the time of the height reversal. But they did share an important feature: they filled their military ranks with volunteers rather than conscripts. Thus the samples height scholars rely on for the reversals are selected, in the sense that they contain only individuals who decided to join the army. Elsewhere we have shown that the sample-selection problems in inferring population heights from a group of volunteers can be grave (Bodenhorn, Guinnane and Mroz 2014). The implications of selection bias make the observed "shrinking in a growing economy" less of an anomaly (Komlos 1998a). As the economy grew, the outside option of military service became less attractive, especially to the productive and the tall. Military heights declined because tall people disproportionately chose non-military employment. Thus, we cannot really say whether population heights declined; perhaps only the heights of those willing to enlist in the military declined. Below we draw on published heights studies to document an important feature of the data: height reversals or "puzzles" emerge only rarely in countries that filled their ranks through (near universal) conscription.

Sample selection can take two forms that the heights literature sometimes confuses. Exogenous selection pertains to sampling on an observable, exogenous characteristic, such as race or gender. Modern surveys often deliberately over-sample on such characteristics. If we know the proportions of such groups in the population, then we can construct and apply weights to obtain estimates that reflect the population's characteristics. Endogenous selection reflects a situation where an individual enters the sample in part because of unmeasured characteristics that also are related to outcomes of interest. That is, a soldier in a volunteer army (for example) made a decision based on his

individual unobserved characteristics. We cannot use simple sampling weights to overcome endogenous selection because the required weights would be individual-specific and depend upon unmeasured characteristics.

Endogenous selection likely afflicts many of the samples cliometricians use; the problem is not limited to the antebellum puzzle. But we focus on that issue. Two points warrant emphasis. First, endogenous selection typically produces samples that have a disproportionate number of people sharing some observable characteristic. But this fact reflects both choices made by individuals and constraints on data availability. The endogenous selection depends on both observable and unobservable characteristics. Unless we are willing to make heroic assumptions, we cannot adjust the selected sample to reflect the population through re-weighting on observed characteristics alone. Second, the mechanism underlying endogenous sampling can change over time, as succeeding cohorts differ and as the opportunities and constraints they face change. Because of the selection mechanism, army recruits in a given year will differ from the population; and in different years they will differ from the true population in a different ways. Researchers cannot assume that biases created by endogenous selection "average out" over time.

Some cliometricians contend that the issues surrounding selection and selection bias are now well understood and that most researchers account for (or at least qualify their conclusions based on) any potential selection bias issues. We disagree. Sample selection problems of the endogenous type tend to be minimized when they are discussed at all.[3] Others make a virtue of the fact that many samples over-represent the poor and

[3] A recent study of Portuguese heights is typical. After a consideration of declining minimum height requirements from 62 *polegadas* (Portuguese inches) between 1763 and 1774 to 56.6 *polegadas* (post-

working classes (Carson 2006, 2008, 2009). But this argument both confuses exogenous with endogenous selection and misses the consequence of endogenous selection. Military and prison samples, for example, over-represent the poor and working classes not because the poor were randomly over-sampled to achieve that result, but because poor and working-class men have unobserved characteristics that made them more likely to find soldiering or crime to be their best option at that moment. Equally important, the poor men who entered the military or the prison are not a random sample of poor men; as we argue below, there is good reason to believe that even for those of a given class background, men in the army are shorter than those who did not join. And this endogenous selection likely changes over time.

Our argument is not what the anthropometricians have labeled the "basketball problem," which "refers to the possibility that the military [or any self-selected group] is drawn from a normal distribution that is not typical of the entire male distribution" because basketball rewards height (Sandberg and Steckel 1987, 103). Endogenous selection is not the basketball problem per se because it is usually presented as one of invariant height-based selection: the mean height of basketball players is presumed to be greater than the mean height of the adult male population by an amount that remains (approximately) constant over time. But as competitive strategies in professional

---

1886), Stolz et al (2013, p. 548-550) note that military recruitment was "extremely unpopular" and that "informal protection networks enabled many ordinary persons as well as many of those in the upper social strata to avoid recruitment altogether." Despite the ability of the rich and the ordinary to avoid recruitment, the authors contend that their sample, which spans birth cohorts between 1730 and 1820, provides a "relatively unbiased sample of recruitment-aged males."

basketball have changed over time, the average height of world-class basketball players has changed. This is an example of endogenous selection.[4]

We are not the first to recognize the problem of endogenous selection in the historical heights literature. We may be the first to systematically explore whether it can explain the most important anomalies in the literature, the most notable being the apparent reversals in average heights observed in the eighteenth and nineteenth centuries and labeled the "antebellum puzzle" in the case of the nineteenth-century US or the "industrialization puzzle" elsewhere.[5] Our discussion proceeds as follows. Section 2

[4] In their study of US baseball players, Saint Onge et al (2008, 487) discuss something close to this when they write: "Changes in height may also be related to different strategies for playing the game over time, and thus, selection for different kinds of players."

[5] In his reanalysis of boys recruited by the Marine Society in eighteenth-century UK, Komlos (1993, 119) argued that Floud, Wachter and Gregory (1990) reported changes in heights that were "unreasonably abrupt and unacceptably large." Komlos attributed the abrupt changes to changing recruitment practices and changes in the types of boys offered for recruitment. Similarly, Johnson and Nicholas (1995) claim that a 2.25-inch decline in the mean height of young adult males entering the military between 1822 and 1847 is too large to be consistent with the nutritional insults suffered by these cohorts. Johnson and Nicholas (1995, 471) attribute some part of that decline to selection bias in military samples. For early statements of the puzzle, see Komlos (1996; 1998a). Using a state of the art, two-step semi-parametric estimator for choice based samples with possible selection on a single event (i.e., in sample vs. not in sample), Zimran (2015) reports that selection issues are important when one examines height data based derived from U.S. military records in the mid1800s. The selection biases he uncovers with this approach are not severe enough to completely overturn the apparent downward trend in heights over this time period. Selection into his sample, however, depends on both a model of occupational choice

describes the puzzle and its empirical basis in more detail. Section 3 reports a meta-analysis of the height literature, documenting the striking fact that the puzzle appears most often in studies that rely on selected or small samples. Section 4 describes a simple economic model of endogenous selection and discusses the few heights studies that appear to have recognized the issue in the past. Section 5 outlines the problem of testing for selection bias in selected samples, and section 6 reports tests for selection bias in a collection of influential U.S. and British sources.

2. The antebellum puzzle

Margo and Steckel (1983) and Fogel (1986) report an anomaly; the heights of adult male native-born Americans began to decline among those born in the 1830s.[6] The downward trend persisted until the 1870s or 1880s and was not reversed until the last decade of the century, after which adult male stature increased at the relatively rapid rate of 1.8cm per decade between 1902 and 1931 (Fogel 1986, 511). The pattern of declining

---

during one's early adult years and a selection model of how one can link successfully these military records to micro-level information from the US. Censuses. As Mroz (2015) demonstrates, a two-step estimator using information from just a single compound event when the underlying selection mechanism depends on multiple underlying events can seldom control adequately for selection biases and can actually exacerbate the severity of the selection biases. Zimran's (2105) study clearly indicates that there are potentially severe sample selection issues.

[6] Throughout the paper, references to dates when discussing heights will pertain to birth year, not observation year unless otherwise noted.

heights at mid-century is "puzzling because according to conventional indicators the American economy was expanding rapidly during the antebellum decades" (Komlos 1987, 898). Ordinarily one expects periods of economic growth to witness rising living standards.

The antebellum puzzle is all the more puzzling because, among the early industrializers, only England, Sweden and Austria-Hungary appear to have experienced a puzzle. Figure 1 reproduces Fogel's (1986) graph of US adult male heights. The data underlying this graph form the foundation of what has come to be labeled the "antebellum puzzle." We superimpose on Fogel's graph the mean heights reported for several other countries. The heights of the Dutch, Swedes, Italians, and French traced out long secular growth paths that, while country-specific, demonstrated no sharp reversals. The heights of Russians, Bulgarians, and Japanese (not shown) similarly all increased without reversal between the mid-nineteenth and early twentieth century. Only three other countries – Britain, Sweden and Austria-Hungary -- exhibit a US-like pattern of declining heights during early industrialization. Komlos (1998b, 236; 1998a; 2012) reviews the puzzle literature and concludes that the "pattern has been found repeatedly .... [and] is surely not a statistical artefact [sic]." A similar pattern of height decline appears in samples of military academy students, prisoners, and manumitted slaves, among others. Given its appearance in multiple samples, attention turned to uncovering its sources. Explanations for the puzzle have focused on declines in available foodstuffs (or a rise in their relative price) that led to a long-run decline in net nutrition; to increases in the disease load due to urbanization and a widening transportation network; to increased income inequality, which negatively affected the height of the lower classes more than

secular growth positively affected the height of the middle and upper classes; to increased work intensity, which also would have more negatively affected the height of factory workers relative to farmers (who already worked hard) and white-collar workers; and to increased immigration around mid-century, which would reduce mean adult height if immigrants were shorter than native-born Americans and if the children and grandchildren of immigrants carried the immigrant height disadvantage across generations.[7] The puzzle literature has become part of the continuing debate between "pessimists" who argue that industrialization diminished aggregate well-being in the short run and "optimists" who think the opposite (Feinstein 1998). The finding of declining average height supports the pessimist's case.

2.1 The empirical basis of the "puzzle"

Although broadly accepted as a fact in the heights literature, the puzzle has a surprisingly limited evidentiary basis. In an early summary, Komlos (1998a, 782) discussed 21 heights samples, 14 of which report declining mean height. In three other samples, mean height increased, and one implies no change in mean height. Komlos also reports the approximate turning point in mean height, with dates ranging between the 1780s and the 1840s. Excluding the three studies that report increasing height, the modal downturn begins in the 1830s; the median downturn occurs in the 1820s (Komlos 1998b, 236). Since Komlos's 1998 summary, more than a dozen articles have appeared that document or discuss the puzzle. We collected these studies to summarize results from 36

[7] See Sunder (2004, 76-77) and Komlos (2012) for the details of these and other explanations that have been advanced in the literature..

samples that span all or part of the puzzle era (1830s-1880s). These 19 different studies do not constitute a census of puzzle articles, but they span the range of sources this literature uses, including volunteer armies, students, prisoners, as well as free, enslaved, and emancipated African-Americans.[8] Table 1 summarizes the sources and their results. We list author and publication year, the decades covered, the measured group, the mean height change among that group, and which of the five broad explanations for the puzzle that the authors attribute to the height change when one is invoked.

Even an unsophisticated summary reveals that, taken at face value, the published studies of the heights puzzle do not present an overwhelming case. Column 5 of Table 1 reports the change in mean height reported for each sample. The mean change is -0.44%; the median estimate is -0.5% and the modal estimate is -0.8%. The mean estimate is consistent with a decline of about three-quarters of centimeter if we assume the mean height before the onset of the decline was 172cm.[9] While a 0.75cm decline is nontrivial, it is not consistent with a mid-century nutritional crisis.

This overall estimate may mask more subtle patterns, so we probed a bit further. One concern is that the 36 samples span different decades, so we calculate the average

[8] The studies we summarize sometimes report results from new data, and sometimes rework earlier material. Four, for example, rely wholly or in part on the Union Army samples. But no two samples use the identical data. Haines' (1998) analysis of the Union Army data uses only soldier who enlisted in New York; A'Hearn's (1998) analysis subdivides the sample by occupation and place of residence and we include three of Komlos's (1998) analyses of black soldiers because he uses three alternative methods of dealing with left-tail truncation. All these studies use different methods and report results from different samples than that used in Margo and Steckel's (1982) original analysis.

[9] This 172cm estimate is approximately the mean reported by Fogel (1986, 511) for the 1720s through the 1790s

decline in height per decade. The result is a change in mean height of -0.08% per decade. But this estimate may also mislead, because the puzzle is considered a post-1830 phenomenon. So we recalculate the change assuming that the entire reported change in mean height occurred in whatever decades the sample spans between the 1830s and the 1880s. This calculation yields a much larger change in mean height of -0.23% per decade. One might also worry about the different sample sizes, which range between 454 and 41,173 measured individuals. When we weight the mean change in heights by sample size, the decline for all decades is -0.10% per decade, or -0.14% (=-0.24cm) per decade if we assume that entire decline occurred between the 1830s and 1880s.[10] To sum up: the weighted mean of the three dozen estimates suggests that height declined by 1.48cm (=0.58 inches) over six decades. This nontrivial decline is modest, even if true. The remainder of the current paper explains why we think heights probably did not decline even by this small amount and that the puzzle is, in fact, an artifact.

3. A meta analysis of the historical heights literature

Our analysis of about three dozen antebellum puzzle studies casts some doubts on whether the mid-century decline in heights was a real phenomenon. But these studies alone cannot resolve whether the observed decline in height was a consequence of the selected nature of the samples. Ideally, we would need to compare selected samples to non-selected samples drawn from the same populations. That is not possible, for obvious reasons. Figure 1 does the next-best thing: it compares results from volunteer and

[10] The standard deviation (error) of the estimate is 0.02. If we assume that the sampling distribution of the estimated sample-size weighted per decade percentage changes is normally distributed, the 95% confidence intervals for the 0.14% estimate are bounded by -0.18% and -0.10%.

conscript armies for different countries. Universal conscription, a system under which nearly all young men eligible for military service were called for medical examination to determine their fitness, was common across Europe during the nineteenth and twentieth century. In theory, all men of a certain age were eligible to serve; in practice, of course, some young men were able to avoid service for various reasons. Unless a significant fraction of the eligible young men were able to avoid and that ability was correlated, either positively or negatively, with the young men's height, conscript samples should provide good estimates of the true mean height-at-age of each cohort.

Conscript data have additional appealing features. One advantage is that because men were typically examined at a given age, an age that did not change much over time, the data create snapshots of, say, 20-year olds across several decades. This feature makes inter-temporal comparisons relatively straightforward. A second advantage is that many countries called young men when they were near their terminal adult height. Third, because the samples are self-weighted, the econometrician does not have to rely on a census to determine the sampling weights needed to reconstruct a population estimate.

Figure 1 reports just a few examples. We expanded the range of examples by drawing on approximately 150 historical heights studies. (The appendix describes our strategy in identifying these studies.)[11] We classify the source according to type of sample (conscript, volunteer, prisoner, students, passport applicant, etc.), the time period

[11] Briefly stated, we used the following procedure to identify and classify articles for inclusion: (1) search for studies that use conscripts; (2) search for studies that use volunteers; (3) identify nonmilitary historical height articles in the principal outlets (i.e., economic history journals, collected essay volumes, and human biology journals) published between 1995 and 2014; (4) read nearly all of the articles identified in steps one and two and read a selection of articles from step three until we reached a total of 150 articles. The resulting sample includes 169 samples drawn from 144 separate sources (some articles, books or chapters report more than one independent sample). Our sample of studies includes 50 conscript samples, 39 volunteers and 80 "other." It also includes a wide range of countries, though it is over-weighted with North America and western Europe, which is also consistent with the larger literature. We also include some pre-1995 studies when, in reading the more recent literature, they seemed particularly salient to our project.

covered, mean height in the first year the series is observed, mean height in the last year the series is observed, and whether the study reported a significant height reversal during any decade. A "reversal" here means a decline in heights that breaks a longer-term increase in heights. We coded a reversal as significant if during any decade mean height declined by one centimeter or more. [12]

The database reports results from 167 samples (some papers report more than one sample). Fifty of these rely on conscripts. The meta analysis includes a fair representation of the other sources commonly used in the historical heights literature: 39 samples of volunteer soldiers; 14 samples of students, including students at military academies; 23 samples of prisoners; and 41 samples that we classify as under the heading of Other, which includes an eclectic mix of runaway and manumitted slaves, indentured servants, immigrant workers, voters, passport applicants, government employees, enrollees in health and life insurance programs, unclaimed Korean corpses, and American baseball players, among others. Only a few of the non-conscript samples are national in scope. Some are very small; one has only 151 observations over 34 years, though the median number of observations in the "Other" group is nearly 3,900 spread over a median period of 40 years. The database includes samples from 40 different countries, including 36 entries for the United States. Table 2 reports the proportion of samples that document a height reversal (a decline of 1cm or more) in any decade (Column 1) or a decline across

[12] Our 1cm standard for identifying a reversal provides a relatively low threshold, but one that still occurs so irregularly that we are not falling into what Komlos (1993a, 130) notes is the trap of attaching "too much significance to slight deviations from the main trend" in heights. Baten (2009, 172) and Baten and Komlos (1998) contend that a 1-1.2cm change in a decade is a biologically (and economically) significant change in mean male adult height.

two or more decades (Column 2). Twenty percent of studies based on conscripts identify at least one reversal across an average sample period of five decades (median of six decades). The other 80% of conscript samples show no evidence of a substantive height reversal in any decade. Just four percent of conscript studies identify at least two puzzle-like reversals. Compare this to the 56% of volunteer samples that identify one reversal over an average sample period of four decades (median of three decades). One-third of volunteer samples identify two or more reversals. More than one-half of studies involving students identify one reversal; 70% of prisoner studies identify a reversal; and, one-third of the catch-all "Other" samples imply reversals.

This meta-analysis shows that the phenomenon associated with the industrialization puzzle appears most often in samples subject to selection bias. We sharpen this conclusion with the assistance of a series of probit regressions in which the dependent variable equals one if the study reports at least one reversal (or, alternatively, two or more reversals), and zero otherwise. This approach allows us to control for other effects. Table 3 reports the marginal effects from two specifications. In each case the difference is relative to a conscript sample. Column 1 provides summary statistics for each variable coded from our reading of each study. The sample has approximately equal representation of conscripts, volunteers, students/prisoners, and others. The studies differ in that some cover less than two complete decades; others span a century or more. They also vary widely in sample size, from less than 200 to more than 10 million. And about two-thirds of the volunteer sample results were reported after making a correction for left-tail truncation.

Columns 2 and 3 report marginal effects for a given type of data on the presence of a reversal (or two). The probability of observing at least one 1cm or more decline in average height in at least one decade is 44 percentage points greater for a sample of volunteer soldiers and 31 percentage points greater for prison samples. The estimated probability was notably higher for samples of students, though the estimates are imprecise. The marginal probability of observing a decline was greater for samples that spanned longer periods, but lower for those that are nationally representative. The puzzle, however, is not about a short-term decline in heights that may in occur in a single decade, but rather a longer term phenomenon. Thus, in Column 3 we estimate the marginal probabilities of observing at least two (not necessarily consecutive) decades of declining mean heights. The results are broadly consistent with the one-decade decline results: samples involving volunteer soldiers and prisoners are more likely to exhibit two reversals; the marginal probabilities are higher for samples that encompass more years and smaller for larger samples.

Although this analysis includes studies far removed from the period that spans the puzzle, it provides an important insight into the literature. The United States did not have military conscription until the twentieth century.[13] Yet that history makes the cliometrician's work more difficult, because U.S. height data before the twentieth century comes almost exclusively from selected samples. Thus the basic constraints of U.S. sources make definitive statements about the antebellum puzzle tenuous.

4. Endogenous selection

[13] Except for a brief period during the Civil War. Most men could avoid that draft (A'Hearn 1998).

Endogenous selection arises when individuals appear in a sample only because they (or possibly someone else) make decisions that reflect unmeasured individual constraints and preferences that are related to key outcomes of interest. The most general case is soldiers joining a volunteer army, but the relevant decision can also be made by someone other than the person whose height is measured: runaway or transported slaves, for example. To make this problem concrete, we summarize the Roy-style model described in detail in Bodenhorn, Guinnane and Mroz 2014. The Roy (1951) model is a workhorse tool in labor economics and other areas, applied to situations where individuals make a binary choice.[14] The model assumes that each individual decides whether to work in the civilian or military sector (that is, join the army), and only the latter appear in the height sample. Each individual has a prospective civilian and military wage; at least one of these wages reflects, in part, their height.

Wages could be correlated with height under two different interpretations. The first is that the army or some civilian occupations might reward height itself. Promotion might come faster to taller soldiers because of their height, and in some military and civilian occupations a person might be more productive because he is tall. A second interpretation seems more consistent with the basic tenets of the heights literature. Height is correlated with a person's (mostly unobserved) health human capital and thus their productivity (Schultz 2002). The model also assumes that each person receives

[14] Our model draws on Roy (1951), and resembles Heckman and Sedlacek's (1985) two-sector occupational choice model.

individual-specific shocks to their civilian and military wages, and has individual–specific preferences over civilian versus military life.[15]

The decision to join the army reflects individual shocks and preferences, as well as individual height and the return to height. The distribution of heights for men who join the army can then be written as the product of the population height distribution (say, f(h)) and a second expression (Z(h)) that summarizes the decisions made by individuals of different heights. Z(h) is a function of all model parameters, most importantly, the return to height or its unobserved correlates in the civilian sector (β_C) and the military sector (β_M). If $\beta_C > \beta_M$ then the heights of volunteer soldiers will understate the population height. We assume that in general the military rewards individual traits such as height less than the civilian sector. The degree of height understatement in the selected sample depends critically on the relative sizes of β_C and β_M (as well as the other parameters); as the parameters change over time (or space), the size of the selection bias will change. The simulations reported in Bodenhorn, Guinnane and Mroz (2014) demonstrate that small variations in the incentives to join the army can generate empirically important variations in the height of the resulting sample, even when we hold constant the heights of the underlying population.

The model also suggests that a sample selected endogenously can look like it was selected exogenously. Suppose that a sample has proportionally more common laborers than are known to exist in the population. This might reflect exogenous selection: the sampling procedure took a random sample of workers, but for some reason over-

[15] The simulations reported in BGM use parameter values that imply that height only "explains" a small proportion of an individual's height.

weighted the laborers. If so, then it is simple to re-weight the sample to obtain correct population estimates. More likely, the over-representation of laborers reflects the fact that laborers were more likely to find their best option in the military. If this is the case then we cannot simply re-weight the sample to obtain population estimates, because the correct weights would be individual-specific and are unknown. The necessary weights would need to reflect the individual's height in addition to their productivity and taste parameters. There is a second, related problem. If we had too many laborers because of careful exogenous sampling, then the sample height of laborers would be a good estimate of the heights of laborers in the population. Such is not the case for endogenous selection. The Roy model implies that the laborers who joined the army would be on average shorter than the laborers who did not, with the degree of shortfall varying directly with civilian-sector opportunities.

The heights literature contains numerous examples of studies that suffer from some type of endogenous selection. Komlos (1994a), for example, compares trends in the heights of upper-class French students and French conscripts. His sample of students enrolling in the Ecole Polytechnique between the 1780s and 1860s shows sharp declines in the 1820s (Parisian students) and 1830s (provincial students). He argues that the student sample reflects a type of exogenous selection, and interprets his results as showing a height reversal even for the comparatively well-off portions of French society. The French series in our Figure 1 come from Weir (1997), who shows that the heights of French conscripts rose continuously throughout the nineteenth century. Komlos' findings more likely reflect changes in the endogenous selection of French university students. Komlos's sample of 18-year old students are always taller than the general population,

but over time they become relatively shorter compared to the population of 20-year old French men called up for military medical examination. This is one of the few cases where data exist to compare a selected sample to one that is more nationally representative. The selected sample's misrepresentation of the population trends closely tracks the predictions of the Roy model. Conclusions drawn on the basis of the selected, student sample would lead to incorrect conclusions about trends in France's economy or the well-being of French men.[16]

Volunteer soldiers, militia men, National Guardsmen, prisoners, runaway servants or manumitted slaves: all were measured only because of something they or someone else did. We are not the first to recognize the problem of endogenous, as opposed to exogenous, selection. Brennan, McDonald and Shlomowitz (1994a; 1994b; 1997) raise this issue in their study of Indian workers who migrated to Fiji. They acknowledge that temporal changes in the mean observed height may be driven, at least in part, by changes in the underlying demand and supply of eligible migrants due to changes in macroeconomic conditions. Similarly, in his discussion of the declining heights of London boys recruited into the Marine Society in the eighteenth and nineteenth centuries, Komlos (1993, p. 122) discusses what he labels the "offer curve" of recruits, which may

[16] Komlos (2008) compares U.S. military volunteers born in 1940 and 1950 to random draws from the U.S. population. The relevant enlistment period spans the height of the Vietnam War. The military sample shows a decline and then recovery in average heights not reflected in the random sample. These differences between population heights and volunteer soldier heights cast doubt on any conclusions that may be drawn considering only the heights of volunteer soldiers, a fact that Komlos (2008, 447) endorses, when he writes: "these [military] data have their own limitations insofar as they pertain to those who selected into the US military at moderate wages."

have shifted over time in response to changing economic conditions. He does not, however, push this idea. Ricardo Salvatore (1998, 105), in describing selection issues surrounding his sample of Argentine soldiers between the 1780s and 1830s, writes that the shift from a (mostly) volunteer army in the 1810-1829 period to a (mostly) mercenary army in the 1850-1865 period "made the quality of recruits more dependent on labor market conditions," which he acknowledges may have been responsible for some of the observed changes in average height over time. In the end, however, Salvatore dismisses the potential importance of this process. In their study of English prisoners transported to New South Wales, Nicholas and Steckel (1991, 949) raise the possibility of endogenous selection (they labeled it a "period effect"), which they defined as the possibility that crime became more concentrated among "poorer and shorter men" over time.[17]

The clearest earlier discussion of endogenous selection appears in Mokyr and O'Grada's (1994; 1996) two articles investigating the heights of Irish recruits into the Royal Navy and the English East India Company. They find that Irish recruits were (surprisingly) taller than English recruits and attribute the finding to relative Irish poverty. A man of given height in Ireland had fewer non-military opportunities than a similarly situated Englishman of equal stature and was, therefore, more inclined to join the military. (In the Roy model, the finding is consistent with the idea that the Irish and

[17] Nicholas and Steckel conclude this was not the case because their Jarque-Bera tests failed to reject normality for those transported before and after 1833. Bodenhorn, Guinnane and Mroz (2014) show that the standard tests for normality are unlikely to reveal changes in height-based selection, because the tests have very low power so that even samples known to be selected can appear normal. Baten et al (2012) also allude to something like endogenous selection.

English have identical population height distributions, but that $\beta_C - \beta_M$ was smaller in Ireland than in England.) Remarkably, many heights papers report Mokyr and O'Grada's "tall but poor Irish" result *without* any discussion of the fact that they attribute the apparent anomaly to differential selection on height (see, for example, Nicholas and Steckel 1991; A'Hearn 1998; Salvatore 2004; Morgan 2004; Canfield and Inwood 2011; Riggs and Cuff 2013).

Gallman (1996, 194), too, raises concerns with both exogenous and endogenous selection in his critique of Komlos's (1987) study of West Point cadets. West Point cadets are interesting only if they can be taken to represent "a random sample of some larger, more interesting group – say all young white men in the United States." But they are not. Not everyone was eligible for West Point, and those who applied were primarily interested in pursuing a military or engineering career. Gallman notes that the sample might still be useful if the pool of candidates from which the cadets were drawn "retained an unchanging character." He suspected they were not. The fact that the post-Civil War army was full of former brigadiers, colonels, majors and captains with battlefield and command experience "surely caused more than a few potential soldiers to seek other careers" (Gallman 1996, 194-195).

5. Selection bias and identification

Few historical data sets were collected as random samples from the populations of birth cohorts, and for some important countries such as the U.S., we have none at all. Can we estimate trends in historical heights from biased samples? In the simplest of all possible cases, exemplified by the classic description of enlistment in the military (e.g.,

Trussel and Wachter, 1982), researchers would observe only a random sample of the population for those whose height exceeded some minimum height requirement. Three prominent methods have been discussed in the literature to account for such simple sample selection issues: reduced sample maximum likelihood estimation or truncated regression (Wachter and Trussel 1982), quantile bend (Wachter 1981), and Komlos-Kim regression (1990). Unfortunately, the methods devised to deal with this simple selection mechanism must assume that any height-based selection implied by the Roy model not take place. That is, all of these approaches require that an individual's propensity to enter the military or some other sample be independent of height (above the minimum height threshold, if there is one). The Roy model implies that we should expect this assumption to be violated, so long as individuals from better socioeconomic backgrounds or who have better civilian labor market opportunities are less likely to enlist in the military. Uncovering temporal trends in mean height requires the anthropometrician to use multiple birth cohort samples, where changing economic and military conditions could lead to across cohort samples exhibiting non-constant degrees of selection biases.

To address the general problem, we separate the height distribution in a birth-cohort from the process of selection into an observed sample. The former is the object of interest, and the latter is the source of possible selection. The distribution of observed heights is a convolution of these two functions: the distribution of observed (sampled) heights depends on both the parameters of the birth-cohort height distribution and the parameters of the selection function. Consider a simple world where individuals born at date b grow until just before they turn age w. At date b+w, all have reached their height potential. Immediately when they turn age w they make a decision that affects whether or

not they enter the observable sample.[18] The distribution of observed height in this example depends on two distinct sets of factors: those that affect growth, dated from time period b to time period b+w-1, and those that influence the decision to join the sample, observed at date b+w. The distribution of height for the entire birth cohort b depends only on the collection of factors that affect growth from date b to b+w-1. We denote these factors by e(b). Let $v(t_b)$ denote the factors that could affect the chance that a birth cohort's member is in the observable sample. The mean height for birth cohort b would then be a function of e(b) alone, while the mean height for the those selected into the sample would depend on both e(b) and $v(t_b)$. Denote the expected height of birth cohort b by $E_b^T(H)=f(e(b))$. Then let the expected height among those selected into the observable sample be $E_b^0(H) = h(g(e(b), v(t_b)) = f(e(b)) + g(e(b), v(t_b))$. This expression equals the true height of the birth cohort plus a bias term represented by g(). One can obtain a consistent estimator of the mean height in the selected sample. Without further assumptions, however, it is not possible to decompose this estimate into the part of the observable mean height that is due to the cohort's growth environment (which is what we want) and the part due to the selection into the observed sample (which is just bias).

Now consider a comparison of individuals born in different years: those born in year b (as above) and those born in year c. Can we reliably estimate the trend in heights from selected samples? The mean height of those born in year c and of those born in year c who were selected into observable sample at date t_c are $E_c^T(H)=f(e(c))$ and $E_c^0(H) = h(g(e(c), v(t_c)) = f(e(c)) + g(e(c), v(t_c))$. To assess changes in heights over time we want

[18] There could be minimum or maximum height requirements. These are assumed built into the selection function.

estimates of f(e(c))-f(e(b)). This expression relies on random samples from the two populations which we do not typically have. Thus most studies use the expected observed heights to proxy for the change in height across birth cohorts: $E_c^0(H) - E_b^0(H) = [f(e(c)) - f(e(b))] + [g(e(c), v(t_C)) - (g(e(b), v(t_b))]$. The second term in square braces captures the bias from using the observed heights.

We observe only a single difference in expected mean height. But we need to de-convolute this difference into two changes: the change in the cohort-specific height and the change in the difference of the bias terms. This is a classic identification problem. There are three possible ways to identify the difference. First, if we had true random samples from each of the birth cohorts, then each component of the bias would be zero. Second, we could assume that the biases in each cohort are identical. This requires, however, the unlikely claim that selection does not depend on any temporal changes in macroeconomic or environmental conditions between dates b+w and c+w.

Komlos and Kim (1990) propose a third approach that does not yield consistent estimators of the difference in mean heights, but allows one to infer the sign of the differences in mean height. Their approach, however, requires that any height-based selection involves only the very tall or very short. Their approach also requires that the considerations that led men in two successive cohorts to join the army not change over time. The effect of labor-market conditions on the decision to join the army, for example, must be identical across cohorts. These are strong assumptions.

5.1 An indirect test of selection bias

We cannot estimate the true height from a selected sample, but we can nonetheless detect the presence of selection itself. Suppose we have two different dates when members of a birth cohort could be selected into a sample. These could be two ages at which an individual can enlist. Call these ages t_b and $t_b + 1$. In the absence of height-based selection, the distribution of height for those born in cohort b and "observed" at date t_b should be identical to those born in the same year and "observed" at date $t_b + 1$. If there is no selection, then an individual from a given cohort should not be shorter or taller because he joined at one date rather than another. A rejection of the null hypothesis of equal mean height in different "observation" years for the same birth cohort, provided all members are sufficiently old to have reached full height, would be evidence of height-based selection. A simple regression model can easily carry out this test.

Evidence of this type of height-based selection, however, still faces an identification problem. This is due to the classic perfect collinearity of cohort, age, and calendar time. What could appear to be selection biases due to calendar year effects from tests described above could instead be described by just age and cohort effects. Suppose, for example, that 21 year olds from any birth cohort who join the army are always on average .5 cm taller than 20 year-olds from the same cohort who join the army. Then the difference in the average height between birth cohorts, with or without adjusting for age dummy variables, would reflect true differences in mean height in the birth cohorts' population mean height provided the relative fractions enlisting at ages 20 and 21 are the same across cohorts. The test described above, however, would indicate biases due to height-based selection.

This example demonstrates that not all samples exhibiting height-based selection will provide biased estimates of how mean height varies in the population across birth cohorts. The cliometrician could always make the implausible argument that there are no calendar year effects indicating age-based selection, and instead attribute all such variations about birth cohort specific means to age effects. This argument becomes implausible if we compare cross-cohort differences in the posited age-specific mean height differentials. That is, suppose the difference in height between age 20 and age 21 for birth cohort b is different than it is for birth cohort (b+1). Then then one would need to make a compelling argument for why focusing on height differentials at one particular age, say age 21 rather than age 20, would capture the true difference in mean population height across birth cohorts. One might consider taking some average of the age-specific differentials to measure the birth cohort height differential. Positing that any selection biases over time would "cancel out" without specifying the reasons for the across age differences, however, would be unwise. Variations in age specific height differentials across cohorts need not be due to cohort specific differences in selection on unobservables, but the presumption should be that they do unless one has specific knowledge about why they vary.

A simple way to address much of the above discussion would be to assume that a random sample from each birth cohort is required at each calendar date at which the cohort could enter an observed sample. This is in itself a strong assumption, and any violation of the assumption would become exacerbated as a cohort ages. In military samples, for example, height is typically measured at age of enlistment; entering the military at age 20 typically precludes one from being in the population at risk to enter the

military at age 21 or later. Consequently, if there were any height-dependent selection into a sample at younger ages, then this would affect the at risk population's distribution of height at older ages.

5.2 Recruitment-year effects in the heights literature

Most heights studies use cohort dummies or similar controls. The justification for including birth cohorts is well understood, and is cogently articulated by Mironov (1999, 3-4). The forgoing discussion, however, suggests the inclusion of a set of recruitment-year (or observation-year) effects. This has been much less common in the literature. Brennan, McDonald and Shlomowitz (1997, 199) recognize that the inclusion of recruitment-year effects will capture "varying recruitment conditions …. [and] is the preferred specification because it ... allow[s] for variation in the recruiting environment" (Brennan, McDonald, and Shlomowitz 1997, Table 10, 200).[19] In his study of early-modern French volunteers, Komlos (2003a, 167) reports the results of some preliminary regressions that include *only* enlistment-decade effects; he finds statistically significant coefficients of relatively large magnitude (-0.88cm to 2.52cm). But because he observes an inconsistent pattern between youth (less than 23 years) and adults (23 to 49 years) during a given decade he is reluctant to attribute the estimated effects to either changes in measurement techniques or to changes in the supply or demand for height. "While it is imaginable," Komlos (2003a, 167) writes, "that somewhat taller men entered the army during economic downturns, the inconsistencies across age groups lead us not to attribute

[19] See also, Brennan, McDonald and Shlomowitz (1997, Table 12 and 14) for instances in which the inclusion of recruitment-year effects reduces the magnitude and significance of birth-year effects for Indian recruits emigrating to Mauritius.

much significance to this result." His finding strongly suggests a form of endogenous selection.

Only a handful of heights studies control for both birth-year and recruitment-year effects. They include some research on Union Army soldiers (Margo and Steckel 1983; A'Hearn 1998; Haines 1998). In each paper, the inclusion of recruitment-year effects reduces the magnitude and statistical significance of the birth-year effects. A'Hearn (1998) notes that the recruitment-year effects reflect selection.

6. Testing for selection in U.S. sources

In this section we apply the selection diagnostic described above to some of the height samples that form the backbone of the antebellum puzzle. The puzzle first appeared in volunteer military samples and as Komlos (1996, 2020) notes, "subsequent research has reproduced these results many times over: among the free blacks of Maryland [and Virginia], among Georgia convicts, … among Amherst students … and among Pennsylvanian [Union Army] soldiers."[20] We use these, or closely related samples to investigate whether these samples exhibit selection and whether selection might be the source of the observed decline in heights in the mid-nineteenth-century United States and Britain.

The discussion above suggests a simple regression-based diagnostic for the presence of height-based selection. Suppose that individuals who are born in the same year are of different full heights depending on the age at which they join the army. Such a

[20] Haines (1998, 156) recites the same list of studies that offer evidence in support of the puzzle, almost verbatim. He then argues that the trend in height of New York recruits into the Union Army exhibit a pattern consistent with the puzzle.

result would imply that the labor-market and other forces obtaining when an individual makes this decision act as the Roy model implies: the army sample would exhibit height-based selection. We use this idea to test for the presence of height-based selection in two different ways. We first ask whether age at recruitment appears to affect the heights of men born in a given year. This test can yield evidence of selection that is equal across cohorts, under the following conditions: for example, suppose that men who join the army at age 21 are always taller than those who join at age 20, and by the same amount across all birth cohorts. Although implausible, if this were true, one could still use such selected samples to examine trends in heights across cohorts. (Under this assumption, however, the height estimated for any particular cohort would not correspond to the actual population). There is a more exacting test that allows all age-at-recruitment effects to vary by birth cohort. This amounts to including in the regression all possible interactions of birth-year and age-at-recruitment dummies. We then test the null hypothesis that any height variations by age at recruitment are constant across birth cohorts. A rejection of this null implies height-based selection that cannot be averaged away in a simple fashion.

The same intuition suggests an alternative specification. We construct parallel tests in which we hold constant birth year, and vary the calendar-year an individual joined the military. Without interactions terms, the estimates of the calendar-year dummies would imply that all individuals joining in a particular year are either shorter or taller than their birth-cohort specific mean. Such a model would ignore the fact that individuals from two different birth cohorts joining in the same year are of different ages. There is once again a more exacting test: we interact dummies for all birth years with

dummies for all calendar years, and test the null that the calendar-year effects do not vary by birth cohort. Once again, rejection of this null generally implies the height-based selection implied by the Roy model.

These tests ask whether there is a type of homogeneity in the age- or calendar-year patterns of recruitment across birth cohorts. While one can imagine selection that would be homogeneous in this sense, the absence of this homogeneity implies selection patterns that are complex and not easily averaged out over successive cohorts. We report parallel tests of these models for two samples of British soldiers, for three subsets of the Union Army data, for free people of color, and for Pennsylvania prisoners. To restrict the estimation sample to those who have reached full height, the models only use observations for men in age ranges specified in the tables below. All models use OLS with standard errors corrected for heteroskedasticity. Table 4 reports the results for the two British Army samples. (The appendix provides sources for all data used in this section.) The data reject the homogeneity hypothesis, with the exception of the last model estimated with the AMD data.

Table 5 reports the results for three subsets of the Union Army. These subsets differ by how they measure the soldier's age at enlistment. The first relies on an integer age variable directly available in the Union Army data set; here, birth cohort is defined as enlistment year minus this reported age at enlistment. The second constructs an (integer) age at enlistment from the difference between the date of enlistment and the birth date. The third subset discards all observations for which these two integer age measures do not agree exactly; the sample size with this restriction is fairly small. As Table 5 shows, the model rejects the null of homogeneity for all but some versions of model 4. The

Union Army results may reflect the fact that the recruitment period spans a just five years. Figure 2 reports a graphical warning based on the Union Army results. A model with birth-year dummies only implies that soldiers born in the 1840s were indeed shorter than those born in the late 1830s. But this result disappears in models that control for recruitment year as well. The figure underscores results such as A'Hearn (1998) and Haines (1998).

Two other types of sources have figured heavily in the puzzle literature: free-born and manumitted African Americans, and prisoners. Virginia's and Maryland's "black code" required all free and manumitted African Americans to register with the local county clerk. Any noncompliant free person risked arrest and jailor's fees, which might be expected to have encouraged near universal registration because the law was enforced, even if unevenly (Komlos 1992, Bodenhorn 1999). But only a fraction of African Americans actually registered. A second feature of Virginia's 1793 act imposed a $5 fine (per act) on any employer who hired a free person of color without a proper registration. This provision might lead to selective registration; most free-born registrants appear in the records between the ages of 17 and 25, when young adults typically enter the paid workforce. The literature on manumission gives good reason to think that freed slaves themselves would not represent a random draw from the population of African-Americans born into slavery, so we have several reasons to think that samples of free blacks would suffer endogenous selection. Table 6 reports tests of homogeneity for free blacks. The model rejects homogeneity for the recruitment-year version of the model, but not for models 1 and 2.

Many studies of the antebellum puzzle use data drawn from the heights of prisoners.[21] We ask whether there is evidence of self-selection in a typical sample of convicts incarcerated during the era of early industrialization, drawing on data from the Pennsylvania penitentiary system between the late 1820s and the late 1870s. Prisoners, especially those confined to state penitentiaries in the nineteenth century, were unlikely to represent random draws from the wider population. Prisoners might not even be representative of criminals. The imprisoned arrived after traversing a criminal process required several decisions by different agents: individuals chose to (allegedly) commit a crime; the police chose whether to arrest and charge the suspect; the prosecutor chose whether to prosecute the case; a judge and jury chose to convict and to impose a sentence of more than one year of incarceration. Ultimately, men committed to the state prisons were those who were convicted of relatively serious crimes. Bodenhorn, Moehling and Price (2012), in fact, show criminals were short relative to their contemporaries and that shorter men entered prison at younger ages. The mean age at admission into the Eastern and Western penitentiaries was 28.5 years, and ages ranged from 11 to 89 years. Criminologists identify the prime offending ages from the mid-teens to the mid-twenties, which is consistent with the historical data as well.[22] Because less-privileged individuals tended to not reach their terminal adult heights until age 20 or later and, because immigrants faced different childhood environments, we limit the sample to native-born

[21] For studies of US prisons, see Komlos and Coclanis (1997); Carson (2009); Maloney and Carson (2008); Tatarek (2006); Sunder (2004).Nicholas and Steckel (1991) and Nicholas and Oxley (1993) investigate heights using prison records from Great Britain. Frank (2012) and Twrdek and Manzel (2012) use heights from Peruvian prisons.

[22] Moehling and Piehl, "Immigration, Crime and Incarceration ."

men between 23 and 50 years. Table 6 reports results that parallel those for free blacks. The model rejects the homogeneity assumption for models based on enlistment year, but not for those based on age.

7. Conclusions

For several decades now, anthropometricians have discussed the industrialization (or "antebellum") puzzle: the apparent finding that human heights declined during periods of rising real incomes. The industrialization puzzle has achieved the status of stylized fact in its depiction of the nature of economic growth, modernization, and urbanization in the mid- to late nineteenth-century United States. The findings for the U.S have led heights researchers to look for similar patterns in other countries. The core issues this literature discusses are central to understanding the process of modern economic growth; the "standard of living debate" gets to the heart of how economic growth affects human welfare.

Unfortunately, the heights literature has relied heavily on sources that likely reflect various forms of endogenous sampling. Volunteer militaries are the most common source. The decision to join the army reflects an individual's evaluation of his best prospects in life. Those prospects depend on unobserved, individual-specific factors that are a function of the individual's human capital, and thus likely correlated with height. Thus the heights of recruits at any one time cannot yield unbiased estimates of population heights. In addition, the heights of those in the choice-based sample (the army) will react to changing economic conditions in complex ways. The Roy model reported in Bodenhorn, Guinnane and Mroz (2014) implies that improvements in civilian labor

markets will lead to a shorter army. That is, the industrialization is no puzzle once one appreciates the consequences of the endogenous selection process underlying many heights sources.

Direct testing of the selection hypothesis requires matching military records (for example) to some other source, a process that can induce its own selection problems. The heights sources contain internal evidence of sample-selection bias, however. We develop and report a series of tests that rely on the idea that the decision to join the military (for example) reflects conditions at the time one joins, while, under the basic idea of the heights literature, the forces that determine adult height reflect events that occurred long before the age people join the military. The tests look for selection that will not "average out" over a cohort's lifetime; all show that the sources underlying the industrialization puzzle findings cannot yield unbiased estimates of heights or trends in heights.

Despite the claims made by this literature, the direct evidence for the puzzle is less robust than one would want. A meta-analysis of some 167 papers that deal with this subject demonstrates that most findings of a height reversal rely on selected or small samples. In the vast majority of cases where conscription provided something near to a random sample of young men, or the population of young men, heights grew monotonically throughout the nineteenth century. The United States, which is the core example for the puzzle literature, did not have a meaningful draft during the nineteenth century.

So is the industrialization puzzle real? Scholars who believe it is typically point to a range of evidence other than heights to support the findings based on heights alone. Mortality rates remained stubbornly high through the early decades of industrialization,

for example, and in some cases actually increased, as cities became larger and more unhealthy. Real wages rarely fell, but there is reason to doubt that feeble nominal-wage growth protected the lowest strata from the consequences of food-price shocks. The largest standard of living debate continues for the simple reason that there is evidence of both improvement and deterioration of living standards as part of the process of economic growth. If anthropometric evidence is to contribute to this debate, scholars must bear in mind the sample-selection bias demonstrated here.

8. References

A'Hearn, Brian. "The Antebellum Puzzle Revisited: A New Look at the Physical Stature of Union Army recruits during the Civil War." In *The Biological Standard of Living in Comparative Perspective,* 250-267. Edited by John Komlos and Joerg Baten. Stuttgart: Steiner, 1998.

Baten, Joerg. "Protein Supply and Nutritional Status in Nineteenth Century Bavaria, Prussia and France." *Economics and Human Biology* 7, no. 2 (2009): 165-180.

Baten, Joerg, and John Komlos. "Height and the Standard of Living." *The Journal of Economic History* 58.03 (1998): 866-870.

Bodenhorn, Howard. "A Troublesome Caste: Height and Nutrition of Antebellum Virginia's Rural Free Blacks." *Journal of Economic History* 59, no. 4 (1999): 972-996.

Bodenhorn, Howard, Timothy Guinanne, and Thomas A. Mroz. "Sample Selection Bias in the Historical Heights Literature." Cowles Foundation Working Paper, Yale University, 2014.

Bodenhorn, Howard, Carolyn Moehling and Gregory N. Price. "Short Criminals: Stature and Crime in Early America." *Journal of Law and Economics* 55, no. 2 (2012): 393-419.

Brennan, Lance, John McDonald and Ralph Shlomowitz. "The Heights and Economic Well-Being of North Indians under British Rule." *Social Science History* 18, no. 2 (1994a): 271-307.

Brennan, Lance, John McDonald and Ralph Shlomowitz. "Trends in the Economic well-Being of South Indians under British Rule: The Anthropometric Evidence." *Explorations in Economic History* 31, no. 2 (1994b): 225-260.

Brennan, Lance, John McDonald, and Ralph Shlomowitz. "Toward an Anthropometric History of Indians under British Rule." *Research in Economic History* 17 (1997): 185-246.

Carson, Scott Alan. "The Biological Living Conditions of Nineteenth-Century Chinese Males in America." *Journal of Interdisciplinary History* 37, no. 2 (2006): 201-217.

Carson, Scott Alan. "The Effect of Geography and Vitamin D on African American Stature in the Nineteenth Century: Evidence from Prison Records." *Journal of Economic History* 68, no. 3 (2008): 812-831.

Carson, Scott Alan. "African-American and White Inequality in the Nineteenth Century American South: A Biological Comparison." *Journal of Population Economics* 22, no. 3 (2009): 739-755.

Cranfield, John and Kris Inwood. "Stayers and Leavers, Diggers and Canucks: The 1914-1918 War in Comparative Perspective." Working paper, University of Guelph (2011).

Deaton, Angus. *The Great Escape: Health, Wealth and the Origins of Inequality*. Princeton: Princeton University Press, 2013.

Feinstein, Charles. ""Pessimism Perpetuated: Real Wages and the Standard of Living in Britain during and after the Industrial Revolution." *Journal of Economic History* 58, no. 3 (1998): 625-658.

Floud, Roderick, Robert W. Fogel, Bernard Harris, and Sok Chul Hong. *The Changing Body: Health, Nutrition, and Human Development in the Western World since 1700*. New York: Cambridge University Press, 2011.

Floud, Roderick, Kenneth Wachter, and Annabel Gregory. *Height, Health and History: Nutritional Status in the United Kingdom, 1750-1980*. New York: Cambridge University Press, 1990.

Fogel, Robert William. "Nutrition and the Decline in Mortality since 1700: Some Preliminary Findings." In *Long-Term Factors in American Economic Growth,* 439-556. Edited by Stanley L. Engerman and Robert E. Gallman. Chicago: University of Chicago Press (1986).

Gallman, Robert E. "Dietary Change in Antebellum America." *Journal of Economic History* 56, No. 1 (1996): 193-201.

Haines, Michael R. "Height, Nutrition and Mortality." In *The Biological Standard of Living in Comparative Perspective,* 155-180. Edited by John Komlos and Joerg Baten. Stuttgart: Steiner, 1998.

Heckman, James J. and Guilherme Sedlacek. "Heterogeneity, Aggregation and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market." *Journal of Political Economy* 93, no. 6 (1985): 1077-1125.

Johnson, Paul and Stephen Nicholas. "Male and Female Living Standards in England and Wales, 1812-1857: Evidence from Criminal Height Records." *Economic History Review* 48, no. 3 (1995): 470-481.

Komlos, John. "The Height and Weight of West Point Cadets: Dietary Change in Antebellum America." *Journal of Economic History* 47, no. 4 (1987): 897-927.

Komlos, John. *Nutrition and Economic Development in the Eighteenth-Century Habsburg Monarchy*. Princeton: Princeton University Press (1989).

Komlos, John. "Toward an Anthropometric History of African-Americans: The Case of Free Blacks in Antebellum Maryland." In *Strategic Factors in Nineteenth American Economic History: A Volume to Honor Robert W. Fogel*, pp. 297-329. Edited by Claudia Goldin and Hugh Rockoff. Chicago: University of Chicago Press, 1992.

Komlos, John. "The Secular Trend in the Biological Standard of Living in the United Kingdom, 1730-1860." *Economic History Review* 46, no. 1 (1993a): 115-144.

Komlos, John. “A Malthusian Episode Revisited: The Height of British and Irish Servants in Colonial America.” *Economic History Review* 46, no. 4 (1993b): 768-782.

Komlos, John. “The Nutritional Status of French Students.” *Journal of Interdisciplinary History* 24, no. 3 (1994a): 493-508.

Komlos, John. “Anomalies in Economic History: Toward a Resolution of the ‘Antebellum Puzzle’.” *Journal of Economic History* 56, no. 1 (1996): 202-214.

Komlos, John. “Shrinking in a Growing Economy? The Mystery of Physical Stature during the Industrial Revolution.” Journal of Economic History 58, no. 3 (1998a): 779-802.

Komlos, John. “On the Biological Standard of Living of African Americans: The Case of the Civil War Soldiers.” In *The Biological Standard of Living in Comparative Perspective,* 236-249 Edited by John Komlos and Joerg Baten. Stuttgart: Steiner, 1998b.

Komlos, John. “Stagnation of Height among Second-Generation US-Born Army Personnel.” *Social Science Quarterly* 89, no. 2 (2008): 445-455.

Komlos, John. “A Three Decade ‘Kuhnian’ History of the Antebellum Puzzle: Explaining the Shrinking of the U.S. Population at the Onset of Modern Economic Growth.” Working paper, University of Munich (2012).

Komlos, John, and Joo Han Kim. "Estimating Trends in Historical Heights."*Historical Methods: A Journal of Quantitative and Interdisciplinary History* 23, no. 3 (1990): 116-120.

Komlos, John, and Peter Coclanis. "On the Puzzling Cycle in the Biological Standard of Living: the Case of Antebellum Georgia." *Explorations in Economic History* 34.4 (1997): 433-459.

Maloney, Thomas N. and Scott Alan Carson. “Living Standards in Black and White: Evidence from the Heights of Ohio Prison Inmates, 1829-1913.” *Economics and Human Biology* 6, no. 2 (2008): 237-251.

Margo, Robert A. and Richard H. Steckel. “Heights of Native-Born Whites during the Antebellum Period.” *Journal of Economic History* 43, no. 1 (1983): 167-174.

Moehling, Carolyn, and Anne Morrison Piehl. "Immigration, Crime, and Incarceration in Early Twentieth-Century America." *Demography* 46, no .4 (2009): 739-763.

Mokyr, Joel, and Cormac O'Grada. "The Heights of the British and the Irish c. 1800-1815: Evidence from Recruits to the East India Company's Army." In *Stature, Living Standards, and Economic Development*. Edited by John Komlos. Essays in Anthropometric History, 39-59. Chicago: The University of Chicago Press, 1994.

Mokyr, Joel, and Cormac Ó Gráda. "Height and Health in the United Kingdom 1815–1860: Evidence from the East India Company Army." *Explorations in Economic History* 33, no. 2 (1996): 141-168.

Morgan, Stephen L. "Economic Growth and the Biological Standard of Living in China, 1880-1930." *Economics and Human Biology* 2, no. 2 (2004): 197-218.

Nicholas, Stephen, and Deborah Oxley. "The Living Standards of Women during the Industrial Revolution, 1795-18201." *Economic History Review* 46, no. 4 (1993): 723-749.

Nicholas, Stephen and Richard H. Steckel. "Heights and Living Standards of English Workers during the Early Years of Industrialization, 1770-1815." *Journal of Economic History* 51, no. 4 (1991): 937-957.

Riggs, Paul. "The Standard of Living in Scotland, 1800-1850." In *Stature, Living Standards, and Economic Development: Essays in Anthropometric History*, 60-75. Edited by John Komlos. Chicago and London: University of Chicago Press, 1994.

Riggs, Paul and Timothy Cuff. "Ladies from Hell, Aberdeen Free Gardeners, and the Russian Influenza: An Anthropometric Analysis of WWI-Era Scottish Soldiers and Civilians." *Economics and Human Biology* 11, no. 1 (2013): 69-77.

Roy, A. D. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3, no. 2 (1951): 135-146.

Saint Onge, Jarron M., Patrick M. Krueger, and Richard G. Rogers. "Historical Trends in Height, Weight, and Body Mass: Data from U.S. Major League Baseball Players, 1869-1983." *Economics and Human Biology* 6, no. 3 (2008): 482-488.

Salvatore, Ricardo D. "Heights and Welfare in Late-Colonial and Post-Independence Argentina." In *The Biological Standard of Living in Comparative Perspective*, 97-

121. Edited by John Komlos and Joerg Baten. Stuttgart: Franz Steiner Verlag. 1998.

Salvatore, Ricardo D. “Status Decline and Recovery in a Food-Rich Export Economy: Argentina, 1900-1934.” *Explorations in Economic History* 41, no. 3 (2004): 233-255.

Sandberg, Lars G. and Richard H. Steckel. “Heights and Economic History: The Swedish Case.” *Annals of Human Biology* 14, no. 2 (1987): 101-110.

Schultz, T. Paul. “Wage Gains Associated with Height as a Form of Health Human Capital.” *American Economic Review* 92, no. 2 (2002): 349-353.

Sunder, Marco. “The Height of Tennessee Convicts: Another Piece of the ‘Antebellum Puzzle’.” *Economics and Human Biology* 2, no. 1 (2004): 75-86.

Tatarek, Nancy E. “Geographical Height Variation among Ohio Caucasian Male Convicts Born 1780-1849.” *Economics and Human Biology* 4, no. 2 (2006): 222-236.

Twrdek, Linda, and Kerstin Manzel. "The Seed of Abundance and Misery: Peruvian Living Standards from the Early Republican Period to the End of the Guano Era (1820–1880)." *Economics & Human Biology* 8, no. 2 (2010): 145-152.

Voth, Hans-Joachim and Tim Leunig. “Did Smallpox Reduce Height? Stature and the Standard of Living in London, 1770-1873.” *Economic History Review* 49, no. 3 (1996): 541-560.

Wachter, Kenneth W. "Graphical Estimation of Military Heights." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 14, no. 1 (1981): 31-42.

Wachter, Kenneth W. and James Trussell. “Estimating Historical Heights.” *Journal of the American Statistical Association* 77, no. 378 (1982): 279-293.

Weir, David. "Economic welfare and physical well-being in France, 1750-1990."In *Health and Welfare during Industrialization*, 161-200. Chicago: University of Chicago Press, 1997.

Zimran, Ariel. “Does Sample Selection Bias Explain the Industrialization Puzzle? Evidence from Military Enlistment in the Nineteenth Century United States. Working paper, Northwestern University (2015).

Table 1 Summaries of antebellum puzzle studies									
Author(s)	**Pub year**	**Period**	**Sample**	**Height change**	**Nutrition**	**Disease**	**Work intensity**	**Immigration**	**Inequality**
Margo and Steckel	1982	1790s-1830s	Slaves	+1.4%					
Margo and Steckel	1982	1810s-1830s	Black recruits	-0.5%	?		?		
Margo and Steckel	1983	1810s-1830s	Union army	0.0% - +0.3%	?	?	?	?	
Fogel	1986	1820s-1890s	Military	-2.3%	?	Y	?	Y	?
Komlos	1987	1820s-1860s	West Point	-0.8%	Y	N	?	N	
Komlos	1992	1800s-1840s	Black women	-1.4%	Y				
Komlos	1992	1800s-1840s	Black men	-0.5%	Y				
Steckel and Haurin	1994	1840s-1890s	National guard	-2.2%	?				
Gallman	1996	1820s-1860s	West Point	-0.8% - +0.1%	N	?			
Komlos	1996	1820s-1870s	West Point (middle-class)	+1.1%	Y	N			Y
Coclanis and Komlos	1997	1860s-1880s	Citadel (18yrs)	-0.5%	Y	Y			Y
Coclanis and Komlos	1997	1860s-1880s	Citadel (17yrs)	+0.6%	Y	Y			Y
A'Hearn	1998	1810s-1830s	Union army-farmers	+0.2%	Y	?			
A'Hearn	1998	1810s-1830s	Union army-rural	+0.1%	Y	?			
A'Hearn	1998	1810s-1830s	Union army-urban	+0.1%	Y	?			
Craig and Weiss	1998	1810s-1840s	Union army	na	Y				
Haines	1998	1810s-1840s	Union army - NY	-0.8%	Y	Y			?
Komlos	1998	1810s-1840s	Union army-blacks (RSML)	-0.6%					
Komlos	1998	1810s-1840s	Union army-blacks (TOLS)	+0.2%					
Komlos	1998	1810s-1840s	Union army-black (Komlos-	0.0%					

			Kim)						
Bodenhorn	1999	1800s-1830s	Free-born black men	-0.8%	Y				
Bodenhorn	1999	1800s-1830s	Manumitted men	+1.5%	Y				
Bodenhorn	1999	1800s-1830s	Free-born women	-1.3%	Y				
Bodenhorn	1999	1800s-1830s	Manumitted men	-1.6%	Y				
Margo	2000	1820s-1830s	Convicts	-0.3%		Y			
Margo	2000	1840s-1850s	National guard	-0.8%		Y			
Haines et al	2003	1830s-1860s	Union army	-1.7%	Y	Y			
Sunder	2004	1830s-1850s	Prisoners-white	+0.4%	Y				
Sunder	2004	1830s-1850s	Prisoners-black	+2.4%	Y				
Cuff	2006	1810s-1840s	Union army -Penn	-1.4%	Y	Y			
Tatarek	2006	1790s-1840s	Prisoners-white	-0.6%					
Maloney and Carson	2008	1780s-1880s	Prisoners-white	-0.3%					
Maloney and Carson	2008	1800s-1880s	Prisoners-black	-1.7%					
Bodenhorn	2010	1790s-1840s	NY legislators	-2.7%					
Hiermeyer	2010	1860s-1880s	West Point	+0.8%	Y	Y			
Zehetmayer	2011	1840s-1890s	US army	-0.6%	Y	Y			Y

Notes: decades reported may not represent all the decades included in the original study; only those relevant to the puzzle period. Y = some evidence provided in support of hypothesis; N = cause rejected as unlikely, either directly or by inference; ? = discussed as likely cause, but with qualifications and/or without supporting evidence.

Table 2 Sample type and fraction demonstrating a 1cm or more reversal		
	One reversal	Two or more reversals
Conscripts	0.20	0.04
Volunteers	0.56	0.33
Students	0.57	0.15
Prisoners	0.70	0.35
Other	0.34	0.07
Notes: see text for explanation of sample types. **Sources: see online appendix table.**		

Table 3			
Probit regressions results – observation of 1cm or more decline in any decade			
	Summary statistics	At least one 1cm decline	At least two 1cm declines
Conscript	0.265	<reference>	<reference>
Volunteer	0.247	0.442 (0.148)**	0.283 (0.177)
Prisoner	0.146	0.313 (0.151)*	0.310 (0.177)*
Student	0.082	0.250 (0.171)	0.029 (0.177)
Other	0.259	0.054 (0.139)	0.024 (0.104)
ln(Years)	3.799 (0.526)	0.373 (0.105)**	0.183 (0.058)**
ln(Observations)	8.945 (2.128)	-0.090 (0.031)**	-0.042 (0.019)*
National	0.228	-0.114 (0.108)	0.066 (0.064)
Shortfall correction	0.171	-0.101 (0.141)	0.121 (0.109)
N	158	158	158

Notes: columns 2 and 3 report marginal effects. dependent variable in column 2 = 1 if study reports a 1cm or greater decline in height during any decade, =0 otherwise; dependent variable in column 3 =1 if study reports more than one 1cm or greater decline in mean height during any decade, =0 otherwise. Other category includes indentured servants, slaves, migrant workers, baseball players and sundry other groupings of individuals. See appendix Table A.x for details. ln(Years) variable calculated as difference between first year and last year included in the study. If the study reported dates as, say, 1820s – 1880s, the first year is taken to be 1820 and the last year is 1889. National =1 if the study uses nationally representative sample. Shortfall correction=1 if study reports period mean heights only after correcting for left-tail shortfall using any of the standard methods: RSMLE, QBE, or Komlos-Kim.
Sources:

Table 4: British Army

Summary of selection diagnostic tests using cohorts and observation years

	British Army		AMD
Ages included	23-27		23-25
Observations	5879		71109
Model 1: Adding age dummies to a model with only birth-cohort dummies			
F-stat: 4 age dummies	3.83	2 dummies	6.32
P-value	0.0041		0.0018
DF	[4, 5759]		[2,71064]
Model 2: Adding year dummies to a model with only birth-cohort dummies			
F-stat: 92 year dummies	2.4		2.1
P-value	0	34 dummies	0.0002
DF	[92, 5671]		[34,71032]
Model 3: Adding age by observation year dummies to a model with birth cohort and age dummies			
F-stat: 92 age by observation dummies	2.4	66 dummies	1.56
P-value	0		0.0025
DF	[228, 5531]		[66,70998]
Model 4: Adding birth by enlistment year dummies to a model with birth-cohort and enlistment years			
F-stat: 4 age dummies	1.41	34 dummies	1.27
P-value	0.0012		0.1367
DF	[140, 5531]		[34,70998]

Table 5: Union Army
Tests for non-homogenous heights: Ages 23-30
With 3 different age definitions for the Union Army Data

	Reported age at enlistment	Constructed age at enlistment	Reported and constructed ages agree
Sample Size	11,372	4,315	2,227
Model 1:Adding age dummies to a model with only birth cohort dummies			
F-Statistic: 7 age dummies	16.92	4.76	2.06
P-value	<0.0001	<0.0001	0.0445
DF	[7,11353]	[7,4296]	[7,2208]
Model2: Adding enlistment year dummies to a model with only birth cohort dumies			
F-Statistic: 4 year dummies	33.9	9.15	3.32
P-value	<0.0001	<0.0001	0.0101
DF	[4,11356]	[4,4299]	[4,2211]
Model 3: Adding age by enlistment year dummies to birth and age at enlistment dummies			
F-Statistic: 21 unique dummies	1.84	2.55	1.57
P-value	0.0106	0.0001	0.0473
DF	[21,11332]	[21,4275]	[21,2187]
Model 4: Adding birth by enlistment year dummies to birth cohort and enlistment years			
F-Statistic: 24 unique dummies	0.9	2.07	1.37
P-value	0.606	0.0017	0.1068
DF	[24,11332]	[24,4275]	[24,2187]

TABLE 6

Tests for non-homogenous heights: Ages 23-30

PA Prisoners and Free Black Samples

	PA Prisoners	Free Blacks
Sample Size	1,800	4,797
Model 1:Adding age dummies to a model with only birth cohort dummies		
F-Statistic: 7 age dummies	1.81	1.39
P-value	0.0812	0.2032
Degrees of freedom	[7,1744]	[7,4725]
Model 2: Adding observation year dummies to a model with only birth cohort dumies		
F-Statistic:	0.94	0.86
P-value	0.5929	0.7563
Degrees of freedom	[44,1707]	[57,4675]
Model 3: Adding age by observation year dummies to birth cohort and age dummies		
F-Statistic:	1.25	1.19
P-value	0.0242	0.015
Degrees of freedom	[157,1587]	[308,4417]
Model 4: Adding birth cohort by enlistment year dummies to cohort and enlistment years		
F-Statistic:	1.22	1.2
P-value	0.0606	0.0182
Degrees of freedom	[120,1587]	[258,4417]

Figure 1: Mean heights of volunteer soldiers in the United States and in selected countries with conscription

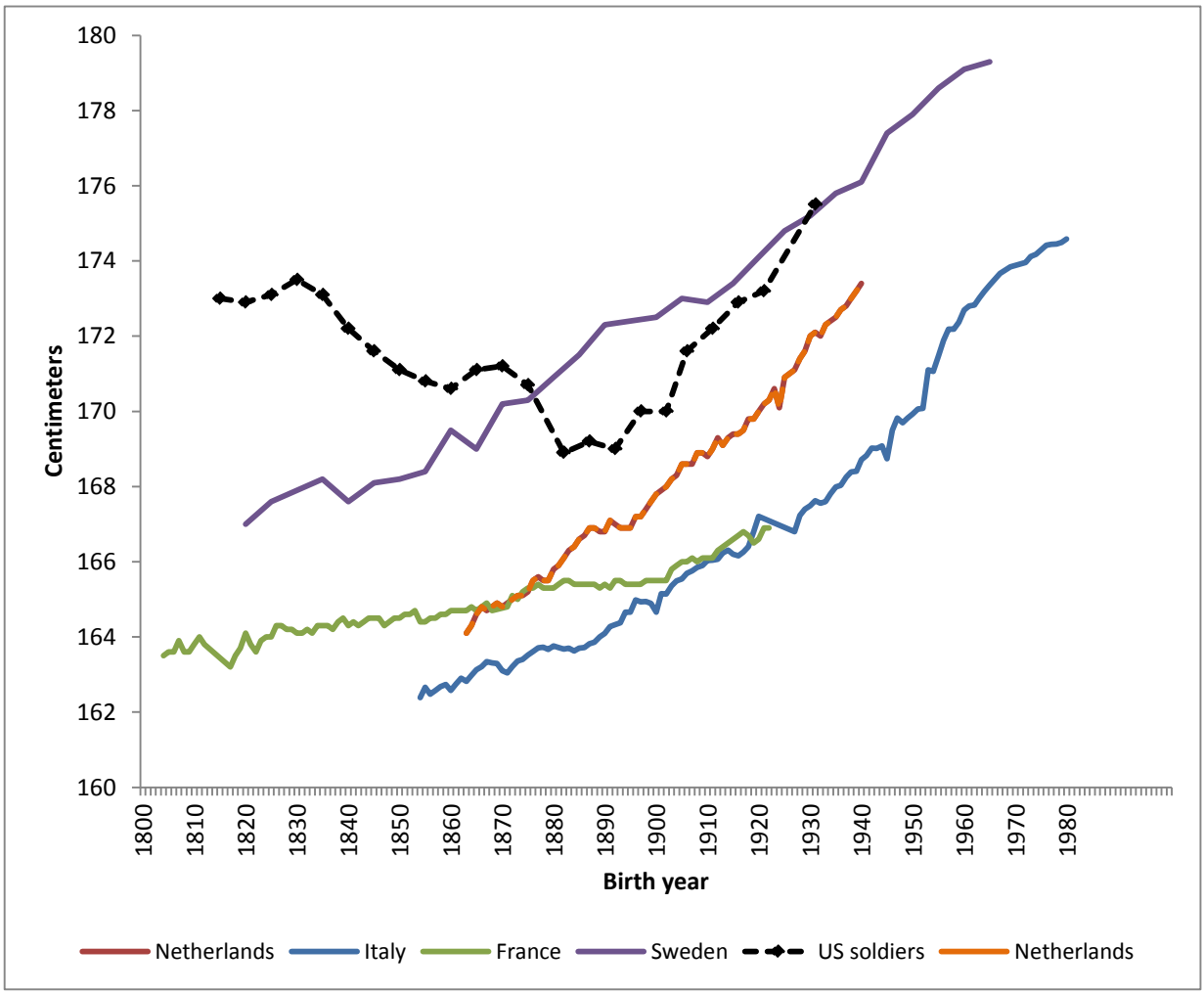
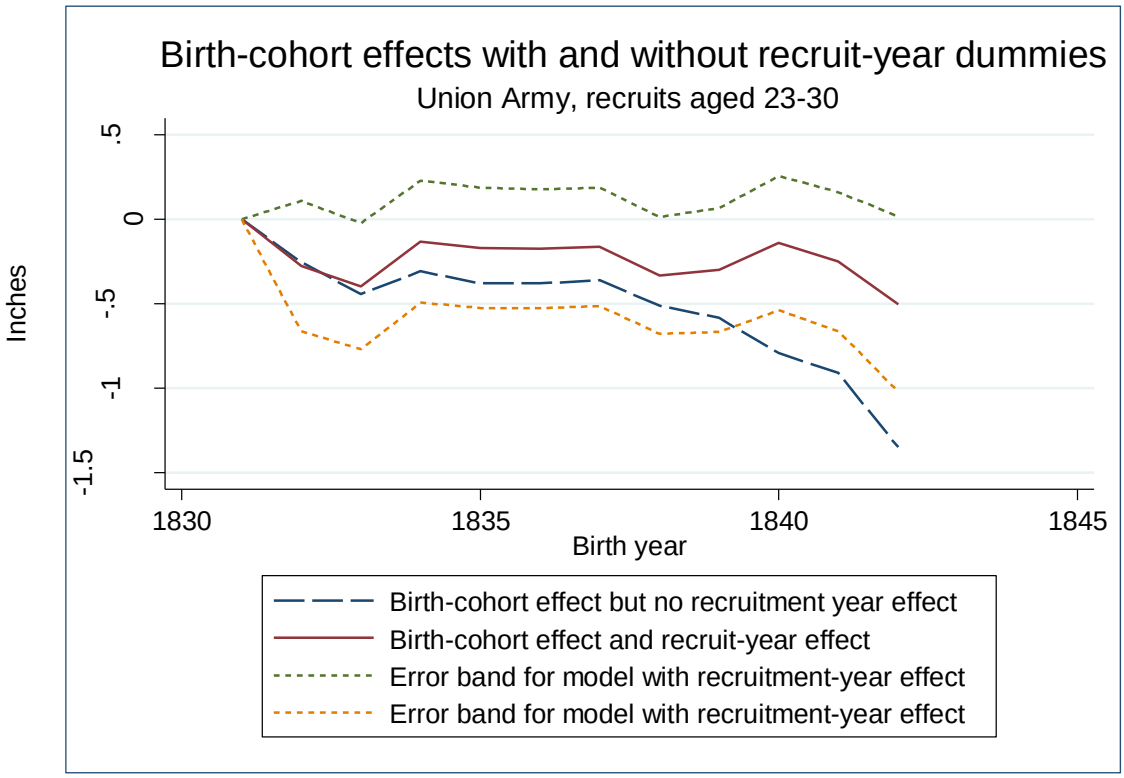


Figure 2



Note: the figure plots the birth cohort effects estimated by Model 3 reported in Table 5.