B Additional Tables

Table B.1: Binary choice model estimation

	(1)		(2)	
	(a)	(b)	(a)	(b)
Cohorts	’32–’46	’47–’60	’32–’46	’47–’60
Variables	UA	RA	UA	RA
Individual or Household Variables				
Household Owns Property	0.081***	0.189***	0.232***	0.453***
	(0.017)	(0.019)	(0.066)	(0.048)
Household Real Property (1,000)	−0.015***	−0.046***	−0.051***	−0.120***
	(0.002)	(0.004)	(0.006)	(0.012)
Related to Head of Household	0.092***	−0.062***	0.243***	−0.189***
	(0.019)	(0.018)	(0.052)	(0.070)
Household Size	0.017***	−0.000	0.059***	−0.001
	(0.003)	(0.003)	(0.009)	(0.007)
Attended School	−0.177***	−0.148***	−0.641***	−0.501***
	(0.016)	(0.015)	(0.049)	(0.065)
Household Occupation (Unproductive excluded)				
Farmer	−0.341***	−0.223***	−1.179***	−0.565***
	(0.038)	(0.024)	(0.109)	(0.085)
Professional	−0.286***	−0.106***	−0.810***	−0.209***
	(0.056)	(0.036)	(0.110)	(0.060)
Clerical	−0.361***	−0.154***	−1.013***	−0.290***
	(0.046)	(0.030)	(0.084)	(0.039)
Skilled and Artisan	−0.316***	−0.097***	−0.941***	−0.224***
	(0.038)	(0.021)	(0.075)	(0.045)
Semi-Skilled and Clerical	−0.322***	−0.133***	−0.908***	−0.262***
	(0.047)	(0.028)	(0.085)	(0.043)
Unskilled	−0.305***	−0.101***	−0.856***	−0.212***
	(0.042)	(0.022)	(0.073)	(0.042)
Farm Labor	−0.187*	−0.126***	−0.557	−0.236***
	(0.102)	(0.033)	(0.342)	(0.040)
Birth Region (South excluded)				
Midwest	0.209***	−0.054*	0.947***	−0.134*
	(0.033)	(0.031)	(0.158)	(0.078)
Northeast	0.102***	−0.046	0.343***	−0.124
	(0.033)	(0.030)	(0.112)	(0.090)
County Variables				
Fraction Urban	−0.123***	0.049	−0.420***	0.128*
	(0.039)	(0.031)	(0.121)	(0.072)
Wheat Bushels per capita	0.013***	−0.004***	0.043***	−0.009***
	(0.001)	(0.001)	(0.004)	(0.003)
Milk Cows per capita	0.010	−0.066	0.033	−0.173
	(0.038)	(0.070)	(0.130)	(0.190)
Swine per capita	0.075***	−0.076***	0.256***	−0.198***
	(0.011)	(0.017)	(0.032)	(0.043)
Value of Agricultural Production per capita (1,000)	−0.046	−0.042	−0.157	−0.109
	(0.472)	(0.432)	(1.483)	(1.106)
Lincoln Vote Share (1860)	0.544***	0.127**	1.853***	0.332**
	(0.064)	(0.054)	(0.181)	(0.140)
Observations	13,683	11,271	13,683	11,271

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Notes: Column (1) presents estimates of the coefficients β and δ from the binary choice model. Column (2) presents the average semi-elasticity of the impact of each variable on enlistment probability as implied by the estimates of column (1). All specifications include cohort indicators. Standard errors are clustered at the county level. UA denotes Union Army. RA denotes Regular Army.

Table B.2: Height regressions

	(1)	(2)	(3)	(4)
Variables	Corr	Not	Corr	Not
Individual or Household Variables				
Household Owns Property	0.092	0.104	0.100	0.104
	(0.094)	(0.111)	(0.096)	(0.111)
Household Real Property (1,000)	−0.011	−0.006	−0.012	−0.006
	(0.012)	(0.008)	(0.012)	(0.008)
Related to Head of Household	0.245**	0.300**	0.245**	0.300**
	(0.104)	(0.131)	(0.104)	(0.131)
Household Size	0.025*	0.023	0.026*	0.024
	(0.014)	(0.015)	(0.014)	(0.015)
Attended School	0.038	0.044	0.026	0.038
	(0.077)	(0.084)	(0.079)	(0.084)
Household Occupation (Unproductive excluded)				
Farmer	0.208	0.055	0.180	0.030
	(0.134)	(0.123)	(0.140)	(0.125)
Professional	0.234	0.039	0.214	0.019
	(0.207)	(0.236)	(0.207)	(0.235)
Clerical	0.011	−0.260	−0.012	−0.279
	(0.177)	(0.197)	(0.182)	(0.198)
Skilled and Artisan	−0.128	−0.313**	−0.147	−0.331**
	(0.144)	(0.148)	(0.147)	(0.149)
Semi-Skilled and Clerical	0.080	−0.125	0.061	−0.140
	(0.185)	(0.216)	(0.188)	(0.216)
Unskilled	0.250	0.091	0.228	0.067
	(0.173)	(0.187)	(0.178)	(0.189)
Farm Labor	−0.205	−0.479*	−0.222	−0.496*
	(0.230)	(0.283)	(0.232)	(0.285)
Birth Region (South excluded)				
Midwest	0.034	0.042	−0.070	−0.136
	(0.163)	(0.169)	(0.173)	(0.198)
Northeast	−0.244	−0.322*	−0.362**	−0.524**
	(0.162)	(0.181)	(0.181)	(0.216)
County Variables				
Fraction Urban	−0.553***	−0.613**	−0.556***	−0.630***
	(0.202)	(0.238)	(0.202)	(0.238)
Wheat Bushels per capita	−0.001	0.004	−0.001	0.003
	(0.007)	(0.006)	(0.007)	(0.006)
Milk Cows per capita	0.008	0.080	−0.016	0.048
	(0.206)	(0.196)	(0.212)	(0.205)
Swine per capita	0.077	0.064	0.094	0.087
	(0.051)	(0.051)	(0.057)	(0.054)
Value of Agricultural production per capita (1,000)	−3.964*	−3.758*	−4.305**	−4.284*
	(2.047)	(2.257)	(2.089)	(2.320)
Lincoln Vote Share (1860)			0.292	0.469
			(0.300)	(0.315)
Observations	7,249	6,873	7,249	6,873

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Notes: Standard errors in parentheses. Dependent variable is height, measured in inches. All specifications include age-of-measurement, year-of-birth, and household occupation indicators. The selection-corrected specifications, indicated by the column header Corr, also include the selection-correction function $\Omega(\cdot)$. The uncorrected specifications, indicated by the column header Not, correct for truncation with a truncation point of 64 inches. Standard errors are clustered at the county level. The difference in sample sizes between columns is the result of the need to drop heights below 64 inches in the truncation-corrected regressions when not correcting for sample-selection bias.

C Correcting for Selection: Formal Arguments

The model developed in section 2 points out the need to correct for selection into military service on the basis of both observable and unobservable characteristics in order to estimate $E(h_{it}|t)$ for each t—that is, the trend in average heights over birth cohorts. In this Appendix, I formally develop the weighting approach to correct for selection on observables and the control function approach to correct also for selection on unobservables.

C.1 Selection on Observables

By the law of iterated expectations, the object of interest can be written as

$$\begin{aligned} E(h_{it}|t) &= \int_{\mathfrak{X}} E(h_{it}|\mathbf{x}_{it}, \mathbf{z}_{it}; t) f(\mathbf{x}_{it}, \mathbf{z}_{it}|t) \, \mathrm{d}\mathbf{w}_{it} \\ &= \int_{\mathfrak{X}} E(h_{it}|\mathbf{x}_{it}; t) f(\mathbf{x}_{it}, \mathbf{z}_{it}|t) \, \mathrm{d}\mathbf{w}_{it}, \end{aligned} \tag{C.1}$$

where $\mathbf{w}_{it} = [\mathbf{x}'_{it}, \mathbf{z}'_{it}]'$,[36] $\mathfrak{X}$ is the support of $\mathbf{w}_{it}$, $F(\cdot)$ is its distribution function, and $f(\cdot)$ is its density. Equation (C.1) follows from the assumption—implicit in equations (1) and (2)—that height is uncorrelated with $\mathbf{z}_{it}$ conditional on $\mathbf{x}_{it}$. If there were no self selection, either on the basis of observables or unobservables, the left hand side of equation (C.1) could be computed trivially from the data; however, the researcher observes $E(h_{it}|y_{it} = 1; t)$ and not $E(h_{it}|t)$ in a selected sample. Moreover, the components of the right-hand side of equation (C.1) cannot, in general, be directly computed from a sample consisting solely of military enlisters—the researcher observes $E(h_{it}|\mathbf{x}_{it}, y_{it} = 1; t)$ and not $E(h_{it}|\mathbf{x}_{it}; t)$; but if the selection is only on observables (i.e., ε_{it} and u_{it} are uncorrelated), the assumptions discussed above imply that

$$E(h_{it}|\mathbf{x}_{it}, y_{it} = 1; t) = E(h_{it}|\mathbf{x}_{it}; t) = \gamma_t + \mathbf{x}'_{it}\theta. \tag{C.2}$$

What remains as the main pitfall is that selection into military service on the basis of observables implies that $f(\mathbf{x}_{it}, \mathbf{z}_{it}|t) \neq f(\mathbf{x}_{it}, \mathbf{z}_{it}|y_{it} = 1; t)$. That is to say, simply averaging the observed heights within each birth cohort will not yield consistent estimates of the true heights because the weighting is based on the distribution of covariates in the selected sample, which differs from that in the population. However, Bayes's

[36] Although consideration of the exclusion-restriction variables $\mathbf{z}_{it}$ is unnecessary for the correction of this type of selection, I include them as they are required for identification in the correction for selection on unobservables.

Theorem implies that

$$f(\mathbf{x}_{it}, \mathbf{z}_{it} | y_{it} = 1; t) = \frac{P(y_{it} = 1 | \mathbf{x}_{it}, \mathbf{z}_{it}; t) f(\mathbf{x}_{it}, \mathbf{z}_{it} | t)}{P(y_{it} = 1 | t)}, \tag{C.3}$$

so that

$$f(\mathbf{x}_{it}, \mathbf{z}_{it} | t) = \frac{f(\mathbf{x}_{it}, \mathbf{z}_{it} | y_{it} = 1; t) P(y_{it} = 1 | t)}{P(y_{it} = 1 | \mathbf{x}_{it}, \mathbf{z}_{it}; t)} \propto \frac{f(\mathbf{x}_{it}, \mathbf{z}_{it} | y_{it} = 1; t)}{P(y_{it} = 1 | \mathbf{x}_{it}, \mathbf{z}_{it}; t)}. \tag{C.4}$$

Substituting expressions (C.2) and (C.4) into equation (C.1) gives

$$E(h_{it} | t) = \int_{\mathfrak{X}} E(h_{it} | \mathbf{x}_{it}, y_{it} = 1; t) f(\mathbf{x}_{it}, \mathbf{z}_{it} | y_{it} = 1; t) \frac{k_t}{P(y_{it} = 1 | \mathbf{x}_{it}, \mathbf{z}_{it}; t)} \, \mathrm{d}\mathbf{w}_{it}, \tag{C.5}$$

where k_t is the normalizing constant for cohort t. Note that if the researcher were simply to take the (unweighted) average height for each birth cohort from a selected sample, he would estimate

$$E(h_{it} | y_{it} = 1; t) = \int_{\mathfrak{X}} E(h_{it} | \mathbf{x}_{it}, y_{it} = 1; t) f(\mathbf{x}_{it}, \mathbf{z}_{it} | y_{it} = 1; t) \, \mathrm{d}\mathbf{w}_{it} \tag{C.6}$$

from its sample analog $\frac{1}{N_t} \sum_{i \in t} h_{it}$, where N_t denotes the number of individuals in the sample belonging to birth cohort t and $i \in t$ denotes the members of cohort t; because expression (C.6) is equivalent to the right-hand side of expression (C.5) save for the inclusion of the weights $\frac{k_t}{P(y_{it}=1|\mathbf{x}_{it}, \mathbf{z}_{it}; t)}$, it is natural to estimate expression (C.5) by its sample analog

$$\hat{h}_t = \frac{\hat{k}_t}{N_t} \sum_{i \in t} \frac{h_{it}}{P(y_{it} = 1 | \mathbf{x}_{it}, \mathbf{z}_{it}; t)} = \frac{\hat{k}_t}{N_t} \sum_{i \in t} \frac{h_{it}}{G(\alpha_t + \mathbf{x}_{it}' \beta_k + \mathbf{z}_{it}' \delta_k)}. \tag{C.7}$$

It can be shown that expression (C.7) is precisely the estimated coefficient on the year-of-birth indicator for cohort t when observed heights are regressed on birth-cohort indicators (and no constant), weighting by inverse conditional enlistment probabilities, thus providing a method to perform this correction.

C.2 Selection on Unobservables

When selection on unobservables is admitted alongside selection on observables (i.e., correlation is permitted between ε_{it} and u_{it}), the arguments made in equations (C.3) and (C.4) continue to hold, as they did not rely on uncorrelatedness of ε_{it} and u_{it}, but rather were simply an application of Bayes's theorem. However, expression (C.2) is no longer true. Instead,

$$\begin{aligned} E(h_{it}|\mathbf{x}_{it},\mathbf{z}_{it},y_{it}=1;t) &= \alpha_t + \mathbf{x}_{it}'\theta + E(\varepsilon_{it}|\mathbf{x}_{it},\mathbf{z}_{it},y_{it}=1;t) \\ &= E(h_{it}|\mathbf{x}_{it};t) + \Omega(\alpha_t + \mathbf{x}_{it}'\beta_k + \mathbf{z}_{it}'\delta_k), \end{aligned} \tag{C.8}$$

where (C.8) follows from the assumptions regarding u_{it} and ε_{it} and equations (1) and (2). Thus, it is possible to write the missing piece of information in calculating expression (C.1) as a function of data and an unknown (but possible to estimate) object:

$$E(h_{it}|\mathbf{x}_{it};t) = E(h_{it}|\mathbf{x}_{it},\mathbf{z}_{it},y_{it}=1;t) - \Omega(\alpha_t + \mathbf{x}_{it}'\beta_k + \mathbf{z}_{it}'\delta_k).$$

The analog of equation (C.5) is then

$$\begin{aligned} E(h_{it}|t) = \int_{\mathfrak{X}} &\left[E(h_{it}|\mathbf{x}_{it},\mathbf{z}_{it},y_{it}=1;t) - \Omega(\alpha_t + \mathbf{x}_{it}'\beta_k + \mathbf{z}_{it}'\delta_k)\right] \\ &\times f(\mathbf{x}_{it},\mathbf{z}_{it}|y_{it}=1;t)\frac{k_t}{P(y_{it}=1|\mathbf{x}_{it},\mathbf{z}_{it};t)}\,\mathrm{d}\mathbf{w}_{it}, \end{aligned} \tag{C.9}$$

which can also be estimated by its sample analog:

$$\hat{h}_t = \frac{\hat{k}_t}{N_t}\sum_{i\in t}\frac{h_{it} - \Omega(\alpha_t + \mathbf{x}_{it}'\beta_k + \mathbf{z}_{it}'\delta_k)}{G(\alpha_t + \mathbf{x}_{it}'\beta_k + \mathbf{z}_{it}'\delta_k)}. \tag{C.10}$$

D Constructing Weights

This appendix describes the computation of the weights for estimation of the binary choice model. In order to compute the fraction of the relevant population serving in the Union Army, I consulted two sources. The first was Gould (1869, p. 28), who reports that 1,660,068 native-born men served in the Union Army. Next I consulted the 1860 census, finding that there were 3,720,008 native-born men aged 15–45 in the portions of the United States that did not secede. Thus, for the Union Army population

$$Q_1^{\mathrm{UA}} = \frac{1,660,068}{3,720,008} = 0.446.$$

To determine the value of Q_1 for the Regular Army, I again consulted two sources. First, I determined from the 1870 census that the total native-born white male population in non-seceding areas born between 1847 and 1860 was 3,467,695. Second, I collected the index of the *Register of Enlistments* for the native-born, and removed duplicate entries of name, birth year and state of birth.[37] This procedure yielded 77,836 distinct enlisters for the 1847–1860 cohorts, in each case restricting attention to those born in non-seceding areas. The estimate of Q_1 is thus

$$Q_1^{\mathrm{RA}} = \frac{77,836}{3,467,695} = 0.022.$$

[37] The removal of duplicates is necessary because individuals could enlist multiple times. Dropping duplicate appearances of name, birth year, and state of birth is an imperfect way of addressing this possibility. It is simultaneously too restrictive—there may have been two enlisters with the same name born in the same state in the same year—and too loose—slight deviations in the spelling or abbreviation of names, or misreporting of birth years would allow multiple enlistments by one individual to survive the removal of duplicates and be counted. Given that the actual value of Q_1 does not seem particularly important in practice, I do not investigate this potential problem further. An alternative method is to estimate the fraction of enlistments that are repeat enlistments using information from the *Register of Enlistments*; however, this information is not readily available for those in the earlier birth cohorts.

E Linkage

E.1 Procedure to Link Regular Army Enlisters to the US Censuses

I use standard census linking techniques developed by Ferrie (1996) to link the Regular Army samples to the US Censuses.

Procedure E.1. The procedure for linking the 1847–1860 cohorts to the censuses is as follows.

1. I obtained a 100 percent index of the 1880 census from Ruggles et al. (2015).

2. On the basis of first name, last name, state of birth, and year of birth (± 4), I linked the 1880 census sample to itself. As above, no name standardization is made, but inexact matches are permitted. Any individual for whom another individual similar on these identifying characteristics existed was removed from the sample.

3. I obtained a 100 percent sample of individuals born 1847–1860 listed in the *Register of Enlistments* from Ancestry.com (2007).

4. Using the same identifying information and criteria described in step 2, I linked individuals in the *Register of Enlistments* to the remaining individuals from the 1880 census. Due to the possibility of multiple enlistments in the lifetime, I permit several individuals in the *Register of Enlistments* to match to one individual in the census. However, in the event that several individuals in the 1880 census are matched to a single individual in the *Register of Enlistments*, I drop all concerned individuals.

5. I match individuals in the *Register of Enlistments* who are matched to the 1880 census to the *Register of Enlistments* once more (using stricter criteria), thus bringing in additional enlistments. Multiple matches from the 1880 census to the *Register of Enlistments* are once again omitted.

6. I then link the 1880 individuals who were linked to the *Register of Enlistments* to the 1860 and 1870 censuses (gathered from Ancestry.com, 2009a,b) using the same information and criteria.

Table E.1 presents the numbers of individuals included in the sample at each stage. Note that I did not restrict attention during linkage to residents of non-seceding areas.

Table E.1: Sample sizes at each stage of linking for 1847–1860 cohorts

(1)	(2)	(3)		(4)	
		Sample Size		% of Previous	
		(a)	(b)	(a)	(b)
Step No.	Description	Census	Enlistments	Census	Enlistments
3	Enlist Full		93, 085		
4 & 5	1880-Enlistments Link	14, 343	18, 802		20.20%
6	1860 Link	3, 133	4, 611	21.84%	24.52%
6	1870 Link	3, 129	4, 632	21.82%	24.64%

E.2 Representativeness of the Linked Samples

The military height data used in this paper differ from those typically used in the anthropometric history literature. While this literature generally takes random samples of the height data available in any particular source (leaving, in a military enlistment sample, the military enlistment decision as the only relevant point of sample selection), I limit my samples to the subset of these records that could also be linked to census records. The introduction of this second selection mechanism is necessary in order to gather covariates for comparison to the population as a whole and thus for estimation of the sample-selection model; but it may induce additional and potentially problematic bias.

Bias from non-representative linking can take two forms. First, I determine whether correcting for sample-selection bias has a meaningful effect on the trends in stature by comparing the trends corrected for selection on both observable and unobservable characteristics (which should represent the population trend in heights) to those corrected only for selection on observable characteristics (representing the conventional approach in the anthropometric history literature). If the sample-selection model corrects for selection both into military service and into census linking, then any difference in trends may be due to selection into census linking and not into military enlistment. Such bias would not be present in most studies of historical heights (because they do not use linked samples), but I would erroneously conclude that sample-selection bias existed in the height samples. Second, it is possible that the sample-selection correction would not properly correct for selection both into military enlistment and census linking. Mroz (2015) shows that studying two types of selection in a single index model has the potential to exacerbate any sample-selection bias, leading to estimates with even greater bias than those from a naive approach that ignores selection altogether.

Due to the possibly severe consequences of selection into census linkage, it is important to determine empirically whether such selection is likely to be present. As with the sample-selection issue that motivates this paper, selection into linkage is only problematic if it varies over cohorts. Fortunately, unlike the sample-selection problem, in which the outcome of interest is observed only for the selected sample, it is possible to directly test for selection of this type because the outcome variable (height) can be observed for both the selected (successfully census-linked) and unselected (failed to link to the census) samples. In order to test for sample-selection bias induced by selection into census-linking, I collected data on a random sample of Regular Army enlisters for the 1847–1860 cohorts without any attempt at linkage to the census. Similarly, I collected from the Union Army project information on enlisters without regard to linkage. Comparing the distributions and trends in heights of the linked sample and the unlinked sample (which represents the population of military enlisters as a whole rather than only those who could not be linked) makes it possible to determine whether problematic sample-selection bias is likely to exist. Throughout this analysis, I do not restrict attention to those living in non-seceding regions because this restriction was not imposed in linkage.

Table E.2 presents regressions comparing the trends in heights of the linked and unlinked (that is, representative of the whole enlisting population regardless of census linking) samples. Each column of this Table presents the results of two specifications. The first regresses heights on birth year indicators, measurement age indicators, and an indicator for being in the linked sample. The coefficient on the linked indicator is presented in Table E.2. This tests whether the linked and unlinked trends differ in level. The second regression adds interactions of the linkage indicator and the cohort indicators. The results of a χ^2-test of joint significance of these trends—which is a test of whether the trends in height of the linked and unlinked differ from one another—are also presented in Table E.2. Results of these regressions show statistically significant evidence of positive selection into linkage on the basis of height. There is not, however, any indication of a statistically significant difference in trends. Figure E.1 replicates this analysis graphically by plotting the trends in average heights over birth cohorts in both the linked and the unlinked groups. While differences in levels are evident between the linked and unlinked trends, the trends themselves are visually quite similar.

It therefore appears that the only difference between the linked and unlinked trends is in level, with the linked taller than the population of military enlisters as a whole; that is, there exists selection into linkage, but it is cohort-invariant. The presence of positive selection into census linking on the basis of height is unsurprising, as census linkage is likely to favor those who provide accurate information in a number of sources, and who are therefore likely to be better educated (Ferrie, 1996), and thus also likely to be taller. This difference suggests that one should be cautious in interpreting the level of the trends corrected for selection on both observable and unobservable characteristics. However, I find no reason to believe that the trend itself is unrepresentative of the population of military enlisters, and therefore conclude that the

Table E.2: Regressions of selection into linkage

	(1)	(2)	(3)
Cohorts	'32–'46	'47–'60	'32–'60
Variables	UA	RA	UA & RA
Linked	0.149***	0.192***	0.169***
	(0.049)	(0.074)	(0.049)
Observations	11,729	5,135	16,864
χ^2-Test of Birth Year FE × Linked	10.99	12.72	23.11

Significance levels: *** p<0.01, ** p<0.05, * p<0.1
Notes: Dependent variable is height, measured in inches. Truncated regression is performed to account for minimum height requirements with a truncation point of 64 inches. All specifications include measurement-age and birth-year dummy variables. Standard errors are clustered by image for the unlinked sample. The sample includes linked and unlinked members of the Regular Army and Union Army. UA denotes the Union Army. RA denotes the Regular Army. The coefficients on linked are from a regression without interactions. The statistics on the interactions are from a separate regression with interactions.

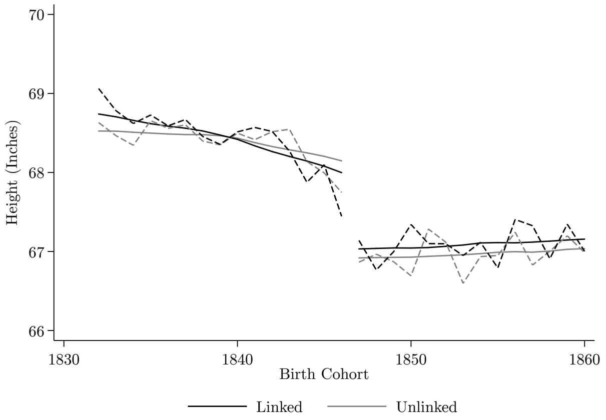


Figure E.1: Height trends of linked and unlinked samples

Note: These trends compare enlisters who could be linked to census data (the "Linked") to a random sample collected without respect to linking (the "Unlinked"). Tests for the statistical significance of differences between the trends are presented in Tables E.2.

correction for sample-selection is informative regarding the trend in heights of the population. To put it briefly, any bias from census linking should be captured by the intercept.

To further explore differences between the linked and unlinked military samples, I gathered a number of covariates from the military enlistment records for both the linked and unlinked. As these are taken from military records rather than from census records located through linkage (as are the covariates used in the main analysis), these covariates are not necessarily comparable to those discussed in the main text. The covariates collected are region of birth, year of birth, year of enlistment, and occupation (categorized using the same categories as above). I also created measures of name complexity and length separately for first name and surname. Name complexity is measured by the scrabble score, which is increasing in the length and complexity of a name (Biavaschi, Giulietti, and Siddique, 2017). These measures are included in order to capture the fact that individuals with unique names are generally easier to match.

In Table E.3 I study whether the sample is balanced on the basis of these covariates. In particular, I present the results of a number of regressions of the form

$$x_i = \varsigma_0 + \varsigma_1 \ell_i + \nu_i,$$

where x_i is some covariate and ℓ_i is an indicator for being in the linked sample. Cells of Table E.3 present estimates of ς_1, which is the degree to which a particular covariate is overrepresented in the linked sample relative to the population of enlisters as a whole. For example, Northeasterners make up 3 percentage points less of the linked Union Army sample than the population of the Union Army. Overall, there are statistically significant differences between the linked and unlinked samples on the basis of name length and complexity, in terms of region of birth (generally under-representing the Northeast and over-representing the Midwest), and on the basis of occupation.

To correct for these imbalances, I compute weights in order to correct for selection into census linkage on the basis of these observable characteristics. In particular, I estimate a probit model for selection into linkage using Cosslett's (1981) likelihood, and use the results to compute inverse conditional linkage probabilities by which to weight.[38] I first reproduce the above analysis of the trends in heights of the linked and unlinked, weighting the linked samples by the inverse linkage probability. The results are presented in Table E.4 and Figure E.2. The differences in level between the heights of the linked and unlinked are smaller than in the unweighted equivalents, suggesting that some of the level differences are due to differences in these

[38] I omit the top one percent and bottom one percent of the sample, in terms of weights, in order to avoid having the results be driven by such outliers receiving too much weight.

Table E.3: Balancing tests for selection into linkage

	(1)	(2)
Cohorts	1832–1846	1847–1860
Dep. Variable	Union Army	Regular Army
Name		
Surname Scrabble Score	0.023	0.036
	(0.068)	(0.128)
First Name Scrabble Score	−0.052	−1.922***
	(0.073)	(0.107)
Surname Length	0.008	0.124***
	(0.029)	(0.047)
First Name Length	0.023	−0.806***
	(0.032)	(0.047)
Region		
Northeast	−0.030***	−0.038**
	(0.009)	(0.015)
Midwest	0.010	0.061***
	(0.009)	(0.013)
South	0.017***	−0.024**
	(0.005)	(0.010)
Birth Year	−0.346***	0.623***
	(0.068)	(0.161)
Enlistment Year	0.003	−0.000
	(0.024)	(0.238)
Occupation		
Farmer	0.022***	0.031***
	(0.008)	(0.012)
Professional	−0.001	0.010***
	(0.003)	(0.003)
Clerical	0.001	0.026***
	(0.003)	(0.009)
Skilled and Artisan	−0.006	0.048***
	(0.006)	(0.013)
Semi-Skilled and Operative	−0.002	−0.089***
	(0.003)	(0.010)
Unskilled	−0.011**	−0.035**
	(0.005)	(0.014)
Unproductive	−0.002	0.009***
	(0.002)	(0.002)
Observations (Linked)	5,412	2,340
Observations (Unlinked)	7,134	2,956

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Notes: Each cell represents the coefficient from a regression of the dependent variable in the first column on an indicator for linkage. The "unlinked" group is not composed only of the unlinked, but is a random sample of the population of enlisters, so the coefficients are to be interpreted as the difference in the dependent variable between the linked and the population of enlisters. All regressions cluster standard errors on image, and are weighted to account for stratification; for the Regular Army, weighting is also performed to make the enlistment years of the whole population similar to that of the linked.

observable characteristics. Importantly, any differences in trend between the linked and unlinked are small and statistically insignificant. In all analyses (except where indicated otherwise), I weight by this inverse linkage probability in order to correct for non-representative linkage.

Table E.4: Selection into linkage, weighted by inverse conditional linkage probability

	(1)	(2)	(3)
Cohorts	'32–'46	'47–'60	'32–'60
Variables	UA	RA	UA & RA
Linked	0.120**	0.080	0.089*
	(0.049)	(0.084)	(0.054)
Observations	11,624	5,064	16,688
χ^2-Test of Birth Year FE × Linked	11.17	12.56	22.50

Significance levels: *** p<0.01, ** p<0.05, * p<0.1
Notes: Dependent variable is height, measured in inches. Truncated regression is performed to account for minimum height requirements with a truncation point of 64 inches. All specifications include measurement-age and birth-year dummy variables. Standard errors are clustered by image for the unlinked sample. The sample includes linked and unlinked members of the Regular Army and Union Army. UA denotes the Union Army. RA denotes the Regular Army. The coefficients on linked are from a regression without interactions. The statistics on the interactions are from a separate regression with interactions.

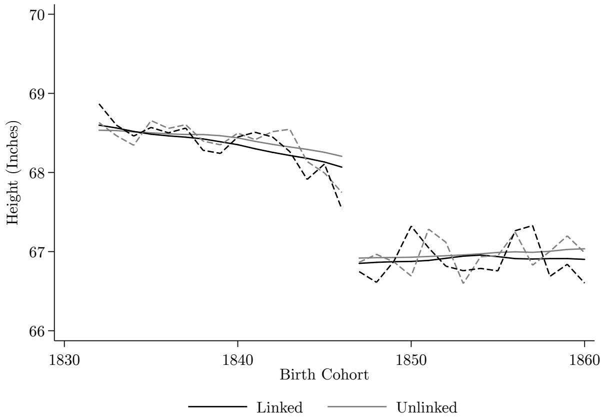


Figure E.2: Height trends of linked and unlinked samples, weighting by inverse conditional linkage probability

Note: See Figure E.1. These graphs are weighted by inverse conditional linkage probability.

F Estimation Details

F.1 Adapting Klein and Spady's (1993) Estimator

Klein and Spady (1993) develop a method to estimate a model of the form of equation (3) in a simple random sample of a population. The sample used in the present research differs from such a sample in two ways. First, the sample is a choice-restricted sample with a supplementary sample (as discussed by Cosslett, 1981), rather than a simple random sample. Second, the sample is composed of two distinct subsamples that are sampled separately: the 1832–1846 cohorts and their supplementary sample, and the 1847–1860 cohorts and their supplementary sample.

Klein and Spady's (1993) estimator is a maximum likelihood estimator, where the likelihood function takes the usual form for binary choice models, and where the function $G(\cdot)$ is estimated using the leave-one-out Nadaraya (1964) and Watson (1964) (NW) estimator. To adapt this estimator to the sample available in the present context, it is useful to define the variable s_{it} as

$$s_{it} = \begin{cases} 0 & \text{for members of the 1847–1860 cohorts} \\ 1 & \text{for members of the 1832–1846 cohorts} \end{cases}$$

and $\psi_{it} = \hat{\alpha}_t + \mathbf{x}'_{it}\hat{\beta}_k + \mathbf{z}_{it}{}'\hat{\delta}_k$. Let $\lambda(\cdot)$ denote the probability density function of ψ_{it}. I then use Bayes's Theorem and the law of total probability to write $P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{z}_{it}; t)$ in terms of objects that can be learned from the available sample:

$$\begin{aligned}
P(y_{it} = 1|\psi_{it}) &= P(y_{it} = 1|\psi_{it}, s_{it} = 0)P(s_{it} = 0|\psi_{it}) + P(y_{it} = 1|\psi_{it}, s_{it} = 1)P(s_{it} = 1|\psi_{it}) \\
&= \frac{\lambda(\psi_{it}|y_{it} = 1, s_{it} = 0)P(y_{it} = 1|s_{it} = 0)}{\lambda(\psi_{it}|s_{it} = 0)}P(s_{it} = 0|\psi_{it}) \\
&\qquad + \frac{\lambda(\psi_{it}|y_{it} = 1, s_{it} = 1)P(y_{it} = 1|s_{it} = 1)}{\lambda(\psi_{it}|s_{it} = 1)}P(s_{it} = 1|\psi_{it}) \\
&= \frac{\lambda(\psi_{it}|y_{it} = 1, s_{it} = 0)P(y_{it} = 1|s_{it} = 0)}{\lambda(\psi_{it}|s_{it} = 0)}\frac{\lambda(\psi_{it}|s_{it} = 0)P(s_{it} = 0)}{\lambda(\psi_{it})} \\
&\qquad + \frac{\lambda(\psi_{it}|y_{it} = 1, s_{it} = 1)P(y_{it} = 1|s_{it} = 1)}{\lambda(\psi_{it}|s_{it} = 1)}\frac{\lambda(\psi_{it}|s_{it} = 1)P(s_{it} = 1)}{\lambda(\psi_{it})} \\
&= \frac{\lambda(\psi_{it}|y_{it} = 1, s_{it} = 0)P(y_{it} = 1|s_{it} = 0)P(s_{it} = 0)}{\lambda(\psi_{it}|s_{it} = 0)P(s_{it} = 0) + \lambda(\psi_{it}|s_{it} = 1)P(s_{it} = 1)} \\
&\qquad + \frac{\lambda(\psi_{it}|y_{it} = 1, s_{it} = 1)P(y_{it} = 1|s_{it} = 1)P(s_{it} = 1)}{\lambda(\psi_{it}|s_{it} = 0)P(s_{it} = 0) + \lambda(\psi_{it}|s_{it} = 1)P(s_{it} = 1)}.
\end{aligned} \tag{F.1}$$

Every portion of equation (F.1) can either be non-parametrically estimated from the available data, or can be deduced from aggregate statistics. The distribution of the linear index in the military service sample, $\lambda(\psi_{it}|y_{it}=1,\cdot)$, can be learned from each of the choice-restricted subsamples. The distribution of this same index in the population, $\lambda(\psi_{it}|\cdot)$ can be learned from each of the supplemental samples. The aggregate enlistment probabilities, $P(y_{it}=1|\cdot)$ are given in Online Appendix D. Finally, $P(s_{it}=0)$ and $P(s_{it}=1)$ can be learned from aggregate data.

In order to discuss the estimation procedure, it is convenient to define an indicator for whether individual i is a member of the choice-restricted or supplementary sample. Define

$$\tilde{y}_{it} = \begin{cases} 1 & \text{for members of the choice-restricted sample} \\ 0 & \text{for members of the supplementary sample} \end{cases}.$$

Observations for which $\tilde{y}_{it}=1$ make it possible to learn the terms in equation (F.1) that are conditional on $y_{it}=1$, while those for which $\tilde{y}_{it}=0$ make it possible to learn the terms that do not condition on $y_{it}=1$ and which are not learned from aggregate data. I adapt the NW estimator and estimate equation (F.1) with the statistic

$$\hat{G}(\psi_{it}) = \widehat{P(y_{it}=1|\psi_{it})} = \frac{\left\{\begin{array}{r} P(y_{it}=1|s_{it}=0)P(s_{it}=0)\left[\sum_j \tilde{y}_{j\tau}(1-s_{j\tau})\right]^{-1}\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\tilde{y}_{j\tau}(1-s_{j\tau}) \\ +\, P(y_{it}=1|s_{it}=1)P(s_{it}=1)\left[\sum_j \tilde{y}_{j\tau}s_{j\tau}\right]^{-1}\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\tilde{y}_{j\tau}s_{j\tau} \end{array}\right\}}{\left\{\begin{array}{r} P(s_{it}=0)\left[\sum_j (1-\tilde{y}_{j\tau})(1-s_{j\tau})\right]^{-1}\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})(1-s_{j\tau}) \\ +\, P(s_{it}=1)\left[\sum_j (1-\tilde{y}_{j\tau})s_{j\tau}\right]^{-1}\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})s_{j\tau} \end{array}\right\}}, \tag{F.2}$$

where ω is a bandwidth and $K(\cdot)$ is a kernel function. For estimation, I use a Gaussian kernel. I use this statistic rather than estimating the model separately for each of the two subpopulations in order to ensure that the estimates of $G(\cdot)$ and the linear index ψ_{it} created by this procedure are comparable across subpopulations and can thus be used in the selection model. Thus, rather than simply maximizing the likelihood (F.5), defined below, separately for each sample, I maximize the sum of the likelihoods for each of the samples, estimating the choice probabilities by (F.2); to accommodate the separate sampling of the two groups of birth cohorts, I weight the likelihood, so that the final likelihood function becomes

$$\mathfrak{L} = L^{\{s_{it}=0\}}\left(\begin{bmatrix}\alpha' & \beta' & \delta'\end{bmatrix}'\right) \times \frac{P(s_{it}=0)}{\Xi(s_{it}=0)} + L^{\{s_{it}=1\}}\left(\begin{bmatrix}\alpha' & \beta' & \delta'\end{bmatrix}'\right) \times \frac{P(s_{it}=1)}{\Xi(s_{it}=1)}, \tag{F.3}$$

where $L^{\{s_{it}=j\}}(\cdot)$ is the likelihood function (F.5) for sample $s_{it} = j$, $j \in \{0,1\}$, $P(s_{it} = j)$ is the population proportion, and $\Xi(s_{it} = j)$ is the sample proportion.

Klein and Spady (1993) also suggest the use of a trimming function, though they report that the particular function is empirically unimportant. When estimating, I first assume that $G(\cdot)$ is normally distributed and estimate a probit model. I then compute a kernel density estimate of the estimated value of $\alpha_t + \mathbf{x}_{it}\beta_k + \mathbf{z}_{it}{}'\delta_k$, and trim from the estimation (that is, exclude from the likelihood function) individuals for whom the density falls below 0.005.

F.2 Cosslett's (1981) Likelihood

The likelihood function also requires adaptation to the structure of the sample. Before proceeding further, it is useful to introduce some additional notation:

- $S = \{0,1\}$: the set of options for each individual in the sample, where 0 denotes never enlisting and 1 denotes enlisting in the military at some point in the lifetime
- N: the number of individuals in the choice-restricted (military enlister) sample
- N_0: the number of individuals in the supplementary (general population) sample
- $H_s = \frac{N_0}{N}$
- Q_j: the proportion of each choice $j \in S$ in the population
- H_j: the proportion of each choice $j \in S$ in the choice-restricted sample
- $\eta_j = \frac{H_j}{Q_j}$

I denote the choice-restricted sample by $i \in \{1, \ldots, N\}$ and the supplementary sample by $i \in \{N+1, \ldots, N+N_0\}$.

Models of this type are studied by Cosslett (1981), who provides a maximum-likelihood estimator.

$$
\begin{aligned}
\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\delta} \end{bmatrix} &= \underset{[\,\alpha' \ \beta' \ \delta'\,]'}{\arg\max}\, L\left([\,\alpha' \ \beta' \ \delta'\,]'\right) \\
&= \underset{[\,\alpha' \ \beta' \ \delta'\,]'}{\arg\max} \sum_{i=1}^{N} \log \left\{ \frac{\eta_1 P(y_{it}=1|\mathbf{x}_{it},\mathbf{z}_{it};t)}{\left[\sum_{j\in S} \eta_j P(y_{it}=j|\mathbf{x}_{it},\mathbf{z}_{it};t)\right] + H_s} \right\} \\
&\qquad - \sum_{i=N+1}^{N+N_0} \log \left\{ \left[\sum_{j\in S} \eta_j P(y_{it}=j|\mathbf{x}_{it},\mathbf{z}_{it};t) \right] + H_s \right\}. \qquad \text{(F.4)}
\end{aligned}
$$

Since the choice-restricted sample contains only enlisters, $\eta_1 = \frac{1}{Q_1}$ and $\eta_0 = \frac{0}{Q_0} = 0$. Thus, the pseudo-log-likelihood function in expression (F.4) reduces to the following:[39]

$$
\begin{aligned}
&L\left([\,\alpha' \ \beta' \ \delta'\,]'\right) \\
&\quad = \sum_{i=1}^{N} \log \left\{ \frac{\eta_1 P(y_{it}=1|\mathbf{x}_{it},\mathbf{z}_{it};t)}{\eta_1 P(y_{it}=1|\mathbf{x}_{it},\mathbf{z}_{it};t) + H_s} \right\} - \sum_{i=N+1}^{N+N_0} \log \left\{ \eta_1 P(y_{it}=1|\mathbf{x}_{it},\mathbf{z}_{it};t) + H_s \right\}. \qquad \text{(F.5)}
\end{aligned}
$$

Finally, because Q_1 differs between the two groups of cohorts, I maximize the sum of equation (F.5) evaluated separately for the two subsamples, weighting the sums to account for the separate sampling of each group of birth cohorts, as described above.

F.3 The Gradient Matrix

When the Gaussian kernel is used, the gradient matrix of the estimated probability with respect to the first-stage coefficients $\Theta_1 = [\alpha', \beta', \delta']$ is given by

$$
\frac{\widehat{\partial P(y_{it}=1|\psi_{it})}}{\partial \Theta_{1\kappa}} = \frac{\widehat{P(y_{it}=1|\psi_{it})}}{E+F}(A+B-C-D)
$$

[39] Location and scale normalizations are required. To this end, I omit a constant and require that

$$
\begin{bmatrix} \alpha' & \beta' & \delta' \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \delta \end{bmatrix} = 1.
$$

Note that this implies that the coefficients are identified only up to sign; however, the function $G(\cdot)$ adjusts to ensure that the marginal effect of each covariate is of the appropriate sign.

where

$$
\begin{aligned}
A =& P(s_{it}=0)\left[\sum_j (1-\tilde{y}_{j\tau})(1-s_{j\tau})\right]^{-1} \sum_{j\neq i}\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\left(\frac{x_{i\kappa}-x_{j\kappa}}{\omega}\right) \\
&\times K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})(1-s_{j\tau}) \\
B =& P(s_{it}=1)\left[\sum_j (1-\tilde{y}_{j\tau})s_{j\tau}\right]^{-1} \sum_{j\neq i}\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\left(\frac{x_{i\kappa}-x_{j\kappa}}{\omega}\right) K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})s_{j\tau} \\
C =& P(y_{it}=1|s_{it}=0)P(s_{it}=0)\left[\sum_j \tilde{y}_{j\tau}(1-s_{j\tau})\right]^{-1} \sum_{j\neq i}\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\left(\frac{x_{i\kappa}-x_{j\kappa}}{\omega}\right) \\
&\times K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\tilde{y}_{j\tau}(1-s_{j\tau}) \\
D =& P(y_{it}=1|s_{it}=1)P(s_{it}=1)\left[\sum_j \tilde{y}_{j\tau}s_{j\tau}\right]^{-1} \sum_{j\neq i}\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\left(\frac{x_{i\kappa}-x_{j\kappa}}{\omega}\right) \\
&\times K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\tilde{y}_{j\tau}s_{j\tau} \\
E =& P(s_{it}=0)\left[\sum_j (1-\tilde{y}_{j\tau})(1-s_{j\tau})\right]^{-1} \sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})(1-s_{j\tau}) \\
F =& P(s_{it}=1)\left[\sum_j (1-\tilde{y}_{j\tau})s_{j\tau}\right]^{-1} \sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})s_{j\tau}
\end{aligned}
$$

and $x_{it\kappa}$ may be substituted by any element of $\mathbf{x}_{it}$, $\mathbf{z}_{it}$, or a cohort fixed effect (for estimating α_t) as appropriate.

F.4 Computing Partial Effects

The semi-elasticities presented in Table 3 are computed as follows. When a Gaussian kernel is used, differentiating expression (F.2) with respect to $x_{it\kappa}$ yields the partial effect for individual i for cohort t for a continuous covariate κ:

$$
\begin{aligned}
\frac{\partial \log[\widehat{P(y_{it}=1|\psi_{it})}]}{\partial x_{i\kappa}} &= \frac{1}{\widehat{P(y_{it}=1|\psi_{it})}} \\
\times \frac{\beta_\kappa}{\omega} \times & \left[\frac{-\left\{ \begin{aligned} &P(y_{it}=1|s_{it}=0)P(s_{it}=0)\left[\textstyle\sum_j \tilde{y}_{j\tau}(1-s_{j\tau})\right]^{-1} \textstyle\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\tilde{y}_{j\tau}(1-s_{j\tau}) \\ &\quad + P(y_{it}=1|s_{it}=1)P(s_{it}=1)\textstyle\sum_{j\neq i}\left[\textstyle\sum_j \tilde{y}_{j\tau}s_{j\tau}\right]^{-1} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\tilde{y}_{j\tau}s_{j\tau} \end{aligned} \right\}}{\left\{ \begin{aligned} &P(s_{it}=0)\left[\textstyle\sum_j (1-\tilde{y}_{j\tau})(1-s_{j\tau})\right]^{-1} \textstyle\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})(1-s_{j\tau}) \\ &\quad + P(s_{it}=1)\left[\textstyle\sum_j (1-\tilde{y}_{j\tau})s_{j\tau}\right]^{-1} \textstyle\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})s_{j\tau} \end{aligned} \right\}} \right. \\
& \qquad - \widehat{P(y_{it}=1|\psi_{it})} \\
\times & \left. \frac{-\left\{ \begin{aligned} &P(s_{it}=0)\left[\textstyle\sum_j (1-\tilde{y}_{j\tau})(1-s_{j\tau})\right]^{-1} \textstyle\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})(1-s_{j\tau}) \\ &\quad + P(s_{it}=1)\left[\textstyle\sum_j (1-\tilde{y}_{j\tau})s_{j\tau}\right]^{-1} \textstyle\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})s_{j\tau} \end{aligned} \right\}}{\left\{ \begin{aligned} &P(s_{it}=0)\left[\textstyle\sum_j (1-\tilde{y}_{j\tau})(1-s_{j\tau})\right]^{-1} \textstyle\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})(1-s_{j\tau}) \\ &\quad + P(s_{it}=1)\left[\textstyle\sum_j (1-\tilde{y}_{j\tau})s_{j\tau}\right]^{-1} \textstyle\sum_{j\neq i} K\left(\frac{\psi_{it}-\psi_{j\tau}}{\omega}\right)(1-\tilde{y}_{j\tau})s_{j\tau} \end{aligned} \right\}} \right]
\end{aligned} \tag{F.6}
$$

For a discrete covariate, the partial effect for individual i is calculated as

$$
\frac{\Delta \log[\widehat{P(y_{it}=1|\psi_{it})}]}{\Delta x_{it\kappa}} = \frac{1}{\widehat{P(y_{it}=1|\psi_{it})}} \times \left[\widehat{P\left(y_{it}=1|\psi_{it}\right)}\Big|_{x_{it\kappa}=1} - \widehat{P\left(y_{it}=1|\psi_{it}\right)}\Big|_{x_{it\kappa}=0} \right]; \tag{F.7}
$$

that is, the difference between the estimated probability of enlistment for each of the two possible values of $x_{it\kappa}$.

The estimates presented in Table 3 are averages of expressions (F.6) and (F.7) across the supplementary samples, thus representing the average marginal effect of the covariate on enlistment in the whole population.

G Results with Exact Matches Only

To address concerns over the role of false positives in linking, I repeat the main results limiting the sample to matches for which it is possible to be certain or nearly certain that there are no false positives. For the Union Army data, no data must be removed because the hand linking by genealogists removes the concern over false positives. For the Regular Army data, I limit the data to individuals whose census and enlistment characteristics (from the enlistment where height data are taken, which is generally the first enlistment) met the following criteria: absolute difference in age-implied birth years not more than one year, same place of birth, and no difference in name, except for double letters and abbreviations (e.g., "William" and "Wm" were allowed to match). These are much stricter requirements than those used for the main results, which allow up to a four-year difference in the age-implied birth year and allow for matching of similar but not identical first and last names. As might be expected from the more stringent requirements imposed on linking, the sample size for the Regular Army is reduced considerably by these refinements. In particular, the number of enlisters observed in the Regular Army data who are "exact" matches is 896, as opposed to 2,214 satisfying the standard criteria used in the main results.

As shown in Figures G.1 and G.2, the main results are largely unaffected by this restriction. Figure G.1 shows a similar pattern to Figure A.1, with a more negative estimated selection function in the later birth cohorts. Similarly, panel G.2(a) shows patterns similar to those of panel 2(a), with the existence of a decline in the corrected trend that is smaller than the decline in the uncorrected trend. However, the Regular Army data in this panel clearly show the much higher variability induced by the smaller samples. Finally, panel G.2(b) shows the confidence interval of the estimated decline. As in panel 2(b), it is possible to reject the null hypothesis of no decline in stature in the period in question ($\chi^2_{28} = 44.49$, $p = 0.025$). Unlike panel 2(b), it is not possible to reject the null hypothesis of no net decline in stature, as the confidence interval of the decline in stature to 1860 includes zero ($\chi^2_1 = 2.04$, $p = 0.153$); but this is not due to a very different estimate of the magnitude relative to the main results (the net decline is estimated as 0.591 inches in panel G.2(b) as opposed to 0.643 inches in panel 2(b)); the difference is largely due to the much higher standard error of this estimate (0.414 as opposed to the main result's standard error of 0.305). Similarly, although it was possible in the main results to reject the null hypothesis of equality of the estimated decline to 1860 between the fully corrected trends and the trends corrected only for selection on observables, it is not possible to do so in this case. Again, this is due to larger standard errors stemming from the loss of a considerable quantity of data rather than due to fundamental changes in the estimates: the difference in estimated declines was 0.647 with a standard error of 0.276 in the main results; after the limitation to exact matches, it is 0.685

with a standard error of 0.398. Thus, I conclude that the main results are not driven by false positives in linkage.

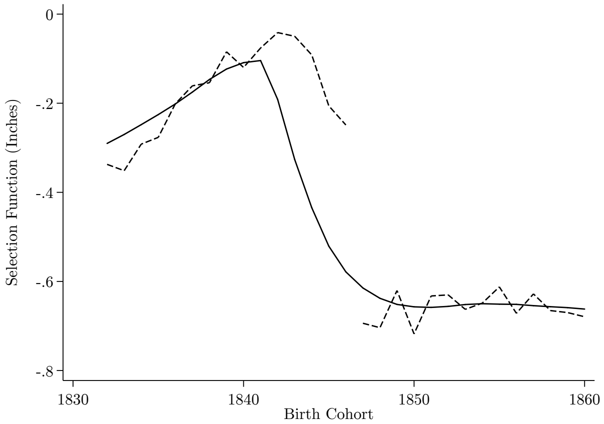
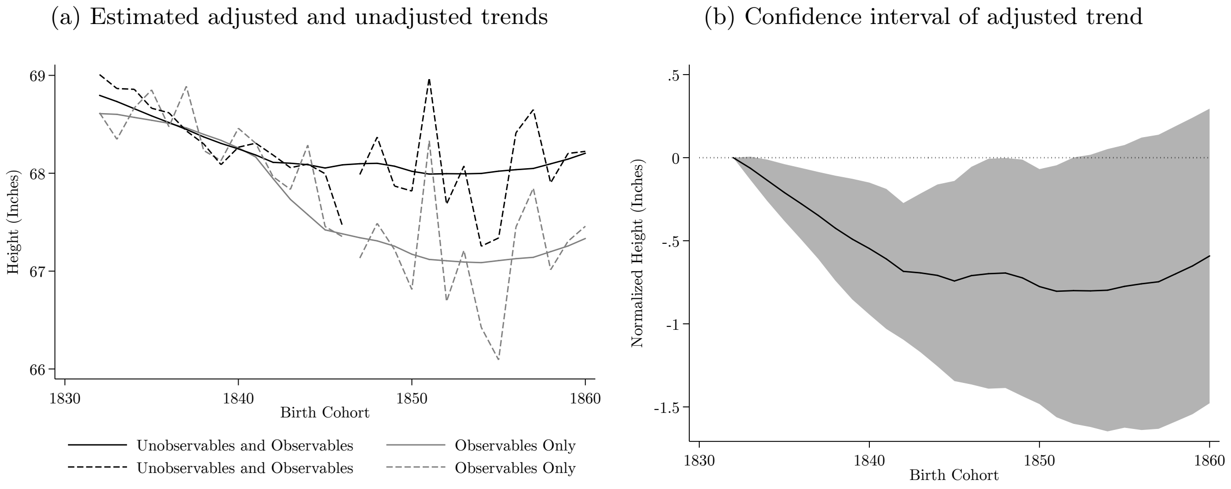


Figure G.1: Estimated $\Omega(\cdot)$ function by birth cohort

Note: This graph plots the coefficients from a regression of the estimated function $\hat{\Omega}(\hat{\alpha}_t + \mathbf{x}_{it}'\hat{\beta}_k + \mathbf{z}_{it}'\hat{\delta}_k)$ on birth year indicators, weighting by inverse enlistment probability (in the dashed line), as well as these coefficients smoothed over birth cohorts (in the solid line).

That the main result that the height decline survives the correction for selection on unobservables is not surprising. This result is largely driven by the Union Army data, which are not in danger of false positives because they are hand matched. Thus, removing the inexact matches from the Regular Army data would be unlikely to affect the results in the Union Army data.


(a) Estimated adjusted and unadjusted trends

(b) Confidence interval of adjusted trend

Figure G.2: Trends in average stature

Note: Panel G.2(a) plots four trends in average height by birth cohort. The first, in solid black (labeled "Unobservables and Observables"), incorporates the correction for selection on both observables and unobservables, and smoothed over birth cohorts; the second, in dashed black, is its unsmoothed analog. The third, in solid gray (labeled "Observables Only"), is corrected only for truncation and selection on observables, and is smoothed over birth cohorts; the fourth, in dashed gray, is its unsmoothed analog. The unsmoothed trends for the 1832–1846 cohorts are based on the Union Army data, while those for the 1847–1860 cohorts are based on the Regular Army data. Panel G.2(b) presents bootstrap 95 percent pointwise confidence intervals clustered at the county level for the smoothed trend in average stature incorporating the correction for selection on both observables and unobservables (the solid black line in panel G.2(a)).

H Results for Alternative Specification and Data Set

In addition to the Regular Army data for the 1847–1860 cohorts, used in the main text of the paper, I also collected Regular Army data for the birth cohorts of 1832–1846. In this Appendix, I present a specification that estimates the model using only Regular Army data (for the 1832–1860 cohorts) instead of the combination of the Union Army and Regular Army data, as used in the main text of the paper. That is, the data for the 1847–1860 cohorts are the same, but the data for the 1832–1846 cohorts are different.[40] These results are referred to throughout this appendix as those with *Different Data.* I also present results using the same data as in the main text, but replacing Lincoln's vote share with Douglas's vote share in 1860 and Buchanan's vote share in 1856. I refer to these results as those with *Different Variables.*

H.1 Summary Statistics for Regular Army, 1832–1846 Cohorts

Table H.1 presents the analog of Table 1 for the 1832–1846 Regular Army data, and Table H.2 presents the analog of Table 2. Note that the supplementary samples in these two tables are identical to those for the 1832–1846 cohorts in the main text because they represent the same birth cohorts. Figure H.1 presents the

[40] Details of the census linkage used for construction of this sample are available on request.

analog of Figure 1, exhibiting a similar pattern of heaping and left-censoring as the Regular Army sample for the 1847–1860 cohorts.

Table H.1: Distribution of observations by census

		1832–1846	
		(1)	(2)
Census	Cohorts	RA (CR)	Supp.
1850	1832–1841	984	5,879
1860	1842–1846	772	2,807
Total		1,756	8,686

Notes: Each cell reports the number of individuals in the sample indicated in the column header with data taken from the census indicated in the row. Samples are restricted to cover individuals with data on all individual-level variables. Abbreviations are as follows: RA is Regular Army, CR is choice-restricted sample, Supp. is supplementary sample.

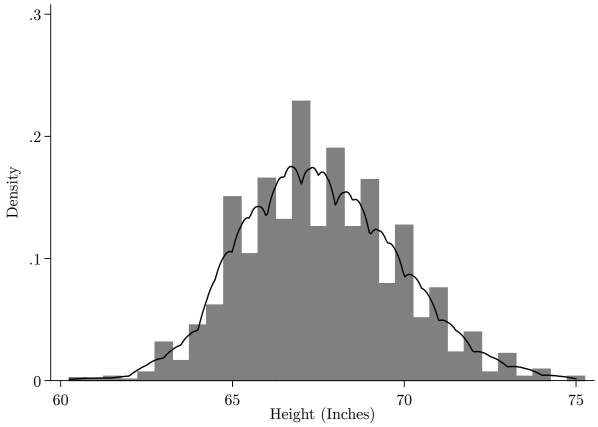


Figure H.1: Height distributions

Note: This figure presents a histogram (with a bin width of 0.5 inches) and a kernel density estimate of the height distribution for the height data for individuals from the 1832–1846 birth cohorts in the Regular Army.

H.2 Selection into Military Service

Table H.3 presents the results of two alternative sets of estimation of the binary choice model (equation 3) in a manner analogous to Table 3. Column (1) presents the estimates of the coefficients β_k and δ_k using the same data as the benchmark results in the main text of the paper but with the Douglas and Buchanan vote shares instead of the Lincoln vote share for identification. Column (2) presents the associated semi-elasticities. This set of results exhibits magnitudes of the relationship of the vote shares for Douglas and Buchanan that are comparable in interpretation (though not in magnitude) to that of Lincoln's vote share

Table H.2: Summary statistics

	1832–1846		
	(1)	(2)	(3)
Variable	RA (CR)	Supp.	Diff.
Individual or Household Variables			
Height (in)	67.496		
Household Owns Property	0.594	0.687	−0.090***
Household Real Property ($1,000)	2.008	2.297	−0.272
Related to Head of Household	0.777	0.863	−0.085***
Household Size	7.242	7.419	−0.177**
Attended School	0.572	0.648	−0.074***
Household Occupation			
Farmer	0.350	0.520	−0.167***
Professional	0.052	0.038	0.015**
Clerical	0.091	0.066	0.024*
Skilled and Artisan	0.239	0.185	0.055***
Semi-Skilled and Operative	0.076	0.058	0.017*
Unskilled	0.095	0.065	0.029***
Farm Labor	0.006	0.006	0.000
Unproductive	0.091	0.062	0.026***
Birth Region			
Midwest	0.213	0.287	−0.073***
Northeast	0.697	0.541	0.162***
South	0.090	0.172	−0.089***
County Variables			
Fraction Urban	0.242	0.158	0.077***
Wheat Bushels per capita	5.650	5.882	−0.232
Milk Cows per capita	0.262	0.282	−0.020**
Swine per capita	0.606	0.972	−0.367***
Value of Agricultural Production per capita ($1,000)	0.045	0.048	−0.003**
Lincoln Vote Share (1860)	0.515	0.455	0.060***
Observations	1,674	8,535	

Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Notes: All Individual or Household Variables are binary unless indicated otherwise. Averages for the choice-restricted samples are weighted to correct for selection into linkage on the basis of observable characteristics. Standard deviations and standard errors are omitted for space. Sample sizes are the minimum of the column with observations for all variables. Abbreviations are as follows: RA is Regular Army, CR is a choice-restricted sample, and Supp. is a supplementary sample. Diff. is a difference.

in Table 3 for the Regular Army. Column (3) presents the estimates of the coefficients β_k and δ_k using the same specification as the benchmark results, but replacing the Union Army data for the 1832–1846 cohorts with the alternative Regular Army data set for the same cohorts. Column (4) presents the associated semi-elasticities. These columns show a small and statistically insignificant role for the vote share in the military enlistment decision;[41] the model is still identified based on the difference in coefficients between the two cohort groups.

H.3 Selection-Corrected Height Regressions

Results of estimation of equation (4) for alternative specifications are presented in Table H.4. Columns (1)–(4) use the Union Army data to represent the 1832–1846 cohorts, as in the benchmark specification, but base identification on the Buchanan and Douglas vote shares rather than on Lincoln's. Columns (5)–(8) base identification on Lincoln's vote share, and use the Regular Army to represent the 1832–1846 cohorts. Results are largely similar to those of the benchmark sample in Table 4. A key test provided in this Table is the overidentification test of columns (3) and (7). In the specification using Douglas's and Buchanan's vote shares, the vote share for Douglas enters marginally significantly before correction, but significance is lost and the coefficient decreases in magnitude after correction. The coefficient on Buchanan's vote share also decreases in magnitude after correction. In the specification using the Regular Army sample, excludability of the vote share is also supported by the lack of a statistically significantly coefficient.

H.4 Adjusted Trends in Height

The main results for the alternative specifications are presented in Figures H.2–H.4. When the Douglas and Buchanan vote shares are used instead of Lincoln's, the results are qualitatively identical to those of the benchmark specification. As a result, I do not discuss them further. When the Regular Army enlisters are used to represent the 1832–1846 cohorts, however, differences are present. The first main result—that the Antebellum Puzzle is robust to the correction for sample-selection bias—is present in this specification as well. A decline in average stature is present in the fully corrected trend, and although it is not possible to reject the null of no net decline in stature over the study period, it is possible to reject the null of no decline at all ($\chi^2_{28} = 54.65$, $p < 0.01$). It might be objected that with the Lincoln vote share not entering significantly into the military enlistment decision for this sample, there is insufficient power to correct for

[41] This estimation requires computation of Q_1 for this alternative Regular Army sample. Using methods similar to that for the 1847–1860 cohorts, I compute it as $Q_1 = \frac{90{,}201}{2{,}469{,}663} = 0.037$.

Table H.3: Binary choice model estimation

	Different Variables				Different Data			
	(1)		(2)		(3)		(4)	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
Cohorts	’32–’46	’47–’60	’32–’46	’47–’60	’32–’46	’47–’60	’32–’46	’47–’60
Variables	UA	RA	UA	RA	RA	RA	RA	RA
Individual or Household Variables								
Household Owns Property	0.092***	0.195***	0.254***	0.455***	0.146***	0.272***	−0.274***	0.451***
	(0.017)	(0.020)	(0.054)	(0.065)	(0.039)	(0.030)	(0.052)	(0.039)
Household Real Property (1,000)	−0.016***	−0.049***	−0.050***	−0.128***	0.006**	−0.019***	−0.008**	−0.035***
	(0.002)	(0.004)	(0.005)	(0.012)	(0.003)	(0.003)	(0.003)	(0.004)
Related to Head of Household	0.103***	−0.061***	0.263***	−0.187***	0.209***	−0.085***	−0.332***	−0.174***
	(0.022)	(0.019)	(0.054)	(0.069)	(0.049)	(0.029)	(0.064)	(0.064)
Household Size	0.018***	0.001	0.056***	0.002	0.014**	−0.001	−0.020**	−0.002
	(0.003)	(0.003)	(0.010)	(0.007)	(0.007)	(0.004)	(0.009)	(0.007)
Attended School	−0.187***	−0.157***	−0.634***	−0.540***	0.073**	−0.195***	−0.144***	−0.420***
	(0.017)	(0.016)	(0.048)	(0.067)	(0.036)	(0.022)	(0.050)	(0.052)
Household Occupation (Unproductive excluded)								
Farmer	−0.310***	−0.214***	−0.994***	−0.520***	0.155**	−0.176***	−0.313***	−0.321***
	(0.038)	(0.024)	(0.100)	(0.094)	(0.065)	(0.042)	(0.093)	(0.072)
Professional	−0.256***	−0.093**	−0.704***	−0.189***	−0.070	−0.006	0.103	−0.011
	(0.062)	(0.036)	(0.139)	(0.052)	(0.093)	(0.062)	(0.126)	(0.119)
Clerical	−0.343***	−0.147***	−0.927***	−0.279***	0.027	−0.056	−0.031	−0.097
	(0.046)	(0.028)	(0.081)	(0.041)	(0.082)	(0.048)	(0.104)	(0.075)
Skilled and Artisan	−0.287***	−0.089***	−0.819***	−0.206***	0.048	0.020	−0.069	0.038
	(0.040)	(0.020)	(0.082)	(0.043)	(0.068)	(0.038)	(0.091)	(0.071)
Semi-Skilled and Clerical	−0.300***	−0.127***	−0.817***	−0.249***	0.056	−0.019	−0.077	−0.034
	(0.050)	(0.025)	(0.099)	(0.039)	(0.088)	(0.045)	(0.109)	(0.079)
Unskilled	−0.282***	−0.096***	−0.767***	−0.202***	−0.075	0.042	0.119	0.081
	(0.046)	(0.024)	(0.086)	(0.043)	(0.081)	(0.043)	(0.113)	(0.087)
Farm Labor	−0.184	−0.126***	−0.520	−0.233***	−0.009	−0.041	0.015	−0.072
	(0.120)	(0.030)	(0.335)	(0.042)	(0.233)	(0.059)	(0.276)	(0.086)
Birth Region (South excluded)								
Midwest	0.391***	0.009	1.622***	0.024	−0.121	−0.141***	0.169	−0.251***
	(0.036)	(0.025)	(0.134)	(0.067)	(0.097)	(0.047)	(0.123)	(0.082)
Northeast	0.319***	0.015	1.010***	0.039	−0.081	−0.087*	0.121	−0.162*
	(0.037)	(0.020)	(0.084)	(0.057)	(0.094)	(0.048)	(0.128)	(0.087)
County Variables								
Fraction Urban	−0.166***	0.040	−0.529***	0.105	0.056	0.133***	−0.082	0.247***
	(0.051)	(0.031)	(0.158)	(0.086)	(0.096)	(0.049)	(0.121)	(0.080)
Wheat Bushels per capita	0.015***	−0.003**	0.048***	−0.009**	−0.002	−0.003	0.003	−0.005
	(0.002)	(0.001)	(0.004)	(0.004)	(0.003)	(0.002)	(0.004)	(0.003)
Milk Cows per capita	0.060	−0.062	0.190	−0.162	0.092	−0.029	−0.133	−0.054
	(0.064)	(0.067)	(0.209)	(0.180)	(0.144)	(0.100)	(0.196)	(0.188)
Swine per capita	0.051***	−0.080***	0.162***	−0.209***	0.209***	−0.151***	−0.303***	−0.281***
	(0.011)	(0.018)	(0.032)	(0.044)	(0.040)	(0.024)	(0.037)	(0.041)
Value of Agricultural Production per capita (1,000)	−0.029	−0.038	−0.092	−0.098	−0.022	−0.031	0.032	−0.058
	(0.545)	(0.451)	(1.628)	(1.246)	(1.172)	(0.679)	(1.478)	(1.148)
Lincoln Vote Share (1860)					0.016	−0.010	−0.023	−0.019
					(0.145)	(0.081)	(0.190)	(0.144)
Buchanan Vote Share (1856)	−0.091	−0.157***	−0.289	−0.408***				
	(0.064)	(0.055)	(0.203)	(0.143)				
Douglas Vote Share (1860)	0.113**	−0.026	0.362***	−0.067				
	(0.045)	(0.037)	(0.131)	(0.097)				
Observations	13,570	11,000	13,570	11,000	10,249	11,271	10,249	11,271

Significance levels: *** p<0.01, ** p<0.05, * p<0.1
Notes: Columns (1) and (3) present estimates of the coefficients β and δ from the binary choice model. Columns (2) and (4) present the average semi-elasticity of the impact of each variable on enlistment probability as implied by the estimates of columns (1) and (3), respectively. All specifications include cohort indicators. Standard errors are clustered at the county level. UA denotes Union Army. RA denotes Regular Army.

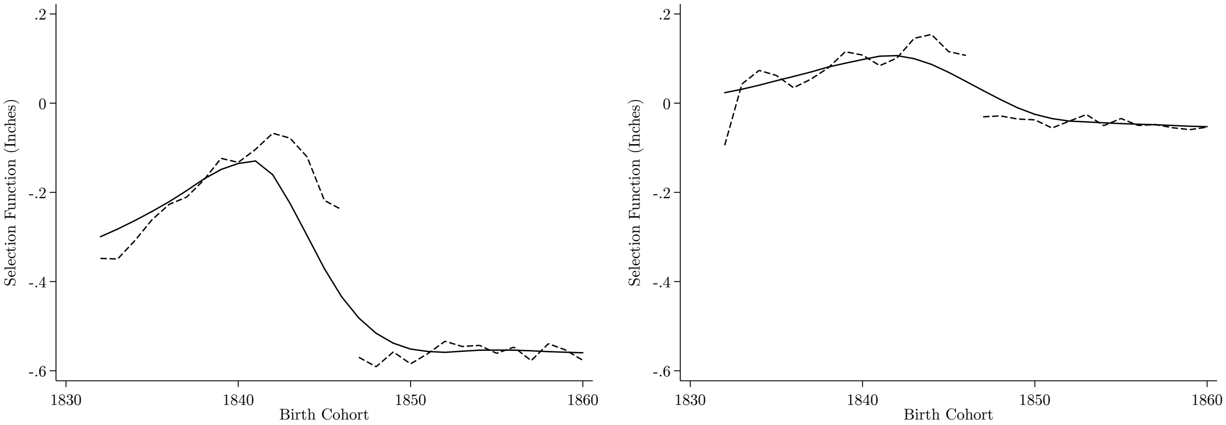
Table H.4: Height regressions

	Different Variables				Different Data			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables	Corr	Not	Corr	Not	Corr	Not	Corr	Not
Individual or Household Variables								
Household Owns Property	0.083	0.084	0.082	0.087	−0.097	−0.064	−0.103	−0.069
	(0.098)	(0.113)	(0.098)	(0.113)	(0.110)	(0.133)	(0.110)	(0.133)
Household Real Property (1,000)	−0.021*	−0.006	−0.020	−0.006	0.003	−0.002	0.004	−0.002
	(0.012)	(0.009)	(0.012)	(0.008)	(0.008)	(0.007)	(0.008)	(0.007)
Related to Head of Household	0.250**	0.305**	0.248**	0.299**	0.283**	0.279*	0.286**	0.278*
	(0.105)	(0.134)	(0.105)	(0.134)	(0.129)	(0.143)	(0.129)	(0.143)
Household Size	0.026*	0.023	0.027**	0.025	0.025	0.016	0.025	0.017
	(0.014)	(0.015)	(0.014)	(0.015)	(0.015)	(0.019)	(0.015)	(0.019)
Attended School	0.040	0.051	0.046	0.058	0.204**	0.230**	0.195**	0.217**
	(0.080)	(0.085)	(0.080)	(0.085)	(0.096)	(0.102)	(0.096)	(0.103)
Household Occupation (Unproductive excluded)								
Farmer	0.166	0.041	0.161	0.024	0.173	0.131	0.175	0.128
	(0.140)	(0.125)	(0.139)	(0.124)	(0.176)	(0.200)	(0.176)	(0.200)
Professional	0.189	0.010	0.183	−0.007	0.081	0.093	0.086	0.100
	(0.211)	(0.238)	(0.210)	(0.238)	(0.230)	(0.271)	(0.230)	(0.272)
Clerical	−0.023	−0.279	−0.034	−0.303	−0.099	−0.286	−0.093	−0.279
	(0.181)	(0.198)	(0.179)	(0.195)	(0.187)	(0.245)	(0.186)	(0.244)
Skilled and Artisan	−0.171	−0.336**	−0.173	−0.345**	−0.300*	−0.415**	−0.301*	−0.414**
	(0.148)	(0.150)	(0.147)	(0.149)	(0.164)	(0.203)	(0.164)	(0.202)
Semi-Skilled and Clerical	0.058	−0.127	0.051	−0.144	−0.093	−0.214	−0.082	−0.202
	(0.189)	(0.218)	(0.188)	(0.216)	(0.202)	(0.252)	(0.202)	(0.252)
Unskilled	0.220	0.081	0.226	0.080	0.048	−0.009	0.042	−0.013
	(0.179)	(0.191)	(0.179)	(0.191)	(0.194)	(0.230)	(0.194)	(0.229)
Farm Labor	−0.277	−0.535*	−0.281	−0.546*	−0.284	−0.436	−0.286	−0.439
	(0.231)	(0.287)	(0.232)	(0.288)	(0.254)	(0.315)	(0.254)	(0.316)
Birth Region (South excluded)								
Midwest	0.031	0.054	−0.038	−0.057	0.259	0.165	0.001	−0.100
	(0.169)	(0.174)	(0.171)	(0.181)	(0.167)	(0.196)	(0.212)	(0.251)
Northeast	−0.223	−0.302	−0.265	−0.380**	−0.022	−0.172	−0.295	−0.460*
	(0.171)	(0.186)	(0.166)	(0.187)	(0.155)	(0.195)	(0.212)	(0.256)
County Variables								
Fraction Urban	−0.530**	−0.617**	−0.569***	−0.691***	−0.546**	−0.595**	−0.533**	−0.586**
	(0.206)	(0.247)	(0.207)	(0.253)	(0.226)	(0.269)	(0.231)	(0.273)
Wheat Bushels per capita	0.003	0.007	0.003	0.007	−0.001	0.008	−0.002	0.007
	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.008)
Milk Cows per capita	0.028	0.077	0.011	0.029	0.035	−0.001	−0.062	−0.103
	(0.209)	(0.206)	(0.212)	(0.215)	(0.352)	(0.415)	(0.361)	(0.431)
Swine per capita	0.094*	0.072	0.083	0.066	0.186*	0.091	0.248**	0.137
	(0.051)	(0.052)	(0.053)	(0.053)	(0.108)	(0.090)	(0.112)	(0.092)
Value of Agricultural production per capita (1,000)	−4.252**	−4.127*	−4.284**	−4.501*	−4.408*	−5.344*	−4.683*	−5.609*
	(2.099)	(2.320)	(2.144)	(2.374)	(2.556)	(3.126)	(2.547)	(3.119)
Lincoln Vote Share (1860)							0.632*	0.653
							(0.341)	(0.399)
Buchanan Vote Share (1856)			−0.069	−0.343				
			(0.297)	(0.360)				
Douglas Vote Share (1860)			0.305	0.450*				
			(0.223)	(0.268)				
Observations	7,102	6,730	7,102	6,730	3,860	3,686	3,860	3,686

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Notes: Standard errors in parentheses. Dependent variable is height, measured in inches. All specifications include age-of-measurement, year-of-birth, and household occupation indicators. The selection-corrected specifications, indicated by the column header Corr, also include selection-correction function $\Omega(\cdot)$. The uncorrected specifications, indicated by the column header Not, correct for trunction with a truncation point of 64 inches. Standard errors are clustered at the county level. Specification (2) covers the 1832–1846 cohorts by the Union Army sample. Specification (3) covers the 1832–1846 cohorts by the Regular Army sample. the difference in sample sizes between columns is the result of the need to drop heights below 64 inches in the selection-corrected regressions when not correcting for sample-selection bias.

sample-selection bias. However, the results for the benchmark specification showed that much of the change in the patterns in stature due to the correction came at the change in samples. With the other source of identification still present in this sample, such changes should have been present but are not.


(a) Different Variables

(b) Different Data

Figure H.2: Estimated $\Omega(\cdot)$ function by birth cohort

Note: Each graph plots the coefficients from a regression of the estimated function $\hat{\Omega}(\hat{\alpha}_t+\mathbf{x}_{it}'\hat{\beta}_k+\mathbf{z}_{it}{}'\hat{\delta}_k)$ on birth year indicators, weighting by inverse enlistment probability (in the dashed line), as well as these coefficients smoothed over birth cohorts (in the solid line). Panel H.2(a) presents the graphs using the Union Army sample to represent the 1832–1846 birth cohorts with identification based on the vote shares for Buchanan and Douglas, while Panel H.2(b) presents the graphs using the Regular Army sample to represent the 1832–1846 birth cohorts and use the vote share for Lincoln for identification; both panels use the Regular Army sample to represent the 1847–1860 cohorts.

The second main result—that correcting for sample-selection bias leads to meaningful and statistically significant changes in the patterns in average stature—is not replicated in the Regular Army-only sample. Indeed, visual inspection of Figures H.2(b) and H.3(b) shows only a small influence of the selection correction. This result is consistent with the different sources of data between this and the benchmark specification. In the benchmark specification, the change in institution with the end of the Civil War was responsible for the change in selection. When the institution remains the same over cohorts no such pattern is present.

The correction for sample-selection bias also contributes to solving another puzzle in the data. In particular, although they are in principle (if issues of selection are ignored) meant to represent the same populations, the sample using the Union Army data to represent the 1832–1846 cohorts and the sample using the Regular Army data to represent these cohorts give very different estimates for the decline in average stature over time before correction. In particular, before correction, the estimated net declines are 1.29 inches in the benchmark sample and 0.94 inches in the alternative sample. After the correction, the estimated net declines are 0.64 inches and 0.42 inches, respectively. The declines to 1846 are 1.24 inches in the benchmark

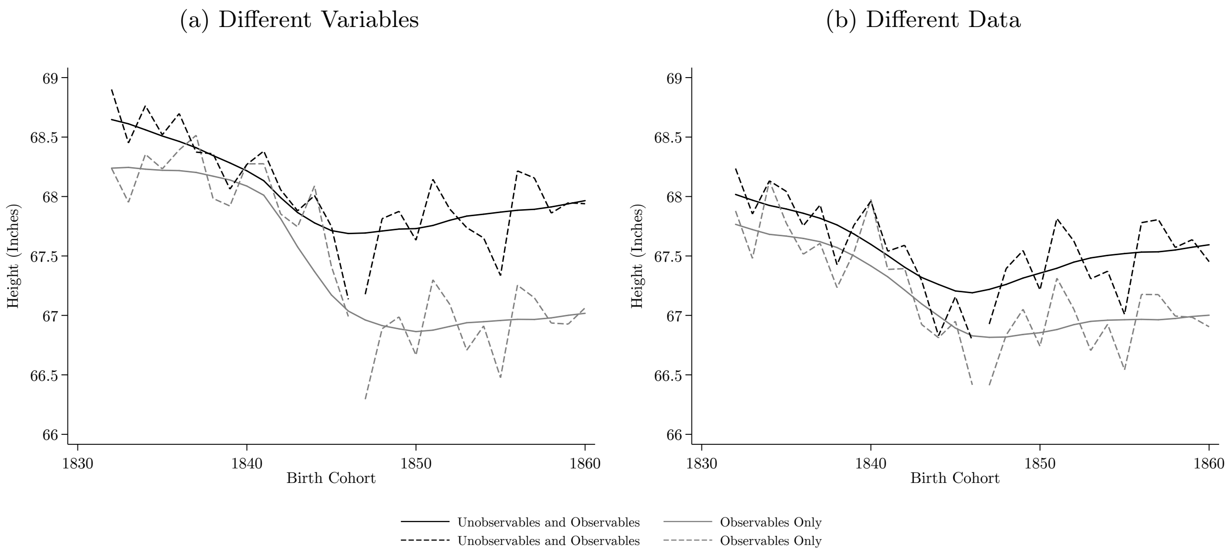


Figure H.3: Estimated adjusted trends by birth cohort

Note: Each graph plots four trends in average height by birth cohort. The first, in solid black (labeled "Unobservables and Observables"), incorporates the correction for selection on both observables and unobservables, and smoothed over birth cohorts; the second, in dashed black, is its unsmoothed analog. The third, in solid gray (labeled "Observables Only"), is corrected only for truncation and selection on observables, and is smoothed over birth cohorts; the fourth, in dashed gray, is its unsmoothed analog. The unsmoothed trends for the 1832–1846 cohorts are based on the Union Army data, while those for the 1847–1860 cohorts are based on the Regular Army data. Panel H.3(a) presents the graphs using the Union Army sample to represent the 1832–1846 birth cohorts with identification based on the vote shares for Buchanan and Douglas, while Panel H.3(b) presents the graphs using the Regular Army sample to represent the 1832–1846 birth cohorts and use the vote share for Lincoln for identification; both panels use the Regular Army sample to represent the 1847–1860 cohorts.

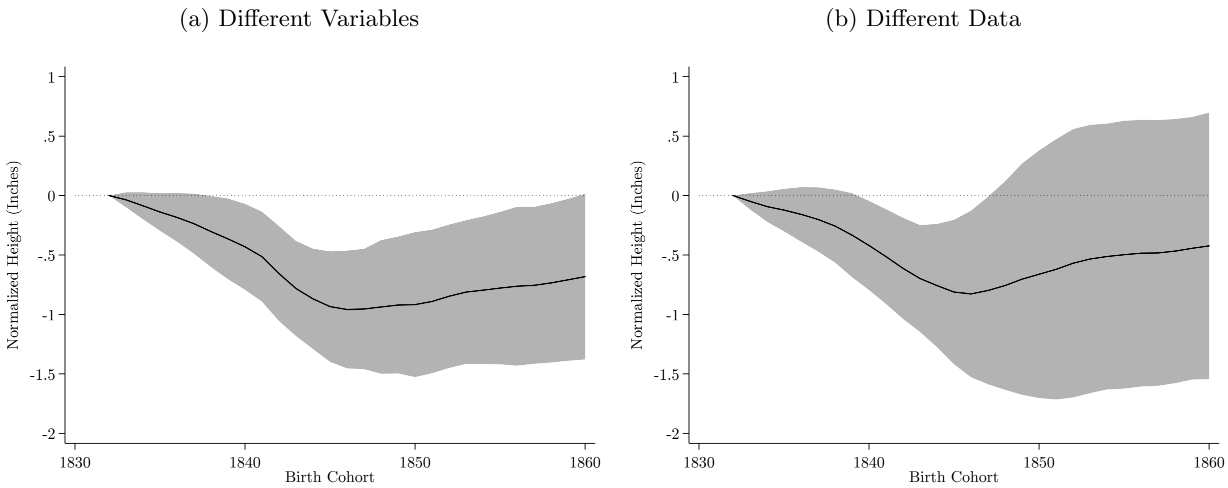


Figure H.4: Confidence intervals of adjusted decline

Note: Each graph presents bootstrap 95 percent pointwise confidence intervals clustered at the county level for the smoothed trend in average stature incorporating the correction for selection on both observables and unobservables (the solid black lines in Figure H.3). Panel H.4(a) presents the graphs using the Union Army sample to represent the 1832–1846 birth cohorts and bases identification on the Buchanan and Douglas vote shares, while Panel H.4(b) presents the graphs using the Regular Army sample to represent the 1832–1846 birth cohorts and bases identification on the vote share for Lincoln; both panels use the Regular Army sample to represent the 1847–1860 cohorts.

sample and 0.94 inches in the alternative sample before correction, and 0.94 and 0.83 inches, respectively, after correction. Thus, I conclude that different sample-selection bias between the different armies is at least partially responsible for the different implications of the two samples.

H.5 Cross-Sectional Patterns

Tables H.5 and H.6 present the results for the cross-sectional comparisons. The results here are similar to those of the temporal trends. Replacing the Lincoln vote share with the Douglas and Buchanan vote shares yields much the same results as does the benchmark specification, though the *p*-values for tests of significance of the effects of correcting for selection are slightly higher. Representing the 1832–1846 cohorts with Regular Army data shows a continued presence of the differences of interest, and only a small effect of correction.

Table H.5: Tests for differences in levels, regional decomposition

	Different Variables			Different Data		
	(1)	(2)	(3)	(4)	(5)	(6)
Region	Northeast	Midwest	South	Northeast	Midwest	South
Panel A: Observables Only						
Northeast	66.714***			66.731***		
	(0.220)			(0.203)		
Midwest	−0.563***	67.277***		−0.532***	67.263***	
	(0.169)	(0.258)		(0.148)	(0.235)	
South	−0.528**	0.034	67.243***	−0.572**	−0.040	67.303***
	(0.241)	(0.290)	(0.306)	(0.250)	(0.271)	(0.325)
Panel B: Unobservables and Observables						
Northeast	67.633***			67.225***		
	(0.442)			(0.483)		
Midwest	−0.386**	68.019***		−0.502***	67.726***	
	(0.175)	(0.454)		(0.157)	(0.487)	
South	−0.376*	0.010	68.009***	−0.417**	0.085	67.641***
	(0.200)	(0.231)	(0.535)	(0.206)	(0.254)	(0.547)
Panel C: $B - A$						
Northeast	0.919*			0.494		
	(0.471)			(0.504)		
Midwest	0.176	0.742		0.030	0.464	
	(0.156)	(0.520)		(0.124)	(0.495)	
South	0.152	−0.024	0.767	0.156	0.125	0.338
	(0.133)	(0.148)	(0.482)	(0.124)	(0.146)	(0.508)
Observations	3,254	3,107	741	2,229	1,189	442

Significance levels: *** $p< 0.01$, ** $p< 0.05$, * $p< 0.1$

Notes: In Panels A and B, the diagonals present the estimated mean heights in each region, corrected for minimum height requirements with a truncation point of 64 inches, for the type of selection in the panel title, for measurement age, and for the separate sampling of the two groups of birth cohorts. The off-diagonals present the differences between the diagonal elements. Panel C presents differences between Panels A and B. In all cases, bootstrap standard errors clustered at the county level are in parentheses. Columns (1)–(3) present results using the Union Army to represent the 1832–1846 cohorts, with identification based on Lincoln's vote share. Columns (4)–(6) present results using the Regular Army to represent the 1832–1846 cohorts, with identification based on the Buchanan and Douglas vote shares. Observation numbers are for the region in the column header for the estimates of Panel B.

Table H.6: Tests for differences in levels, sectoral decomposition

	Different Variables		Different Data	
	(1)	(2)	(3)	(4)
Sector	Urban	Rural	Urban	Rural
Panel A: Observables Only				
Urban	66.764***		66.802***	
	(0.226)		(0.214)	
Rural	−0.563***	67.327***	−0.496***	67.297***
	(0.152)	(0.257)	(0.160)	(0.248)
Panel B: Unobservables and Observables				
Urban	67.695***		67.316***	
	(0.446)		(0.488)	
Rural	−0.331**	68.026***	−0.378***	67.695***
	(0.142)	(0.465)	(0.138)	(0.488)
Panel C: $B - A$				
Urban	0.931**		0.515	
	(0.469)		(0.504)	
Rural	0.233*	0.699	0.117	0.398
	(0.124)	(0.501)	(0.091)	(0.492)
Observations	2,916	4,186	1,532	2,328

Significance levels: *** p< 0.01, ** p< 0.05, * p< 0.1

Notes: In Panels A and B, the diagonals present the estimated mean heights for each sector, corrected for minimum height requirements with a truncation point of 64 inches, for the type of selection in the panel title, for measurement age, and for the separate sampling of the two groups of birth cohorts. The off-diagonals present the differences between the diagonal elements. Panel C presents differences between Panels A and B. In all cases, bootstrap standard errors clustered at the county level are in parentheses. The urban sector is defined as a county with a non-zero urban population. Columns (1)–(2) present results using the Union Army to represent the 1832–1846 cohorts, with identification based on Lincoln's vote share. Columns (3)–(4) present results using the Regular Army to represent the 1832–1846 cohorts, with identification based on the Buchanan and Douglas vote shares. Observation numbers are for the sector in the column header for the estimates of Panel B.

References

Ancestry.com (2007). *U.S. Army, Register of Enlistments, 1798–1914* [database on-line]. Provo: Ancestry.com Operations Inc.

——— (2009a). *1860 United States Federal Census* [database on-line]. Provo: Ancestry.com Operations Inc.

——— (2009b). *1870 United States Federal Census* [database on-line]. Provo: Ancestry.com Operations Inc.

Biavaschi, Costanza, Corrado Giulietti, and Zahra Siddique (2017). "The Economic Payoff of Name Americanization." *Journal of Labor Economics* 35:4, pp. 1089–1116.

Cosslett, Stephen R. (1981). "Efficient Estimation of Discrete-Choice Models." In *Structural Analysis of Discrete Data with Econometric Applications*. Charles F. Manski and Daniel McFadden (ed.). Cambridge: MIT Press, 1990. Chap. 2, pp. 51–111.

Ferrie, Joseph P. (1996). "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 US Federal Census to the 1860 US Federal Census Manuscript Schedules." *Historical Methods* 29:4, pp. 141–156.

Gould, Benjamin Apthorp (1869). *Investigations in the Military and Anthropological Statistics of American Soldiers*. Sanitary Memoirs of the War of the Rebellion. Collected and Published by the United States Sanitary Commission. New York: Hurd and Houghton.

Klein, Roger W. and Richard H. Spady (1993). "An Efficient Semiparametric Estimator for Binary Response Models." *Econometrica* 61:2, pp. 387–421.

Mroz, Thomas A. (2015). "Sample Selection with Multiple Selection Indicators in Two-Step Estimators." Mimeo., Georgia State University.

Nadaraya, Elizbar A. (1964). "On Estimating Regression." *Theory of Probability & Its Applications* 9:1, pp. 141–142.

Register of Enlistments in the U.S. Army, 1798–1914 (n.d.). (National Archives Microfilm Publication M233, 81 Rolls), Records of the Adjutant General's Office, 1780's–1917, RG 94. Washington, DC: National Archives.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek (2015). *Integrated Public Use Microdata Series: Version 6.0* [machine-readable database]. Minneapolis: University of Minnesota.

Watson, Geoffrey S. (1964). "Smooth Regression Analysis." *Sankhya: The Indian Journal of Statistics* 26:4, pp. 359–372.