Online Appendix for
More is Less? The Impact of Family Size on Education Outcomes in the United States, 1850-1940

Hui Ren Tan

1 Age Groupings

This section describes the rationale behind dividing children into three age groups in the baseline analysis. For a start, one might consider all children aged 5-17.[1] However, this is a wide range and the impact of family size, if any, could vary with age. At the same time, it may not be ideal to look at each age separately and sacrifice parsimony or statistical precision. I use three criteria to inform my decision on age groupings: (i) the proportion of children in each age category attending school, (ii) the time series pattern of the OLS family size effect, and (iii) historical policies.

Consider first the proportion of children in each age group attending school. Using nine 1 percent IPUMS census samples (Ruggles et al. 2015), Figure OA1 plots the share of children attending school from 1850-1940 for 2-to-3 year age cells: 5-6, 7-9, 10-11, 12-13, 14-15, and 16-17. Attendance rates decline for all groups after the Civil War, and rise as one moves into the 20th century.[2] Overall, relatively few children aged 5-6 attend school and I thus exclude them from the analysis. The 16-17 year olds have a slightly flatter pattern than the other ages during the 19th century but display a steeper increase in the first-half of the 20th century, corresponding with the high school movement (Goldin and Katz 2008). This distinct trend warrants a separate grouping for 16-17 year olds. The remaining age categories tend to be more similar in levels and trends.

To determine if further divisions among the 7-15 year olds are necessary, I use a second criterion: the evolution of the relationship between family size and school attendance. I run a simple regression of school attendance on family size, and plot the OLS estimates from 1850-1940 in Figure OA2. The solid circle markers and confidence intervals refer to the OLS coefficients controlling only for the child and mother's age, while the open circle markers and dashed intervals incorporate

[1] I base this range on two considerations. The lower bound draws on instructions to enumerators to record schooling for those aged 5-21 in the 1910 census, even though they were also to report schooling if an individual attended school but fell outside this age range (IPUMS-USA 2017). The upper bound follows Goldin and Katz (2008) who study the high school movement using an age range of 14-17.

[2] The spike from 1900 to 1910 may be due to a different way of measuring school attendance in 1900. In 1900, rather than a binary yes or no, those of school-going age with a positive number of months in school were coded as having attended school. However, survey respondents may not have known how many months of school a child attended even if the child did attend school (United States Bureau of the Census 1935). Margo (1990) estimates that while this led to an underenumeration of schooling for southern blacks, the magnitude was not economically significant. Replicating Margo's (1990) procedure for both blacks and whites aged 5-20 across the US, I find that the underenumeration for whites is about 4.18 times that of blacks among 10-14 year olds (based on the 1900 1 and 5 percent IPUMS samples).

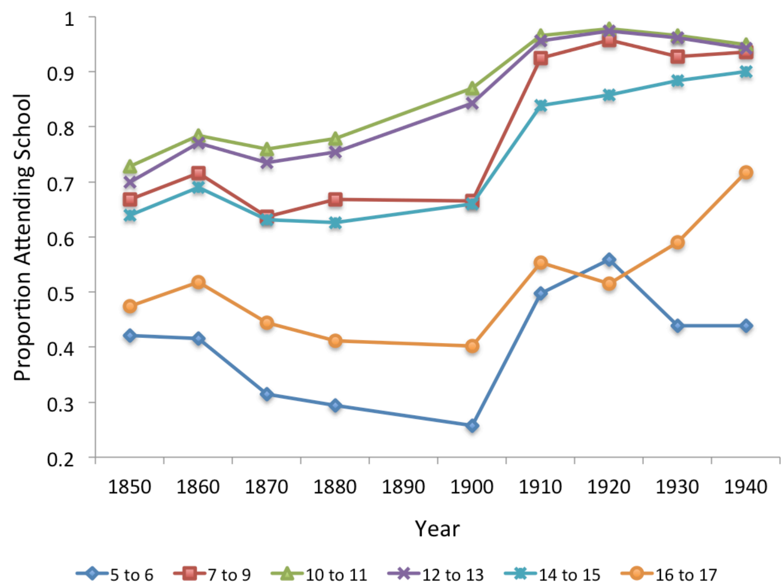

Figure OA1

Proportion of Children Attending School, by Age and Year

Notes: Data are taken from the 1850-1940 1 percent IPUMS samples, excluding 1890. The population of interest is restricted to native-born white children currently residing in one of the 32 states defined in the text. All proportions are adjusted using sample weights.

the full set of controls used in the main text. Consider the three middle groups: 7-9, 10-11 and 12-13. The basic OLS coefficients are negative in the 19th century and fade away over time. The estimates with full controls, on the other hand, are positive initially but also diminish in magnitude with time. This similar pattern makes it reasonable to focus on 7-13 year olds as one category. The 14-15 year olds display a pattern that tends to resemble that of the 16-17 year olds, where the point estimates become negative in the 20th century. I thus consider 14-15 year olds as a separate group.

The third criterion pertains to the rise of compulsory schooling and child labor laws. Insofar as such statutes constrained the ability of parents to adjust their investment in a child's schooling, the age at which the laws were binding may provide useful guidance on reasonable age groupings. Using a dataset of schooling and child labor laws from 1880-1930 compiled by Clay, Lingwall, and Stephens (2016), I find that among states with such legislations, the majority set 14 or 16 years of age as the minimum age a child could leave school and work.[3] Thus, dividing children along the 14 or 16-years threshold fits naturally with the school and child labor policies of the time.

[3] I thank Karen Clay and Jeff Lingwall for generously sharing their data with me.

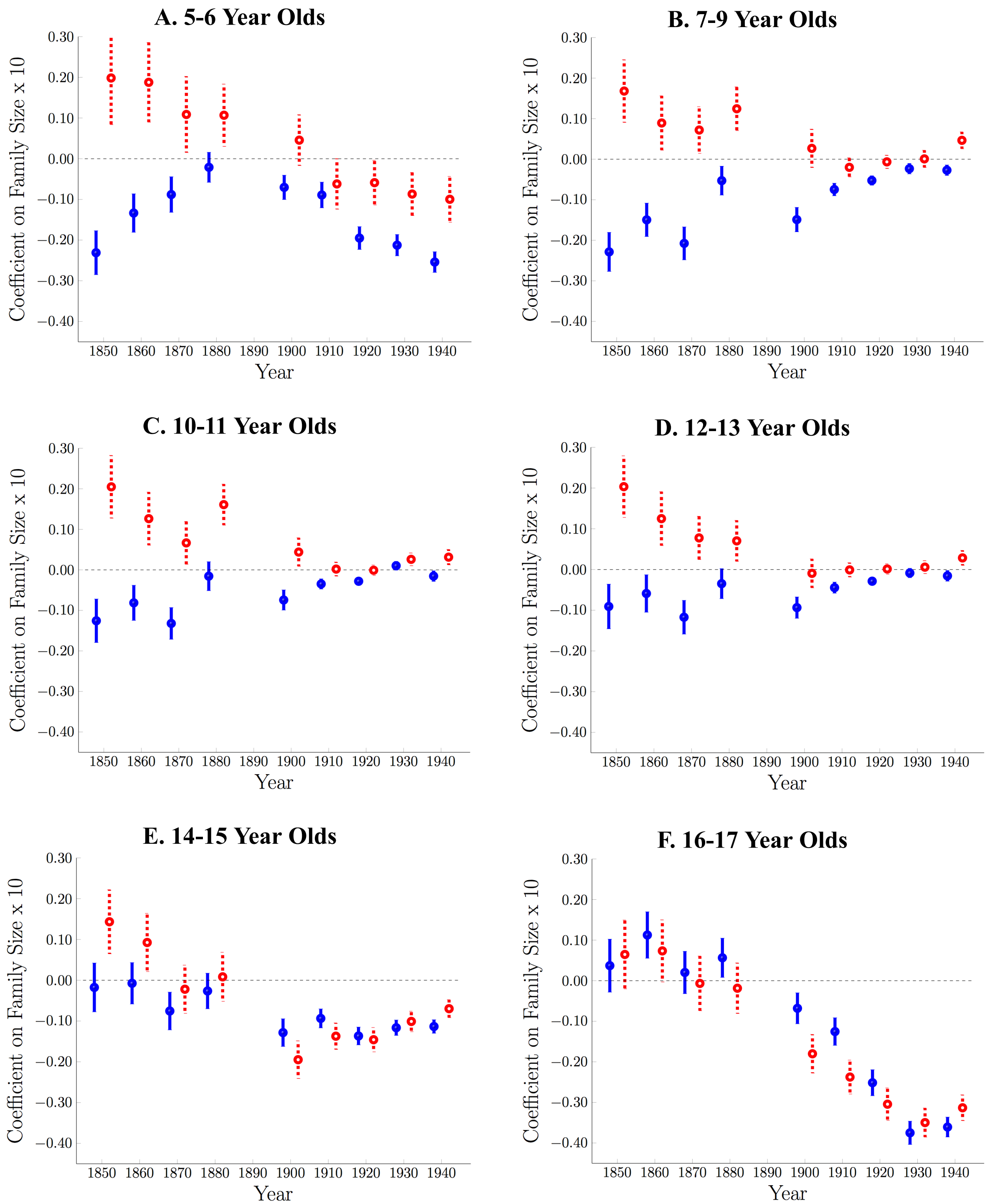

Figure OA2

OLS Relationship Between Family Size and School Attendance Over Time, by Age Group

Notes: Data are taken from the 1850-1940 1 percent IPUMS samples, excluding 1890. The outcome variable is school attendance and the main explanatory variable is family size x 10. The population of interest is restricted to native-born white children currently residing in one of the 32 states defined in the text, who come from male-headed households, who do not come from households with three or more children of the same age or households with multiple twin births, and who are not themselves twins. Solid circle markers refer to the OLS estimates controlling only for the age of the child and mother. Open circle markers refer to the OLS estimates controlling for age, gender, birth order, urban residence, share of girls in the household, characteristics of the father (age, literacy and occupation score), characteristics of the mother (age and literacy), and a full vector of county dummies. 95 percent confidence intervals are shown, based on standard errors clustered at the household level. Sample weights are used.

2 Data Limitations

As alluded to in the main text, the use of historical census data introduces errors in the identification of twins and the measurement of family size. I address both issues in turn, and argue that my main estimates are unlikely to be severely biased toward finding an adverse effect of larger families.

2.1 Accuracy in Identifying Twins

The baseline analysis classified persons who had the same parents and who were of the same age in years as twins. This is necessarily less accurate than if one had either quarter or month of birth to complement age.[4] Such information is not available in any of the full counts data. However, the 1900 IPUMS samples, which were not part of the main analysis because of their smaller sizes, record the birth month for all individuals.[5] This allows me to compare the family size estimates using a "fake" twins instrument based on age alone, and the "true" twins instrument based on both age and birth month. My results indicate that while there is substantial error in identifying twins, the direction of bias may potentially work against finding an adverse effect of family size.

Given the small size of the 1900 IPUMS samples relative to the full counts, I make three adjustments in order to increase statistical power for my analysis.[6] First, I pool together both the 1 percent and 5 percent 1900 IPUMS samples to form a 6 percent sample. Second, I collapse the 14-15 and 16-17 age groups into a common category. Third, instead of running the analysis by separate parities, I focus on the first-born child in all households together, and impose the following assumption: a twin birth at the second, third or fourth parity has the same impact on the likelihood of attending school.[7] This essentially combines three different instruments into one.

The pooled 1900 sample reveals that by using age alone, the number of persons classified as

[4]Angrist and Evans (1998), using the 1980 IPUMS sample, define multiple births as siblings with the same age and quarter of birth. They find that if age alone had been used instead, the twin rate would have increased by 35 percent.

[5]The 1900 full counts have not yet been released at the time of writing.

[6]Had the 1900 complete counts been available, one could simply implement a pure analysis by parity and compare the results when the "fake" and "true" twins instruments are used.

[7]Tables A1-A3 in Appendix 1 reveal that although the 2SLS family size effects are not the same across parities, the estimates are often of similar orders of magnitude, which provides some justification for the structural assumption.

twins is almost twice what it would have been if age and birth month had been used together. 1.97 percent of 7-13 year olds and 1.85 percent of 14-17 year olds are identified as twins using age. This falls to 1.13 and 1.11 percent respectively when both age and birth month are used.[8] The sources of such errors may be a combination of enumeration or transcription mistakes and age heaping.

To what extent will this misclassification bias my results? I make three points here. First, had it been classical measurement error, the usual attenuation bias would have caused my estimates to be less negative than they otherwise would have been. Second, because the twins instrument is binary, measurement error is by definition non-classical, which may offset the attenuation bias.[9] Third, characterizing the bias becomes even more complicated if the likelihood of misclassification varies systematically with individual or household characteristics that are correlated with both family size and school attendance. To illustrate this possibility, I restrict the 1900 sample to individuals identified as twins using age alone. I then regress misclassification status – a dummy variable that equals 1 if a twin is wrongly classified – on the full vector of controls used in the main text as well as family size. Table OA1 presents the coefficients from this linear probability model. The likelihood of misclassification indeed varies with certain individual and family traits. For example, children with younger mothers or who come from larger households are more likely to be incorrectly identified as twins. Insofar as these observable factors are controlled for in the final regression, biases can be addressed. The concern, however, is whether unobservable factors are also important, given that only around 50 to 65 percent of the variation in misclassification can be explained by the observable characteristics.

Table OA2 suggests that my estimates based on the "fake" twins instrument might understate the true impact of family size. Four aspects of the comparison using the "fake" and "true" twins instruments are worth noting. First, in either case, the direction of the family size effects are fairly consistent with the baseline analysis – the OLS results are positive for those aged 7-13 and negative

[8]The computations refer to a sample of native-born white children currently residing in one of the 32 states defined in the text, whose mothers are present in the household and recorded as the second person, who come from male-headed households with at least two children, and who do not come from households with three or more children of the same age or households with multiple twin births. Sample weights are used in the computations.

[9]Those who are "fake" twins can only be wrongly classified as "true" twins, introducing a negative correlation between the true value of the instrument and the measurement error.

Table OA1

Predictors of Twins Misclassification, Linear Probability Model, 1900

	(1)	(2)
	7-13 Year Olds	14-17 Year Olds
Individual characterstics		
Age	0.010	0.036**
	(0.006)	(0.018)
Female	-0.013	0.009
	(0.012)	(0.014)
Household characteristics		
Share of girls	0.112**	0.018
	(0.048)	(0.073)
Urban residence	0.164***	0.100*
	(0.035)	(0.057)
Family size x 10	0.245***	0.379***
	(0.069)	(0.101)
Father's characteristics		
Age	-0.003	0.004
	(0.002)	(0.003)
Literate	0.056	0.045
	(0.044)	(0.081)
Occupation score	-0.002	-0.001
	(0.001)	(0.002)
Mother's characteristics		
Age	-0.011***	-0.012***
	(0.002)	(0.003)
Literate	-0.084**	-0.130**
	(0.040)	(0.061)
Birth order dummies	Y	Y
County dummies	Y	Y
N	8,054	3,512
R^2	0.507	0.645

Notes: Data are taken from the 1 percent and 5 percent 1900 IPUMS samples. The outcome variable is a dummy for misclassified twins and the explanatory variables are: age, gender, share of girls in the household, urban residence, family size x 10, characteristics of the father (age, literacy and occupation score), characteristics of the mother (age and literacy), birth order, and a full vector of county dummies. The population is restricted to native-born white children currently residing in one of the 32 states defined in the text, whose mothers are present in the household and recorded as the second person, who come from male-headed households with at least two children, who do not come from households with three or more children of the same age or households with multiple twin births, and who are themselves classified as twins based on the age criterion. Sample weights are used for all estimates. Standard errors clustered at the household level are in parenthesis. *** 1%, ** 5%, and * 10% significance levels.

for the older 14-17 year olds, but the 2SLS point estimates are all negative. Second, comparing columns (1) and (2), the OLS effects are almost the same regardless of whether twins are identified by age alone or age and birth month.[10] Third, the 2SLS coefficients are always statistically insignificant, which is not surprising given the relatively small samples. In fact, the 2SLS point estimates for the older children are comparable to the full counts analysis for the 20th century – what has changed dramatically are the standard errors. Fourth, comparing columns (3) and (4), the 2SLS results based on the "true" twins instrument tend to be larger in magnitude than those based on the "fake" counterpart. This may suggest that the baseline results serve as lower bounds for the true impact of family size.[11]

Table OA2

Impact of Family Size on First-Born's School Attendance, Fake vs. True Twins, 1900

	OLS		2SLS	
	(1)	(2)	(3)	(4)
	Fake Twins	True Twins	Fake Twins	True Twins
	7-13 Year Olds			
Family size x 10	0.056***	0.056***	-0.007	-0.062
	(0.012)	(0.012)	(0.088)	(0.123)
N	102,404	103,173	102,404	103,173
First stage F-statistic			1,531	801
	14-17 Year Olds			
Family size x 10	-0.216***	-0.217***	-0.175	-0.274
	(0.014)	(0.014)	(0.116)	(0.176)
N	60,159	60,554	60,159	60,554
First stage F-statistic			518	239

Notes: Data are taken from the 1 percent and 5 percent 1900 IPUMS samples. The outcome variable is school attendance and the main explanatory variable is family size x 10. The controls used are: age, gender, urban residence, share of girls in the household, characteristics of the father (age, literacy and occupation score), characteristics of the mother (age and literacy), and a full vector of county dummies. A modified version of analysis by parity is used, where the twins instrument is a dummy that takes the value 1 if the second, third, or fourth birth is a twin birth. The population is restricted to first-born native white children currently residing in one of the 32 states defined in the text, whose mothers are present in the household and recorded as the second person, who come from male-headed households with at least two children, who do not come from households with three or more children of the same age or households with multiple twin births, and who are not themselves twins. Sample weights are used for all estimates. Standard errors clustered at the household level are in parenthesis. *** 1%, ** 5%, and * 10% significance levels.

[10] For the OLS, the sample sizes differ depending on whether twins are identified by age or age and birth month, because twins themselves are excluded from all estimations, as mentioned in the main text.

[11] Readers may wonder if the results presented here are influenced by the way schooling was measured in 1900, which differed from other years as mentioned in an earlier footnote. I argue that the similarity of the OLS results in Table OA2 to the 20th century baseline analysis suggests that any bias introduced may not have been severe.

2.2 Accuracy in the Measurement of Family Size

A further limitation with historical census data is that one cannot observe the true family size since the census only records persons residing in a household at the time of the survey. This misses out on children who have already left home and those who have yet to be born, thus understating the total number of children. The extent of these problems likely depends on the age of the mother. Older mothers may have seen more of their children leave home but they are also more likely to have completed their fertility cycle. I evaluate the resulting bias in two ways. First, I limit my sample to households where none of the children have left home. Second, I restrict the population to children with mothers who are already past childbearing age to minimize errors due to incomplete fertility cycles. The estimates from both subsamples are compared with the full data, and suggest that the baseline results are unlikely to be severely biased.

To address the problem of children leaving home, I focus on the 1910 complete counts. In the 1910 census, enumerators asked ever-married females aged 12 and above for the number of children they ever had who were still alive on census day.[12] I restrict the sample to households where the number of surviving children of a mother matches the actual number of her own children in the household.[13] I then implement the 2SLS analysis by parity for this subset of the population ("no leavers") and compare the results to the full sample. Figure OA3 presents this comparison for the 7-13 and 14-17 year olds separately. Across age groups and parities, the 2SLS point estimates are reasonably similar between the restricted and full populations. The main analysis is thus unlikely to be biased by mis-measurements of family size introduced by children who have already left home.

To handle the issue of incomplete fertility of mothers, I make use of the observation that the likelihood of completed fertility increases with a woman's age. Thus, one exercise that can be conducted is to restrict the population to children with mothers who are past childbearing age and compare the 2SLS estimates by parity to the full data. Specifically, I use 40 years of age as the cut-

[12] This includes children born to different fathers, and is thus not entirely congruent with the baseline analysis where the anchor used in computing family size is the household head. Note that a similar variable is also available in the much smaller 1900 samples.

[13] In practice, this is implemented for mothers who have at most nine children residing with them as the highest category for the variable recording the number of own children in the family is "9+".

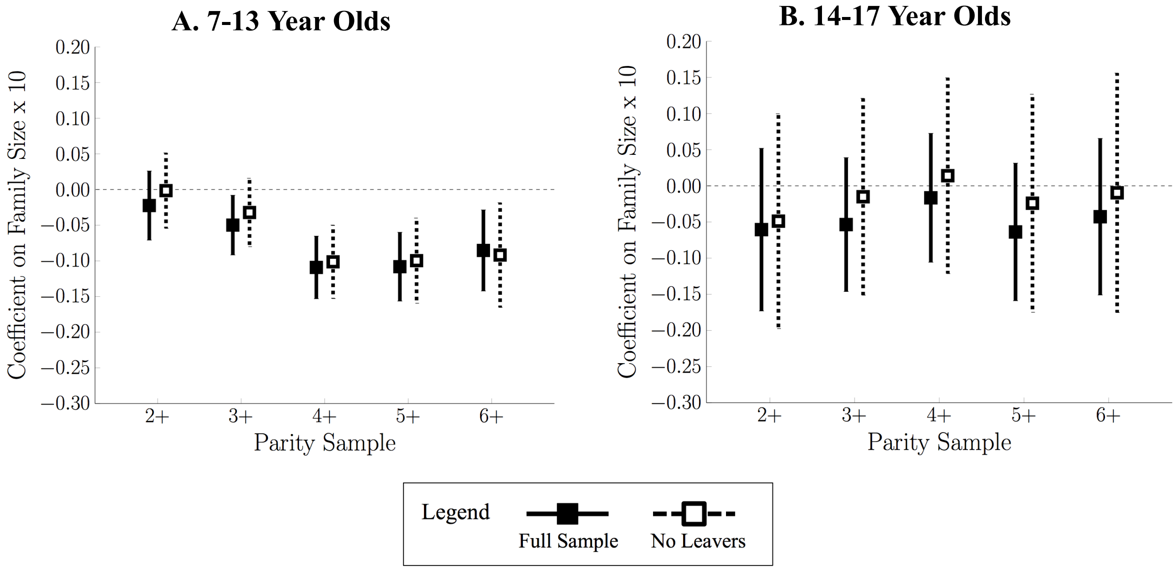

Figure OA3

2SLS Impact of Family Size on School Attendance,
Full Sample vs. Sample with All Children Present, 1910

Notes: Data are taken from the 1910 census full counts. The outcome variable is school attendance and the main explanatory variable is family size x 10. The controls used are: age, gender, birth order, urban residence, share of girls in the household, characteristics of the father (age, literacy and occupation score), characteristics of the mother (age and literacy), and a full vector of county dummies. Each regression is implemented at a specific birth parity as indicated on the x-axis, and the 2SLS coefficient on family size is displayed. The full sample is restricted to native-born white children currently residing in one of the 32 states defined in the text, whose mothers are present in the household and recorded as the second person, who come from male-headed households, who do not come from households with three or more children of the same age or households with multiple twin births, and who are not themselves twins. "No Leavers" refers to the subset of individuals from households where none of the children have left home. 7-13 year olds: the parity sample sizes range from 1,730,051 to 2,761,304 for the full sample, and from 1,152,892 to 2,133,587 for the sample of individuals from households where none of the children have left home. 14-17 year olds: the parity sample sizes range from 1,056,061 to 1,778,593 for the full sample, and from 578,797 to 1,056,439 for the corresponding restricted subset. Solid square markers refer to the 2SLS estimates using the full sample. Open square markers refer to the corresponding 2SLS estimates for the sample of individuals from households where none of the children have left home. 95 percent confidence intervals are shown, based on standard errors clustered at the household level.

off.[14] Table OA3 compares the 2SLS estimates between the restricted and full samples by year, age group and parity, illustrating that the baseline results are generally robust to the mis-measurement of family size due to incomplete fertility. The point estimates are reasonably similar and are typically in the same direction.

Taken as a whole, my checks suggest that the estimates in the main text are unlikely to be severely biased toward finding an adverse family size effect due to errors in identifying twins or in measuring family size.

[14] I use 40 years of age as the cutoff for the following reason. The 1900 and 1910 censuses record the number of children ever born to ever-married women aged 12 and above. I consider the spouses of white male household heads and implement a local linear regression of the number of children ever born to mothers on their age. I find that the fertility of married women begins to plateau some time after age 40.

Table OA3

2SLS Impact of Family Size on School Attendance,
Full Sample vs. Sample of Children with Mothers Past Childbearing Age

	1850		1910		1920		1930		1940	
	Full	Subset	Full	Subset	Full	Subset	Full	Subset	Full	Subset
					7-13 Year Olds					
2+	-0.329***	-0.234	-0.022	-0.078	-0.144***	-0.085	-0.037	0.020	-0.042**	-0.054
	(0.087)	(0.228)	(0.025)	(0.052)	(0.027)	(0.054)	(0.023)	(0.045)	(0.020)	(0.045)
3+	-0.206***	-0.494***	-0.050**	-0.031	-0.111***	-0.120***	-0.046**	-0.051	-0.035*	-0.020
	(0.072)	(0.144)	(0.021)	(0.037)	(0.021)	(0.040)	(0.018)	(0.033)	(0.019)	(0.036)
4+	-0.177***	0.047	-0.109***	-0.075**	-0.104***	-0.079**	-0.007	0.035	-0.011	-0.026
	(0.068)	(0.115)	(0.022)	(0.035)	(0.021)	(0.035)	(0.018)	(0.028)	(0.021)	(0.034)
5+	-0.206***	-0.192*	-0.108***	-0.062*	-0.080***	-0.071**	-0.046**	-0.050	0.009	0.035
	(0.069)	(0.105)	(0.025)	(0.036)	(0.022)	(0.035)	(0.021)	(0.031)	(0.023)	(0.033)
6+	-0.224***	-0.420***	-0.085***	-0.080**	-0.113***	-0.166***	-0.041*	-0.017	-0.011	-0.046
	(0.077)	(0.114)	(0.029)	(0.040)	(0.028)	(0.043)	(0.024)	(0.035)	(0.029)	(0.043)
					14-17 Year Olds					
2	-0.002	-0.004	-0.061	-0.062	-0.220***	-0.234***	-0.147***	-0.173**	-0.140***	-0.030
	(0.121)	(0.171)	(0.057)	(0.080)	(0.058)	(0.073)	(0.050)	(0.068)	(0.040)	(0.056)
3	-0.180*	-0.182	-0.054	0.010	-0.151***	-0.189***	-0.141***	-0.155***	-0.121***	-0.071
	(0.096)	(0.126)	(0.047)	(0.061)	(0.047)	(0.056)	(0.041)	(0.053)	(0.037)	(0.046)
4	0.001	0.077	-0.017	0.037	-0.055	0.008	-0.130***	-0.148***	0.017	0.004
	(0.086)	(0.100)	(0.045)	(0.057)	(0.046)	(0.056)	(0.041)	(0.050)	(0.039)	(0.048)
5	-0.258***	-0.155	-0.064	-0.008	-0.115**	-0.138**	-0.088**	-0.062	-0.108**	-0.075
	(0.089)	(0.110)	(0.049)	(0.059)	(0.047)	(0.059)	(0.044)	(0.055)	(0.044)	(0.055)
6	-0.171**	-0.237**	-0.043	-0.068	-0.135***	-0.140**	-0.154***	-0.145**	-0.168***	-0.149**
	(0.087)	(0.108)	(0.055)	(0.070)	(0.052)	(0.067)	(0.048)	(0.061)	(0.053)	(0.067)

Notes: Data are taken from the 1850, 1910-1940 census full counts. The outcome variable is school attendance and the main explanatory variable is family size x 10. The controls used are: age, gender, birth order, urban residence, share of girls in the household, characteristics of the father (age, literacy and occupation score), characteristics of the mother (age and literacy), and a full vector of county dummies. Each regression is implemented at a specific birth parity, and the 2SLS coefficient on family size is displayed in each cell. The full sample is restricted to native-born white children currently residing in one of the 32 states defined in the text, whose mothers are present in the household and recorded as the second person, who come from male-headed households, who do not come from households with three or more children of the same age or households with multiple twin births, and who are not themselves twins. The restricted sample refers to the subset of children with mothers aged 40 and over. Full sample: sample sizes range from 567,871 to 3,372,152 for the 7-13 year olds, and from 324,158 to 2,350,370 for the 14-17 year olds. Restricted sample: sample sizes range from 59,955 to 831,215 for the 7-13 year olds, and from 131,702 to 1,256,146 for the 14-17 year olds. Standard errors clustered at the household level are in parenthesis. *** 1%, ** 5%, and * 10% significance levels.

3 References

- Angrist, Joshua, and William N. Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review* 88(3): 450-477.
- Clay, Karen, Jeff Lingwall, and Melvin Stephens Jr. 2016. "Laws, Educational Outcomes, and Returns to Schooling: Evidence from the Full Count 1940 Census." NBER Working Paper 22855. Cambridge, MA, National Bureau of Economic Research.
- Goldin, Claudia, and Lawrence F. Katz. 2008. *The Race Between Education and Technology.* Cambridge, MA and London, England: The Belknap Press of Harvard University Press.
- IPUMS-USA. 2017. University of Minnesota. www.ipums.org.
- Margo, Robert A. 1990. *Race and Schooling in the South, 1880-1950*. Chicago, IL and London, England: University of Chicago Press.
- Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2015. *Integrated Public Use Microdata Series: Version 6.0*. Minneapolis, MN: University of Minnesota.
- United States Bureau of the Census. 1935. *Negroes in the United States, 1920-32*. Washington, DC: Government Printing Office.