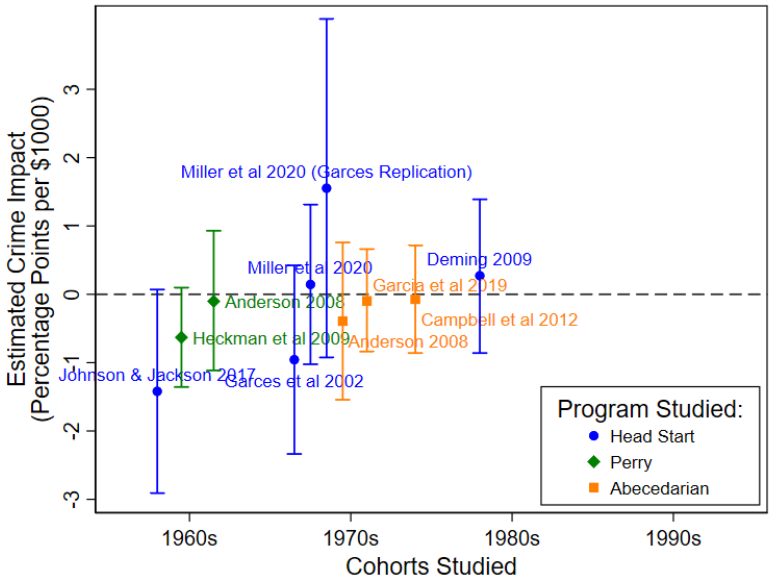
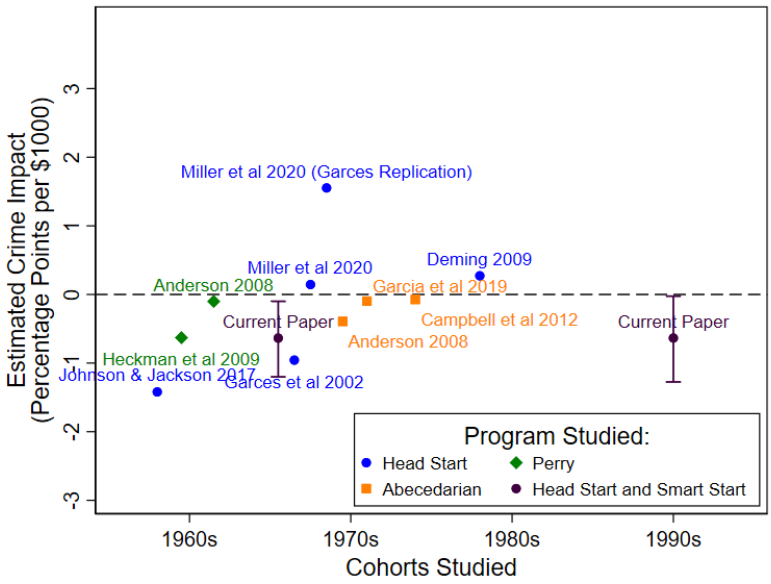
Online Appendix for "The Effect of Early Childhood Education on Adult Criminality: Evidence from the 1960s through 1990s" by John Anders, Andrew Barr, and Alexander A. Smith

Appendix A: Supplementary Figures

Figure A1: Effect Size Comparison



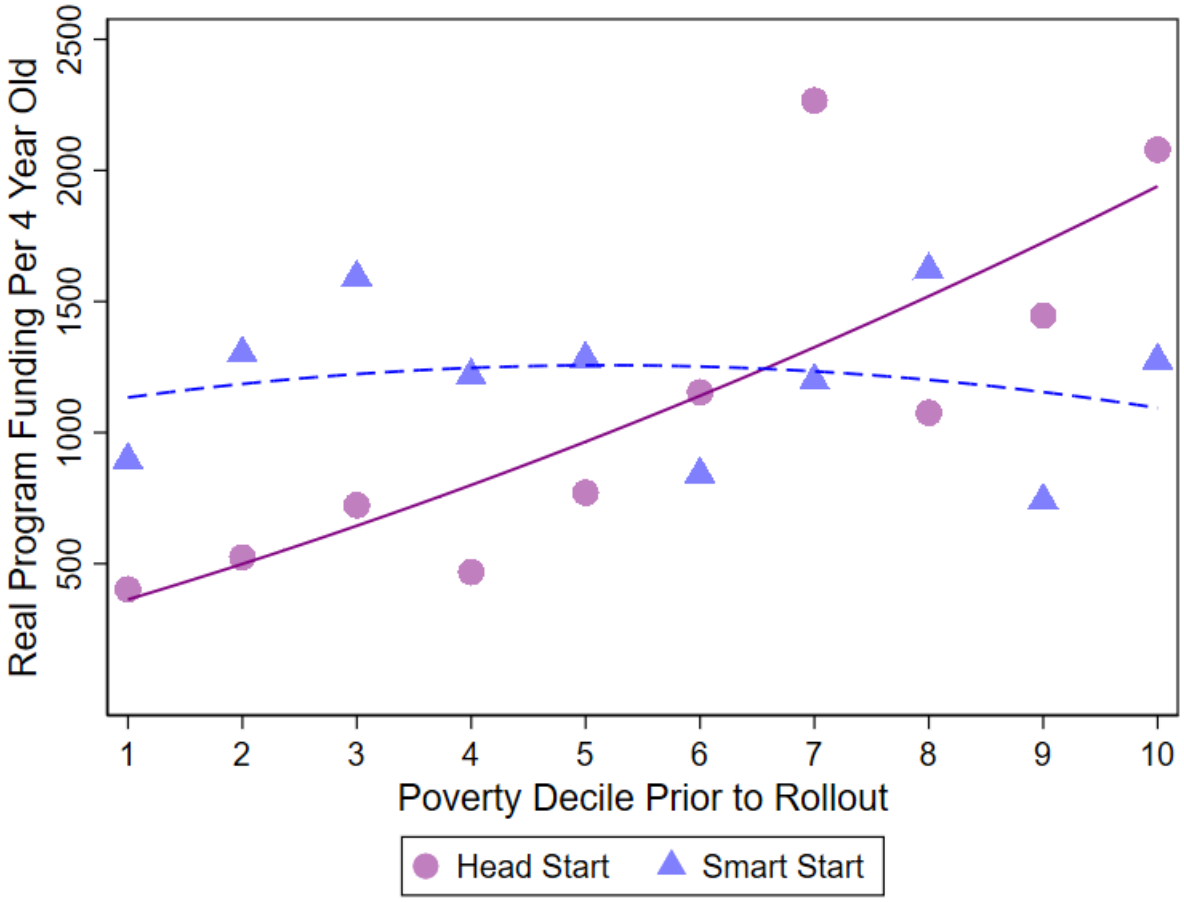
(a) Existing Evidence



(b) New Evidence

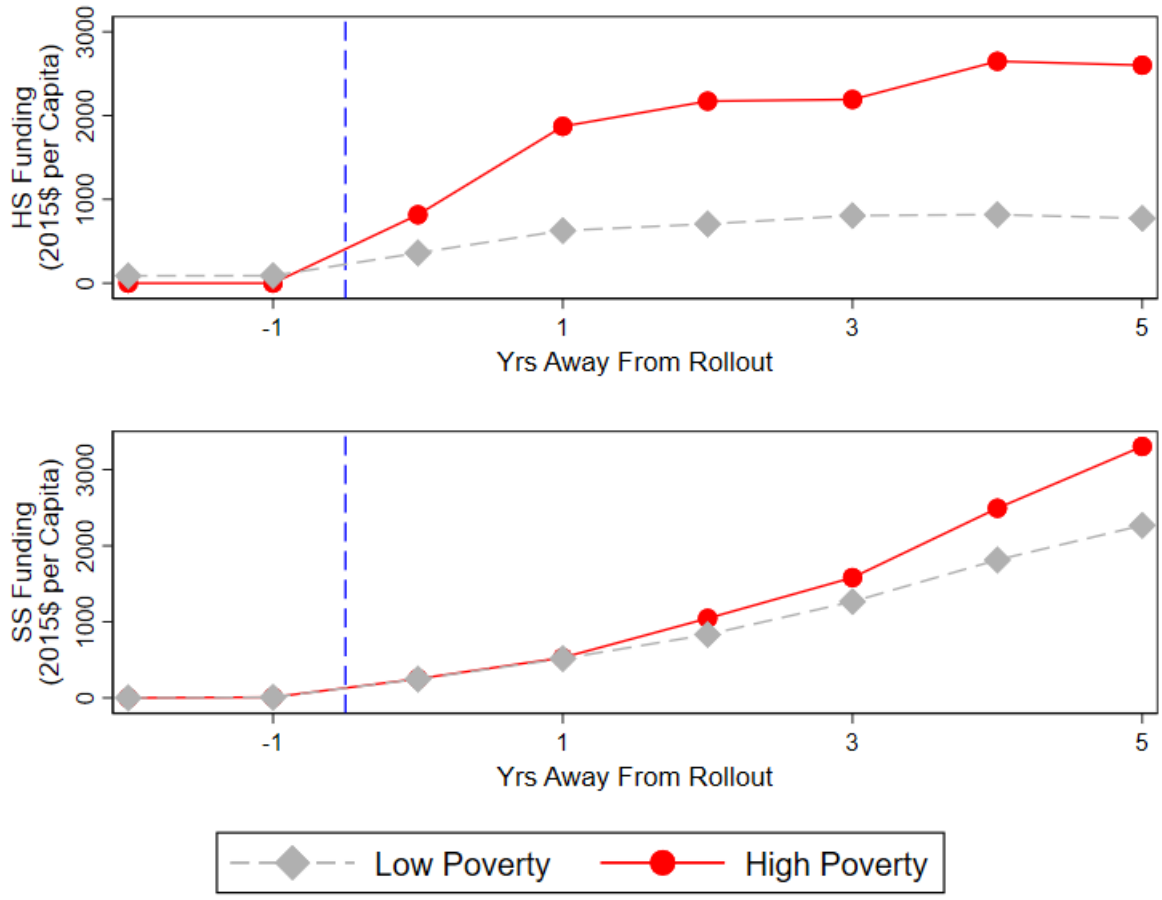
Note: Figure shows estimated crime impact per $1,000 (2015 dollars) and associated 95 percent confidence intervals. Program costs for Perry and Abecedarian are taken from Heckman et al (2009) and Garces et al (2002), respectively, and are $20,648 and $22,687 respectively. Because program funding for Head Start varies across cohorts and different papers study different cohorts, we use the funding reported by each study. For Deming (2009), Garces et al. (2002) and Johnson and Jackson (2017) those numbers are, respectively, $6,981, $5,540, and $6,027. The selected effect estimates are provided in column (4) of Table B2. See Appendix B for additional details on construction.

Figure A2: Targeting of Head Start and Smart Start at Poverty



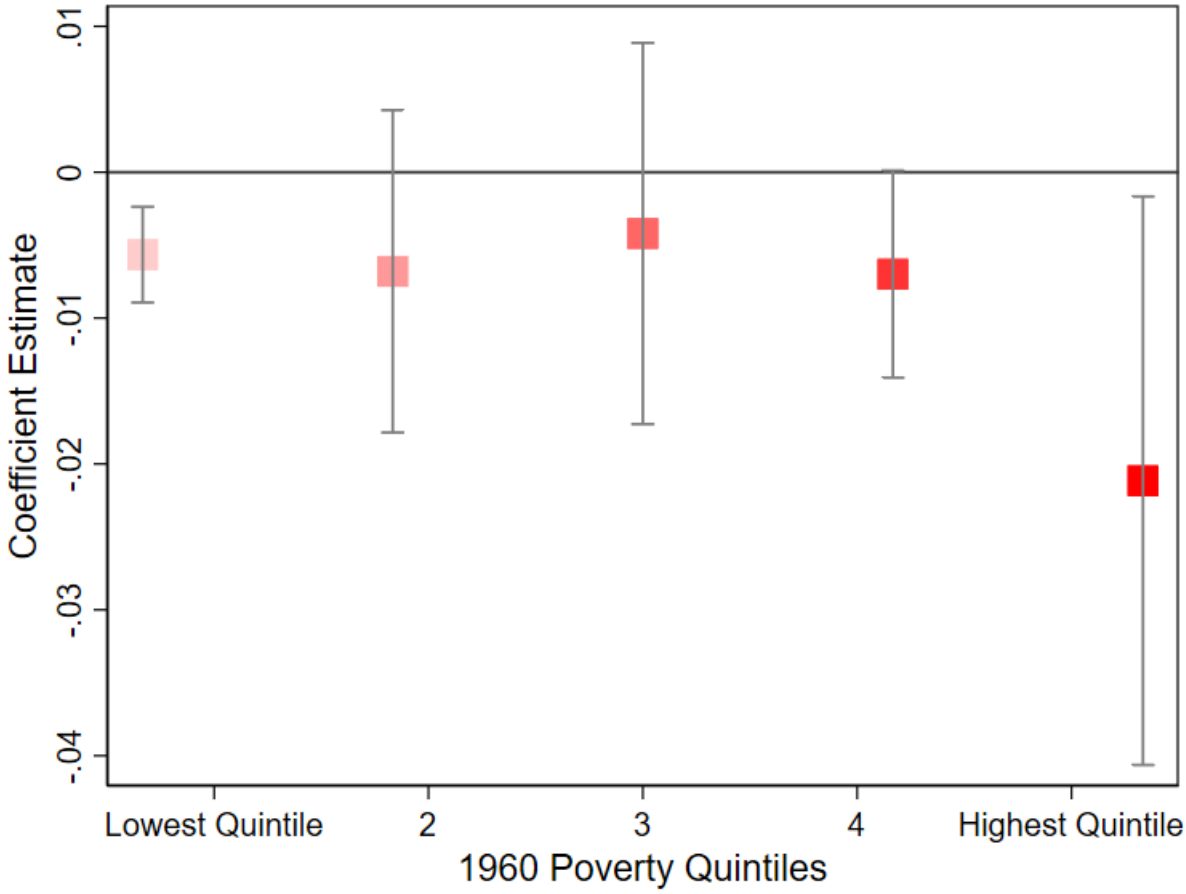
Note: Figure shows per capita county level Head Start and Smart Start funding (given in $ per 4 year olds) by county poverty deciles before program rollout. For Head Start 1960 poverty deciles are used, while for Smart Start 1980 poverty deciles are used. All values are in 2009 dollars. The sample is restricted to counties with nonzero funding amounts.

Figure A3: Funding By County Poverty Level



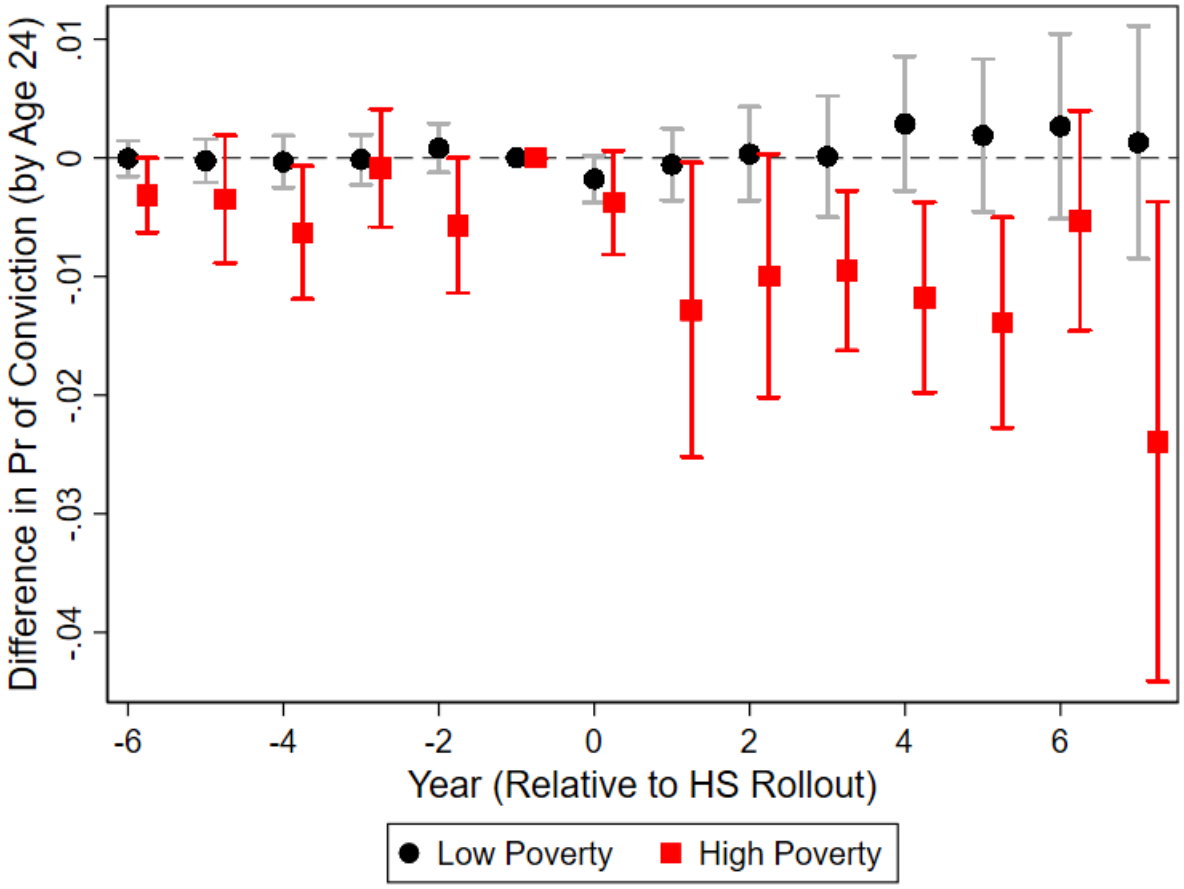
Note: Figure shows per capita county level Head Start and Smart Start funding (given in $ per 4 year olds) separately for high and low poverty counties. In the upper panel which shows Head Start funding, high poverty counties are those counties with a 1960 poverty rate above the median in North Carolina (40.2% poverty), while low poverty are those with a below median 1960 poverty rate. In the lower panel which shows Smart Start funding, high poverty counties are those counties whose poverty rate in 1980 was above the median in North Carolina (17.3% poverty) , while low poverty are those below the median. (There exist small, non-zero funding levels in the year prior to Head Start rollout for two reasons: first, following Barr and Gibbs (2017), county birth cohorts with very low funding levels are treated as not having Head Start availability, and, second, we do not count 1965 as the first year of availability since the Head Start program was introduced only as a pilot program over the Summer in that year.)

Figure A4: Head Start Estimates by Quintiles



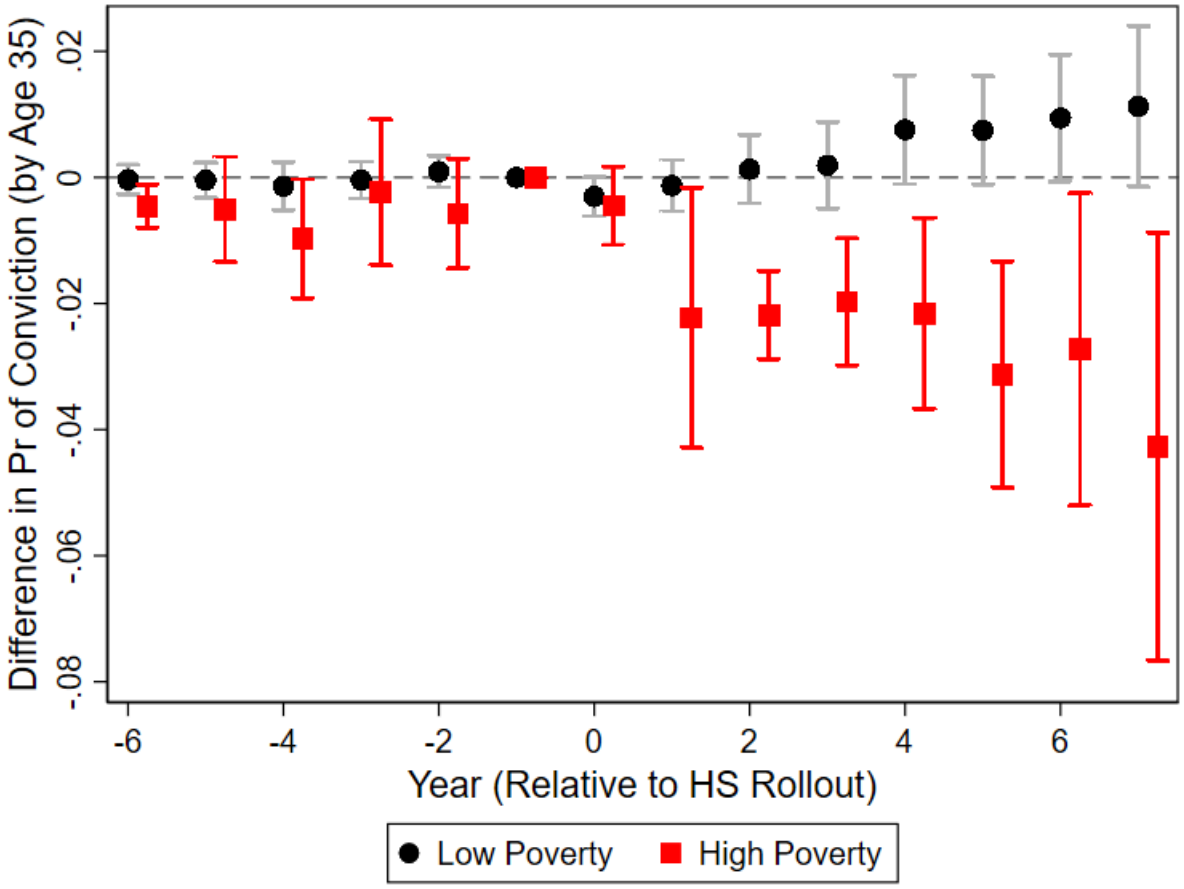
Note: Figure shows the coefficient estimates and 95% confidence intervals from estimating our basic difference-in-differences specification separately for counties in each quintile of the 1960 North Carolina poverty rate. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 35. UCR Part 1 crimes include violent crimes (those in which the description of the offense contains the words "murder", "assault", or "robbery" (rape not being included), and property crimes (those in which the description of the offense contains the words "burglary" or "larceny"). All specifications include birth county and birth-cohort fixed effects as well as 1960 county characteristics interacted with a time trend in birth cohort. 1960 county characteristics include: percent of land in farming, percent of people living in families with less than $3,000, percent of population in urban area, percent black, percent less than age 5, percent greater than age 65, and percent of employment in agriculture. The sample is restricted to counties that ever received Head Start between 1965 and 1976. The sample is further restricted to cohorts who were born between 1955 and 1968.

Figure A5: Event Study of Head Start's Impact on Criminal Conviction (by age 24)



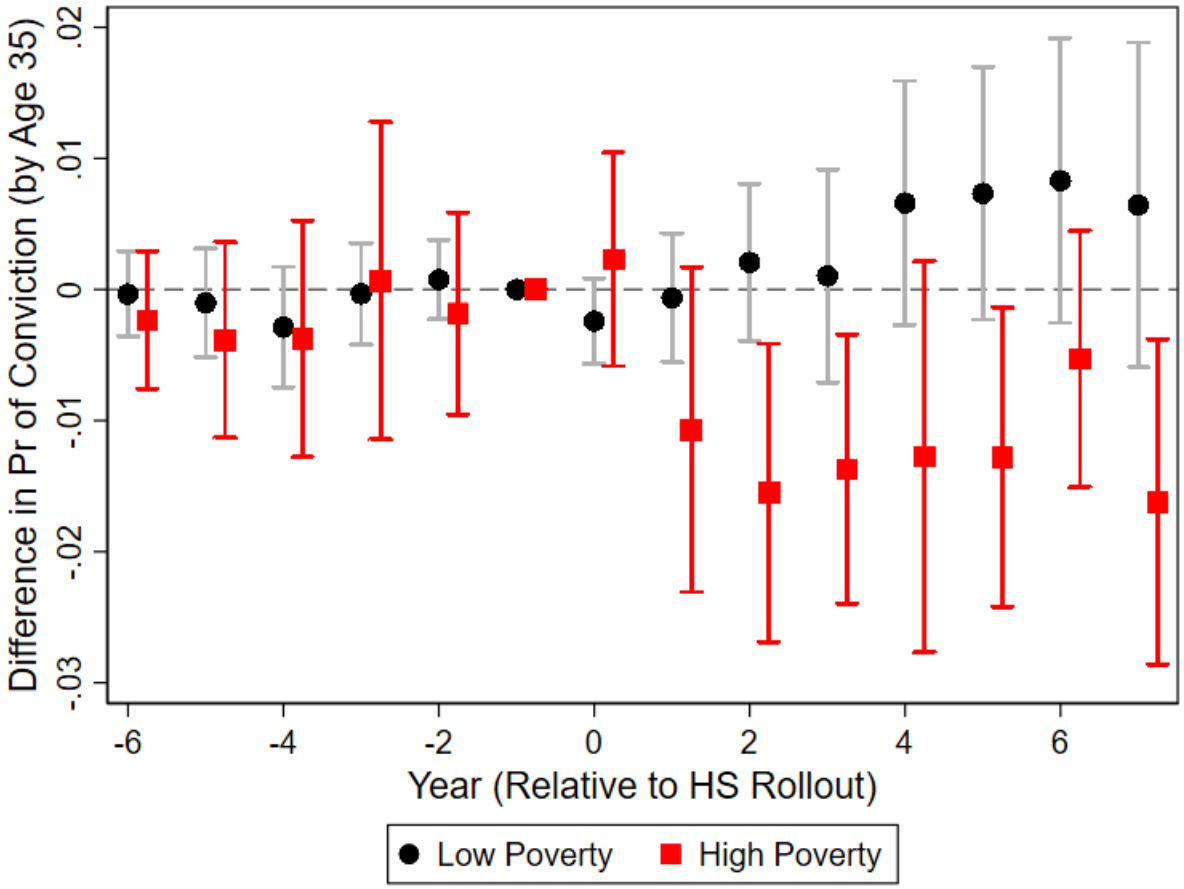
Note: Figure shows the coefficient estimates and 95% confidence interval from estimating Equation 2 separately for high and low poverty counties. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 24. UCR Part 1 crimes include violent crimes (those in which the description of the offense contains the words "murder", "assault", or "robbery" (rape not being included), and property crimes (those in which the description of the offense contains the words "burglary" or "larceny"). All specifications include birth county and birth-cohort fixed effects as well as 1960 county characteristics interacted with a time trend in birth cohort. 1960 county characteristics include: percent of land in farming, percent of people living in families with less than $3,000, percent of population in urban area, percent black, percent less than age 5, percent greater than age 65, and percent of employment in agriculture. Those counties whose poverty rate in 1960 was above the median in North Carolina (40.2% poverty) are called "High Poverty", while those below the median are called "Low Poverty". The sample is restricted to counties that ever received Head Start between 1965 and 1976. The sample is further restricted to cohorts who were born between 1955 and 1968.

Figure A6: Event Study of Head Start's Impact on Criminal Conviction (without trends)



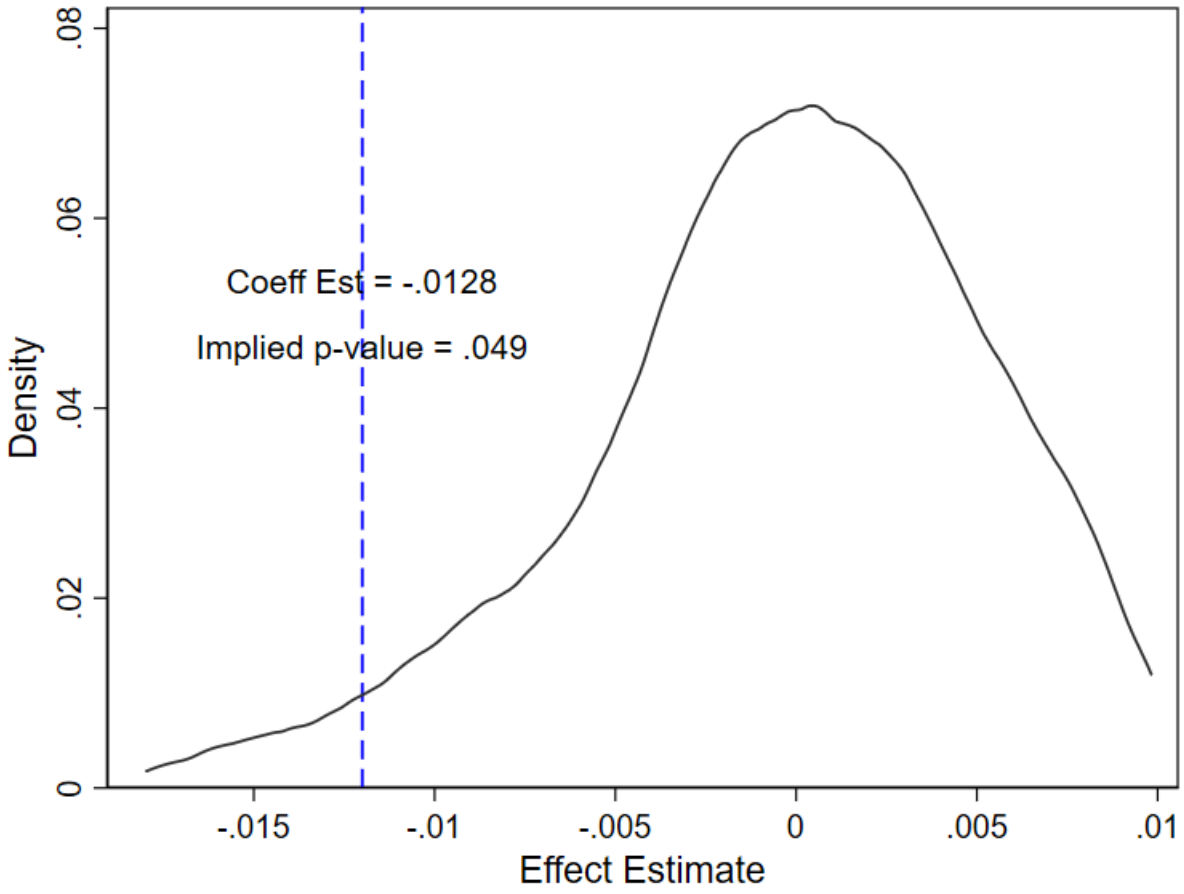
Note: Figure shows the coefficient estimates and 95% confidence interval from estimating Equation 2 separately for high and low poverty counties. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 35. UCR Part 1 crimes include violent crimes (those in which the description of the offense contains the words "murder", "assault", or "robbery" (rape not being included), and property crimes (those in which the description of the offense contains the words "burglary" or "larceny"). All specifications include birth county and birth-cohort fixed effects, but, unlike Figure 2, they do not include 1960 county characteristics interacted with a time trend in birth cohort. Those counties whose poverty rate in 1960 was above the median in North Carolina (40.2% poverty) are called "High Poverty", while those below the median are called "Low Poverty". The sample is restricted to counties that ever received Head Start between 1965 and 1976. The sample is further restricted to cohorts who were born between 1955 and 1968.

Figure A7: Event Study of Head Start's Impact on Criminal Conviction – Robustness to Inclusion of All Counties



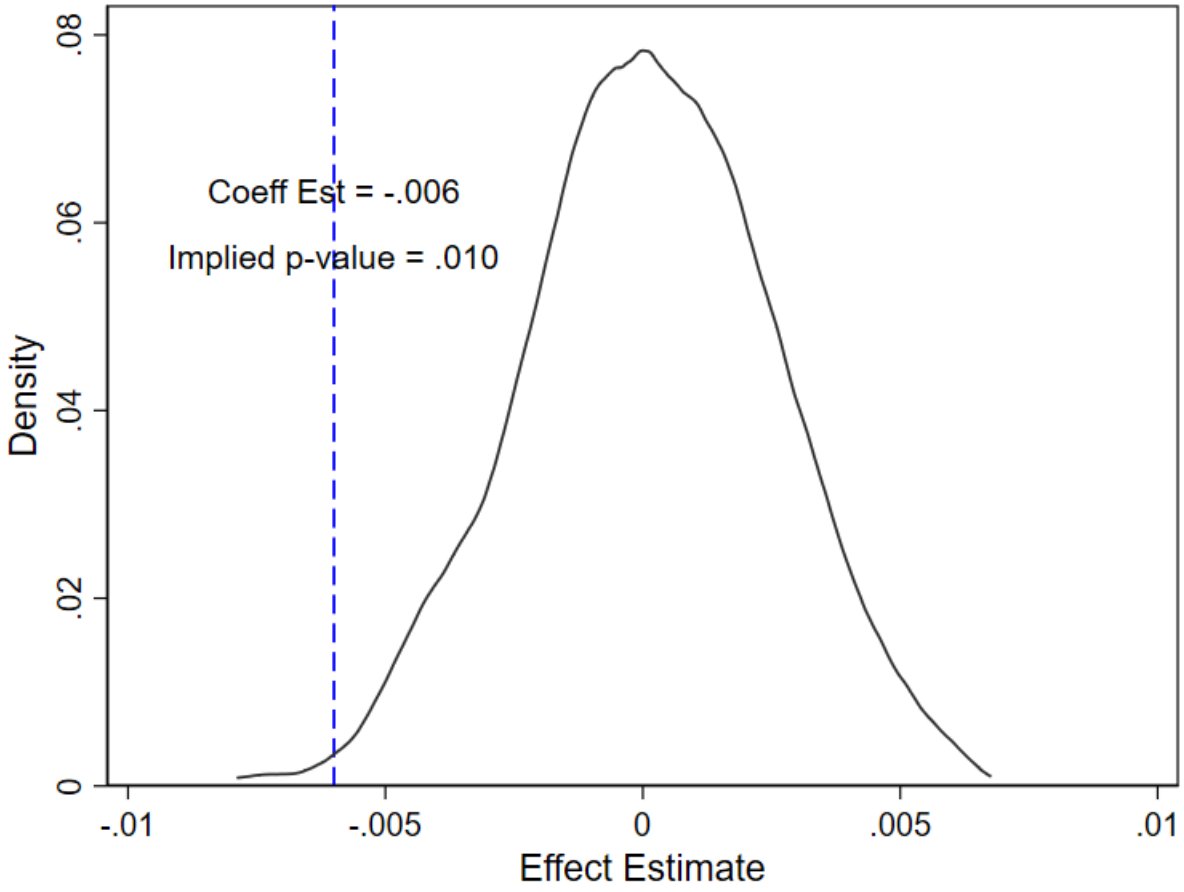
Note: Figure shows the coefficient estimates and 95% confidence interval from estimating Equation 2 separately for high and low poverty counties. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955 multiplied by the inverse propensity to receive Head Start as defined by Stuart (2010). The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 35. UCR Part 1 crimes include violent crimes (those in which the description of the offense contains the words "murder", "assault", or "robbery" (rape not being included), and property crimes (those in which the description of the offense contains the words "burglary" or "larceny"). All specifications include birth county and birth-cohort fixed effects as well as 1960 county characteristics interacted with a time trend in birth cohort. 1960 county characteristics include: percent of land in farming, percent of people living in families with less than $3,000, percent of population in urban area, percent black, percent less than age 5, percent greater than age 65, and percent of employment in agriculture. Those counties whose poverty rate in 1960 was above the median in North Carolina (40.2% poverty) are called "High Poverty", while those below the median are called "Low Poverty". The sample is further restricted to cohorts who were born between 1955 and 1968.

Figure A8: Head Start Randomization Inference (High-Poverty Counties)



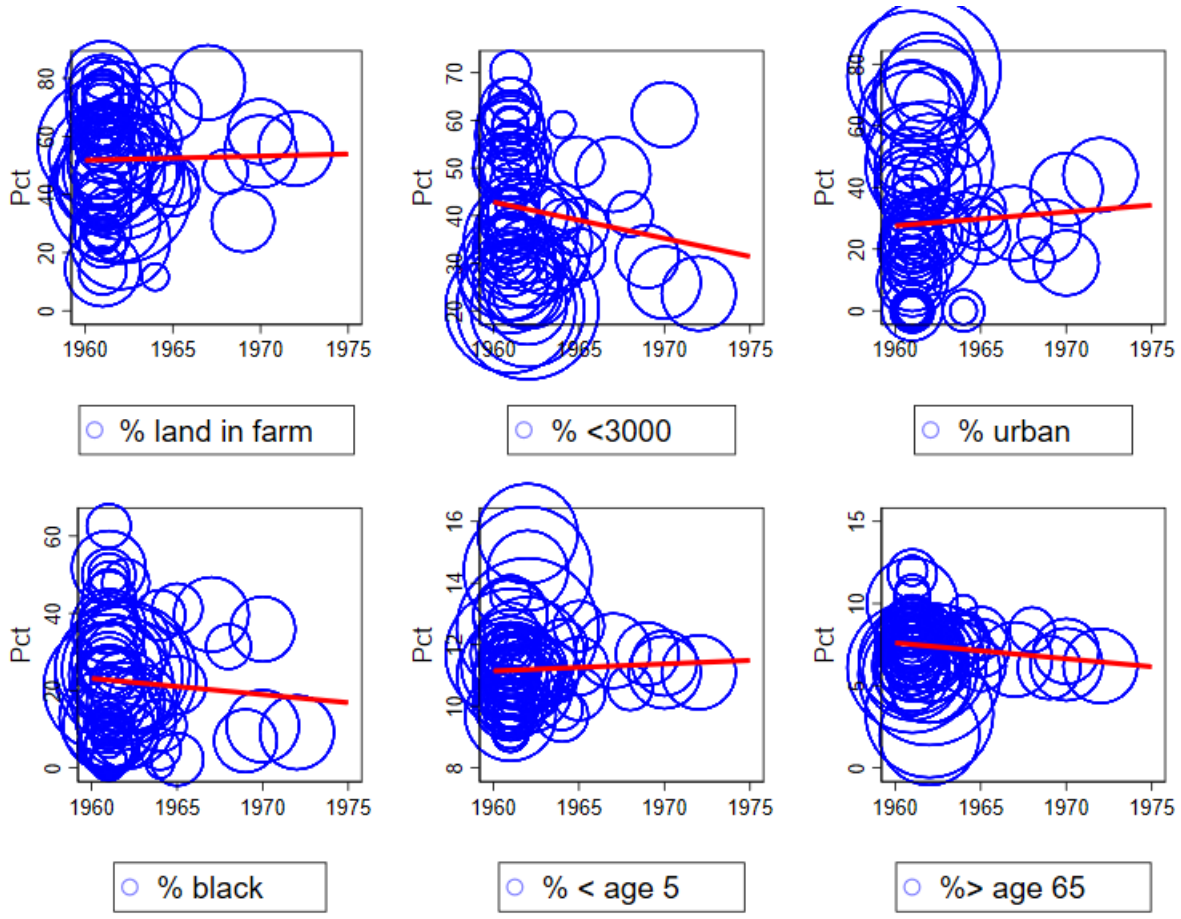
Note: Figure shows the kernel density of coefficient estimates under random assignment of Head Start availability to high poverty counties. 1000 repetitions were performed. The vertical line indicates the coefficient estimate obtained using the actual rollout of Head Start (See Table 2). A two-tailed test statistic is calculated as the share of estimates whose absolute value is greater than or equal to the estimate obtained using the actual rollout. Calculating this statistic gives an implied p-value of .049 as compared with the p-value of .040 given by the standard errors clustered at the county level.

Figure A9: Smart Start Randomization Inference



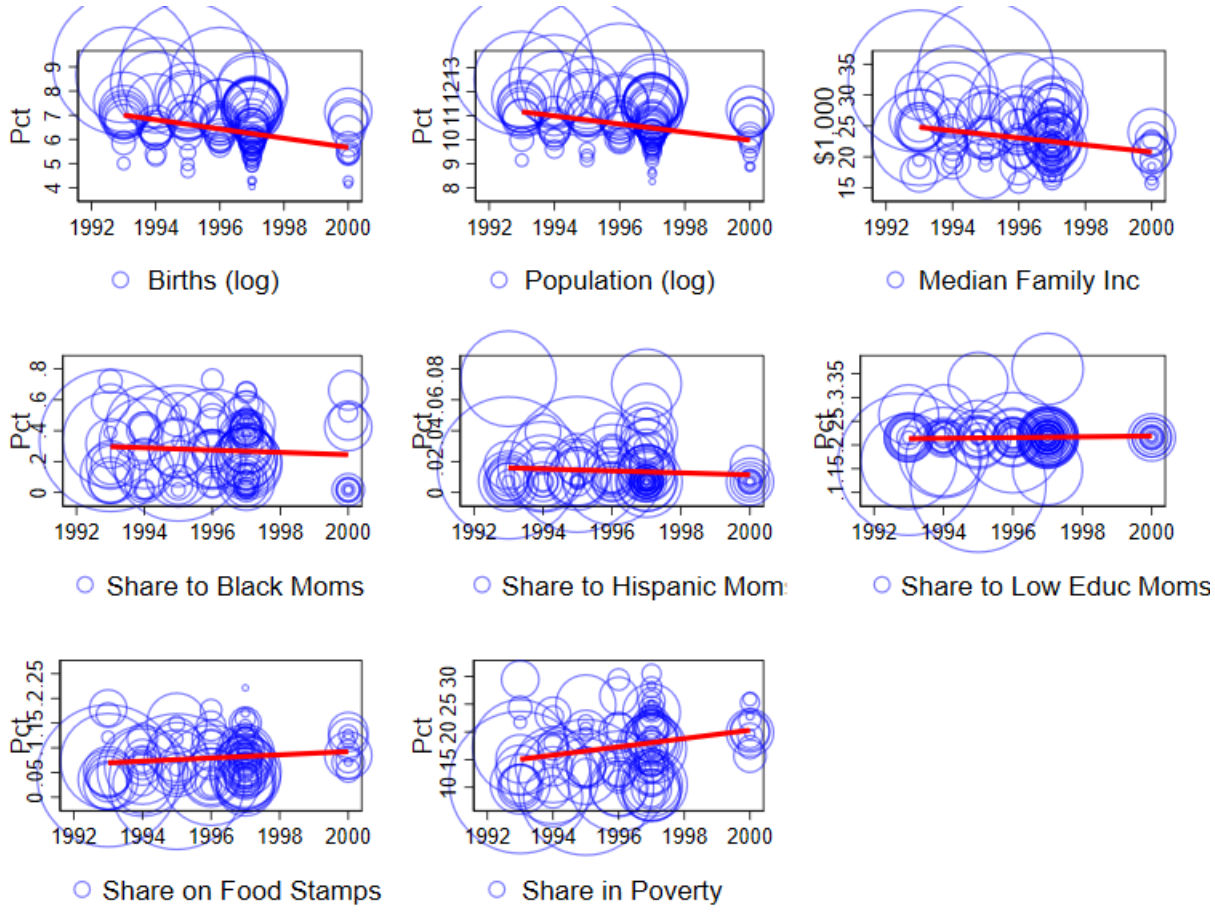
Note: Figure shows the kernel density of coefficient estimates under random assignment of Smart Start availability. 1000 repetitions were performed. The vertical line indicates the coefficient estimate obtained using the actual rollout of Smart Start (See Table 2). A two-tailed test statistic is calculated as the share of estimates whose absolute value is greater than or equal to the estimate obtained using the actual rollout. Calculating this statistic gives an implied p-value of .01 as compared with the p-value of .032 given by the standard errors clustered at the county level.

Figure A10: Relationship between Year of Initial Head Start Funding and Baseline County Characteristics



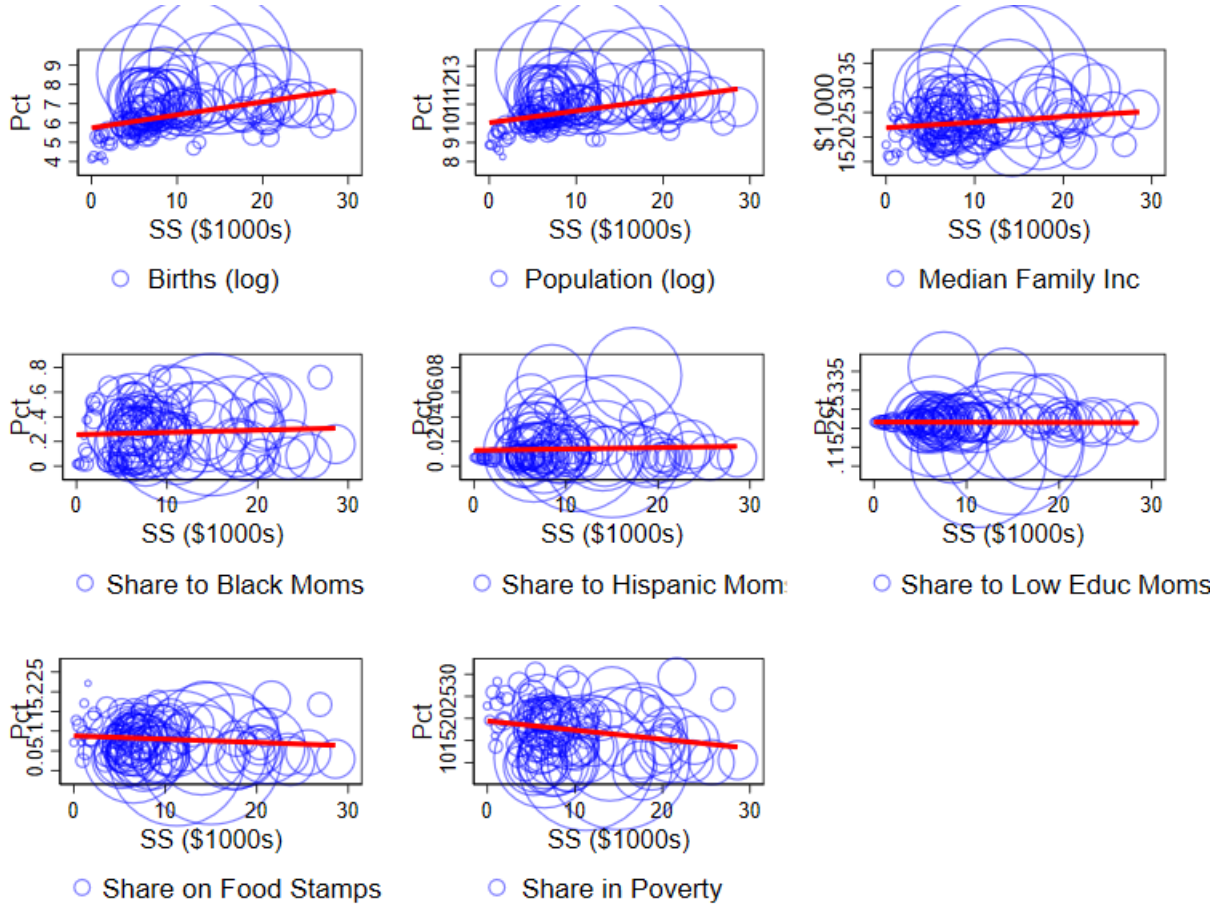
Note: Figure shows population weighted scatterplots of county characteristics against the year in which Head Start first became available in that county. Data are at the county level and weights are defined using 1955 births (represented by circle radius). A flat, horizontal fitted line suggests that the values of a given county characteristic are not systematically connected to the timing of Head Start availability.

Figure A11: Relationship between Year of Initial Smart Start Funding and Baseline County Characteristics



Note: Figure shows population weighted scatterplots of county characteristics against the year in which Smart Start first became available in that county. Data are at the county level and weights are defined using 1980 births (represented by circle radius). A flat, horizontal fitted line suggests that the values of a given county characteristic are not systematically connected to the timing of Smart Start availability.

Figure A12: Relationship between Total Smart Start Funding Levels and Baseline County Characteristics



Note: Figure shows population weighted scatterplots of county characteristics against the total amount of Smart Start funding received in a county over the sample period. Data are at the county level and weights are defined using 1980 births (represented by circle radius). A flat, horizontal fitted line suggests that the values of a given county characteristic are not systematically connected to total Smart Start funding.

Appendix: Supplementary Tables

Table A1: Relationship Between Head Start Ever Available in Sample Period and Baseline County Characteristics

	All	High Poverty	Low Poverty
	(1)	(2)	(3)
Head Start Ever Available In County			
1960 CCDB: % of land in farming	0.00353	0.00782	0.00110
	(0.0177)	(0.0290)	(0.0265)
1960 CCDB: % of people living in families with $\leq$ \$3000	-0.0503	0.0939	-0.381***
	(0.0369)	(0.0945)	(0.123)
1960 CCDB: % of population urban	-0.0297	-0.00259	-0.0712
	(0.0283)	(0.0427)	(0.0621)
1960 CCDB: % of people black	0.00244	0.0233	0.00175
	(0.0259)	(0.0272)	(0.0472)
1960 CCDB: % of people $\leq$ age 5	-0.364	-0.766	0.696
	(0.404)	(0.488)	(0.678)
1960 CCDB: % of people $\geq$ age 65	-0.112	-0.568	1.137
	(0.337)	(0.423)	(1.014)
1960 CCDB: % of employment in agriculture	-9.870	-25.45	16.11
	(14.23)	(20.82)	(20.80)
1960 CCBD: log population	1.720**	0.988	4.548*
	(0.717)	(0.791)	(2.478)
Observations	100	50	50
Mean	0.630	0.440	0.820

Note: Each column reports a separate logistic regression of an indicator for whether a county ever got Head Start by 1976 against the eight county level characteristics recommended in Hoynes and Schanzenbach (2009) and drawn from the 1960 City and County Data Books (CCDB). Observations are at the county level. Those counties whose poverty rate in 1960 was above the median in North Carolina (40.2% poverty) are called "High Poverty", while those below the median are called "Low Poverty". Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$)

Table A2: Relationship Between Year of Initial Head Start Availability and Baseline County Characteristics

	All	High Poverty	Low Poverty
	(1)	(2)	(3)
First Birth Cohort in County To Have Head Start			
1960 CCDB: % of land in farming	0.00735	-0.0396	0.00827
	(0.0234)	(0.0658)	(0.0330)
1960 CCDB: % of people living in families with $\leq$ \$3000	-0.0375	0.0158	-0.138
	(0.0473)	(0.168)	(0.105)
1960 CCDB: % of population urban	-0.0122	-0.00119	-0.000130
	(0.0235)	(0.0358)	(0.0607)
1960 CCDB: % of people black	0.00251	0.0215	-0.0117
	(0.0328)	(0.0756)	(0.111)
1960 CCDB: % of people $\leq$age 5	-0.198	-0.187	0.115
	(0.387)	(1.374)	(0.582)
1960 CCDB: % of people $\geq$age 65	-0.418	-0.358	-0.350
	(0.259)	(0.898)	(0.266)
1960 CCDB: % of employment in agriculture	-4.219	-7.168	1.668
	(13.35)	(34.97)	(19.64)
1960 CCBD: log population	-0.397	1.003	-1.444
	(0.749)	(1.483)	(1.224)
Observations	63	22	41
Mean	0.381	0.0455	0.561

Note: Each column reports a separate OLS regression of the birth year (normalized to 1962) of the first birth cohort in a given county to which Head Start was available against the eight county level characteristics recommended in Hoynes and Schanzenbach (2009) and drawn from the 1960 City and County Data Books (CCDB). Observations are at the county level. Those counties whose poverty rate in 1960 was above the median in North Carolina (40.2% poverty) are called "High Poverty", while those below the median are called "Low Poverty". Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$)

Table A3: Relationship Between Smart Start Penetration in Sample Period and Baseline County Characteristics

	All	High Poverty	Low Poverty
	(1)	(2)	(3)
Total SS Funding ($1000s)			
Share to Hispanic Moms	-0.1437	0.0220	-0.1243
	(0.3496)	(0.3795)	(0.4813)
Share to Black Moms	-0.0118	-0.0246	0.0270
	(0.0280)	(0.0396)	(0.0616)
Share to Low Educ Moms	-0.0417	-0.1474	-0.0221
	(0.0936)	(0.2466)	(0.1180)
Births (log)	1.0846	0.2368	0.1913
	(3.5013)	(3.7876)	(5.7702)
Median Family Inc	-0.0081	-0.1464	-0.1077
	(0.1531)	(0.2513)	(0.2399)
Population (log)	-0.9304	-0.2198	-0.1293
	(3.6591)	(3.7706)	(5.9627)
Share on Food Stamps	0.0369	0.3222	-0.1260
	(0.1644)	(0.2478)	(0.2445)
Observations	100	50	50
Mean	2.0458	1.2738	2.2835

Note: Each column reports a separate OLS regression of total county-level Smart Start penetration funding against selected 1980 county characteristics. Following Ladd et al. (2014), the 1980 county characteristics include the share of births to black mothers, the share of births to Hispanic mothers, the share of births to low education mothers, the share of the population using food stamps, the total number of births, the total population, and the median family income. Observations are at the birth-county level. Those counties whose poverty rate in 1980 was above the median in North Carolina (17.3% poverty) are called "High Poverty", while those below the median are called "Low Poverty". Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$)

Table A4: Relationship Between Baseline and Predicted Crime Changes and Timing and Funding of Program

Crime Index		Program Availability	
	Mean	Coefficient	% of Mean
(1)	(2)	(3)	(4)
Panel A: Head Start			
Head Start Timing			
Adulthood Conviction Rate (Born 1955))	2.7	-0.0004**	1.4
		(0.0002)	
Δ Conviction Rate (1961-1968))	2.1	-0.0005**	2.6
		(0.0003)	
Head Start Funding			
Adulthood Conviction Rate (Born 1955))	2.7	-0.0005	1.7
		(0.0007)	
Δ Conviction Rate (1961-1968))	2.1	0.0008	3.6
		(0.0009)	
Panel B: Smart Start			
Smart Start Timing			
Adulthood Conviction Rate (Born 1980))	6.4	-0.0020	3.1
		(0.0014)	
Δ Conviction Rate (1989-1994))	-1.6	0.0005	2.9
		(0.0003)	
Smart Start Funding			
Adulthood Conviction Rate (Born 1980))	6.4	0.0002	0.3
		(0.0009)	
Δ Conviction Rate (1989-1994))	-1.6	0.0000	0.1
		(0.0003)	

Note: Estimates show the relationship between baseline county characteristics and the program timing and funding, respectively. Each row represents a separate OLS regression, weighted by number of births in the baseline year. An index measuring the conviction rate is the dependent variable in odd rows while an index measuring the trend in the conviction rate is the dependent variable in even rows. In rows 1 and 2 the birth year of the cohort first exposed to Head Start (normed to 1962) is the sole independent variable, while in rows 5 and 6 the cohort first exposed to Smart Start (normed to 1997) is the sole independent variable. In rows 3-4 and 7-8, the county total of Head Start or Smart Start funding per pupil is the sole independent variable. The data are at the county-level and contain the 63 ever exposed counties in rows 1-4, and all 100 counties in rows 5-8. The indexes are constructed by regressing the crime measure on baseline county characteristics and using those coefficient estimates to predict the crime measure for each county. Robust standard errors are in parentheses in column (3). Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$)

Table A5: Effect of Head Start Availability on Criminal Conviction - Robustness of High Poverty Estimates to Inclusion of Counties that Did Not Receive Head Start

	High Poverty			
	Baseline		All Counties	
	(1)	(2)	(3)	(4)
HS Availability	-0.0128**	-0.0128**	-0.0105**	-0.0092**
	(0.0058)	(0.0061)	(0.0040)	(0.0039)
Observations	308	308	700	700
Mean	0.0462	0.0462	0.0429	0.0429
Baseline Chars X Trend		X		X

Note: Each column reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. Observations are at the birth county by birth year level and are weighted by number of births in each county in 1955 multiplied by the inverse propensity to receive Head Start as defined by Stuart (2010). The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 35. The reported variable of interest is an indicator for whether Head Start was available to a given county birth cohort. All specifications include birth county and birth-cohort fixed effects, and, where indicated, 1960 county characteristics interacted with a time trend in birth cohort. 1960 county characteristics include: percent of land in farming, percent of people living in families with less than $3,000, percent of population in urban area, percent black, percent less than age 5, percent greater than age 65, and percent of employment in agriculture. The sample is restricted to those counties whose poverty rate in 1960 was above the median in North Carolina (40.2% poverty), the "High Poverty" counties. In the first two columns, the sample is restricted to counties that ever received Head Start between 1965 and 1976. In the second two columns, the sample includes counties that never received Head Start between 1965 and 1976. See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A6: Head Start Availability and Criminal Conviction - Continuous Measure of Poverty Estimates

	All	
	(1)	(2)
HS Availability	0.0059	0.0045
	(0.0059)	(0.0043)
HS Availability X Poverty	-0.0197*	-0.0188**
	(0.0111)	(0.0082)
Observations	882	882
Mean	0.0469	0.0469
Baseline Chars X Trend		X

Note: Each column reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 35. The reported variable of interest is an indicator for whether Head Start was available to a given county birth cohort interacted with the county poverty rate in 1960. (The reported estimates are also scaled up by a factor of 100.). All specifications include birth county and birth-cohort fixed effects, and, where indicated, 1960 county characteristics interacted with a time trend in birth cohort. 1960 county characteristics include: percent of land in farming, percent of people living in families with less than $3,000, percent of population in urban area, percent black, percent less than age 5, percent greater than age 65, and percent of employment in agriculture. These regressions do not restrict the sample based on the county poverty rate in 1960. See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A7: Smart Start Availability and Criminal Conviction – Binary Availability Measure

	(1)	(2)	(3)
	All	High Poverty	Low Poverty
SS Availability (Binary)	-0.0032*	-0.0067*	-0.0013
	(0.0018)	(0.0038)	(0.0021)
Observations	1500	750	750
Mean	0.0516	0.0512	0.0517

Note: Each cell reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth-cohort fixed effects. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 24. Binary Smart Start availability, the independent variable of interest, is defined as Smart Start penetration level above the 25^{th} percentile of penetration. See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A8: Effect of Smart Start Availability on Criminal Conviction - By Presence of Head Start, Binary Measure

	(1)	(2)	(3)	(4)
	All		High Poverty	
Panel A: Binary Measure				
SS Availability (Binary)	-0.0032*	-0.0099***	-0.0067*	-0.0098**
	(0.0018)	(0.0033)	(0.0038)	(0.0037)
Observations	1500	555	750	435
Mean	0.0516	0.0528	0.0512	0.0532
Head Start	All	No Head Start	All	No Head Start

Note: Each column reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth-cohort fixed effects. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of either UCR Part 1 property crimes or Part 1 violent crimes in North Carolina by age 24. Binary Smart Start availability, the independent variable of interest, is defined as Smart Start penetration level above the 25^{th} percentile of penetration. See the notes to Table 1 for additional sample restrictions and definitions. Columns (2) and (4) further restrict the sample to counties without a Head Start program by 1980. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A9: Smart Start Funding and Criminal Conviction - Continuous Measure of Poverty Estimates

	All	
	(1)	(2)
SS ($1000s)	0.0015	0.0053
	(0.0045)	(0.0040)
SS ($1000s) X Poverty	-0.0508*	-0.0721***
	(0.0266)	(0.0255)
Observations	1500	1500
Mean	0.0516	0.0516
Baseline Chars X Trend		X

Note: Each column reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 24. The reported variable of interest is an indicator for whether Smart Start was available to a given county birth cohort interacted with the county poverty rate in 1980. (The reported estimates are also scaled up by a factor of 100.). All specifications include birth county and birth-cohort fixed effects, and, where indicated, 1980 county characteristics interacted with a time trend in birth cohort. Following Ladd et al. (2014), the 1980 county characteristics include the share of births to black mothers, the share of births to Hispanic mothers, the share of births to low education mothers, the share of the population using food stamps, the total number of births, the total population, and the median family income. These regressions do not restrict the sample based on the county poverty rate in 1980. See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A10: Smart Start Funding and Criminal Conviction – Full Interaction with Presence of Head Start

	(1)	(2)	(3)
	All	High Poverty	Low Poverty
SS ($1000s)	-0.0049	-0.0036	-0.0028
	(0.0030)	(0.0028)	(0.0035)
No Head Start X SS ($1000s)	-0.0062	-0.0101**	-0.0073*
	(0.0046)	(0.0048)	(0.0041)
Observations	1500	750	750
Mean	0.0516	0.0512	0.0517

Note: Each column reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 24. The reported variable of interest is a measure of Smart Start funding penetration for a given county birth cohort, constructed following Ladd et al. (2014). In the second row, the reported variable of interest is the same measure of Smart Start funding penetration interacted with an indicator variable for whether the county was served by Head Start by 1980. All specifications include birth county and birth-cohort fixed effects. See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A11: Head Start Availability and Criminal Conviction – Effects by Conviction Age

	(1)	(2)	(3)
	All	High Poverty	Low Poverty
Conviction By Age 24			
Head Start Availability	-0.0005	-0.0044	0.0009
	(0.0015)	(0.0038)	(0.0015)
Mean	0.0234	0.0229	0.0236
Conviction By Age 30			
Head Start Availability	-0.0017	-0.0101*	0.0016
	(0.0024)	(0.0050)	(0.0025)
Mean	0.0376	0.0373	0.0377
Conviction By Age 35			
Head Start Availability	-0.0017	-0.0128**	0.0026
	(0.0031)	(0.0058)	(0.0033)
Mean	0.0469	0.0462	0.0471
Observations	882	308	574

Note: Each column reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth-cohort fixed effects. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 24, age 30 and age 35, respectively. The reported variable of interest is an indicator for whether Head Start was available to a given county birth cohort. See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A12: Smart Start Availability and Criminal Conviction – Effects by Conviction Age

	(1)	(2)	(3)
	All	High Poverty	Low Poverty
Conviction By Age 24			
SS ($1000s)	-0.0064**	-0.0118**	-0.0030
	(0.0029)	(0.0051)	(0.0035)
Observations	1500	750	750
Mean	0.0516	0.0512	0.0517
Conviction By Age 25			
SS ($1000s)	-0.0059*	-0.0119**	-0.0022
	(0.0032)	(0.0051)	(0.0038)
Observations	1400	700	700
Mean	0.0571	0.0561	0.0575
Conviction By Age 26			
SS ($1000s)	-0.0068*	-0.0138**	-0.0024
	(0.0039)	(0.0057)	(0.0047)
Observations	1300	650	650
Mean	0.0616	0.0603	0.0621

Note: Each cell reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth-cohort fixed effects. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 24, age 25 or age 26, respectively. The reported variable of interest is a measure of Smart Start funding penetration for a given county birth cohort, constructed following Ladd et al. (2014). See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A13: Effect of Early Childhood Education on Criminal Conviction – With Trends

	(1)	(2)	(3)	(4)	(5)	(6)
	All		High		Low	
Panel A: Head Start						
Head Start Availability	-0.0017	-0.0028	-0.0128**	-0.0128**	0.0026	0.0013
	(0.0031)	(0.0031)	(0.0058)	(0.0061)	(0.0033)	(0.0040)
Observation	882	882	308	308	574	574
Mean	0.0469	0.0469	0.0462	0.0462	0.0471	0.0471
Panel B: Smart Start						
SS ($1000s)	-0.0064**	-0.0059**	-0.0118**	-0.0110***	-0.0030	-0.0027
	(0.0029)	(0.0025)	(0.0051)	(0.0040)	(0.0035)	(0.0025)
Observations	1500	1500	750	750	750	750
Mean	0.0516	0.0516	0.0512	0.0512	0.0517	0.0517
Baseline Chars X Trend		X		X		X

Note: Each cell reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. In Panel A, observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 35. The reported variable of interest is an indicator for whether Head Start was available to a given county birth cohort. All specifications include birth county and birth-cohort fixed effects, and, where indicated, 1960 county characteristics interacted with a time trend in birth cohort. 1960 county characteristics include: percent of land in farming, percent of people living in families with less than $3,000, percent of population in urban area, percent black, percent less than age 5, percent greater than age 65, and percent of employment in agriculture. In Panel B, observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 24. The reported variable of interest is a measure of Smart Start funding penetration for a given county birth cohort, constructed following Ladd et al. (2014) All specifications include birth county and birth-cohort fixed effects, and, where indicated, 1980 county characteristics interacted with a time trend in birth cohort. Following Ladd et al. (2014), the 1980 county characteristics include the share of births to black mothers, the share of births to Hispanic mothers, the share of births to low education mothers, the share of the population using food stamps, the total number of births, the total population, and the median family income. See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A14: Effect of Early Childhood Education on Criminal Conviction - Robustness to Inclusion of Time-Varying Covariates

	(1)	(2)	(3)
	All	High Poverty	Low Poverty
SS ($1000s)	-0.0064**	-0.0114**	-0.0029
	(0.0029)	(0.0048)	(0.0032)
Observations	1500	750	750
Mean	0.0516	0.0512	0.0517
Covariates	Yes	Yes	Yes

Note: Each column reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth-cohort fixed effects, as well as a set of county-year level covariates. Following Ladd et al. (2014), the county-year level characteristics include the share of births to black mothers, the share of births to Hispanic mothers, the share of births to low education mothers, the share of the population using food stamps, the total number of births, the total population, and the median family income. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 24. The reported variable of interest is a measure of Smart Start funding penetration for a given county birth cohort, constructed following Ladd et al. (2014) See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A15: Other War On Poverty Programs and Head Start - High Poverty Counties

	High Poverty				
	(1)	(2)	(3)	(4)	(5)
Head Start Availability	-0.0128**	-0.0123*	-0.0122*	-0.0155**	-0.0149**
	(0.0058)	(0.0060)	(0.0062)	(0.0072)	(0.0065)
Observations	308	308	308	308	308
Mean	0.0462	0.0462	0.0462	0.0462	0.0462
Baseline Chars X Trend	X		X		X
WOP Controls	None	FS	FS	FS + Other WOP	FS + Other WOP

Note: Each column reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth-cohort fixed effects. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 35. The reported variable of interest is an indicator for whether Head Start was available to a given county birth cohort. All specifications include birth county and birth-cohort fixed effects, and, where indicated, 1960 county characteristics interacted with a time trend in birth cohort. 1960 county characteristics include: percent of land in farming, percent of people living in families with less than $3,000, percent of population in urban area, percent black, percent less than age 5, percent greater than age 65, and percent of employment in agriculture. In these specifications, controls for exposure to various War on Poverty programs, including the Food Stamp Program (FS) are also included. "Other War on Poverty Programs" are those recommended by Bailey and Goodman-Bacon (2015) and include per capita expenditures on Public Assistance Transfers, Medicaid expenditures, Community Health Centers and Community Action Agencies. See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$)

Table A16: Relationship between Head Start Availability and Possible Confounders

	All		High Poverty		Low Poverty	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: War on Povery Programs						
0-5 Food Stamp Exposure	-0.0429	-0.0429	0.0666	0.0666	-0.0639	-0.0639
	(0.0802)	(0.0802)	(0.1593)	(0.1593)	(0.0950)	(0.0950)
Public Assistance Transfers	6.9467	6.9467	0.7701	0.7701	11.5153**	11.5153**
	(5.7513)	(5.7513)	(6.7353)	(6.7353)	(4.5238)	(4.5238)
Medicaid	9.2430	9.2430	2.2806	2.2806	12.8272**	12.8272**
	(5.7438)	(5.7438)	(5.9320)	(5.9320)	(5.9322)	(5.9322)
Community Health Center Funds	681.9056	681.9056	-82.5776	-82.5776	916.6069	916.6069
	(521.5891)	(521.5891)	(502.3024)	(502.3024)	(720.6964)	(720.6964)
CAP Seniors Program Grant	0.0356	0.0356	0.0297	0.0297	0.0302	0.0302
	(0.0521)	(0.0521)	(0.0407)	(0.0407)	(0.0734)	(0.0734)
Legal Services Program Grant	0.0460	0.0460	-0.0013	-0.0013	0.0557	0.0557
	(0.0333)	(0.0333)	(0.0080)	(0.0080)	(0.0434)	(0.0434)
Panel B: Health						
Adjusted Mortality Rate, All Ages	0.8698	0.8698	19.5563	19.5563	-5.3087	-5.3087
	(9.0852)	(9.0852)	(17.5338)	(17.5338)	(9.3156)	(9.3156)
White, Infant Mortality Rate	0.4678	0.4678	0.2843	0.2843	0.1544	0.1544
	(0.8910)	(0.8910)	(1.3270)	(1.3270)	(1.4135)	(1.4135)
Nonwhite Infant Mortality Rate	-2.2424	-2.2424	-3.6512	-3.6512	-1.4567	-1.4567
	(2.3849)	(2.3849)	(3.8859)	(3.8859)	(3.2687)	(3.2687)
Infant Mortality Rate	-1.1731	-1.1731	-1.1761	-1.1761	-1.5512	-1.5512
	(0.9761)	(0.9761)	(1.1773)	(1.1773)	(1.0473)	(1.0473)
Neonatal Infant Mortality Rate	0.2524	0.2524	0.1171	0.1171	0.1792	0.1792
	(0.8686)	(0.8686)	(1.0237)	(1.0237)	(1.1750)	(1.1750)
Postneonatal Infant Mortality Rate	-1.4255*	-1.4255*	-1.2931	-1.2931	-1.7304**	-1.7304**
	(0.7816)	(0.7816)	(1.3091)	(1.3091)	(0.6643)	(0.6643)
Observations	882	882	308	308	574	574
Baseline Chars X Trend		X		X		X

Note: Each cell reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth-cohort fixed effects. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. In each row the dependent variable is a county-year measure of spending or infant health that could potentially confound our estimates of the impact of Head Start. All dependent variables are taken from Bailey et al (2015). The reported variable of interest is an indicator for whether Head Start was available to a given county birth cohort. All specifications include birth county and birth-cohort fixed effects, and, where indicated, 1960 county characteristics interacted with a time trend in birth cohort. 1960 county characteristics include: percent of land in farming, percent of people living in families with less than $3,000, percent of population in urban area, percent black, percent less than age 5, percent greater than age 65, and percent of employment in agriculture. See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A17: Effect of Early Childhood Education on Criminal Conviction - By Crime Type

	(1)	(2)	(3)
	All	High Poverty	Low Poverty
Panel A: Head Start			
Property Crime			
Head Start Availability	-0.0024	-0.0084***	-0.0000
	(0.0016)	(0.0028)	(0.0016)
Mean	0.0256	0.0255	0.0256
Violent Crime			
Head Start Availability	0.0007	-0.0045	0.0027
	(0.0017)	(0.0032)	(0.0018)
Mean	0.0213	0.0207	0.0215
Observations	882	308	574
Panel B: Smart Start			
Property Crime			
SS ($1000s)	-0.0025*	-0.0048*	-0.0006
	(0.0013)	(0.0024)	(0.0015)
Mean	0.0274	0.0265	0.0278
Violent Crime			
SS ($1000s)	-0.0039**	-0.0070**	-0.0024
	(0.0019)	(0.0030)	(0.0023)
Mean	0.0242	0.0247	0.0240
Observations	1500	750	750

Note: Each cell reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth-cohort fixed effects. Panel A reports results using the Head Start sample. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of either a UCR Part 1 property crime or a Part 1 violent crime in North Carolina by age 35. Panel B reports results using the Smart Start sample. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals in a given birth county and birth year cohort that are later convicted of either a Part 1 property crime or a Part 1 violent crime in North Carolina by age 24. The reported variable of interest is a measure of Smart Start funding penetration for a given county birth cohort, constructed following Ladd et al. (2014). See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A18: Effect of Smart Start Funding on Criminal Conviction - By Race, By Presence of Head Start

	(1)	(2)	(3)	(4)
	All		High Poverty	
White				
SS ($1000s)	-0.0026	-0.0067	-0.0029	-0.0066
	(0.0020)	(0.0048)	(0.0053)	(0.0061)
Observations	1329	470	674	372
Mean	0.0315	0.0300	0.0272	0.0267
Non-White				
SS ($1000s)	-0.0191***	-0.0270***	-0.0216***	-0.0287***
	(0.0064)	(0.0051)	(0.0057)	(0.0055)
Observations	1313	454	662	360
Mean	0.0948	0.0820	0.0805	0.0804
Head Start	All	No Head Start	All	No Head Start

Note: Each cell reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth-cohort fixed effects. Results are reported separately for white cohorts and non-white cohorts. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals of a given race in a given birth county and birth year cohort that are later convicted of either UCR Part 1 crimes in North Carolina by age 24. The reported variable of interest is a measure of Smart Start funding penetration for a given county birth cohort, constructed following Ladd et al. (2014). (Sample sizes are smaller for these specifications because from 1989 to 1993 the natality files for 25% of counties in North Carolina do not have race breakdowns. For these years, race is available only for counties in which 1980 populations for the non-white group formed at least 10 percent of the total population or numbered at least 10,000.) See the notes to Table 1 for additional sample restrictions and definitions. Columns (2) and (4) further restrict the sample to counties without a Head Start program by 1980. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A19: Effect of Early Childhood Education on Criminal Conviction- By Sex

	(1)	(2)	(3)
	All	High Poverty	Low Poverty
Panel A: Head Start			
Male			
Head Start Availability	-0.0039	-0.0225*	0.0034
	(0.0053)	(0.0111)	(0.0053)
Observations	882	308	574
Mean	0.0794	0.0790	0.0796
Female			
Head Start Availability	0.0004	-0.0037**	0.0019
	(0.0014)	(0.0017)	(0.0016)
Observations	882	308	574
Mean	0.0145	0.0139	0.0147
Panel B: Smart Start			
Male			
SS ($1000s)	-0.0105**	-0.0163**	-0.0057
	(0.0045)	(0.0070)	(0.0057)
Observations	1500	750	750
Mean	0.0795	0.0783	0.0801
Female			
SS ($1000s)	-0.0020	-0.0077**	0.0003
	(0.0019)	(0.0035)	(0.0017)
Observations	1500	750	750
Mean	0.0230	0.0237	0.0228

Note: Each cell reports a separate OLS regression with standard errors clustered at the birth county level and reported in parentheses. All specifications include birth county and birth cohort fixed effects. Panel A reports results using the Head Start sample for male and female cohorts separately. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1955. The dependent variable is the fraction of individuals of a given sex in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 35. The reported variable of interest is an indicator for whether Head Start was available to a given county birth cohort. Panel B reports results using the Smart Start sample for male and female cohorts separately. Observations are at the birth county by birth year level and are weighted by the number of births in each county in 1980. The dependent variable is the fraction of individuals of a given sex in a given birth county and birth year cohort that are later convicted of a UCR Part 1 crime in North Carolina by age 24. The reported variable of interest is a measure of Smart Start funding penetration for a given county birth cohort, constructed following Ladd et al. (2014). (Sample sizes are smaller for these specifications because from 1989 to 1993 the natality files for 25% of counties in North Carolina do not have race breakdowns. For these years, race is available only for counties in which 1980 populations for the non-white group formed at least 10 percent of the total population or numbered at least 10,000.) See the notes to Table 1 for additional sample restrictions and definitions. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Table A20: Head Start and Likelihood of Residing in One's State of Birth (Census)

	National				South			
	All		Men Only		All		Men Only	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fraction with HS Avail.	-0.021	0.004	-0.017	0.009	-0.018	-0.009	-0.013	-0.005
	(0.016)	(0.008)	(0.017)	(0.008)	(0.029)	(0.020)	(0.029)	(0.021)
Obs	3,150,292	3,150,292	1,546,355	1,546,355	1,002,875	1,002,875	487,059	487,059
Mean	0.66	0.66	0.66	0.66	0.68	0.68	0.68	0.68
State Linear Trend		X		X		X		X

Note: Each cell represents a separate OLS regression with standard errors clustered at the state of birth level (in parentheses). Observations are at the individual level from the 1990 and 2000 Census. The dependent variable is whether an individual is currently living in his or her state of birth. The key explanatory variables are measures of Head Start availability for a birth cohort in a particular state. This is the weighted average of the Head Start availability variable across counties in a state, where the weights are the number of births in each county in 1960. All specifications include birth state and birth year fixed effects as well as indicators for race, age, and sex. Sample restricted to ages 18-35. Significance levels indicated by: *($p < 0.10$), **($p < 0.05$), ***($p < 0.01$).

Appendix B: Effect Size Comparison Explanation and Backup

Figure A1 consolidates information on effect size and estimate precision from a comprehensive review of studies that contain estimates of the causal effect of an early childhood education program on the likelihood that an individual will become a criminal. These studies are included in Appendix Table B1 and B2 and focus on the evaluation of three separate early childhood education programs: Head Start, Perry Preschool, and the Abecedarian Project.

Perry Preschool and the Abecedarian Project were pilot interventions and enrolled cohorts from the later 1960s and 1970s, respectively. Head Start, the nation's largest early childhood education program, has been in operation since mid-1960s. Two studies that include criminal justice outcomes focus on cohorts from the 1960s and 1970s, while a third focuses on cohorts born in the early 1980s. The period of focus of each study is illustrated by its location on the x-axis in Figure A1.

A common theme among these studies is the relatively small sample size available to the researchers (column (5) of Table B1). In the case of Perry Preschool and the Abecedarian Project, the small-scale single site nature of the interventions resulted in samples of around 100 individuals in each case. In the case of the Head Start evaluations, while the number of program participants was extremely large, researchers were limited by the sample sizes available to them in survey data such as the PSID and CNLSY.

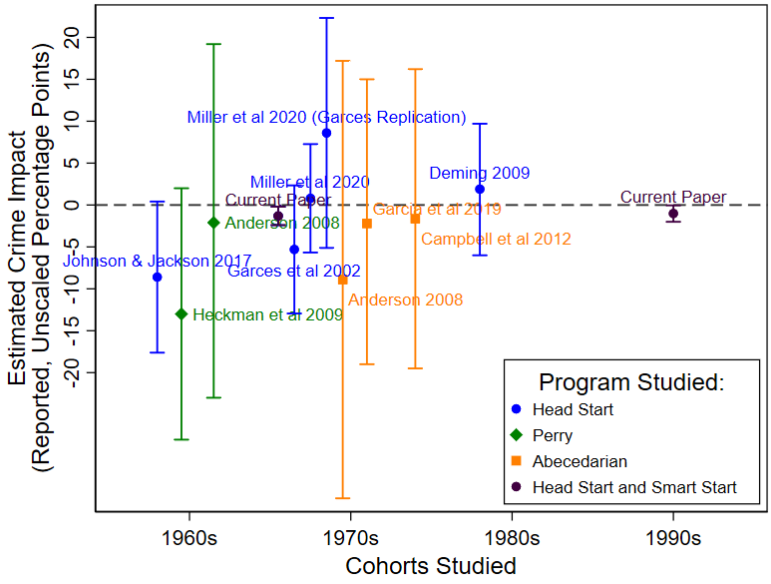
The estimates underlying Figure A1 are displayed in Figure B1 and contained in column (4) of Table B2. Whenever there are several outcomes considered in one study, we display the estimates for outcome measures which are closest to the measures of criminality we study (namely, likelihood of criminal conviction by adulthood). Having said that, it is important to note that some of the reported estimates measure criminality by the likelihood of arrest, some by the likelihood of conviction, and some by the likelihood of incarceration.

In each case we display the estimated effect likelihood that an individual will become a criminal per \$1,000 (2015 dollars).[76] For example, in the case of Johnson and Jackson (2017), we take the estimates from Row 7, Col 10 of Table 2 and scale by the average Head Start funding levels in thousands of 2015 dollars (i.e., 6.027). (The raw reported and unscaled estimates are displayed in Figure B1 below.) The 95 percent confidence intervals are produced similarly and are contained in the final column of the table.

In Table B2 we provide additional details about the estimates used in Figure A1 and Table B2, including the raw estimate, the measure of criminality, the associated mean, whether the criminal behavior was self-reported, and characteristics of the sample (age at observation and subgroup). We also provide additional information on alternative estimates for other criminal measures or subgroups contained in each study.

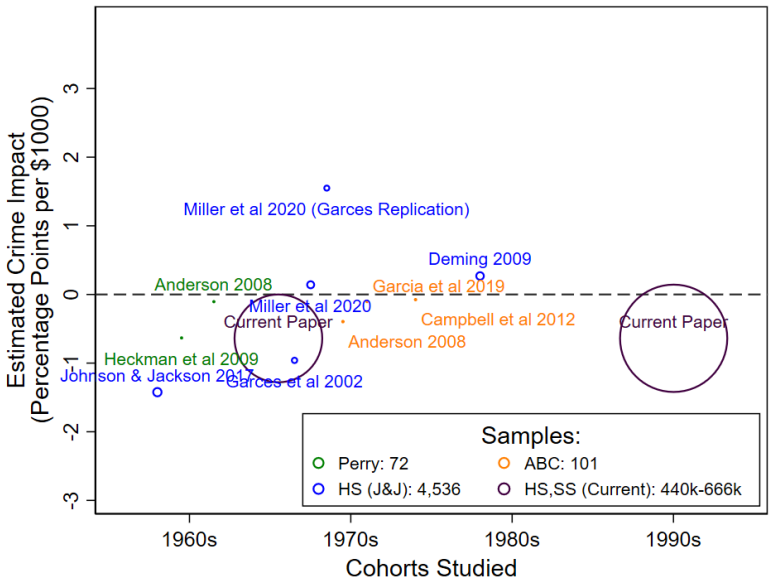
[76] Program costs for Perry and Abecedarian are taken from Heckman et al (2009) and Garces et al (2002), respectively, and are \$20,648 and \$22,687 respectively. Because program funding for Head Start varies across cohorts and different papers study different cohorts, we use the funding reported by each study. For Deming (2009), Garces et al (2002) and Johnson and Jackson (2017) those numbers are, respectively, \$6,981, \$5,540, and \$6,027.

Figure B1: Effect Size Comparison: Reported, Unscaled Estimates



Note: Figure shows estimated crime impact and associated 95 percent confidence intervals. In this figure, as opposed to Figure A1, the estimates displayed are simply the estimates reported in each study with no effort to scale by cost.

Figure B2: Effect Size Comparison: Sample Sizes



Note: Figure shows estimated crime impact per $1,000 (2015 dollars). See Figure A1a and Appendix B for additional details on construction. Bubbles indicate number of individuals represented in each study. Figures presented for Head Start and Smart Start are for high-poverty counties only. Total numbers of individuals are presented in Table 1 and are 1,487,225 (Head Start) and 1,407,042 (Smart Start). See column (4) of Table B1 for exact sample sizes of other studies.

Table B1: Estimates and confidence intervals are reported per $2015 dollars.

Study	Strategy	Program	Cohorts	N	Estimate	CI
Deming 2009	Family FEs	Head Start	1980-1986	3698	.27	[-.85-1.38]
Garces et al 2002	Family FEs	Head Start	1966-1977	1,742	-.96	[-2.34-.42]
Miller et al 2020	Family FEs	Head Start	1966-1977	1535	1.55	[-.92-4.03]
Miller et al 2020	Reweighted FFE	Head Start	1966-1977	3206	.14	[-1.02-1.31]
Johnson and Jackson 2017	DD	Head Start	1950-1976	4536	-1.42	[-2.91-.07]
Precision weighted avg.	-	Head Start	-	-	-.21	[-.82-2.06]
Anderson 2008	RCT	Perry	1962-1967	72	-.10	[-1.11-.92]
Heckman et al 2009	RCT	Perry	1962-1967	72	-.62	[-1.35-.09]
Anderson 2008	RCT	Abecedarian	1972-1977	111	-.39	[-1.54-.75]
Garcia et al 2019	RCT	Abecedarian	1972-1980	143	-.04	[-.79-.66]
Campbell et al 2012	RCT	Abecedarian	1972-1980	101	-.072	[-.85-.71]

Note: Program costs for Perry and Abecedarian are taken from Heckman et al (2009) and Garces et al (2002), respectively, and are $20,648 and $22,687 respectively. Because program funding for Head Start varies across cohorts and different papers study different cohorts, we use the funding reported by each study. For Deming (2009), Garces et al. (2002) and Johnson and Jackson (2017) those numbers are, respectively, $6,981, $5,540, and $6,027. The selected effect estimates are provided in column (4) of Table B2.

Table B2: Estimates and Confidence Intervals: Detailed Breakdowns

Study	Program	Direct Reference	Reported Estimate	Mean	Measure	(Sub)Group	Age	Other Measures	Self Reported?
Current Paper	Head Start	Table 2 Row 1 Col 1	1.3 (.57) pp	4.7	Conviction	Overall	By Age 35	1.3 (.68) pp for Non-White	Administrative
Current Paper	Smart Start	Table 2 Row 2 Col 2	.65 (.30) pp	5.1	Conviction	Overall	By Age 24	2.1 (.59) pp for Non-White	Administrative
Deming 2009	Head Start	Table 5 Row 7 Col 1	1.9 (4.0) pp	NA*	Conviction, Probation, Sentence or Incarceration	Overall	Young Adulthood (Avg. Age: 23)	.2 (5.7) for Women 5.1 (5.0) for Black	Self Reported
Garces et al 2002	Head Start	Table 2 Row 12 Col 4	-5.3 (3.9) pp	10.0	Charge or Conviction	Overall	Young Adulthood (Avg. Age: 24)	-11.7 (5.9) for Black -10.3 (7.0) for Black and No High School	Self Reported
Miller et al 2020	Head Start	Table C7 Row D Col 3	.8 (3.3) pp	10.6	Charge or Conviction	Overall	By 1995 PSID Interview (Avg. Age 23)		Self Reported
" " (Garces Replication)	Head Start	Table D3 Row D Col 4	8.6 (7.0) pp	10.6	Charge or Conviction	Overall	By 1995 PSID Interview (Avg. Age 23)		
Johnson and Jackson 2017	Head Start	Table 2 Row 7 Col 10	-8.56 (4.57) pp	8	Incarceration	Overall	By Last PSID Interview (Up to Age 50)	-2.54 (1.48) pp Without Using IV -14.6 (12.5) for Women	Reported by Family Member*
Anderson 2008	Perry	Table 10 Row 5 Col 9	-2.1 (10.9) pp	71.8	Record	Men	By Age 27	-2.31 (1.50) for Nmb Male Lifetime Arrests	Admin (limitations)
Heckman et al 2009	Perry	Table 2 Row 20 Col 5-6	13 (7.7) pp	95	Arrest	Men	By Age 40	9 (14) pp for Women -10.1 (7.9) for Women	Admin (limitations)
Anderson 2008	Abecedarian	Table 10 Row 1 Col 9	-8.9 (13.3) pp	34.8	Conviction	Men	By Age 21	-11.3 (11.7) for Male Felony Convictions	Self Reported
Campbell et al 2012	Abecedarian	Table 5 Row 1	-1.64 (9.0) pp (*)	28.6	Conviction	Overall	By Age 30		Self Reported
Garcia et al 2019	Abecedarian	Table 1	-2.27 (8.41) pp (*)	51.3 (*)	Arrest or Sentencing	Overall	By Age 34		Administrative

Note: Raw estimates used to construct the estimates listed above in Table B2 and depicted in Figures A1. For each study the estimate referenced in column (3) is reported in column (4) as a percentage point impact with standard errors in parentheses. In column (5) control group means are reported in percentage point terms. * In the case of Deming (2009), the mean was not reported or easily inferred. In Johnson and Jackson (2019), the "ever jail" variable appears to have been constructed using "non-response" reports of incarceration from family members that indicate whether a family member reported another remember being in a prison or jail at the point of the sample interview (which occurred annually until 1999 and roughly every other year thereafter), perhaps combined with retrospective self-reports in a 1995 criminal history survey. In the case of Garcia (2019), authors inferred the estimates and standard errors using reported counts by treatment and outcome status and performing a basic t-test. In the case of Campbell (2012), authors inferred the estimates from reported odds ratios and reported counts of convictions or reported odds ratios to percentage point terms. For both Garcia (2019) and Campbell (2012), standard errors are derived from basic t-tests. In column (9), alternative measures studied in the paper are reported in the same format as in column (4). Column (10) lists whether or not the criminal measure in the study was self-reported by a survey respondent or obtained through administrative data. In both Anderson (2008), Heckman et al (2009), and Garcia et al (2019), limitations on the success of the administrative data linkages are noted. In Garces et al (2002), we use the family fixed effect estimate of Head Start, rather than the difference between Head Start and other preschool attendance, to ease comparison across studies. Lastly, for Miller et al (2020), we report both the authors replication of Garces et al (2002), as well as the authors own estimates using an empirical strategy that reweights the FFE estimate to recover the ATE in the same dataset. In each row, the estimate reported in the figure and listed in this Appendix's Table B2 is calculated by scaling the estimate reported in column (4) of this table by program costs in 2015 dollars.

Appendix C: Simulation Exercises

One of the key advantages of our administrative data relative to commonly used survey data sets (i.e. PSID and NLSY79) is the larger number of individuals supporting the resulting estimates (Appendix Figure B2). In particular, the larger number of individuals within counties (an average of 5000 versus an average of 4-9) generates a significant increase in our ability to identify true effects on criminal behavior as well as the precision of the resulting estimates. We demonstrate the advantages of increasing within county sample size via a series of simulation exercises based on random draws from an augmented version of our actual North Carolina data set.

We begin by taking our actual North Carolina data set of convicted individuals and adding in a row for each individual born in each county and year who was not convicted (backed out of the natality data). We focus on the sample of high-poverty counties, augmenting this data set by adding 50 copies such that the resulting data set contains 1,122 counties, a rough estimate of the number of counties available in the PSID (Johnson and Jackson 2019). We think of the resulting data set as representing the full population of individuals born in these counties in the relevant birth years, with a separate row for each individual indicating their year and county of birth as well as whether or not they were convicted.

To approximate the sampling designs of the PSID and NLSY we then repeatedly draw stratified random samples of size $n \in \mathcal{N} = \{2700, 4536, 10000, 50000\}$ from this data set, drawing from each county of birth in the same proportion as the county populations. The first two elements in $\mathcal{N}$ correspond to the sample sizes used in recent studies of Head Start that rely on rollout designs in the NLSY and PSID, respectively; the larger sample sizes were chosen arbitrarily. The chosen samples resulted in corresponding average sample sizes within county of 2.41, 4.04, 8.91, and 44.56, with significant variation across counties depending on the underlying population. For each sample size we drew 200 stratified random samples from the augmented data set and ran our primary specification using the true assignment of Head Start entry. The resulting distribution of estimates is displayed in red in Figure C1, with the vertical red line to the left of zero indicating the effect size in the population (the augmented data set). At the sample sizes available in the PSID (~ 4536), the distribution of estimated coefficients remains very wide, with reductions in crime of as large as 5.8 percentage points as well as increases in crime of 3.0 percentage points falling within the range of 95% of the estimates.

While the confidence intervals are quite wide for the smaller samples, it is perhaps informative to compare them to the simulated distribution of estimates when the expected effect is zero. To do this, we randomly assigned year of Head Start entry 100 times, drawing 20 stratified samples for each, and estimated our main specification. The resulting distribution of estimates is plotted in grey (Figure C1). As expected, it is centered around zero. While the distribution of true estimates is shifted somewhat to the left of the placebo estimates, it is clear from the figure that for many negative and reasonably sized point estimates our likelihood of obtaining them in small samples is not that much greater when there is no effect as when there is a substantial one.

Figures C2 and C3 perhaps make this point more clearly in terms of obtaining a sta-

tistically significant effect. As expected, the distribution of placebo p-values in Figure C2 is relatively flat under the assignment of placebo entry years, with only 5 percent of the estimates having a p-value of 0.05 or less across sample sizes (as expected). While the the mass of low p-values is larger under the true assignment (when there is an effect), it is only modestly larger for the smaller samples available in the PSID and NLSY79. Indeed, Figure C3 demonstrates that at these low sample sizes and with the true assignment (when there is an effect) we can only reject the null of no effect at the 5 percent level roughly 8 percent of the time.[77] If we assume that the ex-ante odds of there being a true effect or not being an effect are equivalent (i.e., balanced priors), we will falsely claim that a significant effect is a true effect $5/(5+8) \approx 38$ percent of the time. Figure 3 illustrates how this false discovery rate falls as we increase the sample size (and thus the number of individuals within each county).

A second and perhaps obvious point is that in the smaller samples (where the probability of a false positive report is higher), we will be more likely to achieve statistical significance when the estimated effect is large, suggesting an upward bias in the magnitude of *published* effects when there is a preference for statistically significant results.[78] As an illustration of this point, we have estimated our baseline specification in the National Longitudinal Survey of Youth 1979 (NLSY79). We focus on "ever in jail" because conviction information is not available in the data. In Table C1, we present the resulting estimates for the full sample and the set of high poverty counties. The point estimates in the full set of counties are actually positive. Despite containing over 1,500 counties of birth, the standard error of the estimate for the full set of counties here is three times the standard error for the estimates from the North Carolina data (Columns (1) and (2)). The estimate for high poverty counties is negative, but again the standard error is more than four times the standard error for the estimates from the North Carolina data (Columns (3) and (4)).

The key takeaway then is that the use of significantly larger sample sizes (such as those available in our administrative, North Carolina data) reduces the likelihood that a statistically significant estimated effect is a false positive, reduces the expected upward bias in *published* effects, and increases precision. The increase in precision is particularly important in estimating heterogeneity across types of counties. In our context, we can be more confident that early childhood education reduces criminal behavior, this effect is likely considerably smaller than some prior estimates suggest, and the effect appears to be concentrated in counties with high poverty levels.

In contrast, one of the advantages of the smaller survey data sets is additional background information on individuals. This information allows researchers to focus in on subpopulations where we might expect the effect to be larger due to an increased likelihood of program eligibility or participation. However, this focus comes at a cost as analytic samples get even smaller. In Table C2, we reproduce the main estimates for various subgroups with greater eligibility for Head Start.

[77] In other words, statistical power is 0.08.

[78] It is worth noting that the type of potential bias we refer to is likely to occur when there is a general preference among referees and editors for statistically significant effects. It suggests nothing about the internal validity of the underlying research strategies.

Again, the estimates are mixed and imprecise. In no case can we reject the null of no effect. Nor can we reject the effect sizes implied by our estimates using the North Carolina data. The results are largely uninformative as the study is simply underpowered to detect effect sizes of modest magnitude.

Our data have advantages not only over similar rollout approaches in survey data sets, but also over evaluations of small scale experiments and family fixed effect designs. As depicted in Figure A1, we have precision advantages over the small-scale experiments and family fixed effects approaches as well. All of the same relative advantages regarding false positives, an upward bias in reported effects, and precision apply here as well. In Appendix Table C3 we summarize the statistical power of each study to identify the average published effect size, following the guidance provided in the meta-analysis literature.[79] For example, to identify the simple mean effect size per $1,000 in the literature, the statistical power of the Head Start studies ranges from 0.09 to 0.13. Under balanced priors on the likelihood of an effect this implies false discovery rates of 42 to 54 percent. These types of concerns regarding effects on criminality are not limited to the Head Start literature; the randomized evaluations of the more intensive interventions are only modestly better powered. While we cannot provide a concrete example in the context of the small-scale experiments, it is illustrative that different evaluations of the *same* experiment have reached different conclusions regarding the effects on crime (Heckman et al., 2010; Anderson, 2008). Furthermore, the lack of statistical power perhaps explains the apparent lack of crime effects attributed to the Abecedarian intervention.

In the case of family fixed effects, the confidence intervals in Figure A1 illustrate the lack of statistical power of this approach. This problem is emphasized by Miller et al.'s (2020) inability to replicate Garces et al.'s (2002) crime effect estimates for Head Start using the same family fixed effects strategy and same data source. We have estimated similar models in the NLSY79 and again find positive estimates of Head Start participation on the likelihood of incarceration (Table C4).[80] While the outcomes and ages are somewhat different than those used in prior family fixed effect designs, this is quite surprising given prior estimates from the same general time period.[81] However, the inconsistency is resolved when one considers the large confidence intervals associated with this approach in these samples. In Figure C4 we demonstrate visually the relative advantage of our data and strategy over implementing the family fixed effects (FFE) and rollout approach in the NLSY79.

As a concluding point, it is worth emphasizing that all of these studies made large leaps forward in our understanding of the effects of early childhood education, making dramatic improvements in design over prior studies and providing innovative answers to previously unanswered or poorly answered questions. The above discussion is not intended to suggest

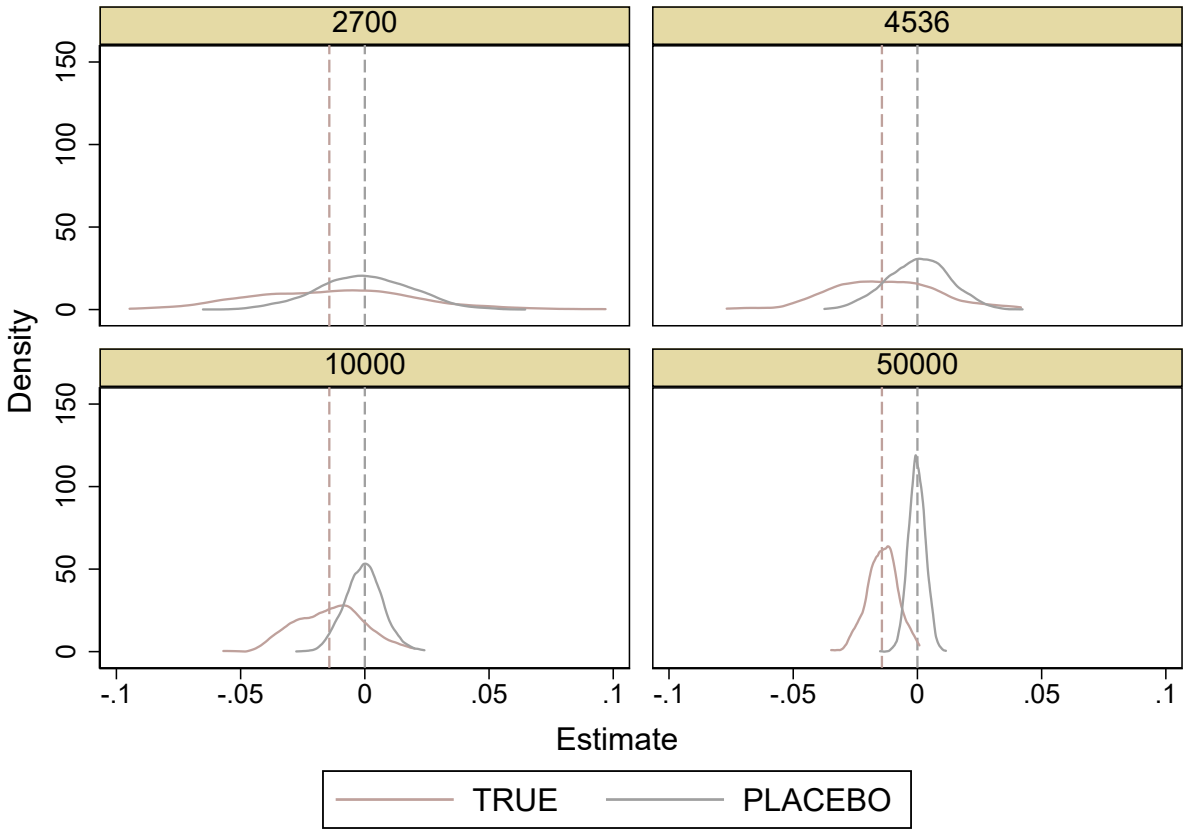
[79] This will be an overestimate of the true average effect size in the presence of a publication preference for statistically significant effects.

[80] We present estimates for blacks separately as earlier evidence suggested larger effects for this subgroup.

[81] Earlier studies use some combination of charges, arrests, and convictions to construct their binary criminal involvement variables. We have run similar specifications with an "ever charged" specification in the NLSY79 and find similar positive coefficients, although this question is only asked to individuals in 1980, when most are relatively young.

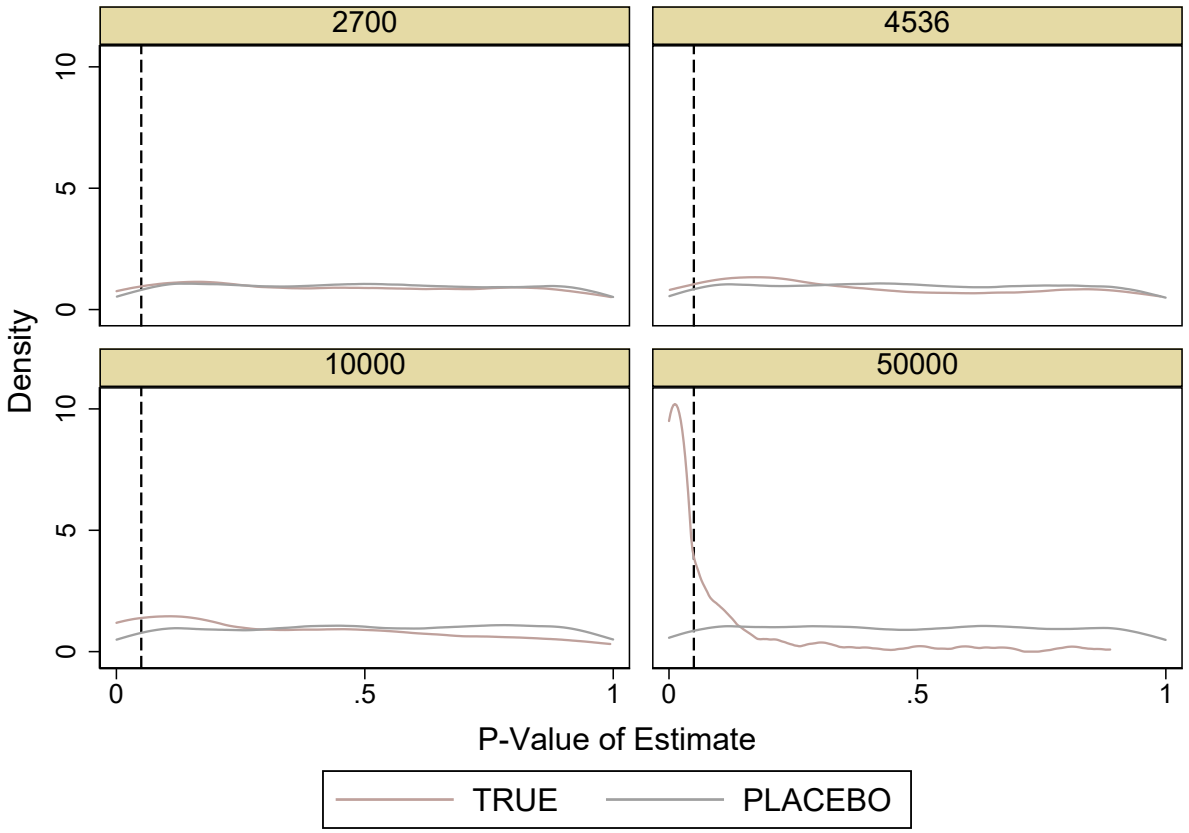
that estimates from prior path-breaking studies are wrong, only to highlight that limitations of the data used in these studies leave substantial remaining uncertainty regarding the true effects of these early childhood education programs on adult criminal behavior.

Figure C1: Distribution of Simulated Estimates by Sample Size



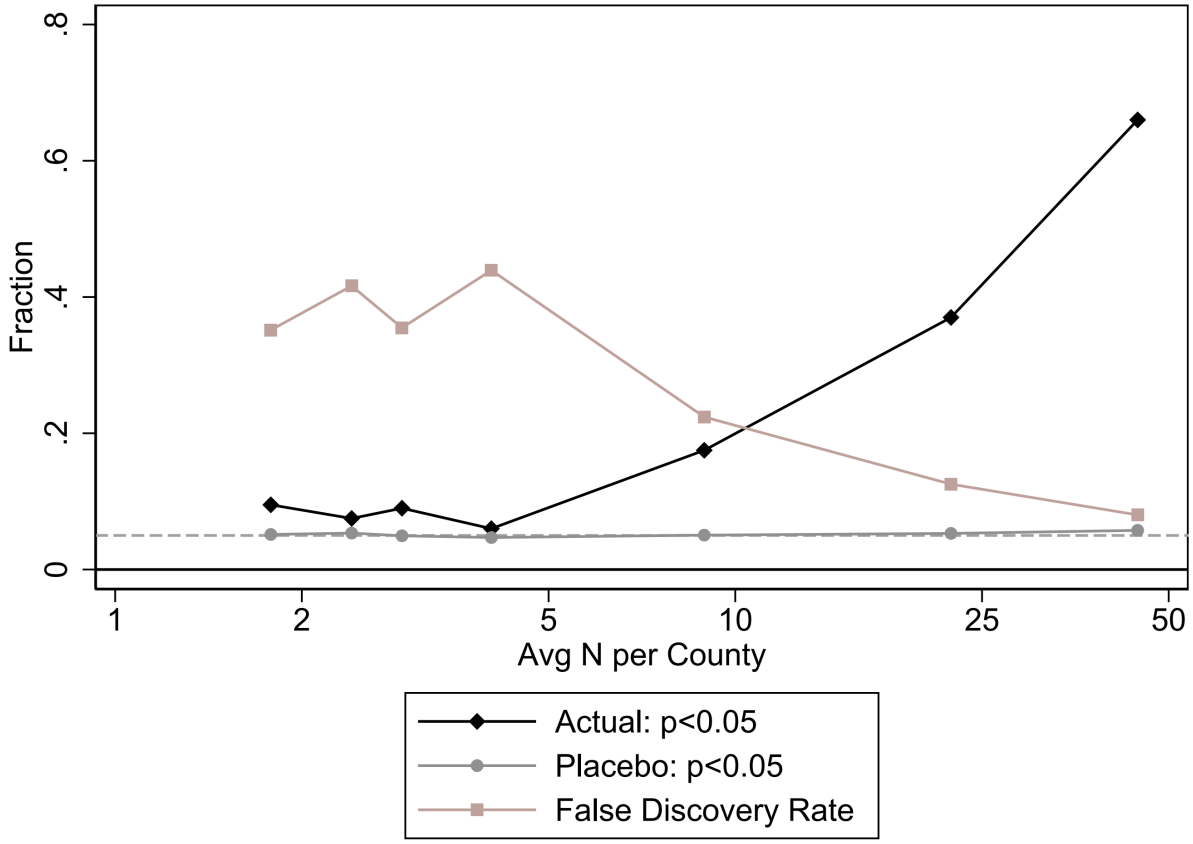
Note: Figure shows a distribution of estimates for our primary specification obtained for the true assignment of Head Start rollout (red) and for placebo assignments of Head Start rollout (grey) from drawing stratified random samples of various sizes from an augmented version of our North Carolina data set. For the true rollout assignment we draw 200 stratified random samples and for the placebo assignments we draw 20 stratified random samples for each of 100 random placebo rollout assignments. The augmented data set includes all convicted individualssin the administrative North Carolina crime dataset born in high poverty counties, uses natality files to add observations for non-convicted births in those counties, and then appends 50 additional copies of the data to increase the number of counties to reflect national survey data. To approximate the sampling designs of the PSID and NLSY, we repeatedly draw stratified random samples of size $n \in \mathcal{N} = \{2700, 4536, 10000, 50000\}$ from this data set, drawing from each county of birth in the same proportion as the county populations. The first two elements in $\mathcal{N}$ correspond to the sample sizes used in recent studies of Head Start that rely on rollout designs in the NLSY and PSID, respectively; the larger sample sizes were chosen arbitrarily. The chosen samples resulted in corresponding average sample sizes within county of 2.41, 4.04, 8.91, and 44.56. Each panel header identifies the overall size of the stratified random sample. Vertical dashed lines indicate the estimated effect for the full populationof the augmented data set.

Figure C2: Distribution of P-values by Sample Size



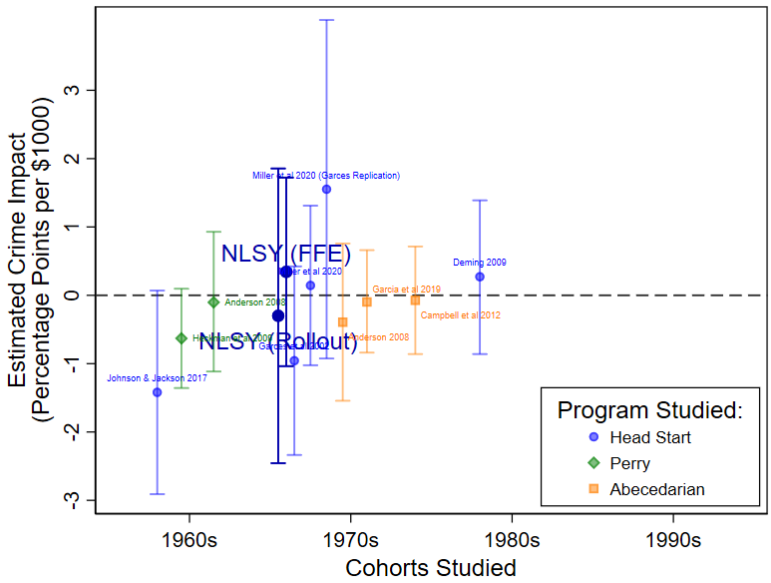
Note: The Figure depicts the distribution of p-values from the estimates in Figure C1. Dashed line shows p-value of .05. See the note to Figure C1 for more details.

Figure C3: Rejection Rates and Implied False Discovery Rate

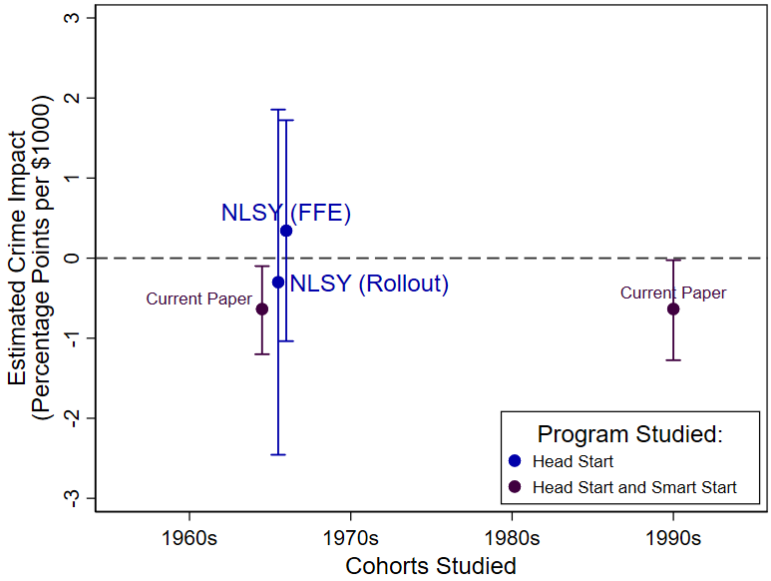


Note:
The Figure depicts the fraction of estimates of our primary specification from a series of random stratified samples of various sizes which are significant at the 5 percent level under the true assignment of Head Start rollout (black), under a random (placebo) assignment of Head Start rollout (grey), as well as the resulting rate of false discovery (red). These fractions are graphed on the y-axis against the average sample size per county for the sample draws that generated the fraction on the x-axis. The false discovery rate is defined as the number of false discoveries divided by the sum of false discoveries and true discoveries. This false discovery rate assumes that the ex-ante odds of there being a true effect and there being no effect are the same (i.e. balanced priors).

Figure C4: Effect Size Comparison: NLSY Comparison



(a) NLSY and Existing Literature



(b) NLSY and Current Paper

Note: Figure repeats Figure A1, this time adding two estimates obtained from using two different identification strategies in the NLSY79 dataset. The estimate labeled "NLSY (FFE)" replicates the family fixed effect strategy in Garces et al 2002 and Miller et al 2020. The estimate labeled "NLSY (Rollout)" replicates the strategy employed in the main analysis of this paper which identifies effects based on when Head Start funding became available to a county birth cohort.

Table C1: Head Start and Ever in Jail (NLSY)

VARIABLES	(1) All	(2) All	(3) High Poverty	(4) High Poverty
HS in County	0.006	0.004	-0.006	-0.013
	(0.009)	(0.009)	(0.022)	(0.024)
Covariates		X		X
Observations	9,617	9,617	1,500	1,500
Mean	0.0597	0.0597	0.0633	0.0633

Note: Table reports estimates of our baseline specification (reported above in main Table 2) using data from the National Longitudinal Survey of Youth 1979 (NLSY79). The outcome variable is an indicator variable for whether the respondent was "ever in jail". (Conviction information is not available in the NLSY79.) We present estimates on the full sample of counties in Columns (1) and (2) and the high poverty counties in Columns (3) and (4). See the note to main Table 2 for more details

Table C2: Head Start and Ever in Jail (NLSY)

VARIABLES	(1) Poverty (79)	(2) Poverty (79)	(3) Mom $\leq$ HS	(4) Mom $\leq$ HS	(5) Mom $<$ HS	(6) Mom $<$ HS
HS in County	-0.001	-0.015	0.004	-0.0001	0.028	0.019
	(0.024)	(0.027)	(0.011)	(0.011)	(0.018)	(0.019)
Covariates		X		X		X
Observations	1,539	1,539	5,086	5,086	2,670	2,670
Mean	0.111	0.111	0.0701	0.0701	0.0855	0.0855

Note: Table reports estimates with the same specification as in Table C1 above, but now broken down for various subgroups known to have greater Head Start eligibility. See the note the Table C1 for more details.

Table C3: Statistical Power and Implied False Discovery Rates to Identify Mean Effect Size

Study	Power	FPR
Deming 2009	0.13	0.43
Garces et al 2002	0.087	0.54
Johnson and Jackson 2017	0.095	0.51
Anderson 2008	0.15	0.40
Heckman et al 2009	0.26	0.28
Anderson 2008	0.13	0.43
Garcia et al 2019	0.26	0.28
Campbell et al 2012	0.20	0.33

Note: Power is calculated to identify the mean effect per $1,000 (2015 dollars) across studies (0.49 percentage points per $1,000). The false discovery rate (FDR) is calculated under balanced priors of the likelihood of an effect.

Table C4: Head Start and Ever in Jail (NLSY)

VARIABLES	(1) All	(2) All	(3) Black	(4) Black
Self Reported HS Participation	0.021	0.019	0.052	0.057
	(0.038)	(0.039)	(0.055)	(0.059)
Self Reported Preschool Participation	0.001	0.002	-0.021	-0.003
	(0.026)	(0.027)	(0.052)	(0.056)
Covariates		X		X
Observations	9,349	9,349	2,614	2,614
Mean	0.0597	0.0597	0.115	0.115

Note: Table reports estimates obtained using a family fixed effect strategy as in Garces et al (2002) and replicated in Miller et al (2020), but in the NLSY79. Each column reports estimates from a separate OLS regression. The outcome variable is an indicator variable for whether the respondent was "ever in jail". The explanatory variables of interest are indicators for whether the respondent reported attending Head Start ("Self-reported HS Participation") and whether the respondent reported attending some other preschool ("Self Reported Preschool Participation"). Estimates are performed separately for all respondents in columns (1) and (2), and for Black respondents in columns (3) and (4).

Appendix D: Implied Treatment on the Treated

These TOT estimates are based on estimated Head Start participation rates in high poverty counties of 15 to 21 percent. The lower bound is based on OEO statistics on state-level North Carolina Head Start enrollment in 1966 and the upper bound is based on author's calculations assuming the national per participant funding level is fixed across North Carolina counties. Our implied TOT effects are between half and two-thirds of the size of effects on somewhat similar measures reported in evaluations of the Perry Preschool program (11 to 12 percentage points on any arrest (or any charges) by age 40).[82,83] As in the Perry evaluation, we find larger effects on property crimes; Head Start access reduces the likelihood of a serious property conviction by 0.8 percentage points, a TOT effect of 4 to 5 percentage points in high-poverty counties (Appendix Table A17). While there is no significant effect on serious violent convictions, the point estimate (0.0046) implies a TOT of 2 to 3 percentage points. In comparison, Schweinhart et al. (2005) find a 16 percentage point reduction in violent arrests by age 40 (32 versus 48 percent) and a 22 percentage point reduction in property arrests by age 40 (36 versus 58 percent) in their evaluation of Perry preschool, four to five times the size of our effects.[84] Perry Preschool enrolled a very particular type of student: extremely disadvantaged, black children in Ypsilanti, Michigan. If we split our property crime estimates by race, we find similar effects for whites and non-whites.

If there are important spillover effects of program availability it is not reasonable to interpret these scaled estimates as TOT effects. In this case, improving the behavioral trajectories of a significant share of a group results in improvements for the group as a whole that are substantially larger than what we might expect to see if an individual was treated in isolation. Especially in high-poverty areas, a substantial fraction of children enrolled in Head Start. As these children interacted with others in their cohort, effects of the program might have spilled over to the other children in a way that would have been unlikely with the smaller treatment and control groups in experimental evaluations of small-scale programs. Particularly when it comes to criminal behavior, it is likely that these spillovers, operating through peer effects, are substantial. We therefore focus our discussion on the estimated effects of Head Start availability rather than participation.

[82]The treatment effect of Perry Preschool on any felony arrest, the definition of which overlaps substantially with Part 1 crimes, is even larger (15 percentage points), but is reported only for males (Heckman et al. 2009).

[83]Our TOT estimates are less than half of the effects estimated for the Nurse-Family Partnership by age 19 (16 percentage points on likelihood of conviction or arrest) (Olds et al. 1998, 2007). Our effect sizes are similar to recent estimates of the effects of early childhood Food Stamp access (Barr and Smith 2018). In contrast to both of these health interventions, which found strong effects on violent criminal behavior, the effects of Head Start access appear to be somewhat stronger on property crimes.

[84]Although we note that these are effects on *any* arrest and thus may not be directly comparable to convictions for a serious violent or property crime. Treatment estimates of Perry Preschool on the *number* of felony arrests indicates no significant difference in the number of serious violent crimes and a 90 % reduction in the number of felony property arrests (0.31 versus 2.91 per individual).

Appendix E: Event Study Heterogeneity

Recent work shows that when there is a staggered rollout of treatment, standard two-way fixed effects regressions provide event study estimates that are a weighted average of the county rollout-cohort specific effects (i.e., the effect from the set of programs adopted at time $t = 1965$, at $t = 1966$, etc.). In the presence of heterogeneous effects, this can lead to "contamination bias", where the coefficient on a given lead or lag can be contaminated by effects from other periods (Sun and Abraham 2020). This can lead to incorrect estimates of the average effect of a program as well as pre-treatment indicators that appear to be zero (i.e., no pre-trends) that are actually non-zero once the contamination of effects from other periods has been accounted for. Guided by the recent econometric and applied literature, we do three things to address the concern of contamination bias in our setting:

1. We implement a specification with a simple and intuitive control group that is not contaminated (i.e., the counties serving as the control have not been treated at the time they are used as control units). This allows us to estimate a treatment contrast without any county rollout-cohort heterogeneity and eliminates concerns about contamination bias.

2. We implement Sun (2020)'s techniques to build "interaction-weights" (IW) by event study indicator (i.e., each lead and lag) and county rollout-cohort groups. We then examine these weights for evidence of heterogeneity following prescriptions in Sun et al (2020). In the absence of heterogeneity, one need not be concerned about contamination bias influencing the estimates produced using the standard two-way fixed effects design.

3. We implement Chaisemartin (2018)'s techniques to produce dynamic treatment effect estimates which accommodate and account for treatment heterogeneity.

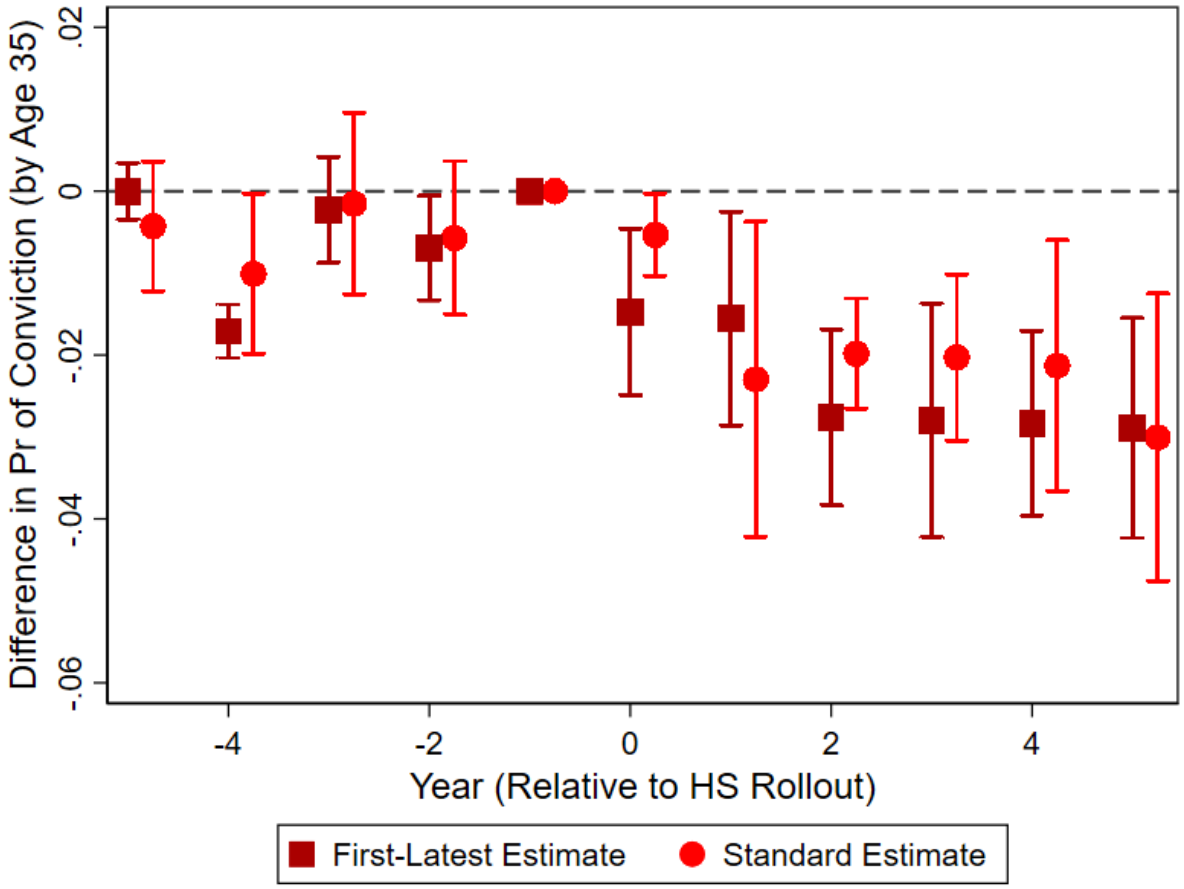
Overall, we find no evidence of contamination bias; if anything, heterogeneity in the county rollout-cohorts is slightly attenuating our standard estimates towards zero.

1. Simple, "Clean" Control Groups

Guided by the recent econometric and applied literature, we begin by presenting estimates from a simple and intuitive control group, which allows us to isolate a treatment contrast using a clean control group and intuitively test whether there are heterogeneous effects across county rollout-cohorts. Specifically, we focus on the first treated rollout-cohort (counties in which the first birth cohort exposed to Head Start was born in 1961) and include as a control group only the latest rollout-cohort (counties in which the first birth cohort exposed to Head Start was born in 1967 or 1970). To avoid contamination bias, we drop birth cohorts born after 1966 to ensure that the control county birth-cohorts are never treated in the estimation sample.

In other words, in the "first-latest" specification, the control group remains untreated throughout the sample time period. As Figure E1 shows, the results are similar to our main

Figure E1: Event Study of Head Start's Impact on Criminal Conviction – "First-Latest" Estimates



Note: "First-Latest Estimates" are the same as our standard estimates in Figure 2 except that they feature a control group consisting of the latest two county rollout-cohorts (1970 and 1967). Thus the control group consists of county-birth-cohort which are never treated in the sample period (but are treated just after the sample ends).

results. While the -4 estimate is statistically significant in this "first-latest" specification, the overall post-period results are slightly larger and suggest a sharp level drop. Overall, the results suggest little difference in the estimated effects when focusing on a clean difference-in-difference estimate of the effect of Head Start for the first county rollout-cohort, which accounts for much of our identifying variation. If anything, the estimates suggest a somewhat larger effect of treatment for the first county rollout-cohort, which would attenuate our main results towards zero.

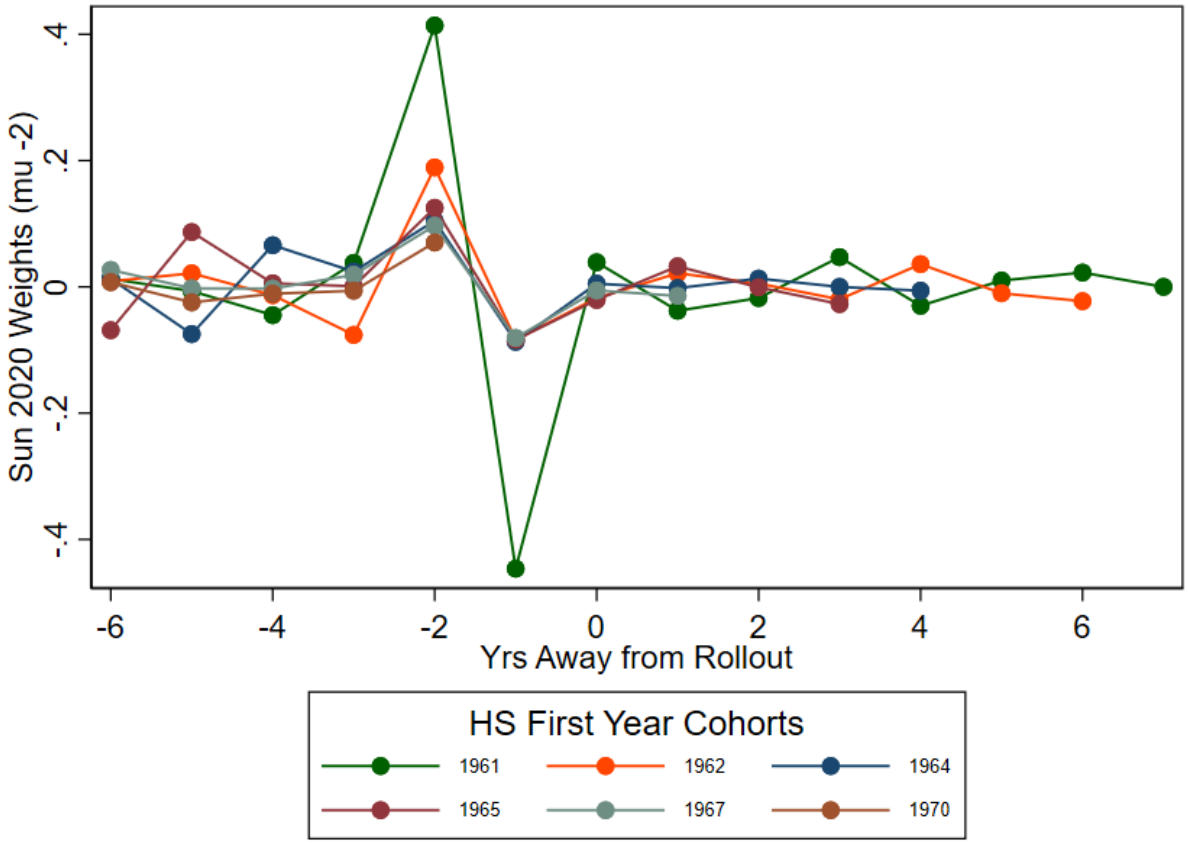
2. Diagnosing Contamination Bias using IW weights

An alternative approach to diagnosing the importance of treatment group heterogeneity is to evaluate the contribution of different county rollout-cohort groups and lead and lag effects to the estimate of each lead and lag coefficient (Sun 2020). Intuitively, if lead and lag effects are contributing in meaningful ways to the identification of other leads and lags, it implies that contamination bias is meaningful. We follow Sun (2020) and produce IW weights for the -2 estimator (two periods before treatment) to investigate the presence of "contamination bias" (Figure E2 below). We then confirm that, for the -2 estimator, the sum of the IW weights over treatment lags is zero for each rollout-cohort (Table E1 below).

The details are as follows: we implement Sun (2020)'s techniques to build "interaction-weights" (IW) by event study indicator and county rollout-cohort. An IW is a measure of a county rollout-cohort's contribution to the estimation of a given coefficient in the standard event study specification. For example, for the -2 estimator (two periods before treatment), we compute an IW weight for the first county rollout-cohort for each event study time period (-6, -5, etc.) and then for the second county rollout-cohort and so on. IW weights could show, e.g., that, in a given event study year, the 1st county rollout-cohort is contributing more to the -2 estimate than the 2nd county rollout-cohort; further, it could show, for example, that event study time periods other than -2 are contributing importantly to the identification of time period -2. We use these IW weights to test whether contamination bias is occurring in our data. If, for example, we see significant weight placed on event study time period 3, 4, or 5 (in identifying the estimate at -2), it would suggest contamination bias.

Figure E2 shows the IW weights estimated using methods in Sun (2020). Following our main specification, we compute the weights for our main sample (which are the ever-treated, high-poverty counties for birth cohorts 1955-1968). In our main specification we pooled all 6 county rollout-cohorts together, but for IW weights we break them apart. Following Sun (2020), we report the weights for the estimate of the indicator variable for two periods before treatment (-2 in our event study time) across rollout-cohorts. First, we note that the weights for each rollout-cohort are largest at event time=-2 with the first rollout-cohort contributing the most weight. Since these are weights for the -2 estimator, we expect each cohort to contribute a positive weight to the estimates of the -2 coefficient, and we expect the first rollout-cohort to have the most weight since more of the treated sample experienced treatment in the first rollout than in any other rollout-cohort. Secondly, we also expect the weights at -1 to be negative, since this is the excluded period in this specification. Finally, we note that the weights on the post-treatment indicators hover around zero and the sum of the post-treatment weights is very nearly zero (i.e. the sum of the IWs from our event study time =0 to =6). As Sun and Abraham (2020) notes, if the "weights are non-negative for lags of treatment", then this might suggest that the standard event study estimate for two periods before treatment is "sensitive to estimates of the dynamics effects ... and does not isolate the pre-trends" (p.33). Fortunately, as Table E1 below shows, our IW weights for lags of treatments are small and sum to $\leq .01$ for each county rollout-cohort. Overall, these IWs suggest that our estimate of the -2 indicator is not contaminated by rollout-cohort heterogeneity.

Figure E2: IW weights following Sun (2020)



Note: IW weights following Sun (2020). -1 is the excluded period.

Table E1: County Rollout-Cohort IW Weight Sums for Treatment Lags

County Rollout-Cohort	Sum of IW weights over Treatment Lags
1st Rollout	-.006
2nd Rollout	.010
3rd Rollout	005
4th Rollout	.005
5th Rollout	-.014

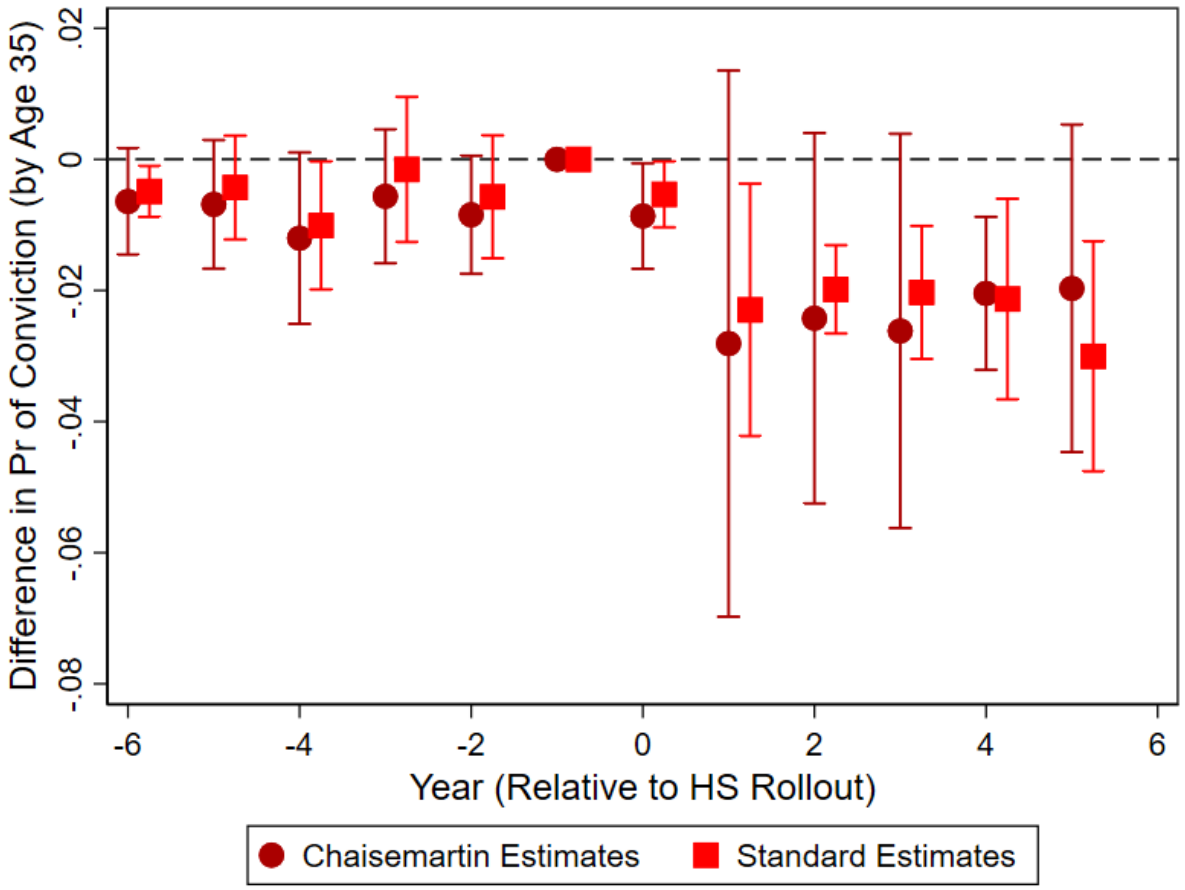
Note: Each row reports the sum of IW weights over treatment lags. Weights are calculated by implementing Sun (2020).

3. Accounting for Treatment Heterogeneity

While we have already provided evidence suggesting the lack of important treatment effect heterogeneity, another approach is to account for this treatment effect heterogeneity directly

in the estimation procedure. To achieve this, we implement the techniques of Chaisemartin (2018) to produce dynamic treatment effect estimates which directly estimate heterogeneous dynamic treatment effects and weights the corresponding effects together into the estimates (Figure E3).

Figure E3: Event Study of Head Start's Impact on Criminal Conviction – Treatment Heterogeneity



Note: Chaisemartin estimates analogous to our standard estimates in Figure 2 except that they account for the possibility of heterogeneous dynamic treatment effects.

Overall, Chaisemartin estimates for pre-treatment periods do not suggest differential movement of treatment and control groups prior to treatment. Period "0", the first period of treatment is when the first significant estimate appears. Furthermore, Chaisemartin post-treatment estimates give estimates very similar to our standard event study estimates. Chaisemartin estimates suggest an overall reduction of .0169, while our estimates suggest a reduction of .0131 (Row 1, Col 2 of Table 2). While the Chaisemartin estimate is not

significantly different from the standard estimate, this further suggests that, if anything, heterogeneity in the county rollout-cohorts is slightly attenuating our standard estimates towards zero.