Appendix

High-Quality Early-Childhood Education at Scale: Evidence from a Multisite Randomized Trial

August 25, 2023

List of Figures

List of Tables

Appendix 1. Appendix to Section 2: Additional Program Details

Table A.1. Treatment and Twinning Distribution by State in the Analysis Sample

		AR	CT	FL	MA	NY	PA	TX	WA	All
Panel a. Ages 0 to 3										
Singletons	Control	68	57	52	85	84	47	80	74	547
	Treatment	43	42	38	41	40	44	43	44	335
Twins	Control	12	9	4	8	8	6	8	6	61
	Treatment	5	4	6	4	6	4	6	7	42
Total		128	112	100	138	138	101	137	131	985
Panel b. Age 5										
Singletons	Control	62	54	43	75	66	42	71	67	480
	Treatment	36	39	26	36	37	36	39	39	288
Twins	Control	12	9	4	8	7	6	8	5	59
	Treatment	5	4	6	4	6	5	6	7	43
Total		115	106	79	123	116	89	124	118	870
Panel c. Age 8										
Singletons	Control	61	56	43	77	65	44	72	61	479
	Treatment	35	41	31	36	37	37	39	39	295
Twins	Control	11	9	4	7	7	6	8	5	57
	Treatment	5	4	6	4	6	5	6	6	42
Total		112	110	84	124	115	92	125	111	873
Panel d. Age 18 (Observe Special Education Outcome Only)										
Singletons	Control	61	55	45	77	67	46	72	64	487
	Treatment	36	41	31	38	37	39	40	40	302
Twins	Control	11	9	4	8	7	5	8	5	57
	Treatment	5	4	6	4	6	4	6	7	42
Total		113	109	86	127	117	94	126	116	888
Panel e. Age 18 (Observe All Outcomes)										
Singletons	Control	50	35	30	38	39	39	39	34	304
	Treatment	31	31	23	17	20	34	20	25	201
Twins	Control	9	9	3	5	7	5	5	2	45
	Treatment	3	2	5	4	3	3	4	3	27
Total		93	77	61	64	69	81	68	64	577

Note: This table shows the joint distribution of treatment and twinning by state in the sample that we analyze. Sample sizes decrease after age 3 because of attrition. The abbreviations in the first row are the state postal codes of the university hospitals and childcare centers.

Figure A.1. Correlation between Twinning and Baseline Characteristics

Joint test: F–statistic = 1.04 (*p*–value = 0.40).
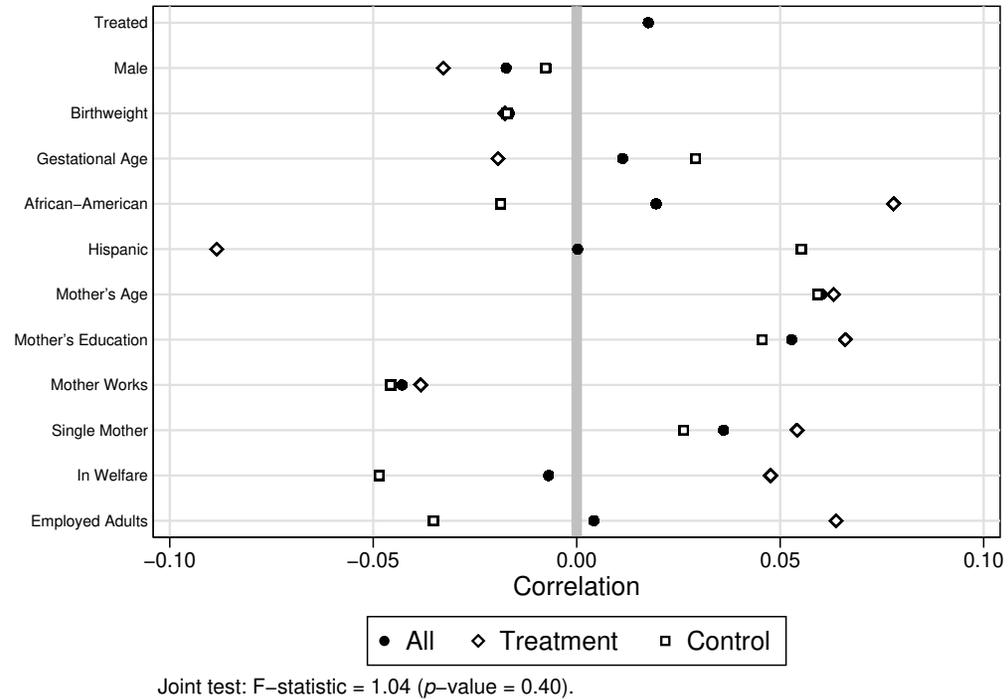
Note: This plot displays the correlation between a twin indicator and each of the baseline characteristics in Table 2 for the treatment, control, and pooled samples. For the pooled sample, it displays a joint test F-test, contrasting a null (intercept only) linear-regression model and a model with all of the baseline variables in the plot. The dependent variable is a twin indicator in both cases. The inference is clustered at the child-participant level.

Appendix 2. Appendix to Section 3: Additional Data Description

Figure A.2. Subscales of the Home Observation Measurement of the Environment

(a) At 12 Months
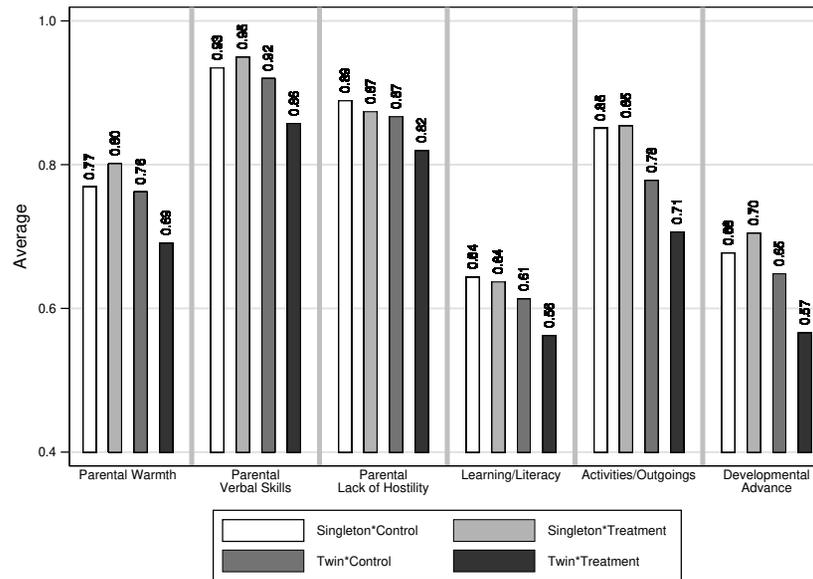
(b) At 36 Months
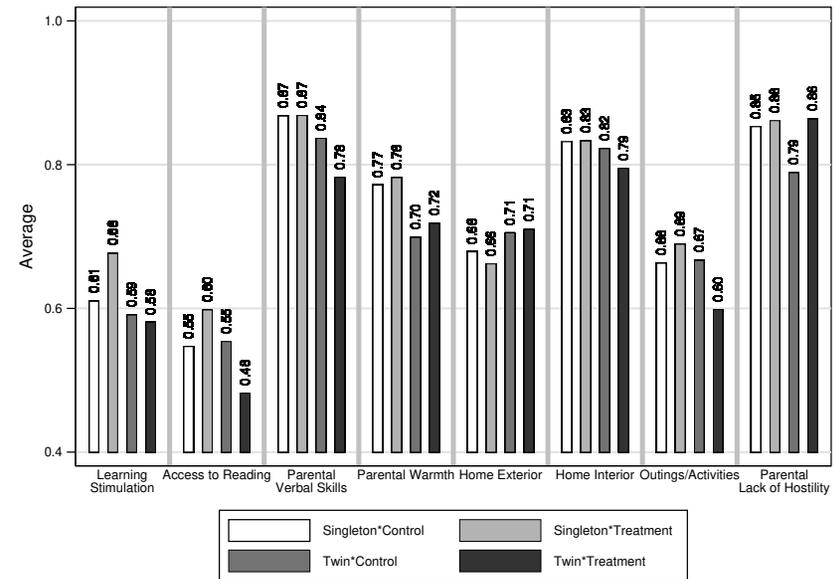
Note: Panel (a) displays the average of the subscales of the Home Observation Measurement of the Environment (HOME) at 12 months by treatment and twinning status. The subscale score is the average within category of indicators of several items (see Section 3.1 for details). Panel (b) is analogous in format to Panel (a) for the HOME observation at 36 months.

Table A.2. Skill Measures by Treatment and Twinning Status (Residualized)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			Singletons				Twins			
	Age	Observations	Control	Treatment	Difference	p-value	Control	Treatment	Difference	p-value
***Panel a.* Cognitive**										
PPVT-R	3	812	85.11	91.39	6.28	0.00	85.24	86.50	1.26	0.73
Stanford-Binet IQ Test		908	81.47	90.81	9.34	0.00	81.29	85.64	4.35	0.26
PPVT-R	5	805	79.98	84.00	4.03	0.02	74.33	72.39	-1.94	0.69
Wechsler PPSI IQ Test		804	91.69	92.60	0.91	0.51	88.71	85.41	-3.30	0.35
PPVT-III	8	863	85.85	85.59	-0.26	0.88	82.07	79.38	-2.69	0.55
Wechsler ISC IQ Test		870	85.65	86.24	0.59	0.78	84.91	79.17	-5.74	0.30
PPVT-III	18	611	95.75	96.50	0.76	0.64	96.67	91.86	-4.82	0.31
Wechsler ASI IQ Test		614	91.53	91.81	0.28	0.85	91.87	89.93	-1.94	0.64
***Panel b.* Non-Cognitive**										
Educational Outcomes										
Never in Special Education		888	0.70	0.74	0.04	0.24	0.67	0.69	0.02	0.80
No Reading Tutor	18	616	0.88	0.95	0.07	0.00	0.98	0.86	-0.12	0.09
No Math Tutor		618	0.76	0.81	0.04	0.22	0.86	0.75	-0.11	0.24
Took SAT or ACT		600	0.56	0.58	0.01	0.74	0.78	0.56	-0.23	0.05
(Residualized) Index: Average of Educational Outcomes		581	-0.11	0.08	0.19	0.03	0.36	-0.00	-0.36	0.08
Behavioral Outcomes										
Not a Smoker		605	0.75	0.79	0.04	0.28	0.85	0.81	-0.03	0.72
Not Idle	18	631	0.06	0.06	-0.00	0.96	0.02	0.07	0.05	0.36
Not in Therapy		622	0.69	0.71	0.02	0.66	0.76	0.54	-0.23	0.05
Not Teen Parent		598	0.86	0.85	-0.00	0.94	0.93	0.85	-0.08	0.29
(Residualized) Index: Average of Behavioral Outcomes		583	-0.05	0.03	0.08	0.40	0.25	-0.13	-0.38	0.07
All Outcomes										
Average of Residualized Indices	18	574	-0.07	0.06	0.13	0.07	0.31	-0.07	-0.38	0.01

Note: Column (1) displays the age of the participants when outcomes are measured. Column (2) reports the corresponding number of observations. Columns (3) and (4) display the average of the outcomes for the singletons in the sample by treatment status. Column (5) displays the difference between Columns (4) and (3). Column (6) displays the p-value of the t-statistic associated with the difference in Column (5). The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. Columns (7) to (10) are analogous in format to Columns (3) to (6) for the twins in the sample. Indices are residualized as explained in footnote 9 and then standardized to an in-sample mean of 0 and standard deviation of 1. The average of indices is the average of the two residualized and standardized indices.

Table A.3. Skill Measures by Treatment Status (Pooling Singletons and Twins)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
					IHDP		
	Age	Observations	Control	Treatment	Difference	S.E.	p-value
Panel a. Cognitive							
PPVT-R	3	812	85.13	90.79	5.66	1.23	0.00
Stanford-Binet IQ Test		908	81.45	90.19	8.73	1.30	0.00
PPVT-R	5	805	79.38	82.48	3.10	1.67	0.06
Wechsler PPSI IQ Test		804	91.37	91.65	0.28	1.28	0.83
PPVT-III	8	863	85.45	84.81	-0.64	1.60	0.69
Wechsler ISC IQ Test		870	85.57	85.35	-0.22	1.94	0.91
PPVT-III	18	611	95.86	95.97	0.11	1.51	0.94
Wechsler ASI IQ Test		614	91.57	91.59	0.02	1.37	0.99
Panel b. Non-Cognitive							
Educational Outcomes							
Never in Special Education		888	0.69	0.73	0.04	0.03	0.24
No Reading Tutor	18	616	0.89	0.94	0.04	0.02	0.04
No Math Tutor		618	0.78	0.80	0.03	0.03	0.45
Took SAT or ACT		600	0.59	0.58	-0.02	0.04	0.71
Index: Average of Educational Outcomes		585	0.74	0.77	0.03	0.02	0.13
Behavioral Outcomes							
Not a Smoker		605	0.77	0.80	0.03	0.03	0.38
Not Idle	18	631	0.06	0.06	0.01	0.02	0.80
Not in Therapy		622	0.70	0.69	-0.01	0.04	0.78
Not Teen Parent		598	0.87	0.85	-0.01	0.03	0.66
Index: Average of Behavioral Outcomes		586	0.59	0.60	0.01	0.02	0.75
All Outcomes							
Average of Indices	18	577	0.67	0.69	0.02	0.01	0.24

Note: Column (1) displays the age of the participants when skills are measured. Column (2) reports the number of observations associated with each skill measure. Columns (3) and (4) display the average of the test scores and behavior measures for the singletons and twins in the sample by treatment status. Column (5) displays the difference between Columns (4) and (3). Column (6) displays the standard error associated with the difference in Column (5). Column (7) displays the p-value of the t-statistic associated with the difference in Column (5). The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level.

Table A.4. Skill Measures by Treatment and Twinning Status (Female Participants)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
			Singletons					Twins				
	Age	Observations	Control	Treatment	Difference	S.E.	p-value	Control	Treatment	Difference	S.E.	p-value
Panel a. Cognitive												
PPVT-R	3	423	84.10	91.78	7.68	1.86	0.00	84.73	86.05	1.31	4.71	0.78
Stanford-Binet IQ Test		463	82.24	93.84	11.60	1.93	0.00	82.40	85.22	2.82	5.45	0.60
PPVT-R	5	407	78.90	86.04	7.14	2.43	0.00	76.22	67.77	-8.45	7.34	0.25
Wechsler PPSI IQ Test		407	91.86	94.83	2.96	1.93	0.13	88.37	84.32	-4.05	5.41	0.45
PPVT-III	8	437	85.20	86.55	1.35	2.39	0.57	78.34	73.87	-4.48	6.63	0.50
Wechsler ISC IQ Test		441	86.32	88.22	1.90	2.85	0.51	83.21	77.04	-6.16	8.05	0.44
PPVT-III	18	325	94.85	97.49	2.64	2.19	0.23	93.57	88.31	-5.25	7.02	0.45
Wechsler ASI IQ Test		327	90.36	93.03	2.66	1.93	0.17	91.30	88.75	-2.55	5.93	0.67
Panel b. Non-Cognitive												
Educational Outcomes												
Never in Special Education		452	0.72	0.80	0.08	0.04	0.05	0.66	0.61	-0.05	0.14	0.73
No Reading Tutor	18	325	0.89	0.97	0.08	0.03	0.00	1.00	0.94	-0.06	0.06	0.30
No Math Tutor		326	0.76	0.81	0.05	0.05	0.32	0.86	0.81	-0.05	0.12	0.68
Took SAT or ACT		321	0.60	0.68	0.08	0.06	0.16	0.83	0.60	-0.23	0.15	0.13
Index: Average of Educational Outcomes		314	0.75	0.83	0.08	0.03	0.01	0.85	0.77	-0.09	0.08	0.29
Behavioral Outcomes												
Not a Smoker		324	0.83	0.79	-0.05	0.05	0.32	0.91	0.73	-0.18	0.13	0.16
Not Idle	18	333	0.05	0.07	0.03	0.03	0.34	0.00	0.06	0.06	0.06	0.30
Not in Therapy		328	0.69	0.75	0.06	0.05	0.29	0.75	0.50	-0.25	0.15	0.10
Not Teen Parent		319	0.84	0.83	-0.01	0.05	0.81	0.91	0.80	-0.11	0.12	0.34
Index: Average of Behavioral Outcomes		314	0.60	0.61	0.01	0.02	0.59	0.65	0.53	-0.12	0.06	0.03
All Outcomes												
Average of Indices	18	310	0.68	0.72	0.05	0.02	0.02	0.76	0.65	-0.11	0.05	0.04

Note: Column (1) displays the age of the participants when skills are measured. Column (2) reports the number of observations associated with each skill measure. Columns (3) and (4) display the average of the test scores and behavior measures for the singletons in the sample by treatment status. Column (5) displays the difference between Columns (4) and (3). Column (6) displays the standard error associated with the difference in Column (5). Column (7) displays the p-value of the t-statistic associated with the difference in Column (5). The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. Columns (8) to (12) are analogous in format to Columns (3) to (7) for the twins in the sample.

Table A.5. Skill Measures by Treatment and Twinning Status (Male Participants)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
					Singletons					Twins		
	Age	Observations	Control	Treatment	Difference	S.E.	p-value	Control	Treatment	Difference	S.E.	p-value
Panel a. Cognitive												
PPVT-R	3	389	86.22	90.98	4.76	1.86	0.01	85.83	87.06	1.23	5.52	0.82
Stanford-Binet IQ Test		445	80.69	87.69	7.00	1.96	0.00	80.11	86.16	6.05	5.42	0.26
PPVT-R	5	398	81.05	81.89	0.84	2.53	0.74	72.28	77.74	5.46	6.14	0.37
Wechsler PPSI IQ Test		397	91.51	90.29	-1.23	1.94	0.53	89.08	86.68	-2.40	4.28	0.58
PPVT-III	8	426	86.51	84.62	-1.89	2.47	0.45	86.07	86.05	-0.02	5.58	1.00
Wechsler ISC IQ Test		429	84.97	84.23	-0.74	3.00	0.81	86.74	81.74	-5.00	7.78	0.52
PPVT-III	18	286	96.76	95.37	-1.39	2.33	0.55	99.78	96.58	-3.20	5.79	0.58
Wechsler ASI IQ Test		287	92.85	90.40	-2.45	2.21	0.27	92.43	91.50	-0.93	5.83	0.87
Panel b. Non-Cognitive												
Educational Outcomes												
Never in Special Education		436	0.67	0.66	-0.01	0.05	0.90	0.68	0.79	0.11	0.13	0.39
No Reading Tutor	18	291	0.87	0.92	0.05	0.04	0.22	0.95	0.75	-0.20	0.13	0.13
No Math Tutor		292	0.76	0.80	0.04	0.05	0.47	0.86	0.67	-0.20	0.16	0.21
Took SAT or ACT		279	0.52	0.45	-0.06	0.07	0.34	0.74	0.50	-0.24	0.17	0.16
Index: Average of Educational Outcomes		271	0.71	0.72	0.01	0.04	0.69	0.82	0.70	-0.11	0.09	0.20
Behavioral Outcomes												
Not a Smoker		281	0.66	0.80	0.14	0.06	0.01	0.78	0.92	0.13	0.12	0.26
Not Idle	18	298	0.08	0.05	-0.03	0.03	0.28	0.04	0.08	0.03	0.09	0.70
Not in Therapy		294	0.69	0.66	-0.03	0.06	0.64	0.77	0.58	-0.19	0.17	0.26
Not Teen Parent		279	0.88	0.89	0.01	0.04	0.83	0.96	0.92	-0.04	0.09	0.66
Index: Average of Behavioral Outcomes		272	0.57	0.59	0.02	0.03	0.45	0.64	0.61	-0.02	0.06	0.72
All Outcomes												
Average of Indices	18	267	0.64	0.65	0.01	0.02	0.65	0.73	0.66	-0.07	0.07	0.30

Note: Column (1) displays the age of the participants when skills are measured. Column (2) reports the number of observations associated with each skill measure. Columns (3) and (4) display the average of the test scores and behavior measures for the singletons in the sample by treatment status. Column (5) displays the difference between Columns (4) and (3). Column (6) displays the standard error associated with the difference in Column (5). Column (7) displays the p-value of the t-statistic associated with the difference in Column (5). The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. Columns (8) to (12) are analogous in format to Columns (3) to (7) for the twins in the sample.

Table A.6. Baseline Characteristics of Non-Responders and Responders of the Age-18 Follow-Up

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
						Infant Health and Development Program						
	Singletons				Twins				All			
	No Response	Response	Difference	p-value	No Response	Response	Difference	p-value	No Response	Response	Difference	p-value
***Panel a.* Children**												
Treatment	0.35	0.40	0.05	0.18	0.50	0.37	-0.13	0.21	0.36	0.39	0.03	0.36
Male	0.54	0.46	-0.07	0.04	0.47	0.47	0.00	0.99	0.53	0.46	-0.07	0.05
Birthweight (grams)	1805.10	1799.47	-5.63	0.86	1629.72	1838.84	209.12	0.01	1789.55	1804.18	14.63	0.63
Low Birth-weight ($\leq$ 2000 grams)	0.61	0.63	0.02	0.49	0.88	0.68	-0.20	0.02	0.63	0.64	0.01	0.86
Gestational Age	33.04	33.04	-0.00	0.99	32.69	33.25	0.56	0.31	33.01	33.06	0.05	0.77
African-American	0.49	0.53	0.04	0.27	0.69	0.50	-0.19	0.07	0.51	0.53	0.02	0.58
Hispanic	0.16	0.08	-0.08	0.00	0.06	0.12	0.06	0.34	0.15	0.08	-0.07	0.00
***Panel b.* Mother at Childbirth**												
African-American	0.49	0.53	0.04	0.27	0.69	0.50	-0.19	0.07	0.51	0.53	0.02	0.58
Age	24.67	24.68	0.01	0.98	26.34	25.59	-0.76	0.57	24.81	24.79	-0.03	0.94
Education	12.02	12.53	0.50	0.00	12.34	12.93	0.58	0.27	12.05	12.58	0.52	0.00
Works	0.31	0.39	0.08	0.02	0.28	0.29	0.01	0.89	0.31	0.38	0.07	0.03
Married	0.46	0.46	0.01	0.89	0.34	0.60	0.26	0.01	0.45	0.48	0.03	0.34
***Panel c.* Household at Childbirth**												
In Welfare	0.44	0.34	-0.10	0.00	0.50	0.29	-0.21	0.05	0.45	0.33	-0.11	0.00
Siblings	0.87	0.68	-0.19	0.01	1.69	1.69	0.00	0.99	0.94	0.80	-0.14	0.06
Employed Adults	0.92	1.07	0.15	0.00	1.00	1.03	0.03	0.84	0.93	1.07	0.14	0.00
***Panel d.* Economy at Childbirth**												
Employment %	93.61	93.18	-0.43	0.00	92.84	93.41	0.57	0.08	93.54	93.21	-0.33	0.00
Median Income in 1000s (2020 USD)	56.79	54.68	-2.11	0.00	53.88	55.10	1.21	0.51	56.53	54.73	-1.80	0.00
Government Expenditure per Capita (2020 USD)	8205.16	8008.10	-197.06	0.02	7927.72	8040.33	112.61	0.65	8180.57	8011.96	-168.60	0.04
***Panel e.* Joint Tests**												
F-statistic	4.00				2.60				3.38			
p-value	0.00				0.00				0.00			

Note: In Panels *a* to *d*, Columns (1) and (2) display the average baseline characteristics of the singleton non-responders and responders of the age-18 follow-up. A person is classified as a responder if the the variable "Average of All Non-Cognitive Outcomes" in Table 3 is observed for them. Column (3) displays the difference between Columns (2) and (1). Column (4) displays the p-value of the t-statistic associated with the difference in Column (3). The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. Columns (5) to (8) are analogous in format to Columns (1) to (4) for the twins in the sample. Columns (9) to (12) are analogous in format to Columns (1) to (4) for the singletons and twins in the sample. Panel *e* presents the F-test corresponding to a joint test of significance in the mean-difference across all of the variables in Panels *a* to *d* for each of the three samples.

Appendix 2.1 Comparison to Parenting Input Measures in the Literature

Chaparro et al. (2020) combine data from the IHDP and other sources to study childcare policies. Their measure of average hours spent in childcare is weekly and represents the average observed between ages 2 and 3, just like in our case. Unlike ours, their measure only includes hours spent in IHDP childcare centers. It does not include hours spent in other childcare centers or nurseries. However, their residual measure of time spent with other caretakers consists of any time spent with caretakers other than mothers or in IHDP childcare centers. Therefore, it includes time spent in any other childcare centers or nurseries. These authors quality-adjust measure of maternal care using a subset of items of the HOME at age 3 (e.g., they select items relevant to the quality of maternal care). Panel *a* of Table A.7 displays their measure, which we aim to replicate using our analysis sample in Panel *b*.[1] Differences between these two panels are minimal. Chaparro et al. (2020) also quantify the average weekly hours spent with mothers. They then calculate the average weekly hours spent with other caretakers as a residual, by assuming a number of hours a week spent awake by children. The average hours a week spent with other caretakers is this assumed number less the average hours a week spent with mothers and at IHDP centers.

The measures in Chaparro et al. (2020) are based on one of the various datasets from the IHDP (i.e., the primary analysis dataset), which include average hours and days in childcare when child participants were 2 and 3 years old. Our weekly measure of hours spent in childcare centers is based on more comprehensive questionnaires containing information about child participants at ages 18, 24, 30, and 36 months old. We summarize this measure in Panel *c* of Table A.7 for singletons and in Panel *d* for twins. The number of observations is greater in our case because we use variables that have higher response rates. The averages are larger. For the control group, the average is larger in our case because we include

---

[1]We use the average number of hours a day spent in IHDP childcare centers and the average number of days a year. We observe these two variables for ages 2 and 3. We create an average number of hours a week by multiplying the average hours a day and average days a year at each age. We sum these products and divide the resulting sum by 104 (the number of weeks in the two years of observation).

all center-based childcare centers. For the treatment group, the average is larger in our case because our measures seem to be more precise—Gross et al. (1997) reports that the treatment-group participants mostly used IHDP childcare centers rather than alternatives. However, the average treatment-control differences are closely aligned. Even when including all center-based childcare centers, control-group children spent a very small amount of their time in them. The measure of time spent with mothers in our analysis sample yields very similar averages to the measure in Chaparro et al. (2020).[2]

We do not need to assume the number of hours that children spent awake because we quantify from data the time spent with other caretakers. We are thus able to calculate the implied average total hours awake. For singletons, this implied total is between 74 and 77. For twins, it is between 71 and 77. The relatively small average treatment-control difference in the implied hours of sleep could be due to time spent napping in either childcare, maternal care, or with other caretakers. In either case, 71 to 77 hours awake per week corresponds to an average of sleep time between 12 and 13.5 hours a day, which is consistent with a recent meta-analysis of infant sleeping time (Galland et al., 2012) and a study of US infants whose families are at socioeconomic disadvantage (Williamson et al., 2021).[3]

---

[2]We calculate hours in maternal care in an identical way as Chaparro et al. (2020). The construction of all other time variables is analogous.

[3]The documentation in Williamson et al. (2021) indicates a daily average of hours of sleep of 11.5 for disadvantaged children, whose hours are lower than those of advantaged children. The daily average of 11.5 hours yields a weekly average close to the average we quantify for control twins than the averages we quantify for treatment and control singletons. The daily average of 11.5 is the preferred assumption of Chaparro et al. (2020), justified using the study of Iglowstein et al. (2003).

Table A.7. Replication and Comparison to Chaparro et al. (2020)

	(1)	(2)	(3)	(4)
	Treatment	Control	Difference	*p*-value
***Panel a.* Chaparro et al. 2020 (as reported in paper, based on singletons)**				
Hours per week of IHDP care	18.18	0.00	18.18	0.00
Hours per week of maternal care	52.40	62.07	-9.67	0.00
Hours per week other caretakers (data)	N/A	N/A	N/A	N/A
Hours per week other caretakers (residual)				
if Total Hours Awake = 87.5	16.92	25.43	-8.52	0.00
if Total Hours Awake = 84	13.42	21.93	-8.52	0.00
if Total Hours Awake = 80.5	9.92	18.43	-8.52	0.00
if Total Hours Awake = 77	6.42	14.93	-8.52	0.00
Number of Observations		880[a]		
***Panel b.* Chaparro et al. 2020 (replication using singletons sample in this paper)**				
Hours per week of IHDP care	17.76	0.00	17.76	0.00
Hours per week of maternal care	47.13	58.01	-10.88	0.00
Hours per week other caretakers (data)	N/A	N/A	N/A	N/A
Hours per week other caretakers (residual)				
if Total Hours Awake = 87.5	22.62	29.49	-6.87	0.00
if Total Hours Awake = 84	19.12	25.99	-6.87	0.00
if Total Hours Awake = 80.5	15.62	22.49	-6.87	0.00
if Total Hours Awake = 77	12.12	18.99	-6.87	0.00
Number of Observations		731		
***Panel c.* This paper (singletons)**				
Hours per week of center-based care	26.34	5.32	21.02	0.00
Hours per week of maternal care	47.31	57.13	-9.82	0.00
Hours per week other caretakers (data)	3.67	11.89	-8.22	0.00
Implied Total Hours	77.32	74.34	2.98	0.05
Number of Observations		779		
***Panel d.* This paper (twins)**				
Hours per week of center-based care	28.88	2.98	25.90	0.00
Hours per week of maternal care	43.48	58.98	-15.51	0.00
Hours per week other caretakers (data)	5.38	8.95	-3.57	0.19
Implied Total Hours	77.74	70.91	6.83	0.10
Number of Observations		99		

Note: Panel *a* reprints from Chaparro et al. (2020) the average hours per week children spent in IHDP center-based care by treatment status between ages 2 and 3, together with the average treatment-control difference and corresponding *p*-value. The null hypothesis is that the difference is 0. The *p*-value is based on robust standard errors clustered at the child-participant level. Panel *a* also reprints the analogous information for time spent with mothers and other caretakers and a residual, which is the rest of the time assuming the total hours children spent awake during a week indicated in the panel. Chaparro et al. (2020) base their analysis on singletons. Panel *b* aims to replicate Panel *a* using the analysis sample in this paper. Panel *c*, based on singletons, is analogous in format to Panels *a* and *b*. Instead of reporting hours spent at IHDP childcare centers, we report hours spent in any center-based childcare center. Instead of assuming a number of hours that children spent awake, we use data on the time spent with caretakers other than mothers. This allows us to provide an implied total of hours, which indicates that children were awake between 71 and 78 hours a week. Panel *d* is analogous in format to Panel *c* but focuses on twins rather than singletons.
a: We thank Juan Camilo Chaparro for providing us with this number of observations.

Appendix 3. Appendix to Section 4:
**Robust to Alternative Estimators and Inferential Methods**

Table A.8. Treatment Effects, Robustness Checks for Estimators and Inference

	Age	(1) Treatment Effect	(2) Child-Clustered	(3) State-Clustered	(4) Child-Clustered	(5) State-Clustered	(6) Treatment Effect	(7) Child-Clustered	(8) State-Clustered	(9) Child-Clustered	(10) State-Clustered
		Singletons					Twins				
		Main Specification	**Analytic p-value**		**Alternative** Bootstrap p-value		**Main Specification**	**Analytic p-value**		**Alternative** Bootstrap p-value	
Panel a. Mean Difference											
Peabody Picture Vocabulary Test	3	6.28	0.00	0.01	0.00	0.02	1.26	0.73	0.63	0.74	0.66
Stanford-Binet IQ Test		9.34	0.00	0.00	0.00	0.01	4.35	0.26	0.14	0.23	0.20
Average (Index) of Educational Outcomes		0.05	0.03	0.04	0.04	0.05	-0.09	0.11	0.27	0.15	0.36
Average (Index) of Behavioral Outcomes	18	0.02	0.35	0.43	0.33	0.39	-0.08	0.07	0.07	0.07	0.12
Average of Indices		0.03	0.04	0.03	0.04	0.03	-0.09	0.03	0.13	0.04	0.20
Panel b. OLS											
Peabody Picture Vocabulary Test	3	7.23	0.00	0.00	0.00	0.00	3.45	0.18	0.13	0.20	0.17
Stanford-Binet IQ Test		9.96	0.00	0.00	0.00	0.00	5.68	0.07	0.02	0.09	0.05
Average (Index) of Educational Outcomes		0.04	0.05	0.03	0.05	0.03	-0.07	0.15	0.33	0.15	0.42
Average (Index) of Behavioral Outcomes	18	0.02	0.24	0.35	0.21	0.34	-0.09	0.04	0.05	0.05	0.10
Average of Indices		0.03	0.06	0.07	0.06	0.10	-0.08	0.03	0.13	0.03	0.22
Panel c. IPW											
Peabody Picture Vocabulary Test	3	6.45	0.00	0.01	0.00	0.02	1.18	0.75	0.65	0.76	0.63
Stanford-Binet IQ Test		9.22	0.00	0.00	0.00	0.01	4.40	0.27	0.17	0.29	0.22
Average (Index) of Educational Outcomes		0.04	0.05	0.07	0.05	0.08	-0.09	0.13	0.29	0.14	0.39
Average (Index) of Behavioral Outcomes	18	0.01	0.50	0.58	0.50	0.55	-0.07	0.10	0.10	0.12	0.14
Average of Indices		0.03	0.09	0.05	0.09	0.05	-0.09	0.04	0.16	0.06	0.22

Note: This table presents estimates of the average treatment effect (ATE) for the age-3 and age-18 outcomes defined in Section 3 by twinning status. We present estimates using the three estimators discussed in this section accompanied by their baseline p-value (analytic and clustered at the child-participant level). We also present alternative p-values (analytic and clustered at the state level, bootstrapped and clustered at the child-participant level, and bootstrapped and clustered at the state level). The null hypothesis for each treatment effect is that it is equal to 0. Analytic clustered p-values are based on t-statistics using standard errors robust to arbitrary forms of heteroskedasticity within clusters, as in Liang and Zeger (1986). Bootstrap p-values are the fraction of non-rejections when the null hypothesis is imposed on the bootstrap distribution of the corresponding estimate. The bootstrap distribution is based on bootstrap re-samplings clustered at the level indicated. The bootstrap procedures that we follow are standard (e.g., Hansen, 2021). The covariates in the OLS and IPW estimators are the variables in Panels *a*, *b*, and *c* of Table 2, as well as site fixed effects.

Appendix 4. Appendix to Section 5:
# Additional Childcare and Parenting Impact Description

Figure A.3. Childcare and Parenting by Treatment and Twinning Status (Non-Standardized Parenting Measure)
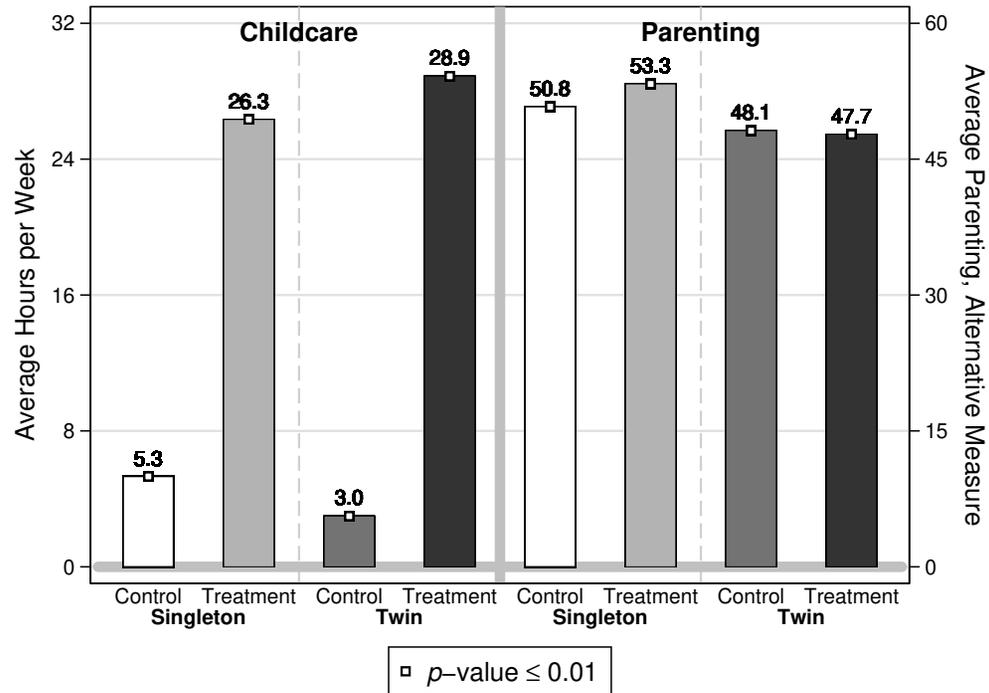

Note: The panel labeled "Childcare" displays the average hours per week in childcare by treatment and twinning status. We first average the observations of average hours per week at ages 18, 24, 30, and 36 months within children. We then average across children to construct the averages for each group in the label. The panel labeled "Parenting" is analogous in format to the panel labeled "Childcare" for the parenting measures observed at ages 1 and 3. The parenting measure displayed is not standardized to an in-sample mean of 0 and a standard deviation of 1 (as it is in the paper). Section 3.1 describes how we measure time spent in childcare and parenting. We label averages according to the p-value associated with their t-statistic. The null hypothesis is that the average is 0. The p-value is based on robust standard errors clustered at the child-participant level. ***: p-value < 0.01. **: p-value < 0.05. *: p-value < 0.10.

Figure A.4. Childcare and Parenting by Treatment and Twinning Status, Ages 1 and 3
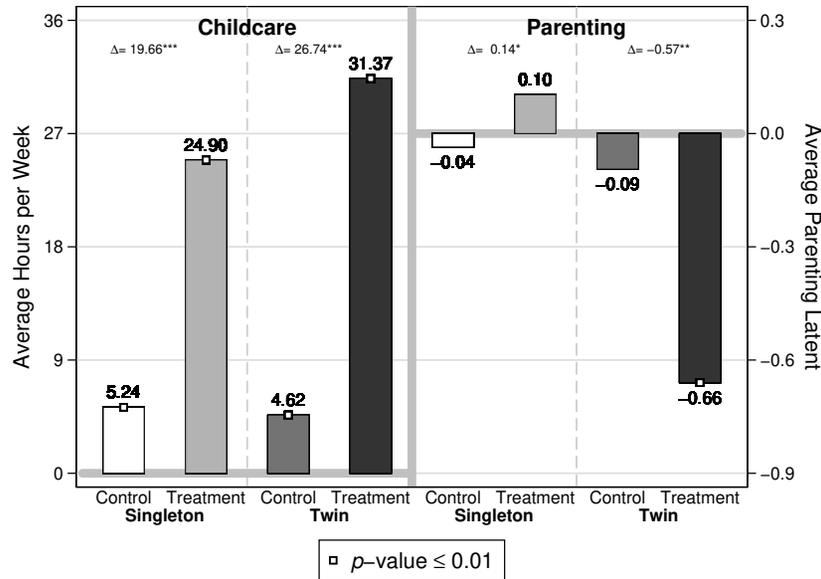
(a) Childcare
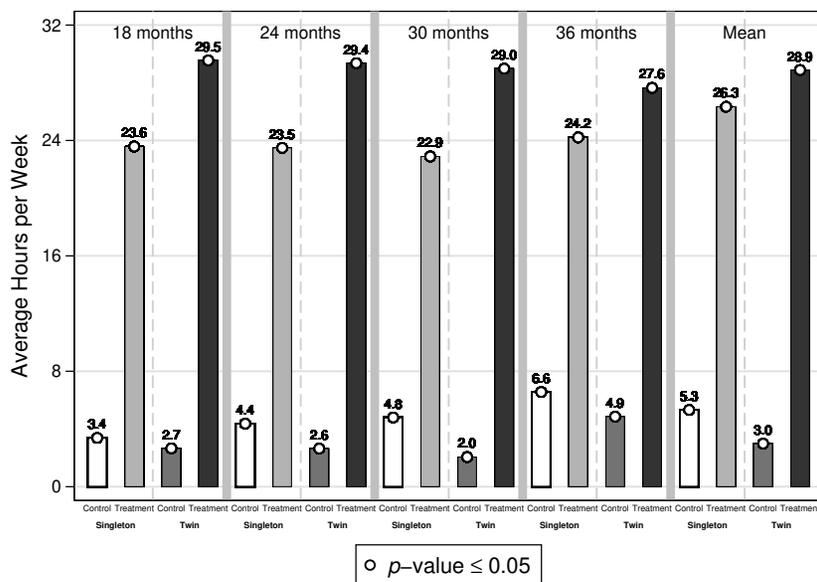
(b) Parenting
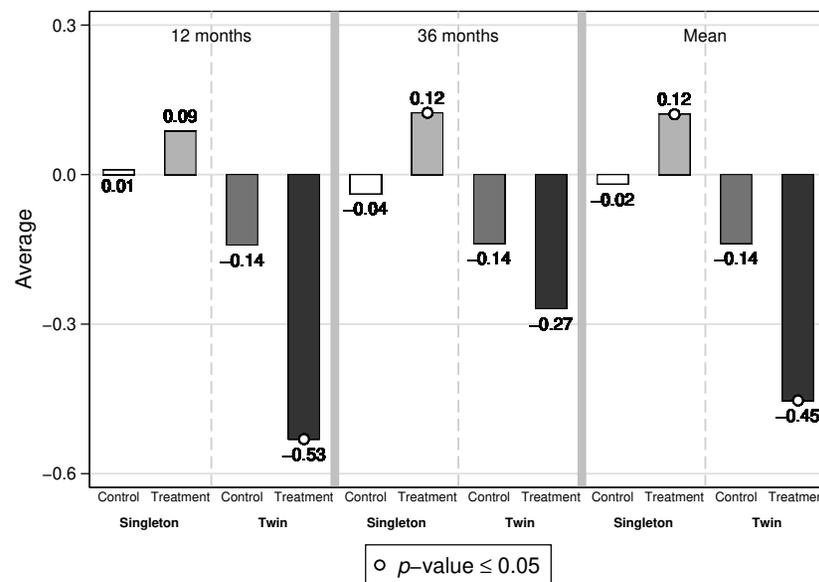
Note: Panel (a) displays the average hours per week in childcare when children were 18, 24, 30, and 36 months old, as well as the mean across all of these ages, by treatment and twinning status. Panel (b) displays the average of the parenting measures at ages 1 and 3 by treatment and twinning status. It also presents the average of the mean of the two measures. The parenting measures are standardized to an in-sample mean of 0 and a standard deviation of 1. Section 3.1 describes how we measure time spent in childcare and parenting. We label averages according to the p-value associated with their t-statistic. The null hypothesis is that the average is 0. The p-value is based on robust standard errors clustered at the child-participant level. ***: p-value < 0.01. **: p-value < 0.05. *: p-value < 0.10.

Figure A.5. Childcare and Parenting by Treatment and Twinning Status and Sex at Birth

**(a)** Female Participants　　　　　　　　　　　　　　　　　　**(b)** Male Participants
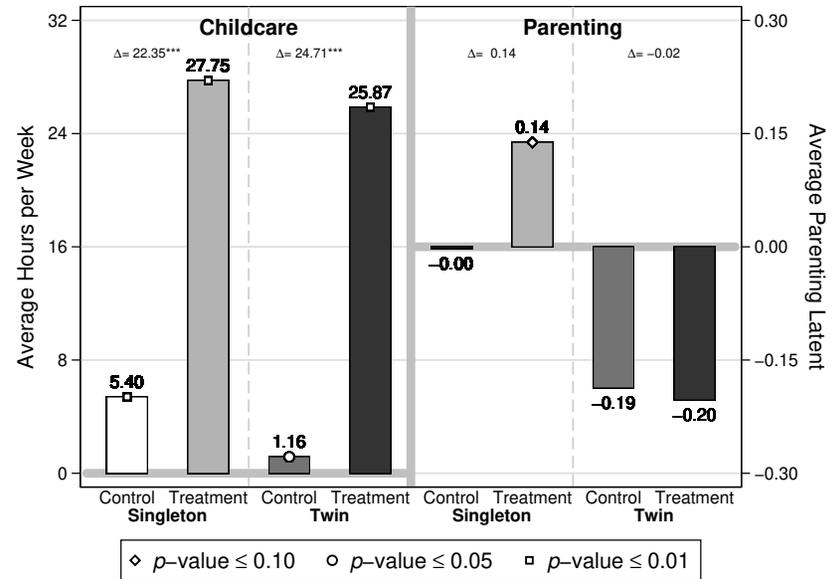
Note: Panel (a) focuses on female participants. The subpanel panel labeled "Childcare" displays the average hours per week in childcare by treatment and twinning status. We first average the observations of average hours per week at ages 18, 24, 30, and 36 months within children. We then average across children to construct the averages for each group in the label. The subpanel labeled "Parenting" is analogous in format to the panel labeled "Childcare" for the parenting measures observed at ages 1 and 3. The parenting measure displayed is standardized to an in-sample mean of 0 and a standard deviation of 1. Section 3.2 describes how we measure time spent in childcare and parenting. Panel (b) is analogous in format to Panel (a) for male participants. Δ is defined as the treatment-control difference in the corresponding averages. We label averages and average differences according to the p-value associated with their t-statistic. The null hypothesis for either the averages or average differences is that they are 0. The p-value is based on robust standard errors clustered at the child-participant level. ***: p-value < 0.01. **: p-value < 0.05. *: p-value < 0.10.

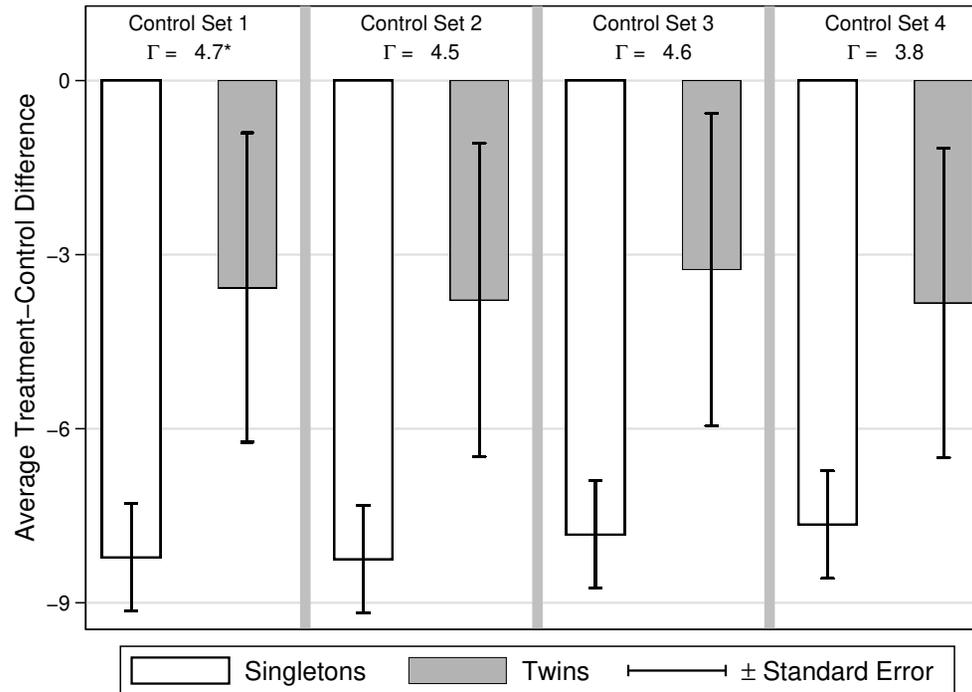**Figure A.6.** Treatment-Control Differences in Maternal Hours Working
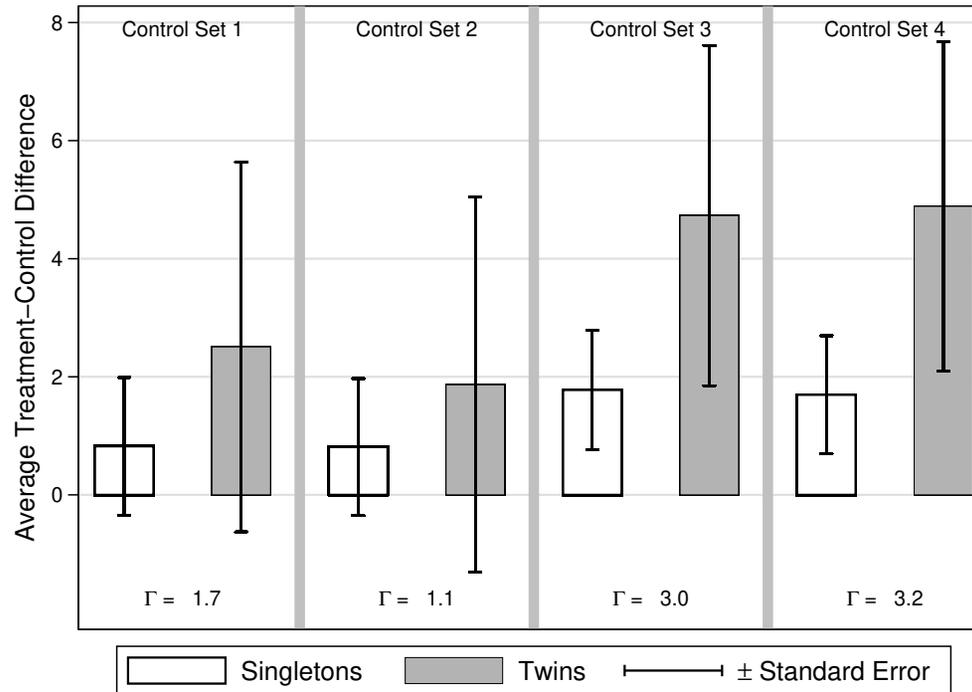
Note: This figure displays the average treatment-control difference in the mean hours a week spent working by mothers of child participants when their children were 18, 24, 30, and 36 months by twinning status. To obtain these average differences we regress mean hours on a constant, a twin indicator, a treatment indicator, an interaction of the twin and treatment indicators, and a control set. The average treatment-control difference for the singletons is the estimated coefficient on the treatment indicator. For the twins, it is the estimated coefficient on the treatment indicator plus the estimated coefficient on the interaction. Control Set 1 is empty. Control Set 2 includes the variables in Panel a of Table 2. Control Set 3 includes Control Set 2 and the variables in Panels b and c of Table 2. Control Set 4 includes Control Set 3 and site fixed effects. The standard errors are clustered at the child-participant level. $\mathbf{\Gamma}$ is defined as the twin-singleton difference in the corresponding average treatment effects. We label the estimates of $\mathbf{\Gamma}$ according to the p-value of the t-statistic associated with the difference. The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. ***: p-value < 0.01. **: p-value < 0.05. *: p-value < 0.10.
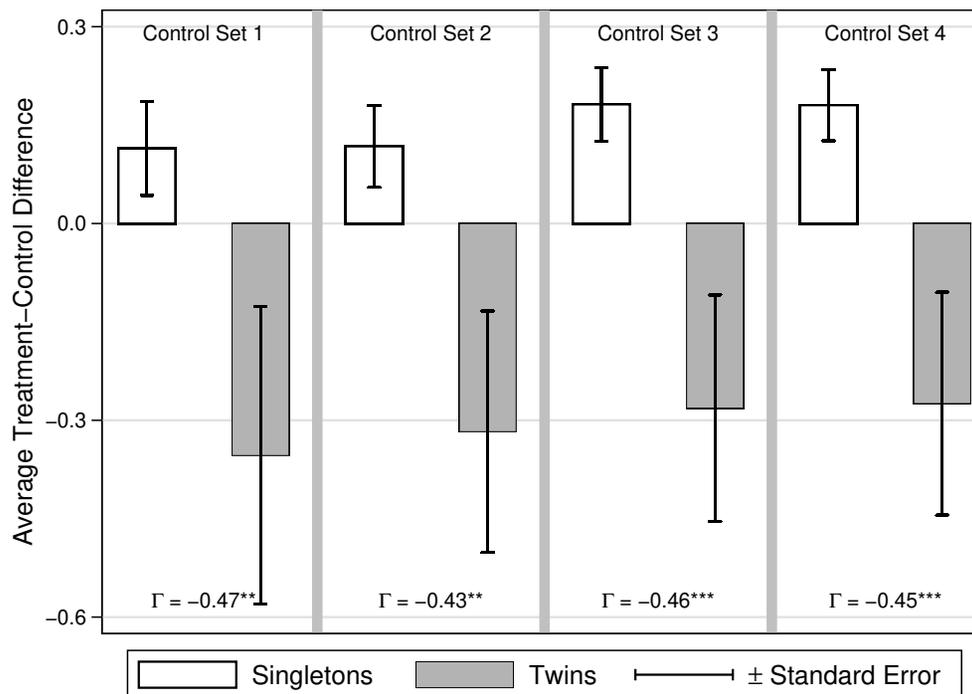
Figure A.7. Treatment-Control Differences in Maternal Care Hours
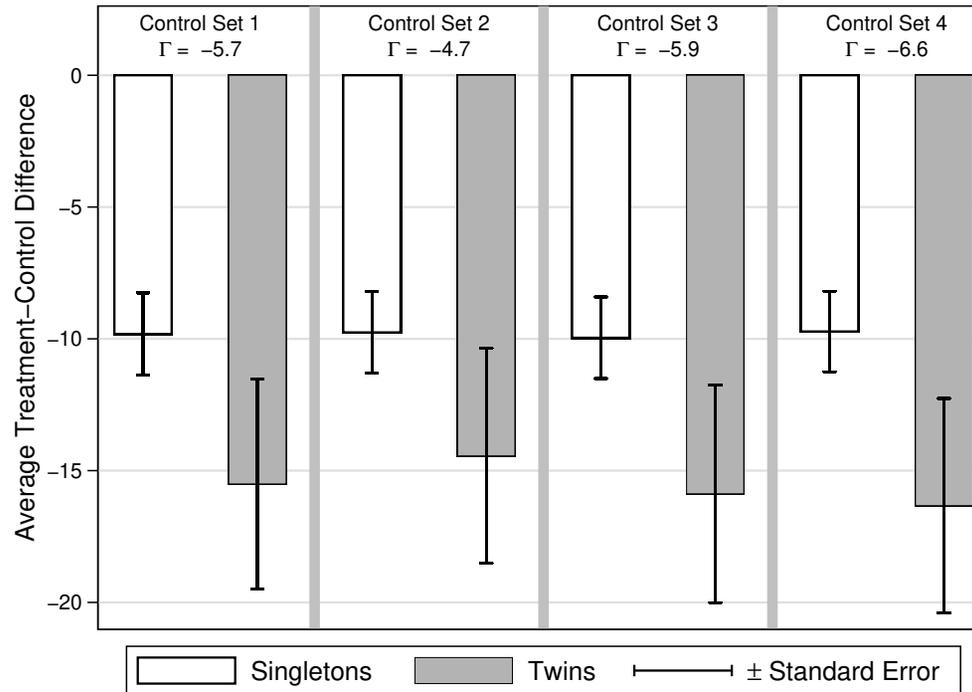
Note: This figure displays the average treatment-control difference in the mean hours a week mothers spent taking care of their child participants when these were 18 and 30 months by twinning status. To obtain these average differences we regress mean hours on a constant, a twin indicator, a treatment indicator, an interaction of the twin and treatment indicators, and a control set. The average treatment-control difference for the singletons is the estimated coefficient on the treatment indicator. For the twins, it is the estimated coefficient on the treatment indicator plus the estimated coefficient on the interaction. Control Set 1 is empty. Control Set 2 includes the variables in Panel *a* of Table 2. Control Set 3 includes Control Set 2 and the variables in Panels *b* and *c* of Table 2. Control Set 4 includes Control Set 3 and site fixed effects. The standard errors are clustered at the child-participant level. $\boldsymbol{\Gamma}$ is defined as the twin-singleton difference in the corresponding average treatment effects. We label the estimates of $\boldsymbol{\Gamma}$ according to the *p*-value of the *t*-statistic associated with the difference. The null hypothesis is that the difference is 0. The *p*-value is based on robust standard errors clustered at the child-participant level. ***: *p*-value < 0.01. **: *p*-value < 0.05. *: *p*-value < 0.10.

Figure A.8. Treatment-Control Differences in Care Hours (Other Place)

Note: This figure displays the average treatment-control difference in the mean hours a week spent by children in care provided in "other places" (not by their mothers or in childcare centers) when their children were 18, 24, 30, and 36 months by twinning status. To obtain these average differences we regress mean hours on a constant, a twin indicator, a treatment indicator, an interaction of the twin and treatment indicators, and a control set. The average treatment-control difference for the singletons is the estimated coefficient on the treatment indicator. For the twins, it is the estimated coefficient on the treatment indicator plus the estimated coefficient on the interaction. Control Set 1 is empty. Control Set 2 includes the variables in Panel a of Table 2. Control Set 3 includes Control Set 2 and the variables in Panels b and c of Table 2. Control Set 4 includes Control Set 3 and site fixed effects. $\mathbf{\Gamma}$ is defined as the twin-singleton difference in the corresponding average treatment effects. We label the estimates of $\mathbf{\Gamma}$ according to the p-value of the t-statistic associated with the difference. The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. ***: p-value < 0.01. **: p-value < 0.05. *: p-value < 0.10.
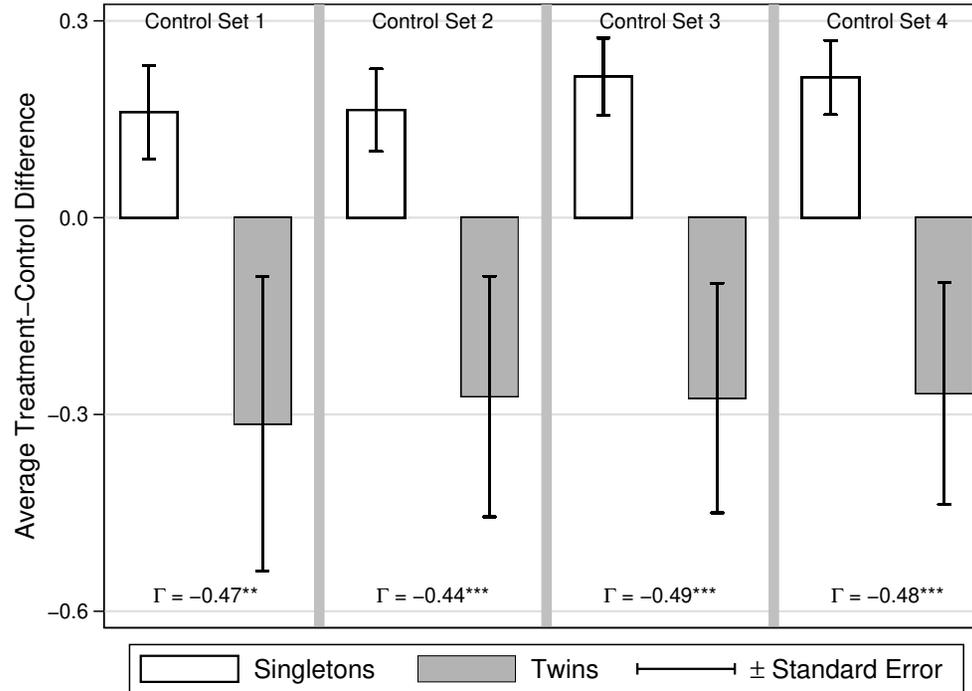
Figure A.9. Treatment-Control Differences in Parenting (Residualized from Maternal Care Hours)

Note: This figure displays the average treatment-control difference in the mean of the parenting measures observed at ages 1 and 3 net (residualized from) of weekly maternal care hours observed at 18 and 30 months by twinning status. To obtain these average differences we regress the residualized parenting measure on a constant, a twin indicator, a treatment indicator, an interaction of the twin and treatment indicators, and a control set. The average treatment-control difference for the singletons is the estimated coefficient on the treatment indicator. For the twins, it is the estimated coefficient on the treatment indicator plus the estimated coefficient on the interaction. Control Set 1 is empty. Control Set 2 includes the variables in Panel *a* of Table 2. Control Set 3 includes Control Set 2 and the variables in Panels *b* and *c* of Table 2. Control Set 4 includes Control Set 3 and site fixed effects. The standard errors are clustered at the child-participant level. $\boldsymbol{\Gamma}$ is defined as the twin-singleton difference in the corresponding average treatment effects. We label the estimates of $\boldsymbol{\Gamma}$ according to the *p*-value of the *t*-statistic associated with the difference. The null hypothesis is that the difference is 0. The *p*-value is based on robust standard errors clustered at the child-participant level. ***: *p*-value < 0.01. **: *p*-value < 0.05. *: *p*-value < 0.10.

Figure A.10. Treatment-Control Differences in Parenting (Residualized from Maternal Hours Working and Maternal Care Hours)

Note: This figure displays the average treatment-control difference in the mean of the parenting measures observed at ages 1 and 3 net (residualized from) of (weekly) maternal hours working observed at 18, 24, 30, and 36 months and weekly maternal care hours observed at 18 and 30 months by twinning status. To obtain these average differences we regress the residualized parenting measure on a constant, a twin indicator, a treatment indicator, an interaction of the twin and treatment indicators, and a control set. The average treatment-control difference for the singletons is the estimated coefficient on the treatment indicator. For the twins, it is the estimated coefficient on the treatment indicator plus the estimated coefficient on the interaction. Control Set 1 is empty. Control Set 2 includes the variables in Panel a of Table 2. Control Set 3 includes Control Set 2 and the variables in Panels b and c of Table 2. Control Set 4 includes Control Set 3 and site fixed effects. The standard errors are clustered at the child-participant level. Γ is defined as the twin-singleton difference in the corresponding average treatment effects. We label the estimates of Γ according to the p-value of the t-statistic associated with the difference. The null hypothesis is that the difference is 0. The p-value is based on robust standard errors clustered at the child-participant level. ***: p-value < 0.01. **: p-value < 0.05. *: p-value < 0.10.

Appendix 5. Appendix to Section 7:
Additional Details on Treatment-Effect Decomposition

Appendix 5.1 Auxiliary Estimates to Baseline Decomposition

Table A.9. Joint Effects on Age-3 Cognition: OLS Estimates by Treatment Status

Panel a. **Control Group**				
Parenting	10.90	8.68	6.30	6.62
(s.e.)	(0.71)	(0.77)	(0.86)	(0.87)
Childcare	2.47	2.06	1.23	1.41
(s.e.)	(0.92)	(0.93)	(0.89)	(0.90)
Observations	339	339	336	336
Panel b. **Treatment Group**				
Parenting	9.11	6.93	5.73	5.54
(s.e.)	(0.95)	(1.00)	(1.16)	(1.25)
Childcare	2.45	2.88	2.99	2.93
(s.e.)	(1.20)	(1.09)	(1.05)	(1.06)
Observations	539	539	536	536
Control Set 1	✓	✓	✓	✓
Control Set 2		✓	✓	✓
Control Set 3			✓	✓
Control Set 4				✓

Note: In Panel *a*, Column (1) displays ordinary least squares (OLS) estimates of the coefficients α_c (childcare) and α_p (parenting) based on Equation (14) using the Stanford-Binet IQ test score at age 3 as dependent variable and Control Set 1. We only use the control-group sample. Columns (2), (3), and (4) are analogous in format to Column (1) for the control set in the label. The standard errors (s.e.) are clustered at the child-participant level. Panel *b*, is analogous in format to Panel *a* but it uses the treatment-group sample.

Table A.10. Joint Effects on Age-3 Cognition: Ordinary Least Squares and Instrumental Variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Control Set 1		Control Set 2		Control Set 3		Control Set 4	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Parenting	10.21	13.54	7.93	16.26	6.35	16.19	6.51	13.73
	(0.58)	(7.18)	(0.62)	(7.71)	(0.69)	(7.29)	(0.72)	(7.09)
Childcare	3.85	5.44	3.91	5.17	4.95	3.90	3.81	5.31
	(0.56)	(0.90)	(0.52)	(0.93)	(1.09)	(0.50)	(0.50)	(1.05)
Observations	878	878	878	878	872	872	872	872
First-Stage Rank Tests								
KP *LM*-statistic		3.98		5.15		6.97		7.11
p-value		0.05		0.02		0.01		0.01
AP *F*-statistic Parenting		4.13		5.30		7.36		7.41
p-value		0.04		0.02		0.01		0.01
AP *F*-statistic Childcare		744.87		674.98		461.85		482.62
p-value		0.00		0.00		0.00		0.00

Note: Column (1) displays OLS estimates of the coefficients α_c (childcare) and α_p (parenting) based on Equation (14) using the Stanford-Binet IQ test score at age 3 as dependent variable and Control Set 1. Column (2) is analogous in format to Column (2) for instrumental-variable estimates. In Column (2), we present standard statistics for testing the first-stage rank—i.e., the Kleibergen-Paap Lagrange-multiplier χ^2 for the joint system of first stages and the Angrist-Pischke (AP) F-statistics for the first stage of each endogenous variable. Columns (3), (5), and (7) are analogous in format to Column (1) for the control set in the label. Columns (4), (6), and (8) are analogous in format to Column (2) for the control set in the label. All specifications include a twin indicator as control. In addition, they include the control set in the label. Control Set 1 is empty. Control Set 2 includes the variables in Panel *a* of Table 2. Control Set 3 includes Control Set 2 and the variables in Panels *b* and *c* of Table 2. Control Set 4 includes Control Set 3 and site fixed effects. The standard errors (in parentheses) are clustered at the child-participant level.

Appendix 5.2 Alternative Identification Strategy

The relationship in Equation (13) has two potentially endogenous variables so identifying α_c and α_p using instrumental variables requires at least two instruments satisfying the rank and exogeneity conditions. We first focus on the rank condition.

Consider the first-stage system with two instruments, z_{i1} and z_{i2}:

$$c_i = \beta_0^c + \beta_1^c \cdot z_{i1} + \beta_2^c \cdot z_{i2} + \boldsymbol{\gamma}^c \cdot \boldsymbol{X}_i + v_i^c \tag{A.1}$$

$$p_i = \beta_0^p + \beta_1^p \cdot z_{i1} + \beta_2^p \cdot z_{i2} + \boldsymbol{\gamma}^p \cdot \boldsymbol{X}_i + v_i^p, \tag{A.2}$$

where we use c and p as superscripts for childcare and parenting and the usual notation applies for coefficients and error terms. The rank condition requires that c_i and p_i are non-trivial functions of z_{i1} and z_{i2}.[4] It is challenging to find instruments satisfying the rank condition when there are two endogenous variables. Conditional on z_{i2} and $\boldsymbol{X}_i$, z_{i1} should non-trivially shift both c_i and p_i. The same should be true for z_{i2}, conditional on z_{i1} and $\boldsymbol{X}_i$.

The findings in Section 5 motivate two instruments satisfying the rank condition: treatment and the interaction of treatment and twinning. Treatment sizably increases time spent in childcare; for twins, the increase is even larger. Conditional on the increase due to treatment for both singletons and twins, being a treated twin increases time spent in childcare. Treatment increases parenting for singletons and decreases it for twins. Thus, conditional on the increase for treated singletons, there is a large decrease in parenting for treated twins.

Panel *a* of Appendix Table A.11 displays OLS estimates of Equation (A.1). In the specification with the broadest control set, assignment to treatment increases time spent in childcare by 1.35 standard deviations for singletons, a magnitude that differs statistically from 0 at usual significance levels. The additional increase for twins of 0.32 standard deviations also differs statistically from 0 at usual levels. Panel *b* of Appendix Table A.11 displays OLS estimates of Equation (A.2); the results are as described in Section 5. The standardization of the childcare and parenting variables aids the interpretation of the magnitude of the substitution away from parenting by parents of treated twins. The most comprehensive specification indicates that parents of singletons increase parenting by 0.20 of a standard deviation, while parents of treated twins decrease it by 0.23 of a standard deviation. The difference between parents of treated singletons and twins is statistically significant at usual levels.

---

[4]Precisely, the rank condition requires that the matrix $\begin{bmatrix} \beta_1^c & \beta_2^c \\ \beta_1^p & \beta_2^p \end{bmatrix}$ is full rank. Tests of this condition are based on measuring the degree to which full rank of this matrix holds (statistically).

Table A.11. Joint Effects on Age-3 Cognition: First Stage and Reduced Form

Panel a. First Stage: Childcare				
1[treatment]	1.347	1.346	1.36	1.35
(s.e.)	(0.061)	(0.061)	(0.061)	(0.061)
F-statistic	482.765	487.854	497.726	480.643
1[twin]	-0.150	-0.140	-0.124	-0.140
(s.e.)	(0.070)	(0.071)	(0.074)	(0.073)
F-statistic	4.566	3.813	2.764	3.698
1[treatment] · **1**[twin]	0.313	0.288	0.286	0.323
(s.e.)	(0.134)	(0.135)	(0.144)	(0.141)
F-statistic	5.464	4.546	3.918	5.212
Panel b. First Stage: Parenting				
1[treatment]	0.140	0.143	0.207	0.205
(s.e.)	(0.071)	(0.063)	(0.057)	(0.054)
F-statistic	3.844	5.220	13.432	14.317
1[twin]	-0.119	-0.091	-0.079	-0.089
(s.e.)	(0.151)	(0.113)	(0.107)	(0.103)
F-statistic	0.628	0.660	0.543	0.753
1[treatment] · **1**[twin]	-0.455	-0.425	-0.448	-0.437
(s.e.)	(0.235)	(0.193)	(0.178)	(0.175)
F-statistic	3.731	4.837	6.370	6.260
Panel c. Reduced Form: Stanford-Binet IQ Score				
1[treatment]	9.222	9.290	10.069	9.963
(s.e.)	(1.402)	(1.229)	(1.159)	(1.138)
F-statistic	43.274	57.113	75.532	76.605
1[twin]	-0.554	0.467	0.861	0.164
(s.e.)	(3.120)	(2.379)	(2.390)	(2.440)
F-statistic	0.032	0.039	0.130	0.005
1[treatment] · **1**[twin]	-4.456	-5.412	-5.835	-4.283
(s.e.)	(4.121)	(3.400)	(3.414)	(3.385)
F-statistic	1.169	2.533	2.921	1.601
Observations	878	878	872	872
Control Set 1	✓	✓	✓	✓
Control Set 2		✓	✓	✓
Control Set 3			✓	✓
Control Set 4				✓

Note: Panel *a* displays OLS estimates of the coefficients β_1^c (**1** [treatment]) and β_2^c (**1** [treatment] × **1** [twin]). It also presents the coefficient estimate associated with a twin indicator (**1** [twin]). The estimates are based on four different specifications of Equation (A.1)—i.e., one specification per column. All specifications include a twin indicator as control. In addition, they include the control set in the label. Control Set 1 is empty. Control Set 2 includes the variables in Panel *a* of Table 2. Control Set 3 includes Control Set 2 and the variables in Panels *b* and *c* of Table 2. Control Set 4 includes Control Set 3 and site fixed effects. The standard errors (s.e.) are clustered at the child-participant level. The *F*-statistics are the squares of the *t*-statistics of the corresponding coefficients. Panels *b* and *c* are analogous in format to Panel *a* for the case when parenting and the Stanford-Binet IQ test score at age 3 are the dependent variables, respectively.

The results in Panels *a* and *b* of Appendix Table A.11 suggest that the proposed instruments shift the inputs independently. We test this formally. Satisfying the rank condition requires a joint test. We report the Kleibergen-Paap Lagrange-multiplier χ^2 statistic together with its *p*-value in Appendix Table A.10. The null hypothesis is that the system formed by Equations (A.1) and (A.2) is underidentified by one degree. The alternative hypothesis is that it is exactly identified. The evidence indicates that we reject the null hypothesis across specifications,[5] so the rank condition plausibly holds.

The exogeneity condition of the instruments requires two subconditions. First, the instruments need to satisfy random assignment, which is satisfied by the program's design. Conditional on twinning, the interaction of treatment and twinning also satisfies random assignment. Second, the instruments need to satisfy the exclusion restriction; they should only impact the relevant skill through childcare and parenting. This is a difficult subcondition to justify because treatment may impact skills through channels other than childcare and parenting for either the singletons or the twins.

To address the possibility that treatment impacts skills through other channels we control for the variables in the control sets discussed so far, which include maternal employment status and indicators of other economic circumstances.[6] Importantly, we also include a twin indicator in all our specifications, which picks up any average differences between singletons and twins. Additionally, our most comprehensive control set includes the number of siblings. The variation in the instrument that interacts treatment assignment and twinning status is thus only based on exploiting access to the program contemporaneous to compressed birth spacing. Finally, treatment may shift unobserved inputs correlated with childcare and parenting, although the evidence indicates that childcare and parenting explain most of the

---

[5]We follow recent studies with similar identification challenges to ours in also reporting the Angrist-Pischke (AP) *F*-statistic (e.g., Harari, 2020). The standard practice is to report one statistic per endogenous variable. The AP *F*-statistic is constructed as follows in the case of childcare. First, parenting is residualized from childcare, the instruments, and the corresponding control set. Second, a regression of residualized childcare on the residualized instruments and controls is estimated. Third, the *F*-statistic from the joint null hypothesis that the coefficients on the residualized instruments in the regression of the second step is computed. This is the AP *F*-statistic associated with childcare. The AP *F*-statistic associated with parenting is computed analogously, residualizing childcare from the variables involved in the first and second steps. While intuitive, the AP *F*-statistic does not account for the sampling variation in the steps required for its computation. Thus, the statistic is usually reported and compared to an arbitrary threshold of 10, which is based on simulation exercises for contexts with a single endogenous variable (Stock and Yogo, 2002). The Kleibergen-Paap Lagrange-multiplier statistic has a precise null hypothesis and proper inference. We base our conclusions regarding the rank condition on it.

[6]Treatment could also shift the relationship between skills and childcare and parenting. As argued before, however, we cannot reject the null hypothesis $\alpha_c^0 = \alpha_c^1 =: \alpha_c$ and $\alpha_p^0 = \alpha_p^1 =: \alpha_p$ when estimating Equation (13) using OLS. Testing this hypothesis using the instruments proposed is impossible (the instruments exploit treatment assignment and we thus cannot use them to estimate Equation (13) by treatment status).

Figure A.11. Decomposition of Treatment Effect on Age-3 Cognition Using Instrumental Variables
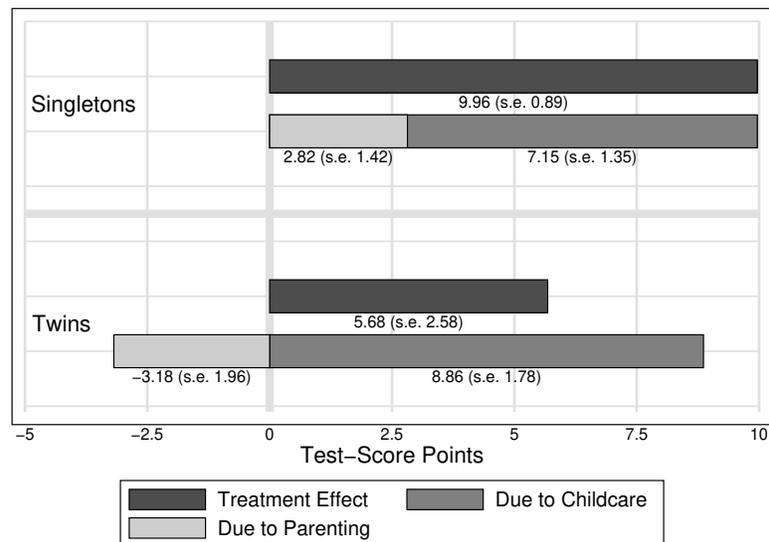
Note: This figure is analogous in format to Figure 7. It displays estimates of the average treatment effect (ATE) decomposition into childcare and parenting based on Equation (14). We use the ordinary least squares (IV) coefficient estimates of α_c (childcare) and α_p (parenting) in Equation (13) reported in Appendix Table A.10. We use the coefficient estimates obtained when using the most comprehensive control set (Control Set 4). The measure of age-3 cognition is the Stanford-Binet IQ. The ATE decomposed is the OLS estimate in Table A.8, also based on the most comprehensive control set. Panel (b) is analogous in format to Panel (a).

ATE. It could also be argued that the program was designed to shift parenting and childcare and not any other inputs. While we are cautious in describing the plausibility of the instruments, the arguments above suggest that results based on them are a valuable complement to the results based on OLS.

Appendix Table A.10 presents the instrumental-variable estimates of α_c and α_p. Appendix Figure A.11 displays the decomposition, which is obtained from Equation (14) by replacing the OLS estimates with the IV estimates.[7]

---

[7]By construction, the decomposition based on instrumental variables has no residual. To see this, consider the case of the decomposition for the singletons. In this case, $\left(\alpha_0^1 - \alpha_0^0\right)$ in Equation (14) is the difference in the constants between the treatment and control-group singletons. The treatment indicator used as an instrument absorbs this difference into the first-stage predictions, therefore absorbing the residual. The analogous case applies to the twins, given that we also use the interaction of being treated and being twin as an instrument.

Appendix 5.3 Decomposition Based on PPVT

Figure A.12. Decomposition of Treatment Effect on Age-3 Cognition Based on the Peabody Picture Vocabulary Test

(a) Ordinary Least Squares **(b)** Instrumental Variables
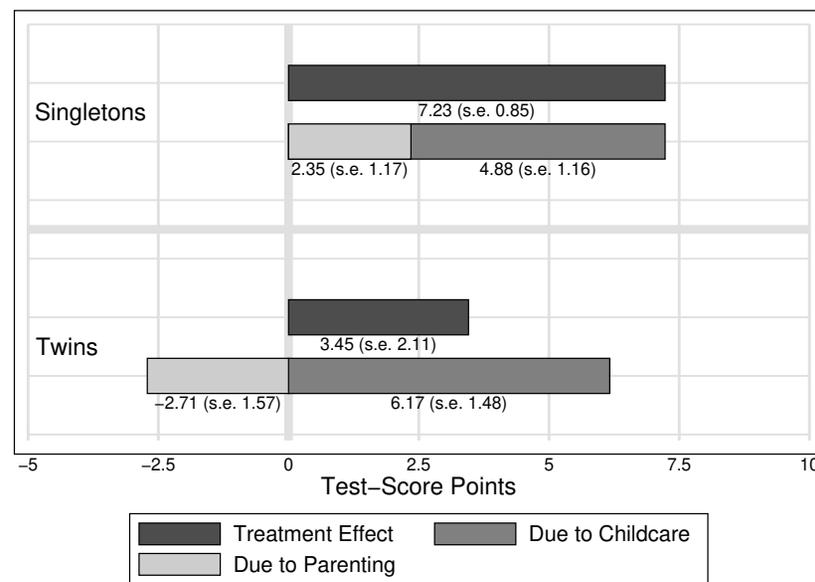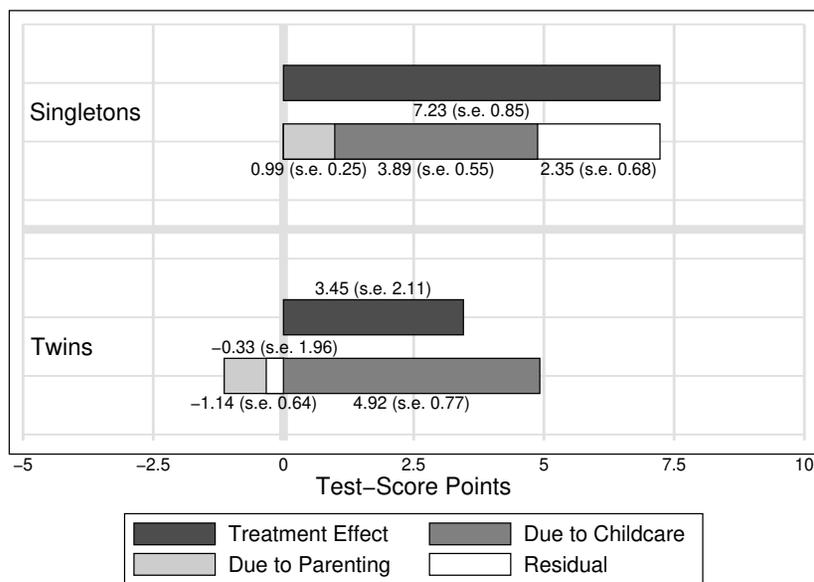
Note: Panel (a) displays estimates of the average treatment effect (ATE) decomposition into childcare, parenting, and an unexplained component based on Equation (14). We use the ordinary least squares (OLS) coefficient estimates of α_c (childcare) and α_p (parenting) in Equation 13. We use the coefficient estimates obtained when using the most comprehensive control set (Control Set 4). The measure of age-3 cognition is the Peabody Picture Vocabulary Test. The ATE decomposed is the OLS estimate in Table A.8, also based on the most comprehensive control set. Panel (b) is analogous in format to Panel (a). All its construction elements are identical, except for the coefficients estimate used in the decomposition are the instrumental variables (IV) estimates, instead of the OLS estimates.

Appendix References

Chaparro, J., A. Sojourner, and M. J. Wiswall (2020). Early Childhood Care and Cognitive Development. NBER Working Paper w26813, National Bureau of Economic Research.

Galland, B. C., B. J. Taylor, D. E. Elder, and P. Herbison (2012). Normal Sleep Patterns in Infants and Children: A Systematic Review of Observational Studies. *Sleep Medicine Reviews 16*(3), 213–222.

Gross, R. T., D. Spiker, and C. W. Haynes (1997). *Helping Low Birthweight, Premature Babies: The Infant Health and Development Program.* Stanford University Press.

Hansen, B. E. (2021). *Econometrics.* [Online; accessed 31-March-2021].

Harari, M. (2020). Cities in Bad Shape: Urban Geometry in India. *American Economic Review 110*(8), 2377–2421.

Iglowstein, I., O. G. Jenni, L. Molinari, and R. H. Largo (2003). Sleep Duration from Infancy to Adolescence: Reference Values and Generational Trends. *Pediatrics 111*(2), 302–307.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika 73*(1), 13–22.

Stock, J. H. and M. Yogo (2002). Testing for weak instruments in linear IV regression. NBER Working Paper w0284, National Bureau of Economic Research.

Williamson, A. A., R. Gould, E. S. Leichman, R. M. Walters, and J. A. Mindell (2021). Socioeconomic Disadvantage and Sleep in Early Childhood: Real-world Data from a Mobile Health Application. *Sleep Health 7*(2), 143–152.