Web Appendix to "New Evidence for an Old Puzzle: The Fadeout of Early Childhood Education Impacts on Cognitive Skills"

Andrés Hojman

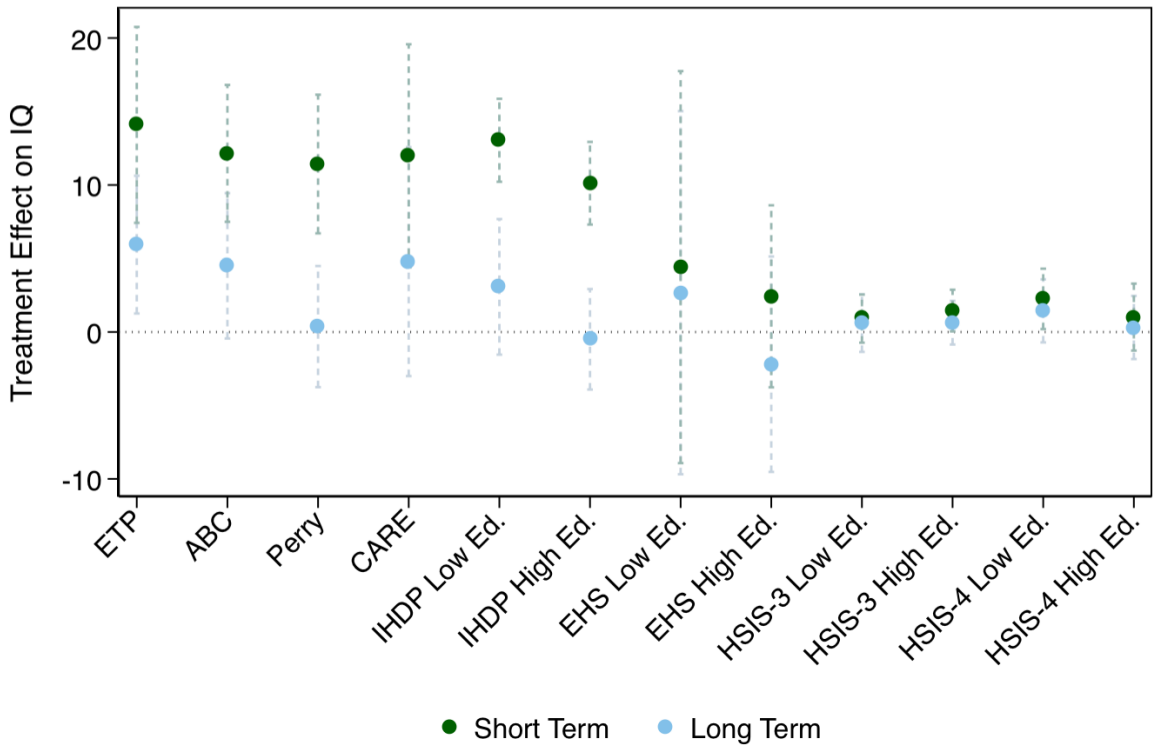
Contents

A Prevalence of Fadeout

This Appendix documents that fadeout is pervasive and compares its magnitude across several experimentally evaluated early education programs. Figure 1 shows 12 estimations of treatment effects from 7 different experimentally assigned EEPs. The dark dots are short-term impacts and the light dots are long-term impacts. In all cases, the short-term impact is measured right after the end of the program, and the long-term impacts are measured between 6.5 and 10 years of age.[1] The figure shows that there is some degree of fadeout in all of the programs and for most of them, even for the ones with strong initial impacts, the long-term impacts on IQ are close to zero.

[1] Using ages 6.5-10 might seem like little time for considering long-term impacts, but in practice, using data up to later ages would show even more fadeout: impacts almost never open up in the data after the gaps have closed.

Figure 1: Short-Term and Long-Term Impacts of Early Education Programs (EEPs)



Note: This chart presents the short-term and long-term impacts of several EEPs on IQ. The impacts are calculated as simple differences in means between the treatment and the control groups, which might underestimate the real impacts of some of the programs. I use standardized scores in this chart, which are normed to have national means of 100 and national standard deviations of 15. For large-scale programs, individuals are grouped by high/low maternal education (Educ.$\leq$ 12) to make samples more comparable, and to observe fadeout in more groups. ETP: Early Training Project; ABC: Abecedarian Carolina Project; Perry: Perry Preschool Project; CARE: Carolina Approach to Responsive Education; IHDP: Infant Health and Development Program; EHS: Early Head Start; HSIS: Head Start Impact Study. Source: Own Calculations.

B Models and Research Questions

In this section, I present two models that can rationalize the pattern of impacts in the data. Both models can explain fadeout, but the intuition and the implications for both are very different. I first present a Differential Gains Model, where fadeout is explained by the treatment group benefiting less than the control group from the first year of school. Then, I present a Depreciation Model, where fadeout is explained by a faster depreciation of the type of cognitive skills gained from participation in early childhood education programs, as compared with the depreciation of other cognitive skills. I introduce one model at a time for clarity, and then I present a combined model that includes differential gains and depreciation at the same time.

B.1 Differential Gains Model

This model allows for the impacts of formal education on cognitive skills to be decreasing in previous exposure to formal education.[2] If that is the case, treated children will benefit from a large initial impact when they first enter the early childhood education programs, but the impacts of additional years of formal education will be decreasing in magnitude. On the other hand, controls first access formal education at school entry. Thus, they will have large impacts from school in that period (that match the treatment group's gains after the start of the early childhood education programs), while the positive impact of entering school for treatment children will be relatively small. As a consequence, the gap between the test scores of the two groups closes and a fadeout pattern is observed in the data.

Under the Differential Gains Model, the cognitive skill in a given period, θ_{it} is given by (i) the persistence of the cognitive skill in the previous period, θ_{it-1}; (ii) participation in formal education during period t, F_{it}; (iii) previous exposure to formal education, F_i^t; (iv) an age fixed effect, ω_t; (v) an individual fixed effect, α_i; and (vi) a random error, η_{it}. F_i^t is defined as $F_i^t = I\left[\sum_{\tilde{t}=0}^{t-1} F_{i\tilde{t}} > 0\right]$. To consider the possibility of differential gains, I allow the

[2] I consider as formal education the participation in early childhood education programs or schools

impacts of F_{it} to be *moderated* by F_i^t.

$$\theta_{it} = \rho_t \theta_{it-1} + \beta_t \left(1 + \tau_t F_i^t\right) F_{it} + \omega_t + \alpha_i + \eta_{it}. \tag{1}$$

A negative coefficient for τ_t implies the existence of lower gains from education for children with previous educational experiences. The presence of differential gains is compatible with different explanations that might be related to fadeout. Their common factor is that fadeout is caused by later schooling. I cannot test those explanations independently, but in Section 7 I discuss how the evidence suggests that diminishing returns to education in the production of test scores seems to be a substantive part of the explanation.

B.2 Depreciation Model

This model allows for the depreciation of the type of cognitive skills gained from early childhood education programs to be faster than the depreciation of other cognitive skills. This can generate the fadeout observed in the data.

Under the Depreciation Model, I assume that the programs teach a particular type of cognitive skills, which we can call K_{it}. They are part of the relevant cognitive skills for this paper, θ_{it} (the skills measured by the tests, as discussed in Section 3).[3] The skill production function for K is:

$$K_{it} = (\rho_t - \delta_t) K_{it-1} + \beta_t F_{it} + \omega_t^K + \alpha_i^K + \eta_{it}^K \tag{2}$$

Where $\omega_t^K, \alpha_i^K, \eta_{it}^K$ are error components in the production function, analogous to the ones in Equation (1). I assume that no proxies for K_{it} are observed, but I can still identify the parameter δ_t, as discussed in Section 5. I assume that in each period, the total cognitive

[3] In Appendix K, I present this model as generated by two mutually exclusive skills that add up to θ. The formulation in the main paper saves some notation.

skills, which are measured by observed tests, can be expressed as:

$$\theta_{it} = \rho_t \theta_{it-1} - \delta_t K_{it-1} + \beta_t F_{it} + \omega_t + \alpha_i + \eta_{it} \tag{3}$$

A positive coefficient for δ_t implies the existence of depreciation. Depreciation can explain the fadeout phenomenon because when children enter schools the skills that the treated children already gained depreciate. This can be true even if there is a gain for both groups of attending school. The usual hypotheses in the literature assume that depreciation will only happen if the subsequent educational environment is poor. I allow for faster depreciation to occur even if the quality of the educational environment is high. In Section 6, I argue this helps the model to better fit the data, but the resulting values of the depreciation parameters seem too high to be plausible.

B.3 Combined Model

I now present a model including differential gains from schooling and depreciation effects. In principle, both effects could coexist. Thus, it is necessary to consider them simultaneously to be able to disentangle their relative importance. I only estimate the Combined Model using two periods of data. Thus, here I develop that case, which is illustrative. In only two periods, the skills for individual i will be given by:

$$\theta_{i2} = \rho_2 \theta_{i1} + \beta_2 F_{i2} - \delta_2 \beta_1 F_{i1} + \beta_2 \tau_2 F_{i1} F_{i2} + \eta^*_{i2} \tag{4}$$

With $\eta^*_{i2} = \omega_t + \alpha_i + \eta_{i2} - \delta_2 \left(\omega^K_1 + \alpha^K_i + \eta^K_{i1}\right) - \delta_2 \left(\rho_1 - \delta_1\right) K_0$

Notice that the depreciation term, δ_2, will be associated to the previous exposure to education, F_{i1}, while the differential gains term, τ_2, will be associated to interaction with education on the current period, $F_{i1}F_{i2}$. It is possible to test whether both models are supported by the data if there is independent variation in F_{i1} and F_{i2}. As I discuss in Section 4, this is not a trivial condition to achieve. In most situations, children that attend

preschool programs keep attending them until they enter school. In that case, $F_{i1}F_{i2} = F_{i1}$, and in empirically relevant cases there will be an observational equivalence between the models.[4] This is the case of Perry and ETP, which is why I need to use IHDP, that counts with independent variation in the attendance to formal education in different periods, to estimate the combined model on it.

In Section 5, I identify the key components of this model using the IHDP dataset. In Section 6, I show that when tested in a model that also allows for differential gains, there is no empirical support for depreciation to be a major contribution to fadeout.

[4]The observational equivalence is discussed in Section 5

C Overview of Cognitive Measures

This Appendix gives information about the intelligence (IQ) and achievement tests used in the Early Training Project (ETP) and the Perry Preschool Project (Perry). The information given below includes (i) definitions of the different types of scores, (ii) general information about each test, (iii) administration and scoring details, (iv) discussion of the standardization protocol given by the tests' creators, and (v) the transformation of the scores in the data to create raw and standardized scores and mental age, if possible.

The tests given for each program are summarized in Table 1. There have been many editions of these tests. Thus, it is important to refer to the version of the test that was used in the original programs to correctly transform the scores in the data. Table 1 lists the editions of the test used for each program as found in the documentation for the programs.

Test		**ETP**	**Perry**
IQ	ITPA	1st Edition (1961)	1st Edition (1961)
	Leiter	—	Revision (1948)
	PPVT	1st Edition (1961); Form B	1st Edition (1961); Form A
	SB	3rd Edition (1960)	3rd Edition (1960)
	WISC	1st Edition (1959) & Revised Edition (1974)	1st Edition (1959)
Achievement	CAT	Revised Edition (1963)	Revised Edition (1963) & 3rd Edition (1970)
	MAT	3rd Edition (1962)	—
	SAT	4th Edition (1966)	—

Table 1: This information on the edition of each test is in the documentation for the programs. The IQ tests for the two programs are: the Illinois Test of Psycholinguistic Abilities (ITPA), the Leiter International Performance Scale (Leiter), the Peabody Picture Vocabulary Test (PPVT), the Stanford-Binet Intelligence Scale (SB), and the Wechsler Intelligence Scale for Children (WISC). The achievement tests for the two programs are: the California Achievement Tests (CAT), the Metropolitan Achievement Test (MAT), and the Stanford Achievement Test (SAT).

C.1 Explanation of Terminology

This section defines the terms that are used across the different tests.

- **Raw Score.** Because these tests are designed to be objective, an individual answers an item either correctly or incorrectly. Let I_j be a particular item j on a test s, with

$j = 1 \ldots J_s$. If the individual answers item j correctly, then $I_j = 1$. Otherwise, $I_j = 0$.

The raw score, $W_{i,s}$, is generally defined for an individual i on a test s[5] as:

$$W_{i,s} = \sum_{j=1}^{J_s} I_{i,j,s} \tag{5}$$

  - **Basal and Ceiling.** Several of the tests define basal and ceiling items so that not every question needs to be asked. The basal item is the first item the individual answers correctly, and is selected based on chronological age. It is assumed that any items that come before this basal item would be answered correctly. Similarly, it is assumed that any item that comes after the ceiling item would be answered incorrectly. Several of the tests given in ETP and Perry use different methods to arrive at the basal and ceiling items, but they all rest on these two assumptions for scoring. Equation 5 is still used to calculate the raw score under this construction.

Different tests have different number of items, and thus the raw scores are not comparable or informative across tests and years.

- **Mental Age.** It can be valuable to understand how an individual performs on a test in a format that is immediately comparable to the individual's age. The mental age provides this measure. It is calculated from the nationally representative sample as the expected age given a raw score (Equation 6), and does not depend on the chronological age of the individual.

$$MA_{i,s} = \mathbb{E}\left[CA | W_{i,s}\right] \tag{6}$$

This correspondence is documented in a table, which will henceforth be referred to as *age norms*. These age norms are needed to convert from raw score to mental age.

[5]The test, s, could also be a subtest. For WISC, the raw score is calculated for the individual subtests. There is no method to calculate total raw score.

Although this method to obtain mental age uses the nationally representative sample, several of the tests used in the programs are designed so that the items match with different mental ages. There is then a different mental age that is constructed from the raw test scores of the child. For example, one question might correspond to 2 months of mental age. To differentiate between this and the mental age defined above in Equation 6, the mental age that comes directly from the child's test score will be notated as $\widetilde{MA}$.

- **Grade Equivalent.** The grade equivalent score (GE) is used in achievement tests to assess how an individual performs on a test in a format that is comparable to the individual's grade. This score is calculated from the national sample and is reported in terms of grade level and months with a decimal number. For example, if GE is 2.3, the digit to the left (2) of the decimal refers to the grade (second grade), and the digit to the right of the decimal (3) refers to the month (three months). GE assumes there are 10 months per school year.

  GE needs to be cautiously interpreted. GE of a raw score represents the grade level at which the average student earns this raw score. If a 2nd grade individual obtains a GE of 3.5 onan achievement test, it does not mean that this individual is able to perform well on an achievement test intended for mid-3rd grade students. Rather, a GE of 3.5 means that this individual can solve 2nd grade problems as well as the average mid-3rd grade student can solve 2nd grade problems.

- **Standardized Score.** The standardized score is equivalent to an intelligence quotient, and uses nationally-representative norms to convert performance on a test to a score that is consistent and comparable across years and individuals. By design, the mean and standard deviation of the population's standardized scores are 100 and 15, respectively.

  Equation 7 shows how to calculate the standardized score from the raw score using

the mean, $\mu^r_{CA_i,s}$, and standard deviation, $\sigma^r_{CA_i,s}$, of raw scores for an individual with a certain chronological age, CA_i.

$$Y_{i,s} = 15\Big(\frac{W_{i,s} - \mu^r_{CA_i,s}}{\sigma^r_{CA_i,s}}\Big) + 100 \tag{7}$$

– **Conventional IQ.** Given the mental age, $\widetilde{MA}_{i,s}$, for a test, s, taken by an individual, i, with a certain chronological age, CA_i, the conventional IQ, $\tilde{Y}_{i,s}$, is an approximation for the standardized score.

$$\tilde{Y}_{i,s} = 100\Big(\frac{\widetilde{MA}_{i,s}}{CA_i}\Big) \tag{8}$$

If an individual has a mental age equal to his chronological age, then the conventional IQ is 100, which is the average by construction. This aligns with the interpretation of mental age as the average chronological age for a given score.

The conventional IQ can be used to calculate the standardized score more precisely. Equation 9 calculates the standardized score given the mean, $\tilde{\mu}_{CA_i,s}$, and standard deviation, $\tilde{\sigma}_{CA_i,s}$, of conventional IQs for individuals with a certain chronological age, CA_i.

$$Y_{i,s} = 15\Big(\frac{\tilde{Y}_{i,s} - \tilde{\mu}_{CA_i,s}}{\tilde{\sigma}_{CA_i,s}}\Big) + 100 \tag{9}$$

– **Stanine.** The stanine evaluates an individual's score with a scale of equal units from 1 (lowest level) to 9 (highest level). It is a standard score with a nine-unit scale that is derived from a nationally representative sample with a mean of 5 and a standard deviation of 2.

C.2 Overview of Available Tests

Tables 2 and 3 show the timing of the tests' administrations for IQ measures. Table 4 shows the availability of different types of test scores. It also provides a summary of what scores are available in the data.

Table 4 also shows what type of intelligence the tests measure. Intelligence is categorized into crystalized and fluid. Crystallized intelligence is demonstrated in measures for which "skilled judgement habits have become crystallized" (Cattell, 1963, p. 2). Measures of fluid intelligence do not depend on crystallized intelligence, and thus demonstrate ability in new situations.

Perry	Age (months)											
	Pre	36	48	60	72	84	96	108	120	132	144	168
Wave 0			■									
Waves 1-4		■	■									
School				■	■	■	■	■	■	■	■	■
Period	1	2	3	4	5	6	7	8	9	10	11	12
ITPA		■	■	■	■	■	■	■				
Leiter		■	■	■	■	■	■	■				
PPVT		■	■	■	■	■	■	■				
SB	■		■	■	■	■	■	■	■	■	■	
WISC												■
CAT						■	■	■	■	■		■

Table 2: This tables shows the timing of the tests's administrations in Perry in relation to timing of the program and school. Table 5 shows detail for which waves were administered which tests. For all CAT administrations, reading, language, and mathematics subtests were given.

Perry had 5 waves of children that were studied. Wave 0 entered the Perry preschool for one year at age 4, whereas the other waves entered the Perry preschool for two years at age 3. There are other differences in the administration of tests for wave 2. Table 5 shows which waves received which tests for Perry. Although ETP also had two cohorts that received the intervention for different durations, all children were administered cognitive measures, even if they were not in the intervention yet.

ETP	Age (months)													
	46	49	58	61	70	73	81	83	85	94	95	117	121	203
Experimental 1		■	■	■	■	■	■							
Experimental 2			■	■	■	■	■							
School								■	■	■	■	■	■	■
Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ITPA						■		■			■			
PPVT	■	■	■	■	■	■		■	■		■		■	
SB	■	■	■	■		■			■		■		■	
WISC					■			■			■			■
CAT								■						
MAT							■			■		■		
SAT											■			

Table 3: This tables shows the timing of the tests' administrations in ETP in relation to timing of the program and school. Only the arithmetic subtest was given for CAT. For MAT, the arithmetic, reading, word discrimination, and word knowledge subtests were given at 81 months; the arithmetic, spelling, reading, word discrimination, and word knowledge subtests were given at 94 and 117 months. For SAT, the arithmetic spelling, paragraph meaning, vocabulary, word reading, and word study skills subtests were administered.

C.3 IQ Scores Available in the Data

C.3.1 Illinois Test of Psycholinguistic Abilities

The Illinois Test of Psycholinguistic Abilities (ITPA) is designed to capture the linguistic abilities of a child based on several dimensions of psycholinguistic ability. The test is composed of nine subtests, each of which assesses a different dimension (McCarthy and Kirk, 1961, p. 4). Seven of the subtests use basal/ceiling scoring. The two other subtests are administered in their entirety (McCarthy and Kirk, 1961, p. 22).

The authors define mental age so that the results from ITPA are comparable to those of other intelligence measures. The overall mental age is the standardized sum of the raw scores of the nine subtests based on the norms (McCarthy and Kirk, 1961, p. 95). Figure 2 shows the relationship between chronological age and raw score of the full test. Norms are also given to standardize the subtests' scores individually to help compare a student's performance on the different subtests (McCarthy and Kirk, 1961, p. 96).

Test		**Mental Age/ Grade Equivalent**	← age norms	**Raw**	→ norms	**Standardized**	**C**	**F**
Perry	ITPA	★	•	■	•	■		
	Leiter	■		■		★		√
	PPVT	■	•	■	•	★	√	
	SB	$★^o$			•	$★^o$		
	WISC	■	•	$★^o$	•	$★^o$	√	√
	CAT			★	•	■	√	
ETP	ITPA	★	•	■	•	■		
	PPVT	★	•	★	•	★	√	
	SB	■			•	★		
	WISC	■	•	★	•	★	√	√
	CAT	★	•	■	•	■	√	
	MAT	★	•	■	•	■	√	
	SAT	★	•	★	•	■	√	

Table 4: The black squares represent a score that was calculated based on the scores given in the datasets (indicated by stars). In general, to go from raw scores to mental age/grade equivalent, age norms are required. Mental age is a measurement for IQ tests and grade equivalent is a similar measurement for achievement tests. A dot (•) indicates that a test has age norms available. Similarly, to go from raw scores to standardized scores, norms are required, and a dot indicates their availability for a particular test. For WISC, raw scores are given by subtests and there is no method to construct an aggregate raw score. For SB, the norms convert from mental age to standardized because raw score is similarly not used. In Perry, the original scans for SB and WISC are given ($★^o$). The right part of the table shows which tests focus on crystallized (C) or fluid (F) intelligence, if applicable.

	Wave																	
	0						1						2-4					
Test	C	I	L	P	S	W	C	I	L	P	S	W	C	I	L	P	S	W
36								●	●	●	●			●	●	●	●	
48		●	●	●	●						●				●	●	●	
60				●	●			●	●	●	●			●	●	●	●	
72		●	●	●	●			●	●	●	●			●	●	●	●	
84	●	●	●	●	●		●	●	●	●	●		●	●	●	●	●	
96	●	●	●	●	●		●	●	●	●	●		●	●	●	●	●	
108	●	●	●	●	●		●	●	●	●	●		●	●	●	●	●	
120	●				●						●		●				●	
132					●		●						●					
144					●						●						●	
168	●					●	●					●	●					●

Table 5: The dots indicate administration of the tests. **C** is CAT; **I** is ITPA; **L** is Leiter; **P** is PPVT; **S** is SB; **W** is WISC.

In the data both ETP and Perry, the score for ITPA is given as mental age. The age norms tables can be applied to convert the mental ages of the children to raw scores. This raw score is the sum of the raw scores of the subtests. The standardized score is obtained from this raw score using the provided norms.

After these calculations, there are some missing data points. Linear extrapolation of the norms is used to fill in these missing values.

C.3.2 Leiter International Performance Scale

The Leiter International Performance Scale (Leiter) is designed to be a non-verbal substitute for Stanford-Binet, and thus the scoring of the two tests is similar (Leiter, 1940, p. 10). The test is comprised of a series of blocks the child places along the notches of a ruler. For example, a child must match green, blue, and red blocks to notches with corresponding colors. The mental age is the basal mental age plus the added credit months for all subsequent, passed tests.

Because the norms for Leiter were not published due to complications in dissemination

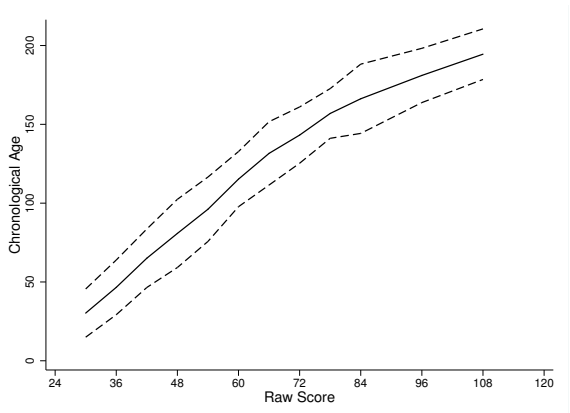


Figure 2: Mean of raw scores for ITPA at different ages. The dashed lines show one standard deviation above and below the mean. Source: McCarthy and Kirk (1961).

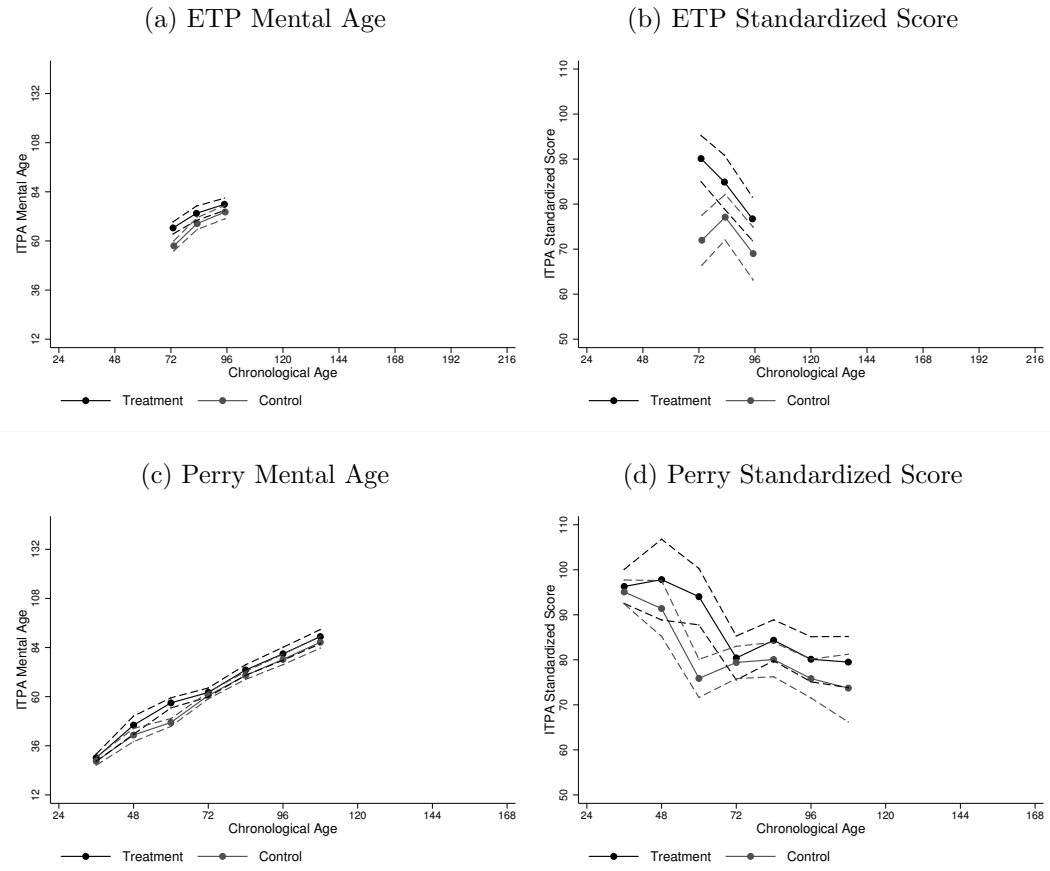
Test	Basal Mental Age
II	30
III	42
IV	51
V	63
VI	75
VII	87
VIII	99
IX	111

Table 6: This table is used to calculate basal mental age. If a test in the left column is passed in its entirety, the basal mental age is the corresponding mental age in the right column. Source: (Arthur, 1952, p. 8).

of testing materials and the arrival of WWII (Leiter, 1952, p. 259), HighScope used the following method to obtain a standardized score. The raw score is transformed to mental age using the age norms in the PPVT manual. Then, the standardized score is found using the equation for conventional IQ ($Y = 100\frac{MA}{CA}$).

In the Perry data, Leiter is given in standardized scores. Using this standardized score and the chronological age, the conventional IQ equation is used to find the mental age. Using the mental age and the PPVT age norms, the raw score is recovered.

Figure 3: Average ITPA scores by treatment status with dashed lines depicting 95% confidence intervals.

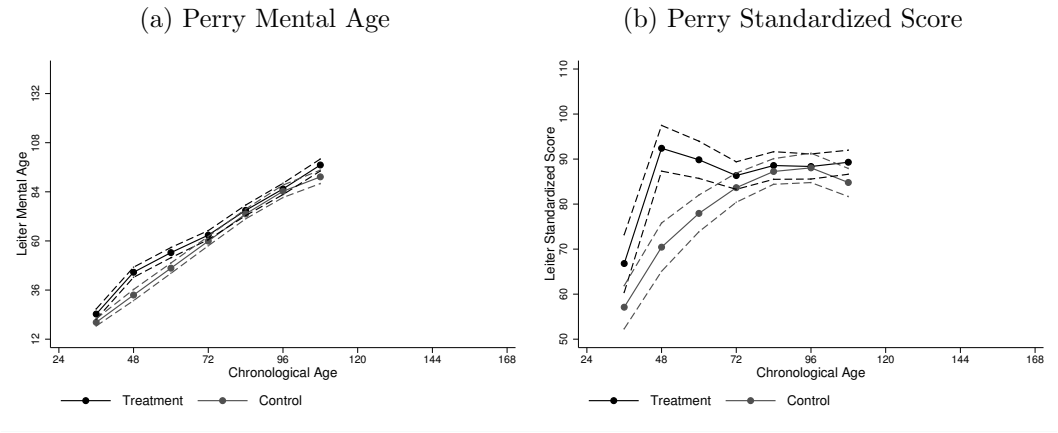


C.3.3 Peabody Picture Vocabulary Test

The Peabody Picture Vocabulary Test (PPVT) measures knowledge of vocabulary words (Dunn, 1965, p. 37). The test is administered for a short duration of time, usually 10 to 15 minutes (Dunn, 1965, p. 5). During this time, the child must identify from four choices the picture that corresponds to a word the test examiner says aloud (Dunn, 1965, p. 7). There are two possible forms, A and B, each of which contains 150 separate vocabulary words ordered by difficulty. Separate norms exist for each form.

The starting item is chosen based on chronological age. After the starting item, the child is asked the next items consecutively until the first error. If this first error occurs after eight items, then the starting item corresponding to the child's chronological age is the basal. If

Figure 4: Average Leiter scores by treatment status with dashed lines depicting 95% confidence intervals.



the child makes an error before eight items have been answered correctly, then the examiner works backwards until the child answers eight consecutive items correctly. After the basal item is determined, the test examiner continues showing plates in order until the child makes six errors in eight consecutive items. The last of these errors is the ceiling item and indicates the end of the test (Dunn, 1965, p. 8).

The raw score for the PPVT is the total number of correct responses including the unasked questions below the basal item and above the ceiling item. The raw score is thus equal to the total number of errors subtracted from the number of the ceiling question (Dunn, 1965, p. 10). Age norms and norms are given in the manual to convert raw scores to mental age and standardized scores. Figure 5 shows the relationship between chronological age and raw scores from the norms for Form A.

In the ETP data, the raw scores, standardized scores, and mental ages are all given. This allows for verification of both the edition and form used. In Perry, only the standardized scores are given. The norms are used to transform from standardized score to raw score. Then, the age norms are used to transform the raw scores to mental age.

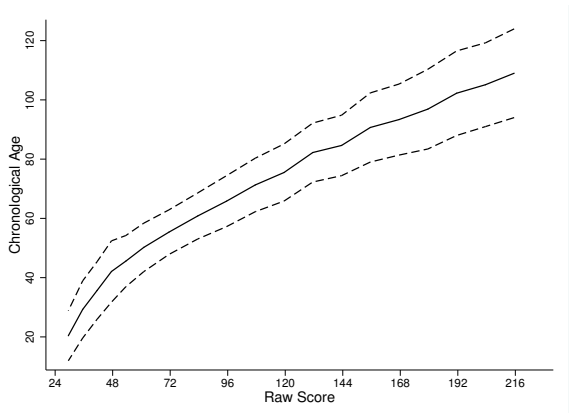


Figure 5: Mean of raw scores for PPVT Form A at different ages. The dashed lines show one standard deviation above and below the mean. Source: Dunn (1965, p.28).

C.3.4 Stanford-Binet Intelligence Scale

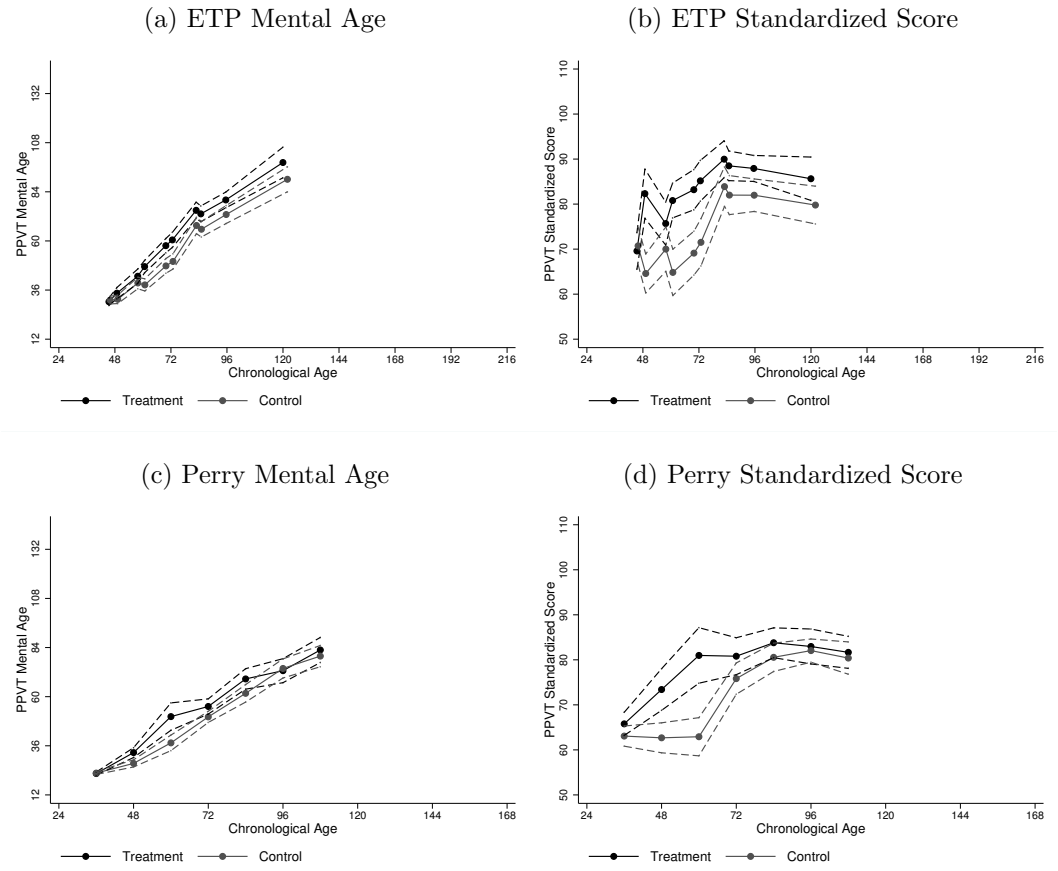
The Stanford-Binet Intelligence Scale (SB) is a revision of the Binet-Simon Scale, the first scale to use the concept of mental age and intelligence quotient as the ratio of mental age to chronological age (Terman and Merrill, 1960, p. 5-6). The test is a collection of tests with items that in aggregate give a measure of general intelligence. The tests are in order of difficulty, and closely correspond to mental ages. Table 7 presents a summary of the scoring structure.

Test	Number of Items	Credits Per Item (Months of Mental Age)	Total Possible Credits
II to IV-6	6	1	6
V to XIV	6	2	12
AA	8	2	16
SA I	6	4	24
SA II	6	5	30
SA III	6	6	36

Table 7: Scoring for the different tests and the relationship to mental age. Source: Terman and Merrill (1960, p.62).

The examiner begins at the test that seems appropriate for the child based on chronological age, grade placement, and presented behavior. For the typical child, the starting test corresponds to one year below the child's chronological age (Terman and Merrill, 1960,

Figure 6: Average PPVT scores by treatment status with dashed lines depicting 95% confidence intervals.



p. 59). The basal mental age is the test for which the child answers all items correctly given he makes at least one error on the proceeding test. If a child answers all items correctly for a test higher than the basal age, the basal mental age does not change. The examiner continues administering the tests to the child until he reaches a test at which none of the questions are answered correctly. This is the ceiling mental age (Terman and Merrill, 1960, p. 60). Because the items correspond to months of mental age, no raw score is calculated for SB.

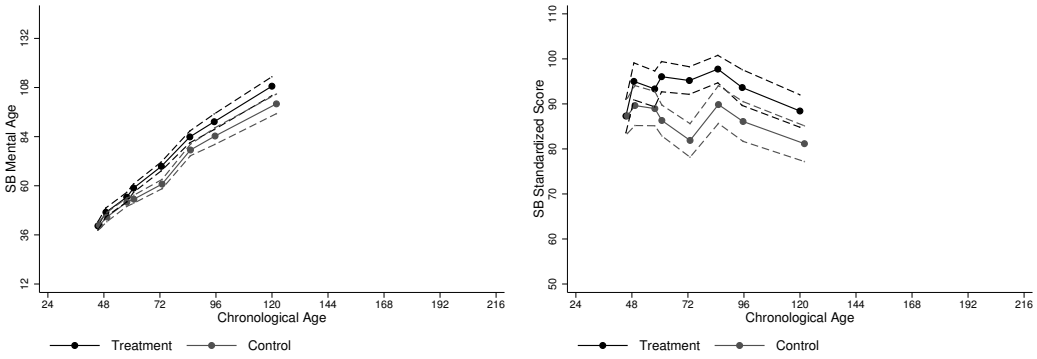
The ETP data include the mental age and standardized scores for SB. In the Perry data, the standardized scores and mental age are available from the scans of the original tests. For any missing values, the conventional IQ is calculated using the conventional IQ norms.

Then, the formula for conventional IQ is used to find mental age.

Figure 7: Average SB scores by treatment status with dashed lines depicting 95% confidence intervals. The graphs for Perry include the WISC scores at 168 months because there is only one observation.

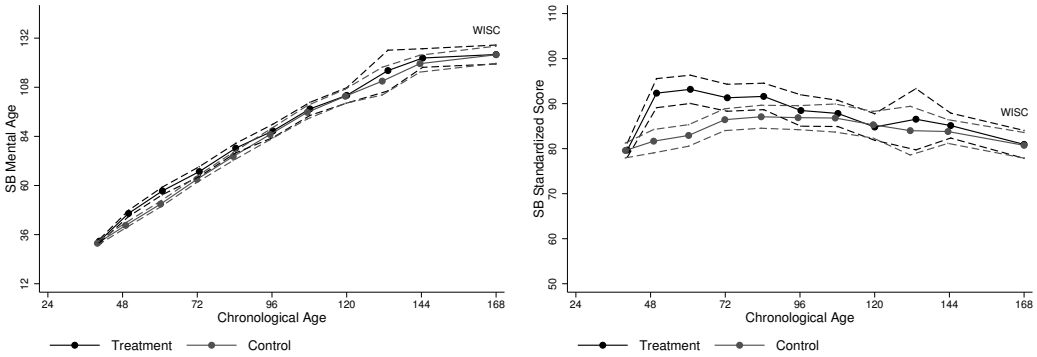
(a) ETP Mental Age

(b) ETP Standardized Score



(c) Perry Mental Age with WISC

(d) Perry Standardized Score with WISC



C.3.5 Wechsler Intelligence Scale for Children

The Wechsler Intelligence Scale for Children (WISC) is built on the theory that intelligence is part of an individual's broader personality. The test is divided into 10 subtests that make up two groups, Verbal and Performance (Wechsler, 1949, p. 5). The raw score for each of the subtests is calculated based on the number of questions answered correctly. These raw scores are then given a standardized score with a mean of 10 and a standard deviation of 2. The standardized scores of the subtests that make up the Verbal group can be summed to get the Verbal score, and the same can be done with the standardized scores of the subtests in the

Performance group.[6] The Verbal score and Performance score added together give the Full score. The Verbal, Performance, and Full scores can then be converted into standardized scores with a mean of 100 and a standard deviation of 15 (Wechsler, 1949, p. 23-24).

Wechsler (1951) provides age norms to convert from raw score to testing age, which is equivalent to mental age. The manual for WISC-R (1973) provides the age norms appropriate for the revision. For both editions, mental age is calculated by taking the mean mental age of the subtests.[7]

The WISC variables in ETP were provided in standardized form. To convert these scores into mental age equivalents, we first use the norms in Wechsler (1951) to convert the standardized scores into raw scores according to chronological age. Because these norms are onto but not one-to-one, we take the floor of the median of the raw score range that corresponds to the single standardized score. We then use the age norms to find the mean mental age using the method described above.

[6]Some individuals in Perry were given subtests in addition to the standard ones either as supplements or substitutions. When they were administered as supplements (i.e., the child has scores for either 11 or 12 subtests instead of the standard 10), the Performance and Verbal scores are prorated by multiplying the sum of scaled scores by $\frac{5}{6}$.

[7]This is the mean mental age. There is also the median mental age, which is the median mental age of the subtests, and the formula mental age, which is calculated given the formula for conventional IQ.

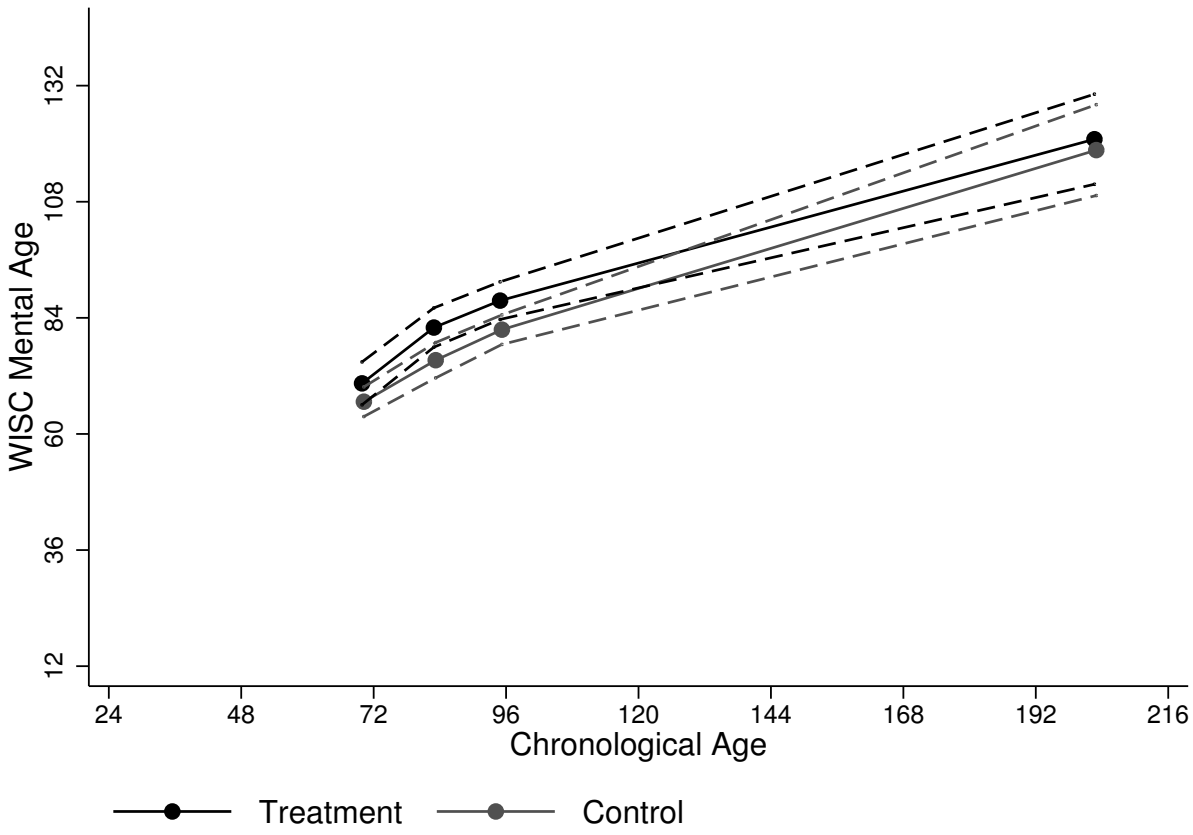


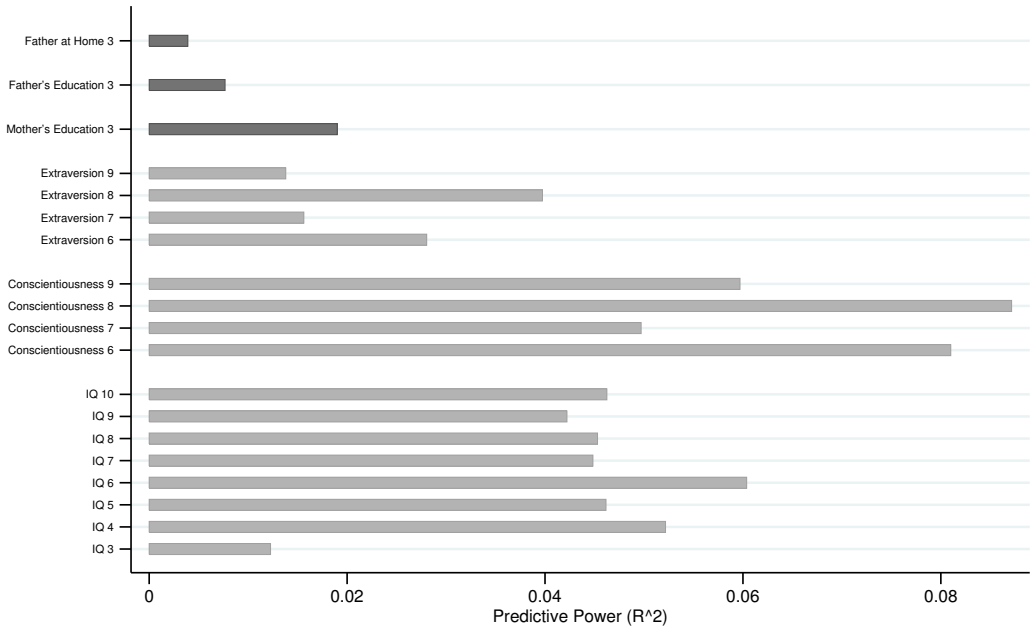
Figure 8: ETP Mental Age

Figure 9: Average WISC scores by treatment status with dashed lines depicting 95% confidence intervals.

D Predictive Power of IQ

This Appendix presents two charts. Both show the predictive power of IQ compared to the predictive power of family characteristics and non-cognitive skills. I include conscientiousness and extraversion as the main non-cognitive measurements, because conscientiousness is known to have an important predictive power, and I include the extraversion is an additional personality measurement for comparison. The first chart presents the predictive power of the age-standardized scores for IQ, and for the raw scores for other skills (non-cognitive skills are typically non-standardized). The second chart presents the predictive power of a ranking measurement of each skill.

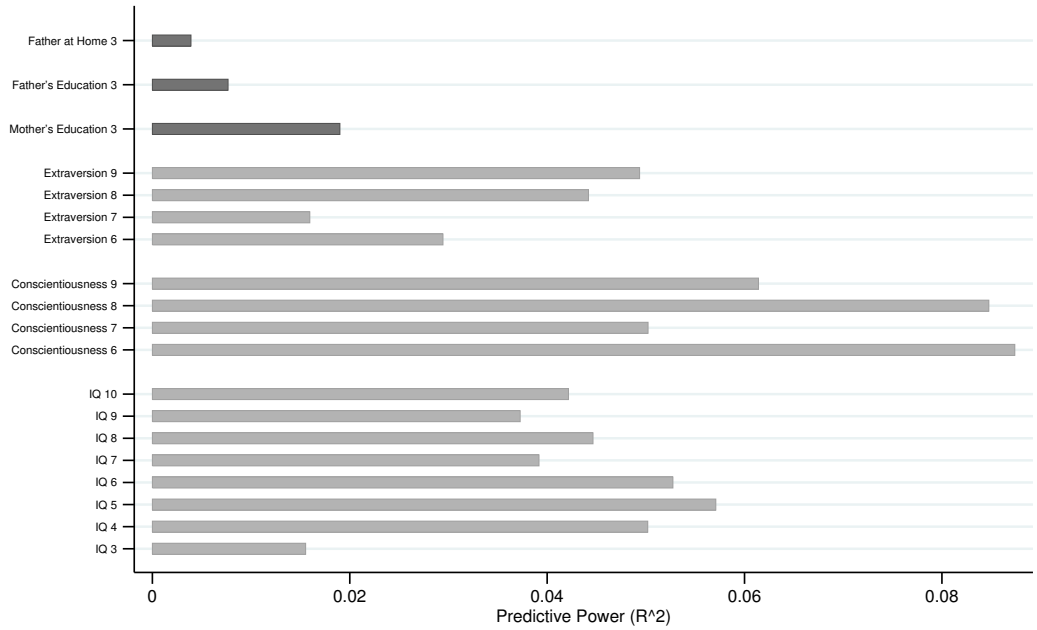
Figure 10: Prediction power of each variable (based on scores)



Note: Horizontal axis shows the average prediction power of each predictor for all adult variables, which are Income at Age 40, Income at Age 27, Years of Education at Age 30, Number of Misdemeanors at Age 40, Number of Felony Arrests at Age 40, Number of Arrests Before Age 27, Number of Arrests After Age 40, Proportion of Year Incarcerated at Age 28, Proportion of Year Incarcerated at Age 40, Graduate from High School at Age 40, Neither Working or Studying at Age 40, Married or In a Serious Relationship at Age 40, Reports Good Health at Age 40, and Parents Before Age 18. The average R^2 of all IQ, conscientiousness, and extraversion predictors (in score) is **0.0485**

.

Figure 11: Prediction power of each variable (based on ranks)



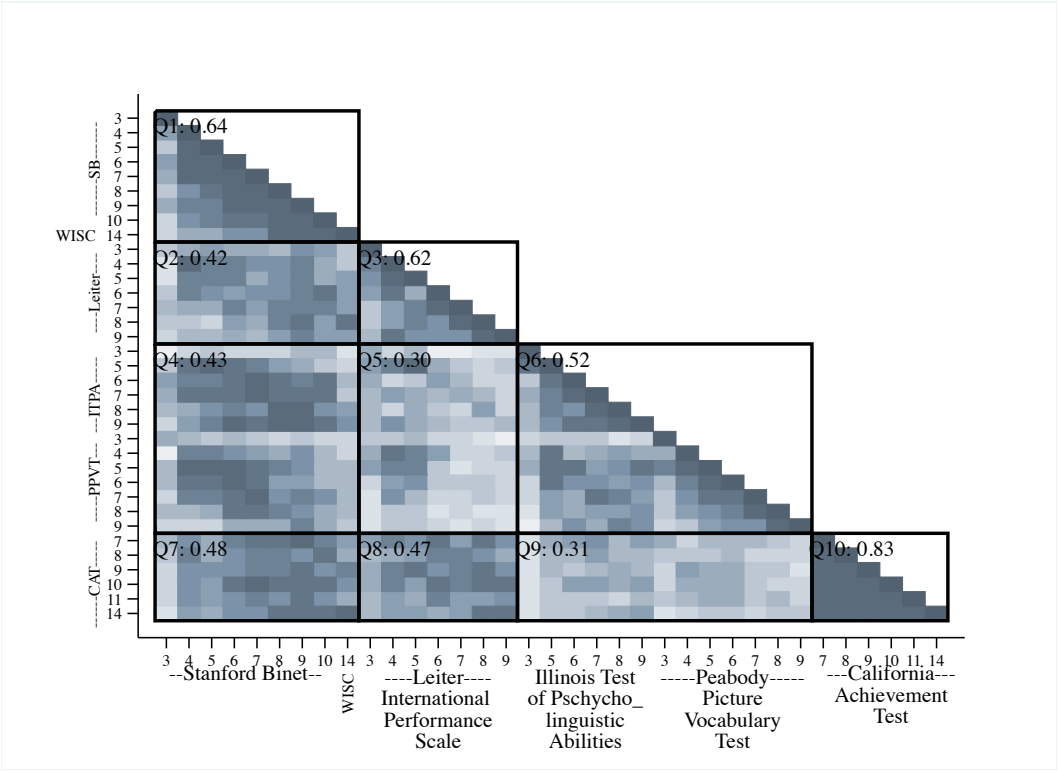
Note: Horizontal axis shows the average prediction power of each predictor for all adult variables, which are Income at Age 40, Income at Age 27, Years of Education at Age 30, Number of Misdemeanors at Age 40, Number of Felony Arrests at Age 40, Number of Arrests Before Age 27, Number of Arrests After Age 40, Proportion of Year Incarcerated at Age 28, Proportion of Year Incarcerated at Age 40, Graduate from High School at Age 40, Neither Working or Studying at Age 40, Married or In a Serious Relationship at Age 40, Reports Good Health at Age 40, and Parents Before Age 18. For measures for IQ, Conscientiousness, and Extraversion, ranks among all subjects are used. The average R^2 of all IQ, conscientiousness, and extraversion predictors (in rank) is **0.0493**

.

E Correlations Between IQ Measures

To graph the degree of correlation among the different instruments that I use, I now present a *Heat Map* of these correlations. I can see from the graph that (i) achievement tests are clearly more correlated with each other than with other tests, forming a separate group, (ii) vocabulary-based IQ tests correlate more with traditional IQ tests than achievement tests, and (iii) correlations between achievement and traditional IQ tests are relatively high.

Figure 12: Convergent Validity of IQ tests: Heat Maps



Notes:
These *Heat Maps* represent the total correlations between every pair of measurements of cognitive ability that I have available for Perry and ABC. The darker the color of the intersection between two measurements, the higher the correlation. The different quadrants illustrate the interactions between specific groups of variables. In each quadrant, I show the average correlation among all the pairs in the quadrant. Notice the Vocabulary IQ tests (PPVT and ITPA) correlate more strongly with regular IQ tests than with achievement tests. I take this fact as support for continuing to use vocabulary-based IQs as part of the IQ series.

F Trajectories of Cognitive Skills Using Ordinal Tests

In this Appendix I document some characteristics of the data using only ordinal properties of the test scores. I present two plots, each using a different type of test (Stanford-Binet and PPVT). Both use Perry data. The plots show the evolution of the cumulative distribution functions of the tests of the treated and the control children across time. I use Mental Age scores, which allows for a clearer interpretation of the findings, and a better visualization of the charts. I only use the initial 5 ages in the data to make the charts more readable (ages 3,4,5,6 and 7). The plots show two facts that are relevant for this paper.

First, the data is compatible with the existence of fadeout in test scores. As shown in Figures 13 and 14, in the first period of the study (baseline) none of the two groups stochastically dominate the other. In the second period, there is stochastic dominance of the treatment group. After that, the distributions cross again (only slightly in the graphs, but cross multiple times at the later ages), so there is no dominance. Moreover, I test for equality of distributions using a version of the Kolmogorov-Smirnov test that is exact in small samples, and I cannot reject the null hypothesis of identical distributions for the test at the baseline period, nor at periods after 72 months of age (84 in the case of SB).

Table 8: p-values of the Kolmogorov-Smirnov Test of Equality of Distributions

Age (months)	36	48	60	72	84	96	108
SB	0.14	0.00	0.00	0.05	0.36	0.85	0.91
PPVT	0.92	0.03	0.00	0.29	0.19	0.64	0.37

Note: The null hypothesis of the test is equality of distributions. This test is exact even in small samples.

Second, from the plots it is also possible to see that the test scores of both groups (treated and control) are increasing throughout the whole period of the study: the distribution of the scores of each group in period t is always dominated by the distribution of that group in period $t+1$. This is suggestive (but weak) evidence against the hypothesis of depreciation discussed in the paper.

To clarify the meaning of those findings in terms of skills, I present a result that implies

that stochastic dominance in test scores imply expected stochastic dominance in skills. Under the most general measurement model for ordinal tests, $M_{it}^m = f^m(\theta_{it}, \varepsilon_{it}^m)$, the interpretation of stochastic dominance in tests is not obvious. Thus, I take a simpler model from Barlevy and Neal (2012):

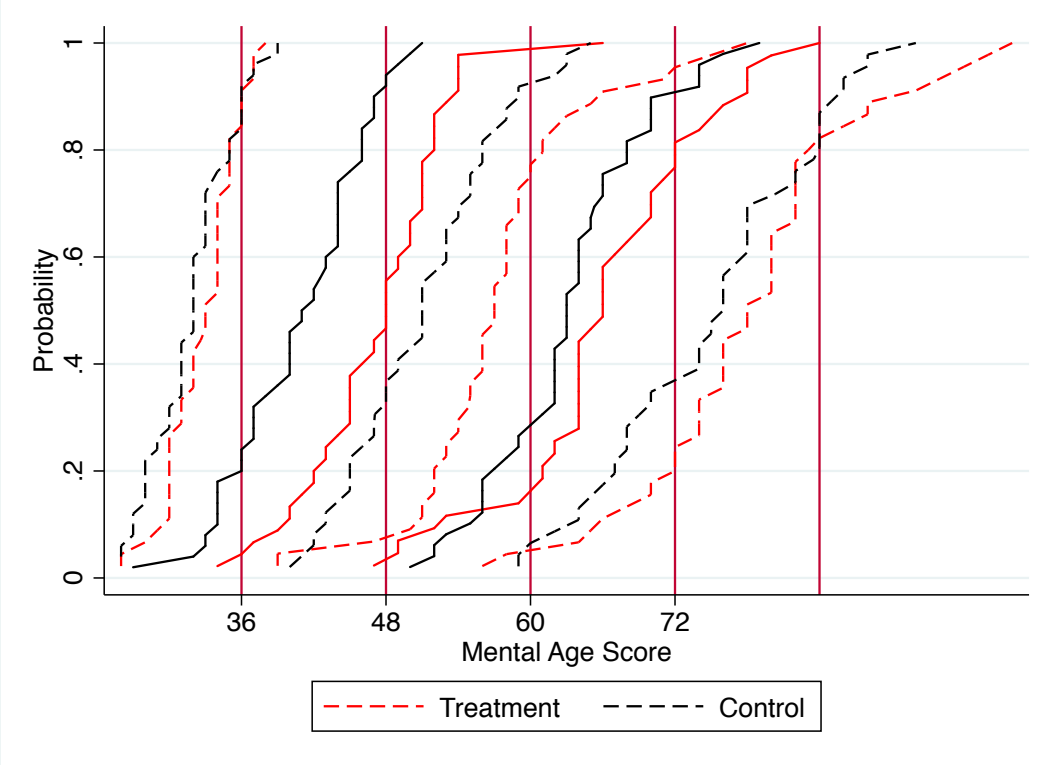
$$M_{it}^m = g^m(\theta_{it} + \varepsilon_{it}^m)$$

Under this model, it is possible to show that stochastic dominance in tests implies a form of expected stochastic dominance in skills:

$$\begin{aligned} F_{M_t|R=1}(M_{it}) &< F_{M_t|R=0}(M_{it}) \quad \forall M_{it} \in supp(M_t) \implies \\ E\left[F_{\theta_t|R=1}(\theta_{it})\right] &< E\left[F_{\theta_t|R=0}(\theta_{it})\right] \quad \forall \theta_{it} \in supp(\theta_t) \end{aligned}$$

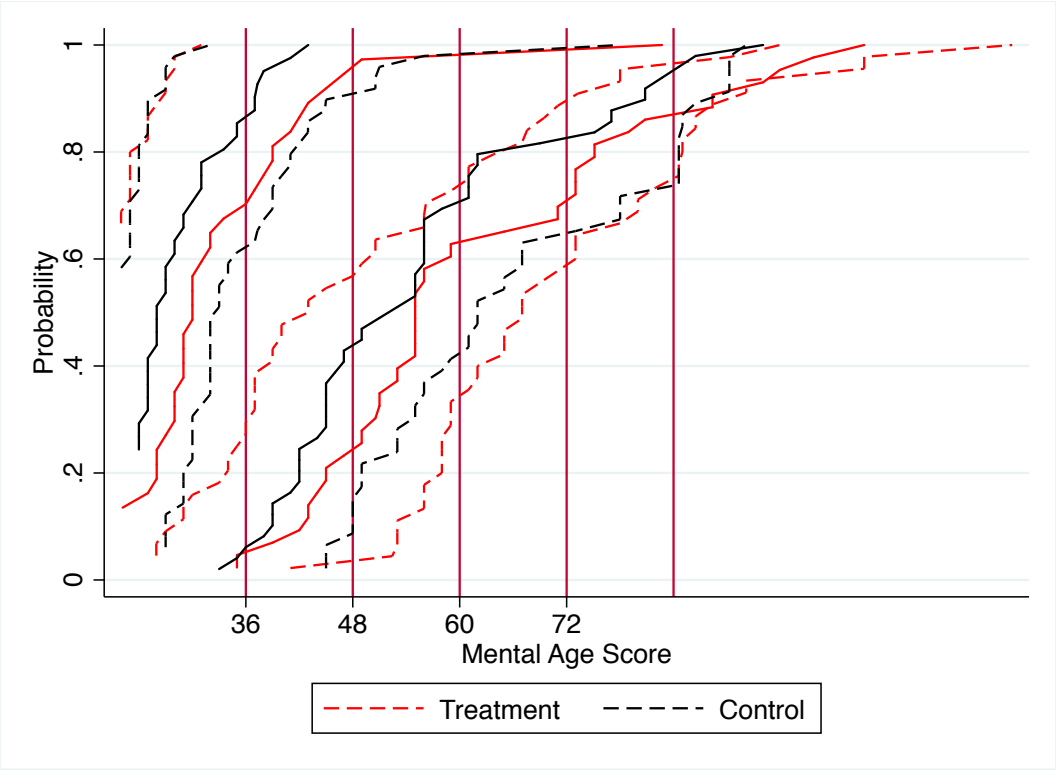
Where the expectations are taken with respect to ε_{it}^m.

Figure 13: Ordinal Comparison: Evolution of CDFs for Stanford-Binet Test



Note: This plot presents, for four different ages, the cumulative distribution function of the Mental Age scores of the Stanford Binet test. For ease of reading, every other year has dotted lines. The distribution of the treatment group is shifted to the right of the control group for all ages after the baseline. The lines represent the ages at which the tests are taken and the expected value of the test for that age.

Figure 14: Ordinal Comparison: Evolution of CDFs for PPVT Test



Note: This plot presents, for four different ages, the cumulative distribution function of the Mental Age scores of the Stanford Binet test. For ease of reading, every other year has dotted lines. The distribution of the treatment group is shifted to the right of the control group for all ages after the baseline. The lines represent the ages at which the tests are taken and the expected value of the test for that age.

G Trajectories of Cognitive Test Scores

In this appendix, I document the trajectories of the test scores of the treated and the control groups. I first document them in Perry, then in ETP and then I present the few scores available in IHDP. For all of these programs, I document the test scores separately for each test available in the data, to make the trajectories as interpretable as possible.

In each program section, I present charts for multiple transformations of the tests: (i) Raw Scores; (ii) Mental Age Scores; (iii) Percentile Ranks Calculated in the Whole Sample; (iv) Age-Standardized Scores; (v) Percentile Ranks Calculated Within an Age; (vi) anchored scores (only for Perry).[8] For each program and Transformation of the tests, I use all tests available in the data (Stanford Binet, PPVT, Leiter and ITPA in Perry; SB, PPVT, ITPA and WISC in ETP; and PPVT in IHDP).

In this section I use 6 transformations of the tests:

1. **Raw Scores** are the simple sum of correct answers in each test. Let the raw scores for a given test taken at age t be be W_{it}

2. **Mental Age Scores** can be calculated in one of two ways: (i) as a simple weighted sum of correct answers that test designers constructed in a way such that the expected score for a child of age t is roughly t; or (ii) as the expected age in a nationally representative sample for a child obtaining a value w on her raw score. In this case, the following function is estimated by the test designers and reported fr each score in the test manual: $g^{MA}(w) = E[t|W = w]$, where the expectation is taken over a nationally representative sample. Then, $MA_{it} = g^{MA}(W_{it})$.

3. **Percentiles Ranks Calculated in the Whole Sample:** are constructed by ranking all children at all ages that answer a specific test (e.g. PPVT) in the sample, and then assigning them numbers from 1 to 100 depending on their ordering in the sample. Let

[8] I do not report anchored scores for ETP because it has no useful adult outcomes and the sample is too small to estimate the relation between the tests and the outcomes reliably. Given that the scores are taken at different ages than in Perry, it is not possible to directly extrapolate the coefficients from Perry to ETP.

F_W be the Cumulative distribution function for the raw scores in our sample. Then, $PCTS_{it} = 100F_W(W_{it})$.

4. **Age-Standardized Scores** are constructed using nationally representative samples to calculate the mean and the standard deviation of a given test score at a given age, t. Then, $STD_{it} = \frac{W_{it}-E[W_t]}{SD[W_t]}$.

5. **Percentiles Ranks Calculated Within an Age:** all children that answered a specific test at a given age (e.g. PPVT) in the sample are ranked, and numbers are assigned from 1 to 100 depending on their ordering in the sample. Let F_{W_t} (notice the double subindex) be the Cumulative distribution function for the raw scores in our sample at age t. Then, $PCTA_{it} = 100F_{W_t}(W_{it})$.

6. **Anchored Tests** are calculated using the relationship between a specific score and an outcome of interest. For this paper, I use Years of Education as that outcome. I construct the anchored measurements following Cunha and Heckman (2008), but unlike them, I assume the outcome depends on a single skill. To anchor the measurement M_{it} to outcome Y_{iT}, I estimate a two-stage least squares regression of Y_{iT} on M_{it}, using alternative measurements of the skill in the same period as instruments to avoid underestimation because of measurement error. Then, I construct the anchored measurements as $ANCH_{it} = \widehat{Y_{iT}}$.[9]

[9] I assume the following relationship of the skills at age t with the outcome:

$$Y_{iT} = \mu_t + \gamma_t \theta_{it} + \nu_{it} \tag{10}$$

And I assume the measurement model:

$$M_{it}^m = a^m + b^m \theta_{it} + \varepsilon_{it}^m \tag{11}$$

Replacing the skill measured by skill 1 in the outcomes equation, we obtain:

$$Y_{iT} = a^* + \frac{\gamma_t}{b^1} M_{it}^1 - \frac{\varepsilon_{it}}{b^1} + \nu_{it} \tag{12}$$

Where $a^* = \mu_t - \frac{a^1}{b^1}$. The coefficient $\frac{\gamma_t}{b^1}$ cannot be estimated directly, because the error ε_{it}^1 is correlated with M_{it}^1. To address this problem, I assume that for other measurements of the same skills in the same period, M_{it}^2, their measurement errors, ε_{it}^2, are independent from ε_{it}^1: $\varepsilon_{it}^1 \perp\!\!\!\perp \varepsilon_{it}^2$.If that is true, those measurements will be valid instruments, and it will be possible to estimate $\frac{\gamma_t}{b^1}$ consistently. Using the estimated parameters

G.1 Discussion of the Findings

G.1.1 Discussion for Perry

I start documenting the trajectories for Perry in Figures 33-20. Given that we are interested in exploring the timing of the changes in the scores, I do not include Wave 1 (24 individuals) in the graphs, because they entered the program at a different age, so the timing of the impacts is different. For all other cohorts, children were tested around age 3 (baseline) using SB, then about two months later they were tested in Leiter and ITPA. At ages 4 and 5 they were tested again, when the treated children were still in preschool. All measurements at ages 6 and later are during elementary school.

Mental Age scores, Raw Scores and Percentiles Calculated in the Whole Sample have very similar trajectories in Perry. For all four types of tests available (SB, PPVT, Leiter and ITPA), they are all strongly increasing with age. In all cases a gap opens after the baseline period, and it persists at least up to age 5, but sometimes as far as up to age 8 (PPVT). For Leiter, a small gap seems to open up at age 9 in some cases.

Standardized scores have strong increases at age 4 for the three tests with available data. However, the scores of the treatment group are harder to interpret after that, as they seem to fall for SB and ITPA, increase for PPVT, and stay constant for Leiter. The relative scores of the control group during preschool age are increasing for SB and Leiter, decreasing for ITPA and stay constant for PPVT. Interestingly, for all tests the relative scores of the control group go up in the school entry year (60-72 months in the charts). This is consistent with the main hypothesis in this paper that schools are especially important for more disadvantaged children. While I explore this hypothesis it in the main paper comparing treatment and control children, these trajectories are suggestive that the same can be true for the comparison between control children and the national population. After school entry, the patterns are again unclear: scores seem to increase for PPVT and Leiter, decrease for ITPA and stay constant for SB. Percentile Scores Calculated Within an Age follow a similar trend

from this equation, I construct the anchored measurements as $ANCH_{it} = \widehat{a^*} + \frac{\widehat{\gamma_t}}{\widehat{b^1}} M_{it}^1$

of opening and closing gaps, but in this case the levels do not give any relevant information.

While the first three transformations analyzed can be clearly categorized as measures of absolute advantage, and the second two tests can clearly be categorized as measures of relative advantage, it is less clear how to categorize the anchoring measurements. In these charts, the measurements are anchored separately in each period, as in Cunha et al. (2010), rather than in a single period, as in Cunha and Heckman (2008). Both procedures are valid, but their interpretations are quite different. If the measurements of a given test at all ages (e.g. PPVT raw score) were anchored to a single price, estimated in a single period, the trajectories of the chart would look just like the trajectories for the original measurements, but displaced to the units of the outcomes. On the contrary, by anchoring in each period, the anchored scores behave more like a measurement of relative advantage. This is partly due to the fact that the anchoring regressions are run within-sample.[10] The anchored measurements in the data are consistent with the pattern of strong gaps opening 2-4 years later.

G.1.2 Discussion for ETP

I now document the trajectories for ETP in Figures 35-25. I only document them for the group that counts with two years of treatment for conciseness. For this groups, children were tested around age 4 (baseline) then about two months later, after they had participated in a first summer school. At age 5 they were tested again, before and after the second summer school. They were tested again right before entering school (at age 6). Then, they were tested again in the next two years, and around ages 10 and 17.

The trajectories of Mental Age Scores, Raw Scores and Percentiles Calculated in the Whole Sample are relatively similar to the ones in Perry. For most tests, a gap opens up between two months and a year of treatment. The gaps generally remain open during 2-4 years. All trajectories are increasing.

[10] A way to illustrate this is that if regressions were used for anchoring, by construction the average of the anchored measurements, which in the univariate case coincides with the forecasted outcome, would be the average of the outcome in the sample.

The trajectories of the standardized scores are again interesting. There are very marked increases in the level of the treatment children between the beginning and the end of the summer schools (3 months). This is true for most summer schools, but especially for the first time they enter. This is suggestive evidence of diminishing returns to education, as mentioned in the paper. Again, the trends during preschool age of the control children are not especially clear. Importantly, for both of the tests available there is a clear and strong increase at school entry (75-85 months in the charts) in the scores of the control children.

G.1.3 Discussion for IHDP

Unfortunately, there are few tests in IHDP that are comparable across time. PPVT raw scores are only available for ages 3 and 5, and PPVT standardized scores are available for ages 3, 5, 8 and 18. In PPVT standardized scores (Figure 37) it is possible to appreciate that very sizable gap between the treatment and the control group is already open at age 3, narrows down at age 5 and closes at age 8. Remarkably, the scores of the control children once again increase at school entry.

G.1.4 Gender Differences

The general trends are very similar for males and females in the data. We still see (i) a pattern of increasing scores for absolute tests; (ii) no clear trends for relative scores; (iii) gaps opening as soon as the treated children enter the education programs; (iv) gaps lasting for 2-4 years; (v) increases in the relative scores of the control group at the school entry ages in all cases. There are no clear differences in the long-term trends between males and females.

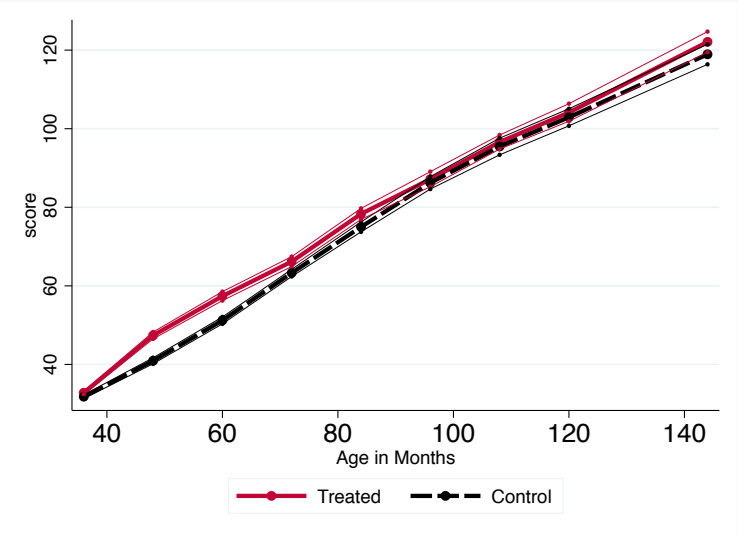
The main difference between males and females is that females appear to have a slightly stronger pattern of impacts. This is hard to see in the raw trajectories for absolute scores, but it can be noticed in relative scores. In particular, out of 9 trajectories of relative scores analyzed, females have stronger impacts or longer-lasting impacts in 7 of them (2/4 in Perry,

4/4 in ETP, 1/1 in IHDP). It is not clear from the data if the reason is larger initial effects or less fadeout for females.

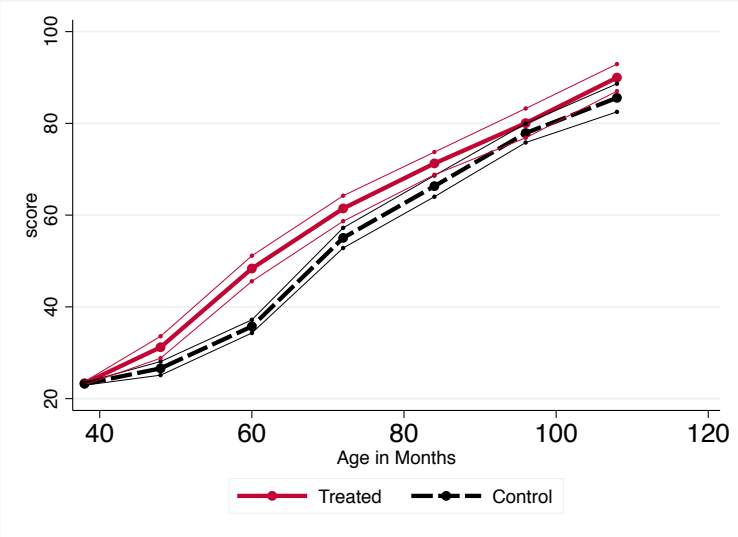
G.2 Trajectories of Cognitive Test Scores for the Pooled Samples

Figure 15: Perry, Mental age scores

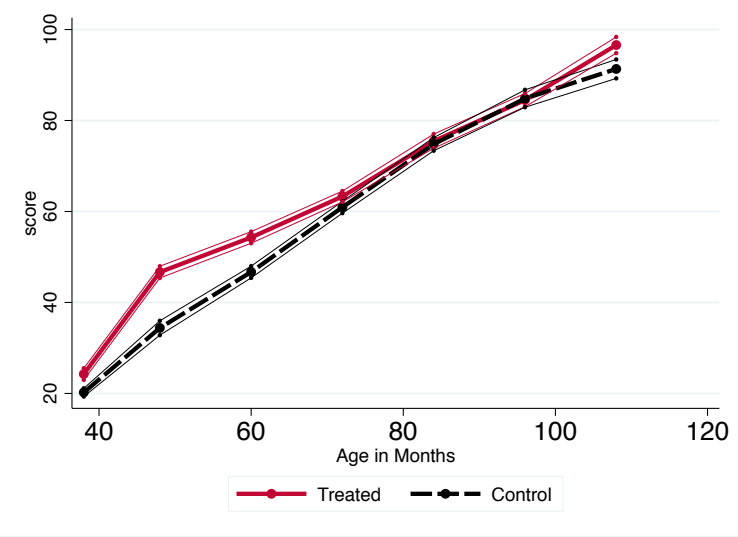
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

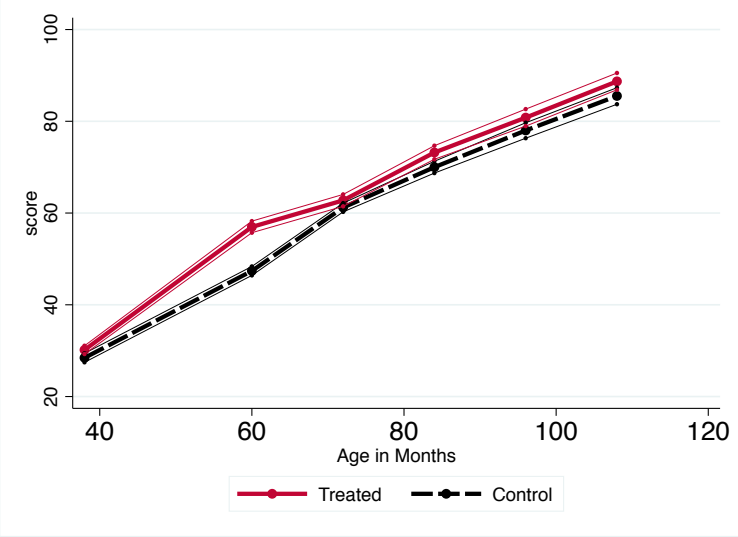
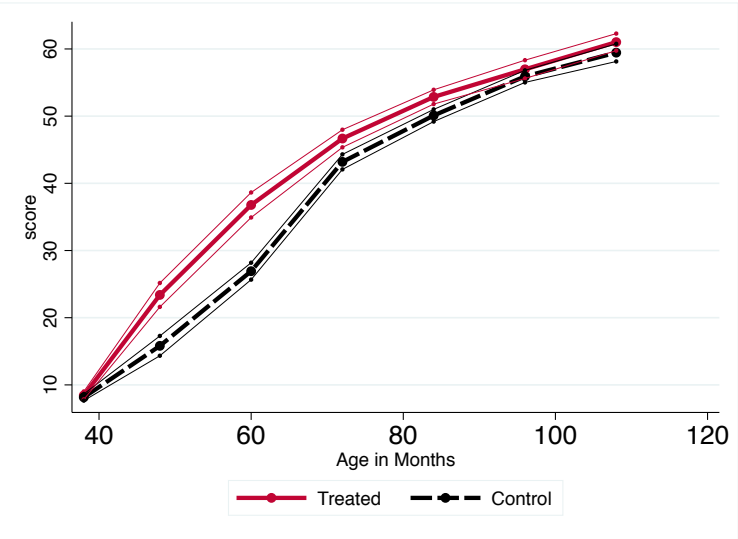
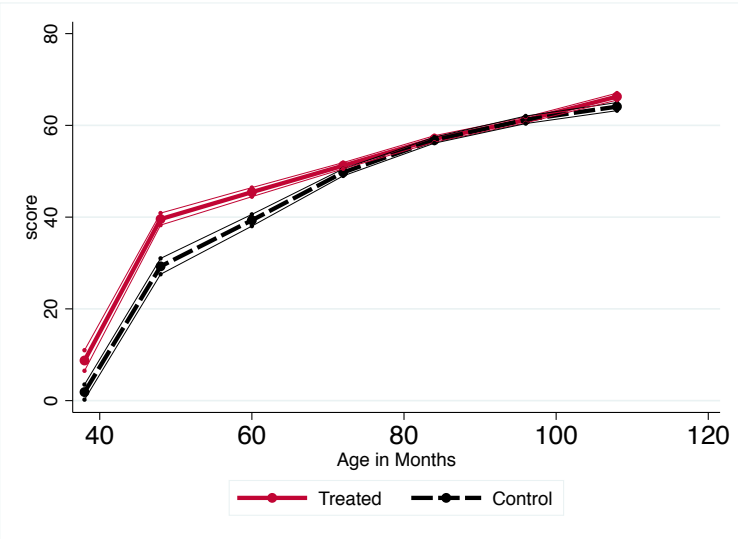


Figure 16: Perry, Raw scores

(a) PPVT



(b) Leiter



(c) ITPA

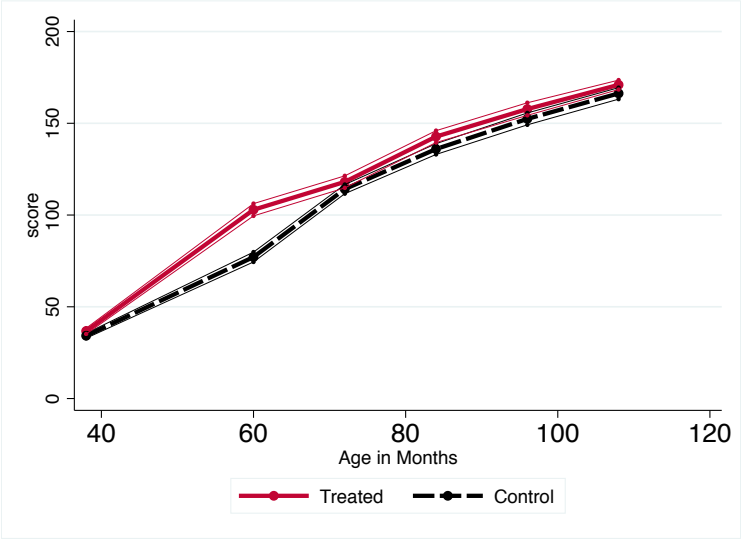
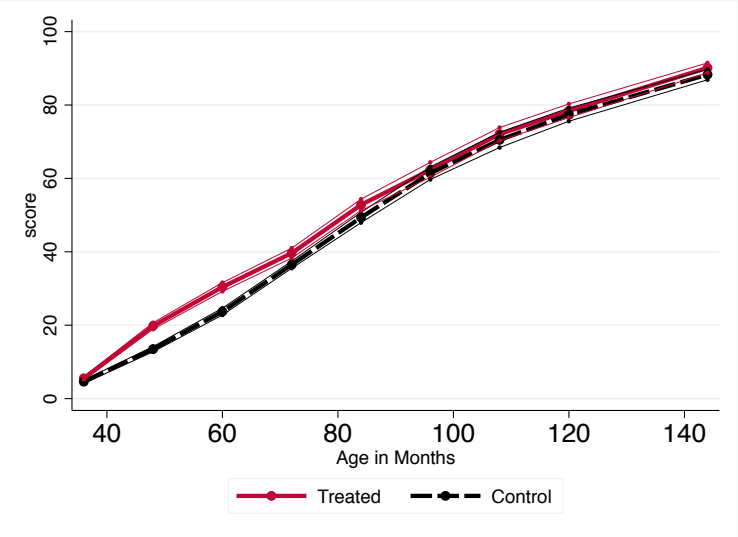
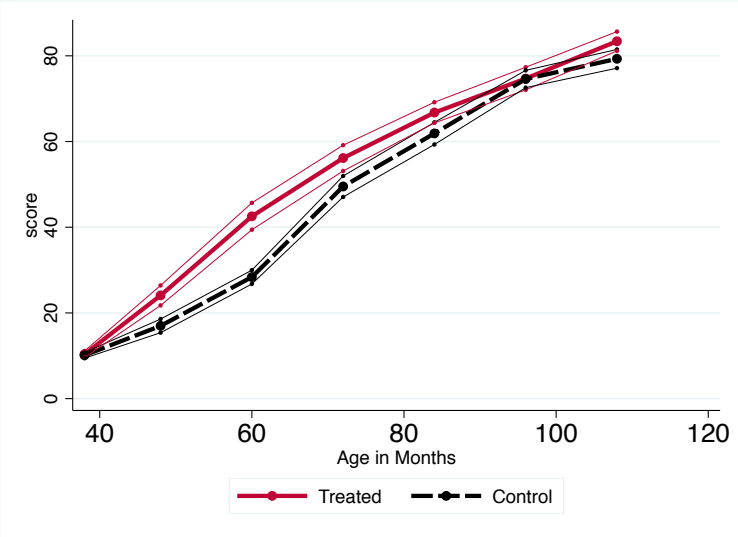


Figure 17: Perry, Percentiles calculated in the whole sample across all ages

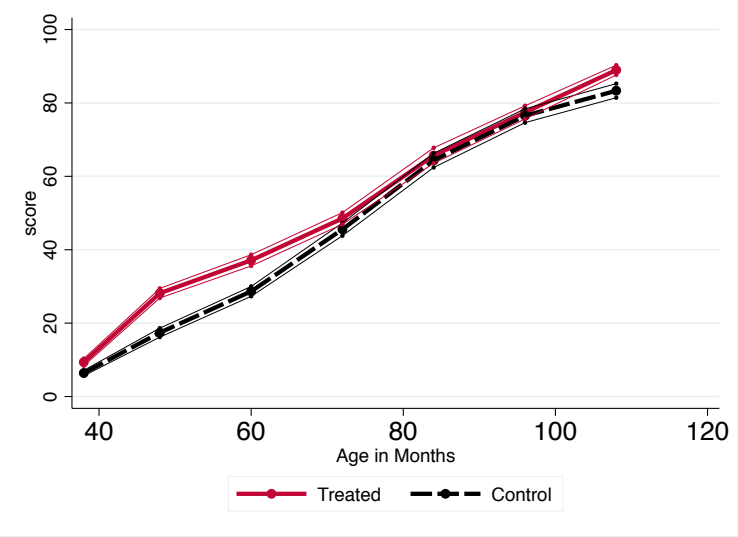
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

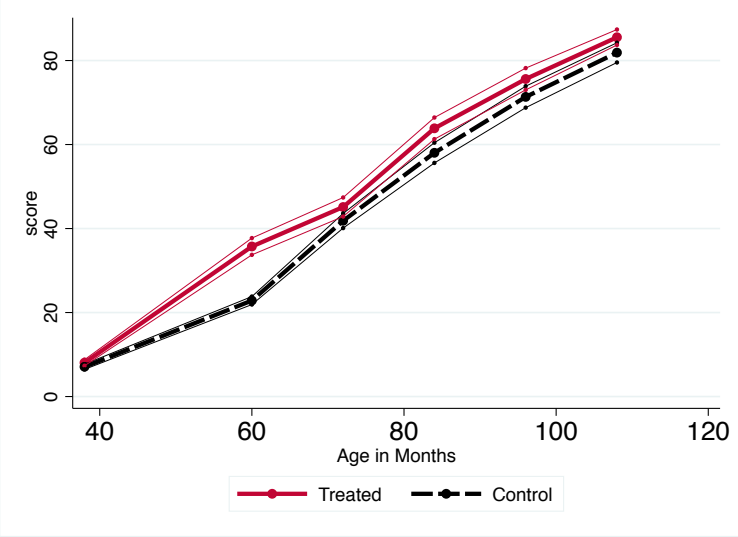
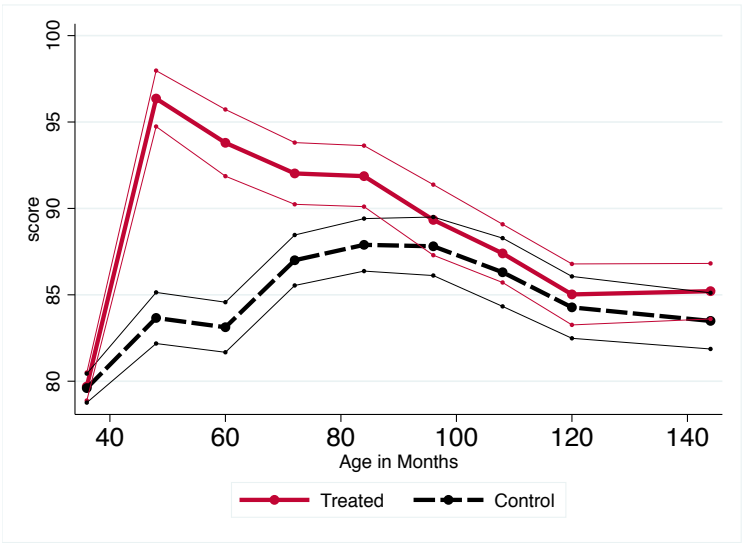
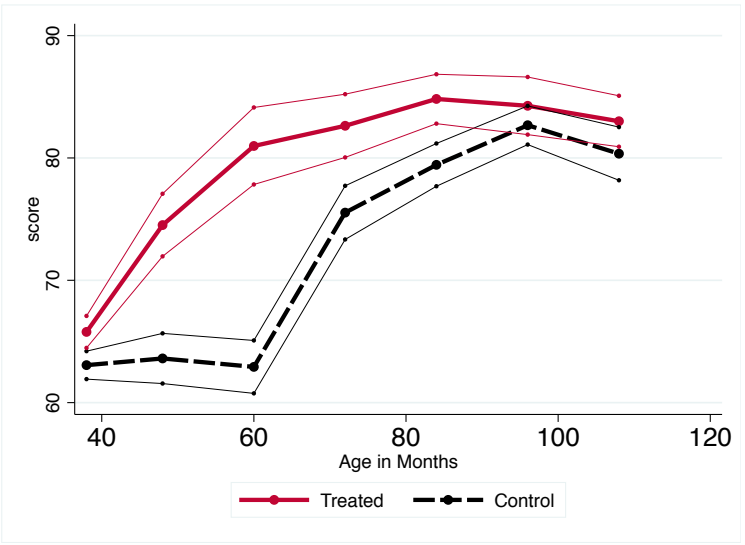


Figure 18: Perry, Standardized scores

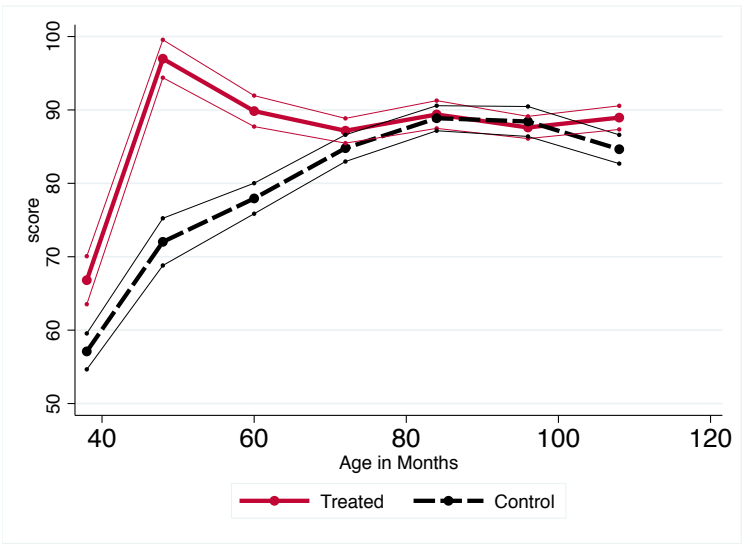
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

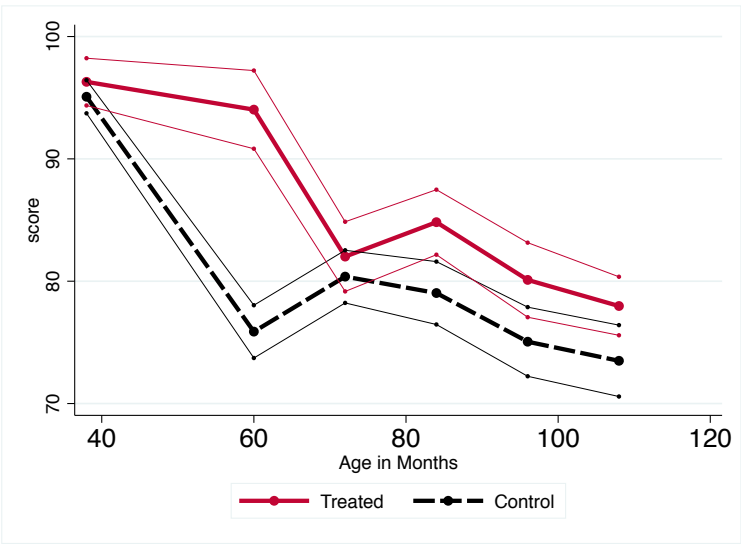
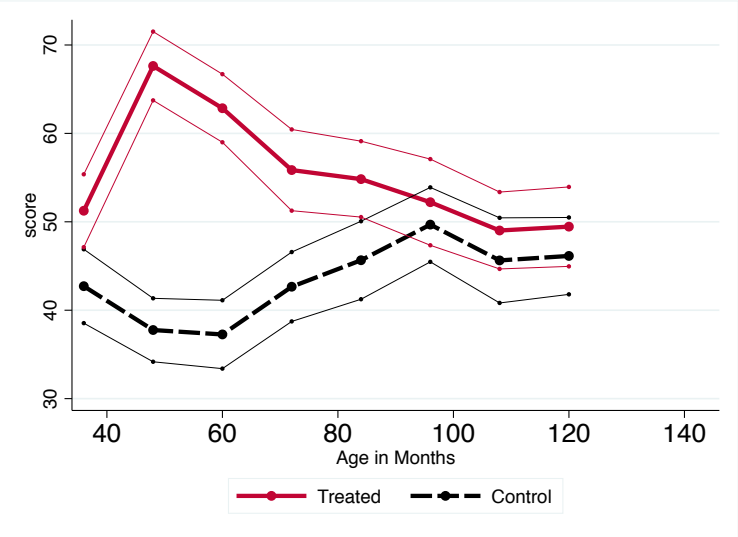
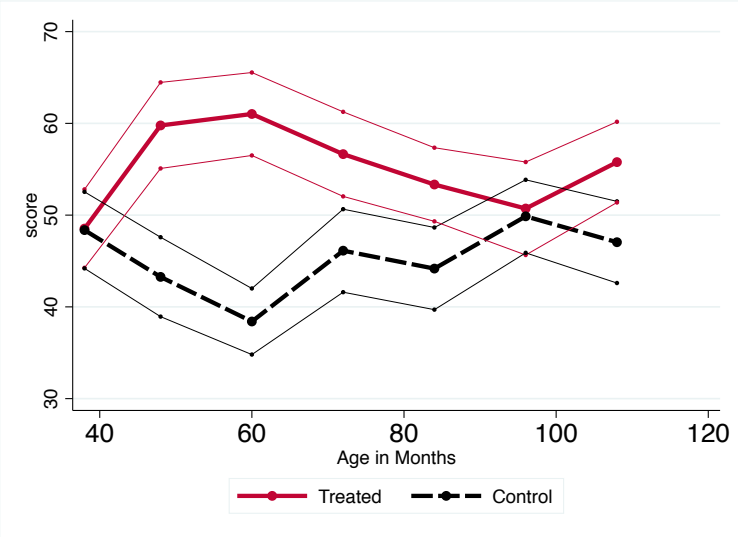


Figure 19: Perry, Percentiles in the sample across the children of a specific age

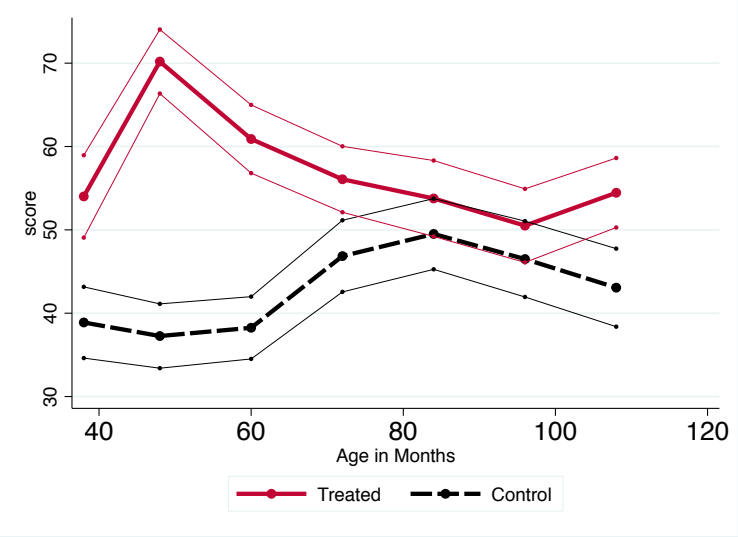
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

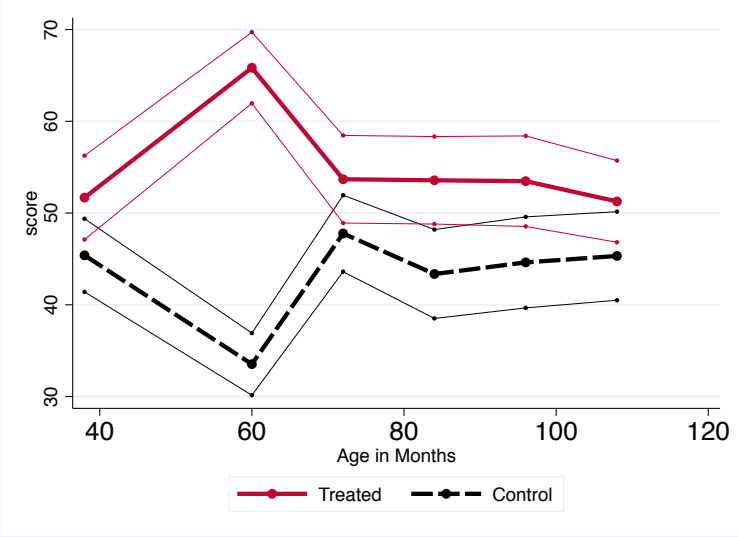
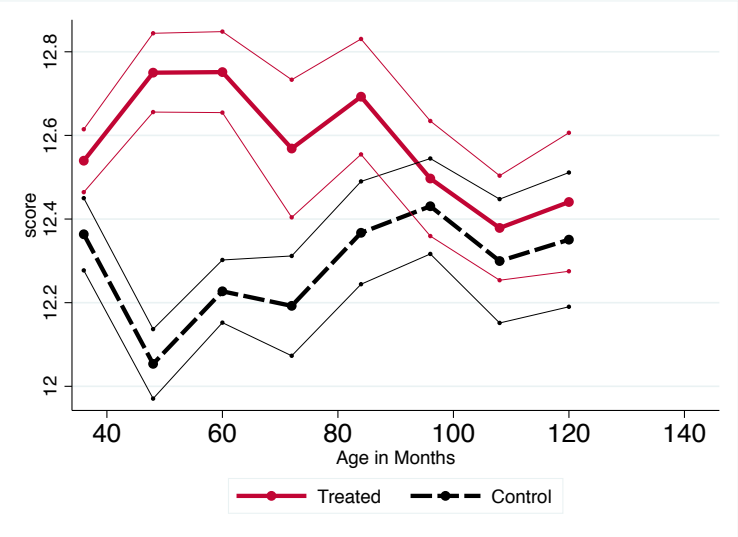
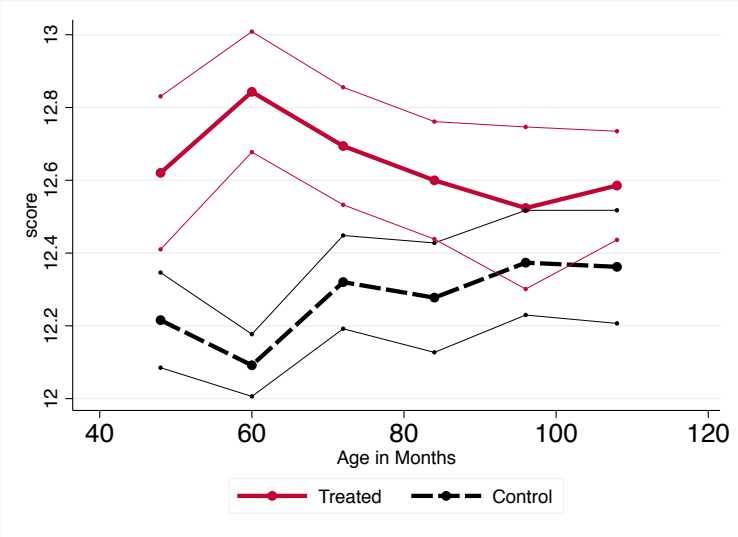


Figure 20: Perry, Anchored scores

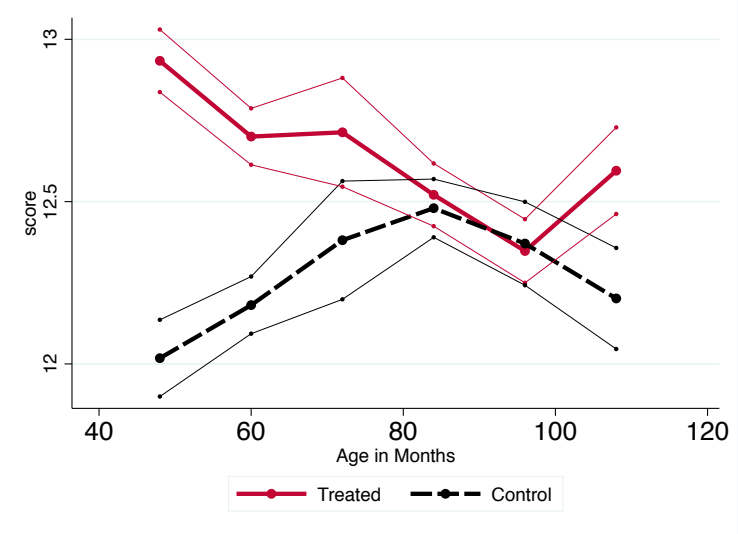
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

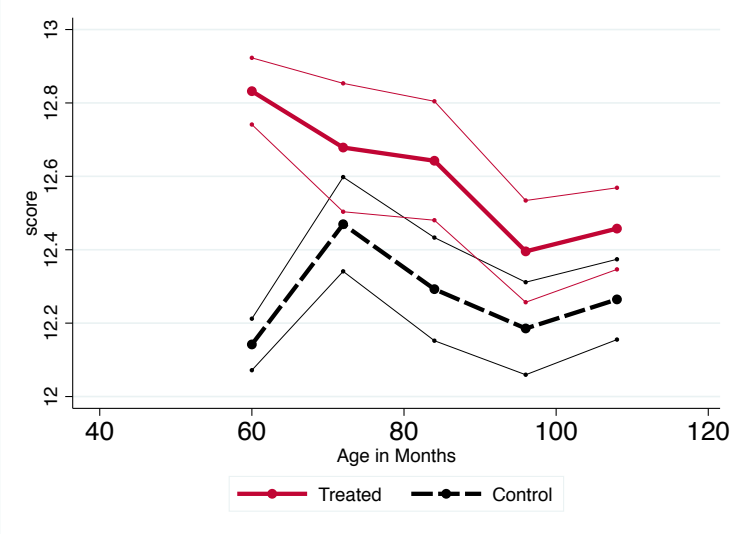
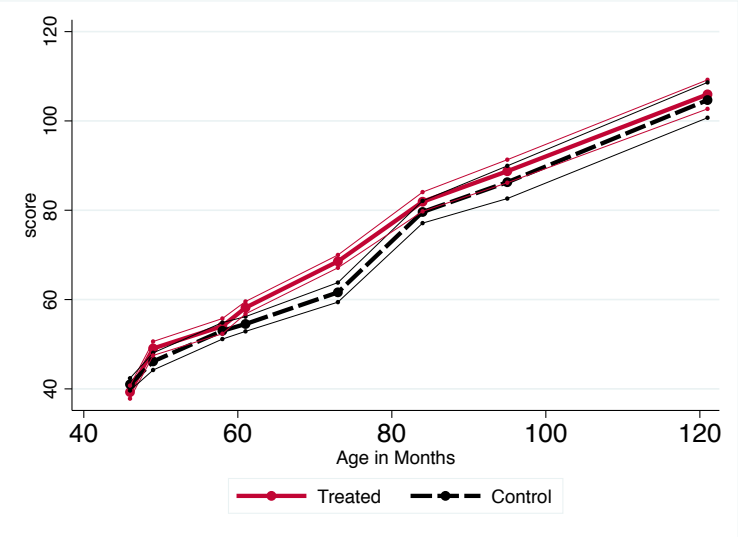
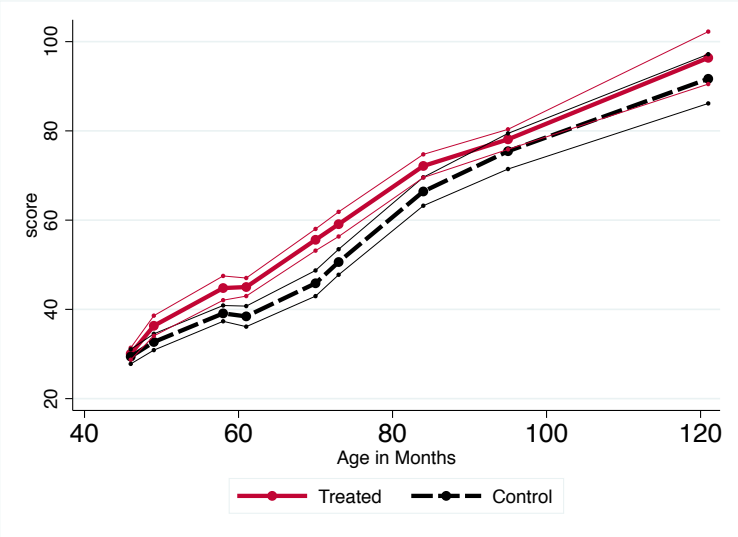


Figure 21: ETP, Mental age scores

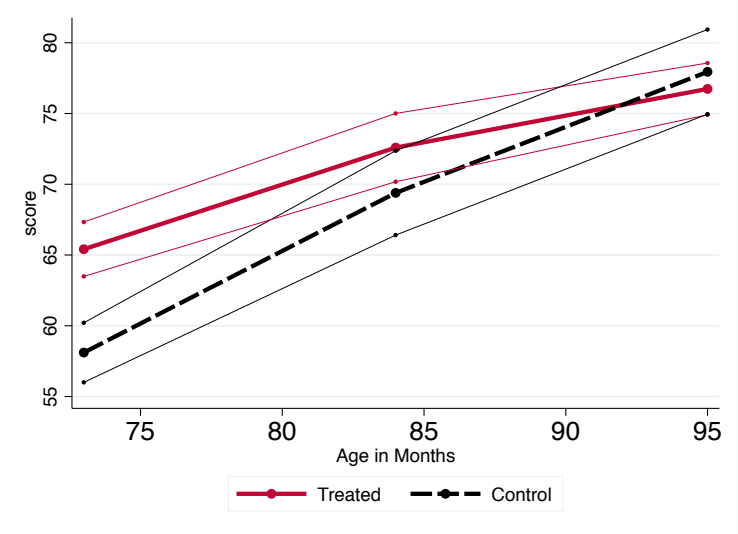
(a) Stanford Binet



(b) PPVT



(c) ITPA



(d) WISC

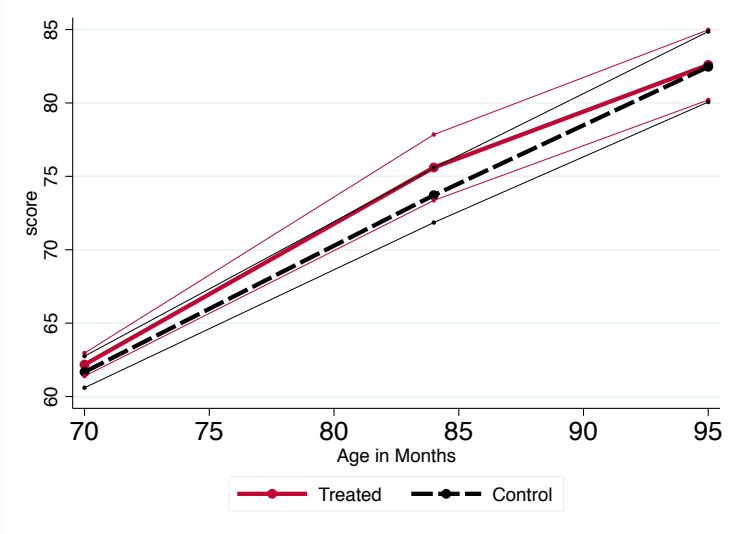
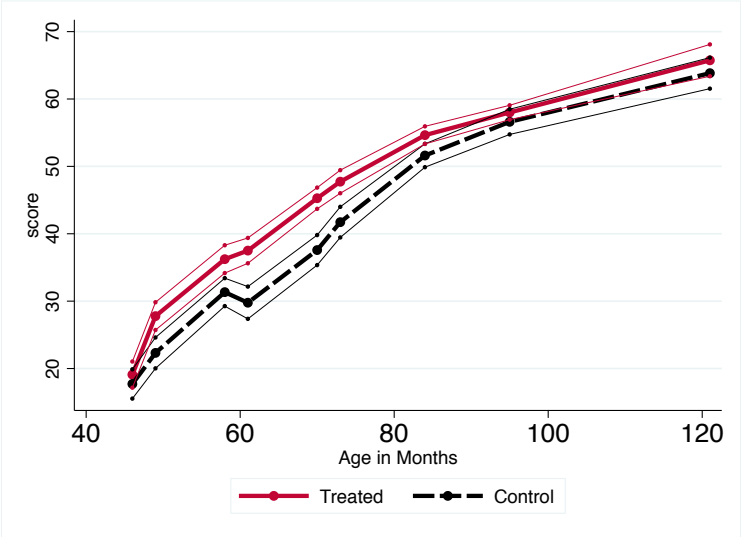


Figure 22: ETP, Raw scores

(a) PPVT



(b) ITPA

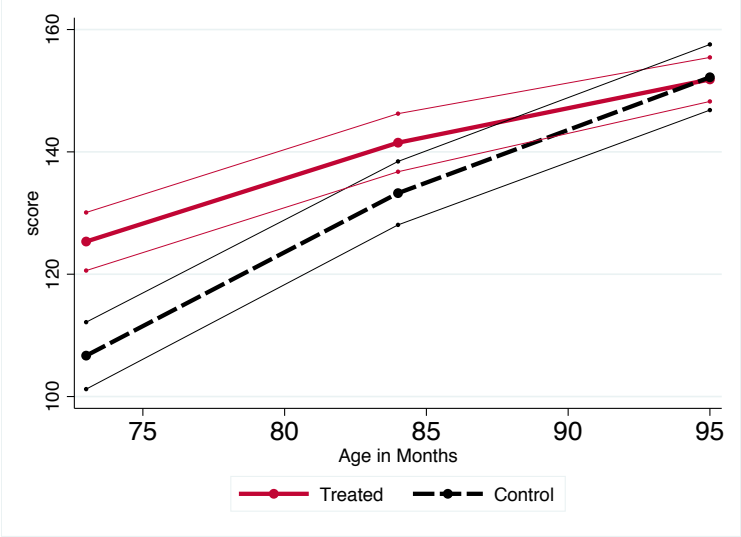
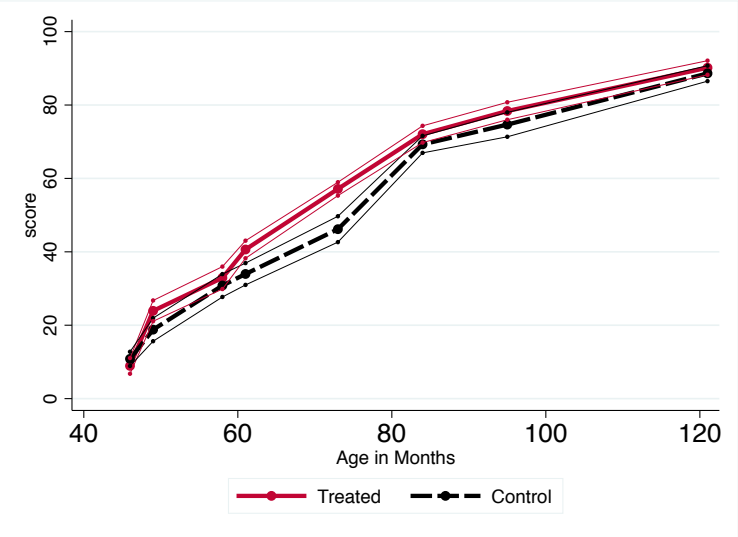
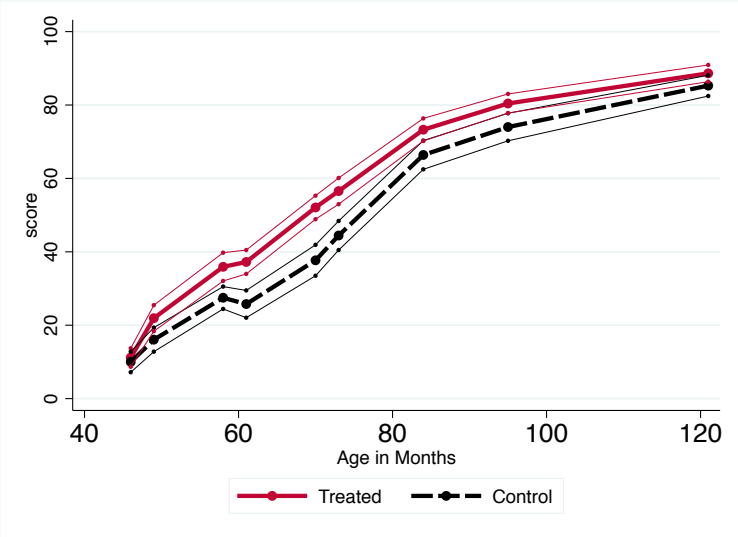


Figure 23: ETP, Percentiles calculated in the whole sample across all ages

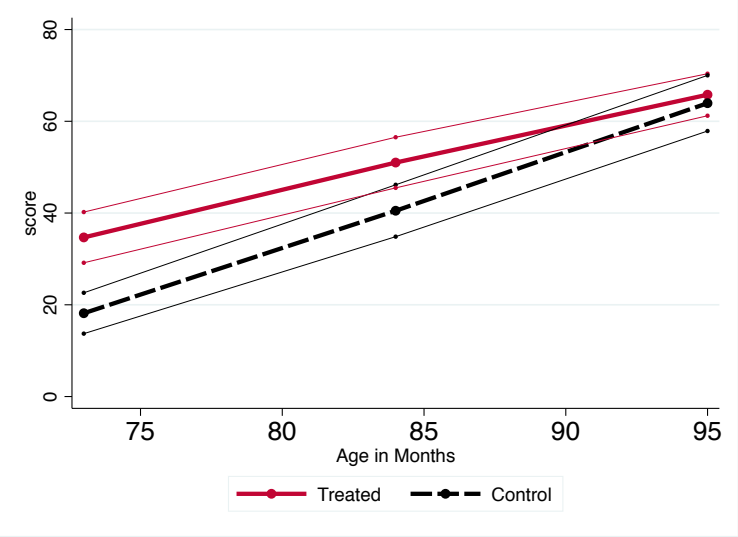
(a) Stanford Binet



(b) PPVT



(c) ITPA



(d) WISC

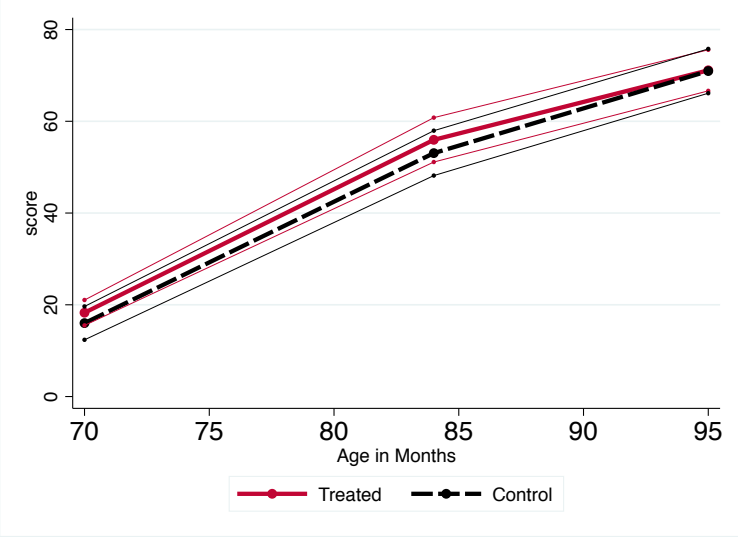
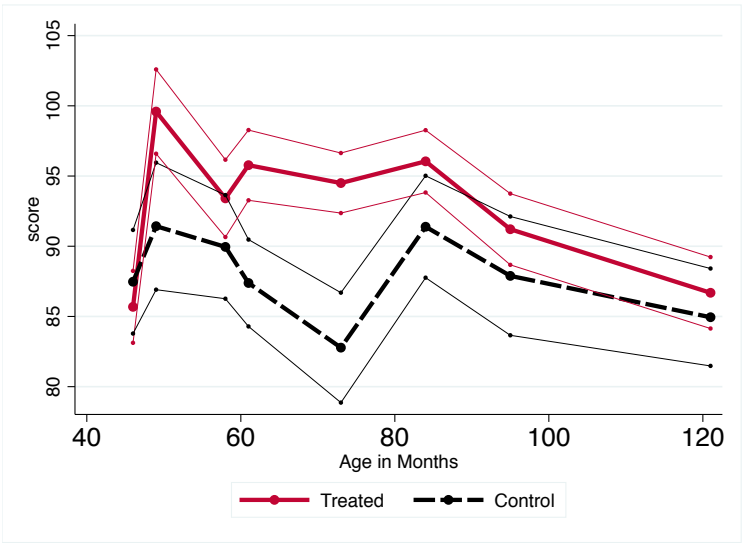
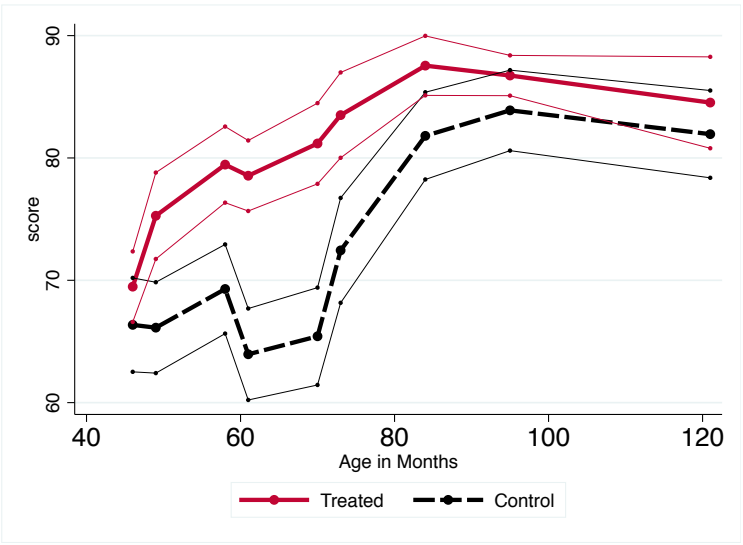


Figure 24: ETP, Standardized scores

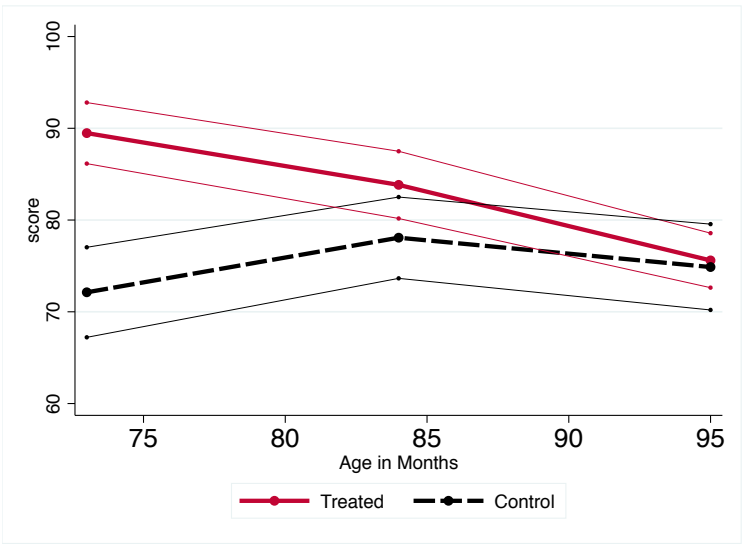
(a) Stanford Binet



(b) PPVT



(c) ITPA



(d) WISC

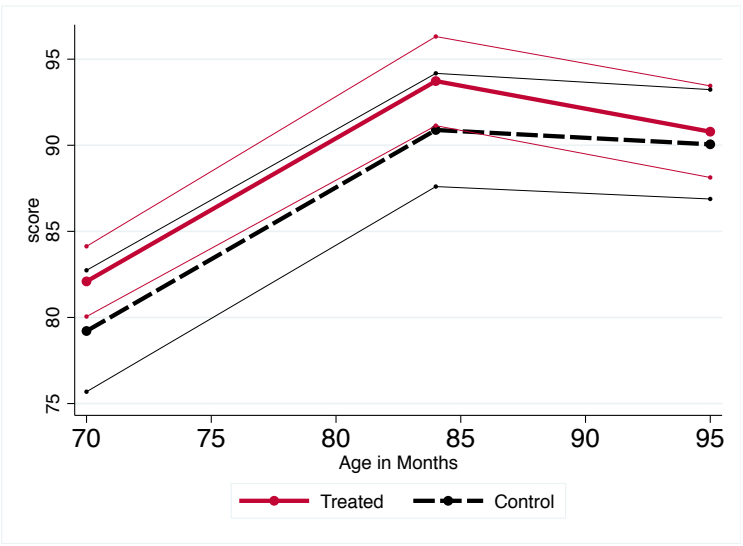
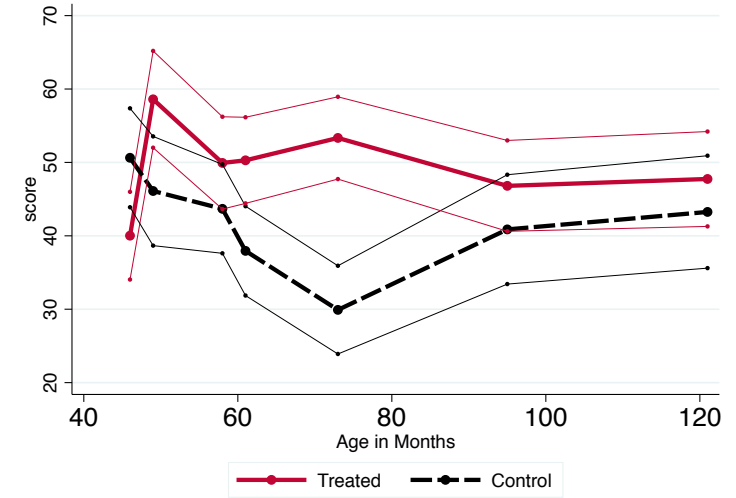
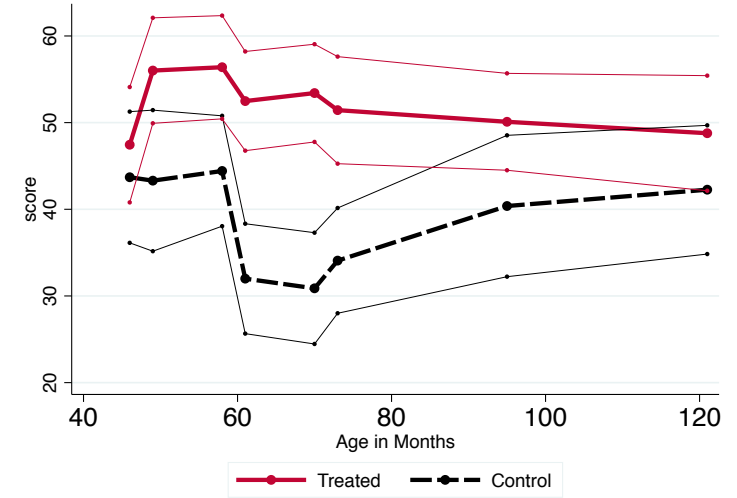


Figure 25: ETP, Percentiles in the sample across the children of a specific age

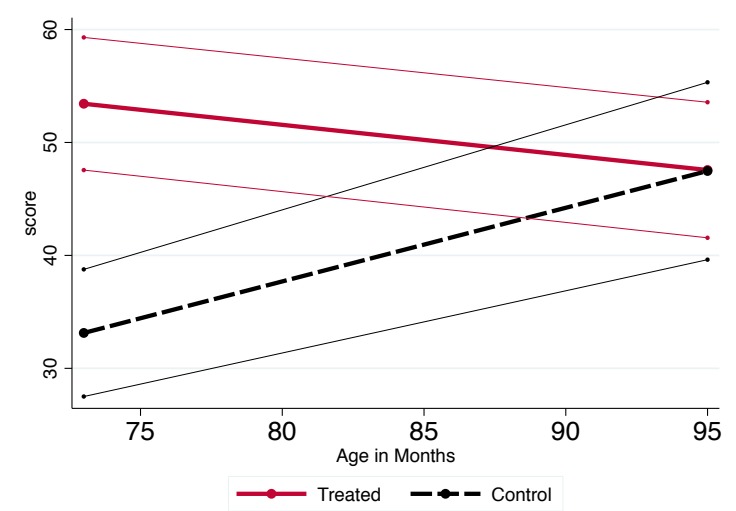
(a) Stanford Binet



(b) PPVT



(c) ITPA



(d) WISC

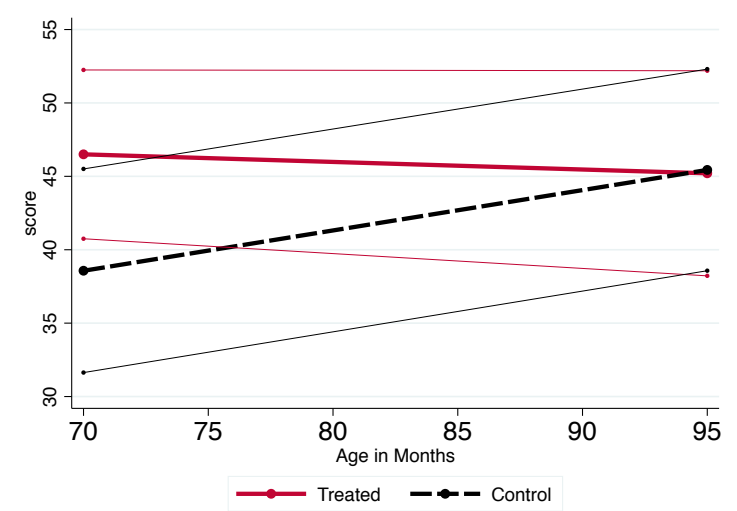


Figure 26: IHDP, Raw scores

(a) PPVT

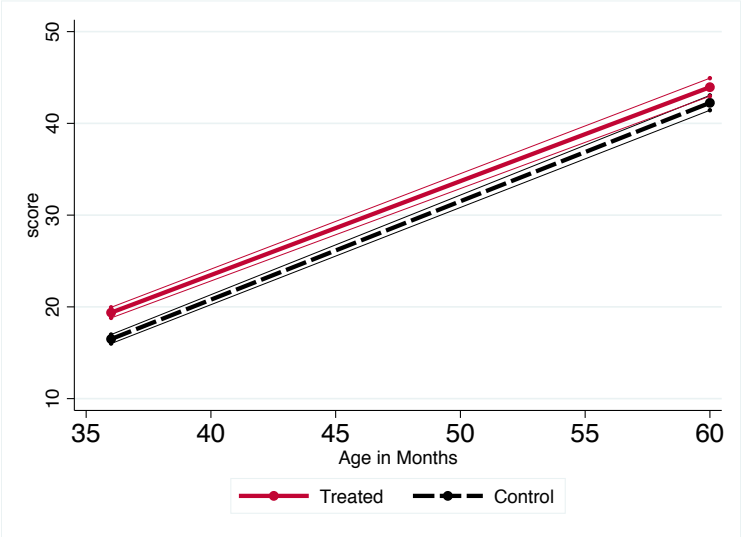
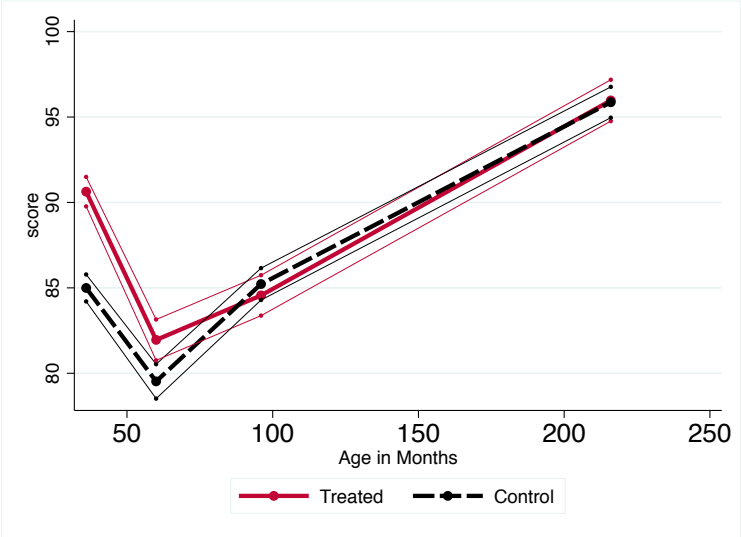


Figure 27: IHDP, Standardized scores

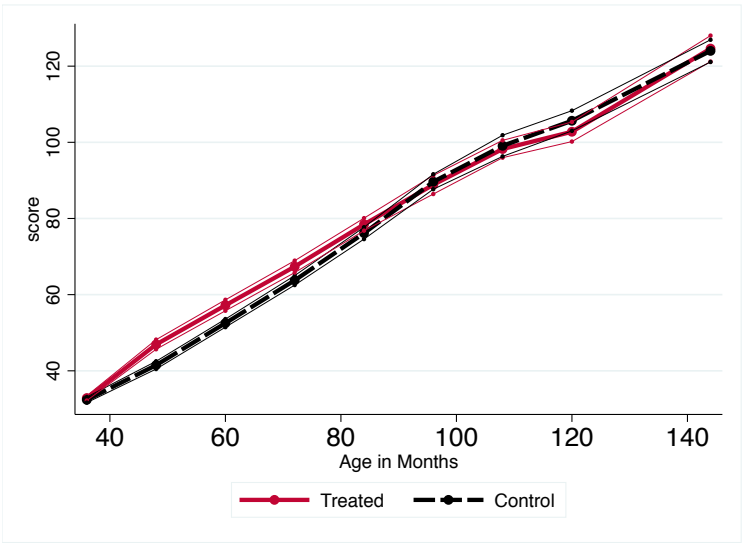
(a) PPVT



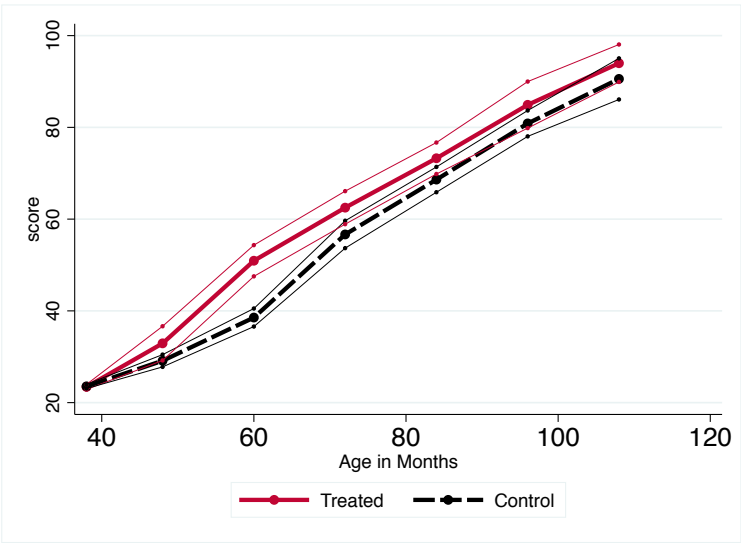
G.3 Trajectories of Cognitive Test Scores for Males

Figure 28: Perry, Mental age scores

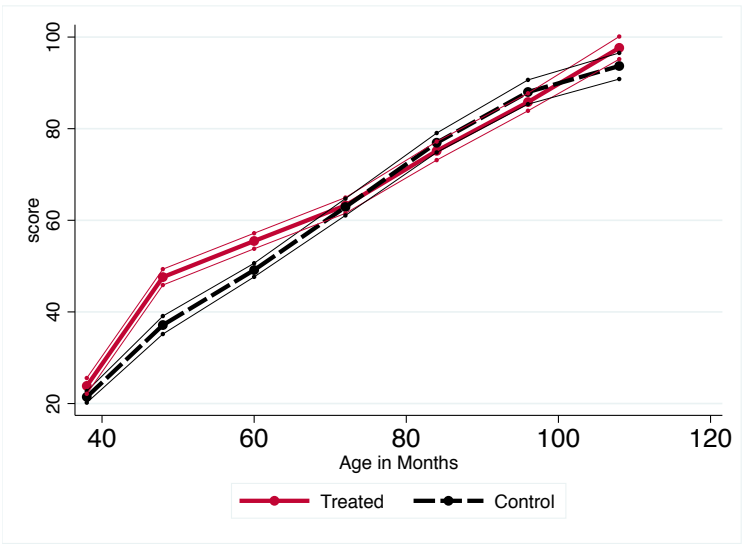
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

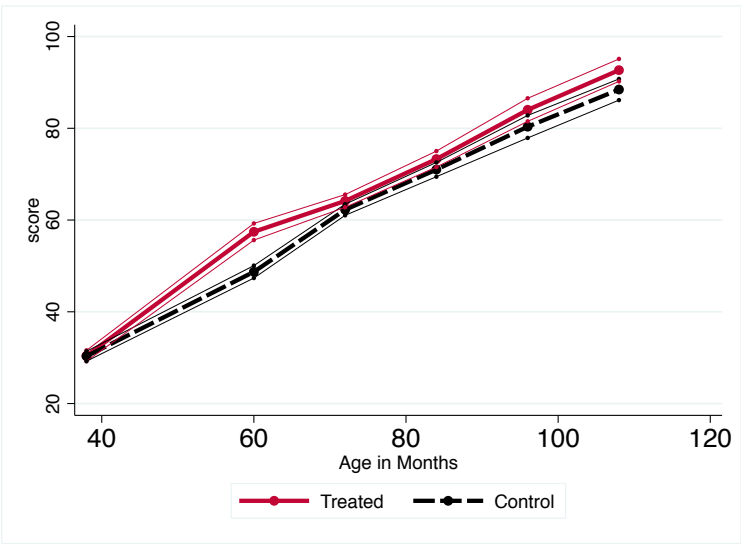
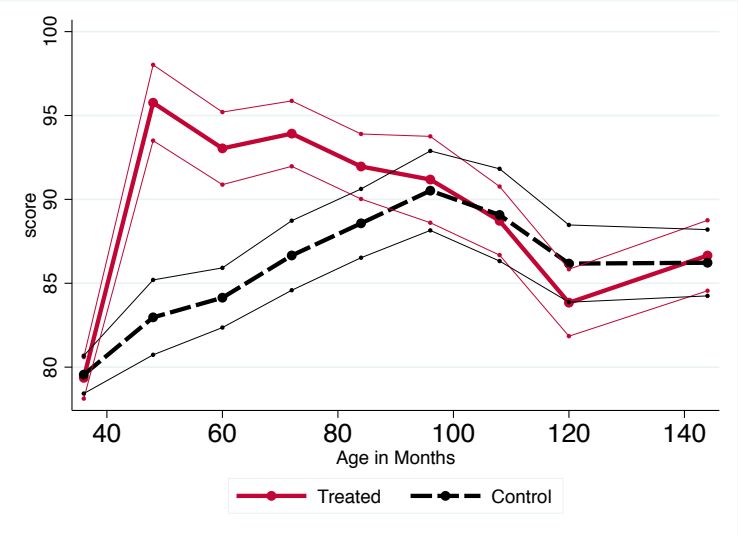
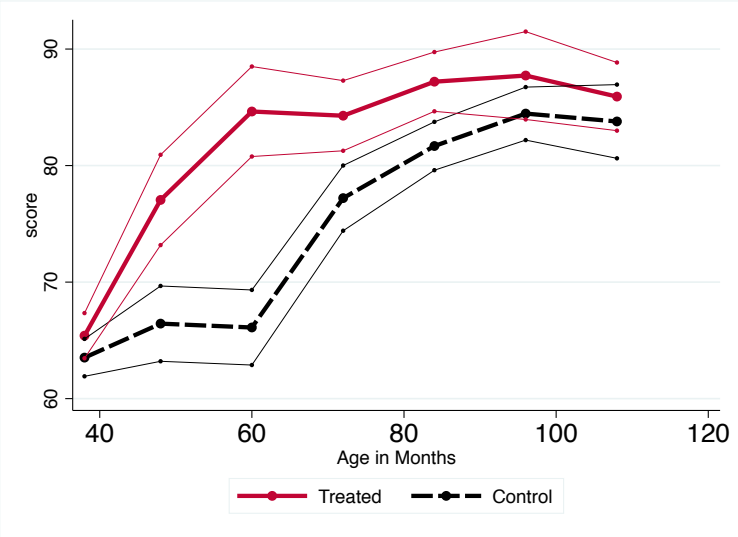


Figure 29: Perry, Standardized scores

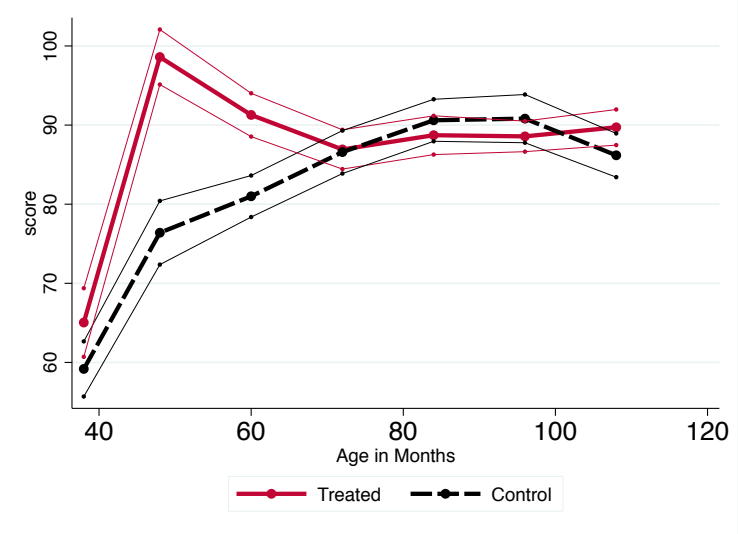
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

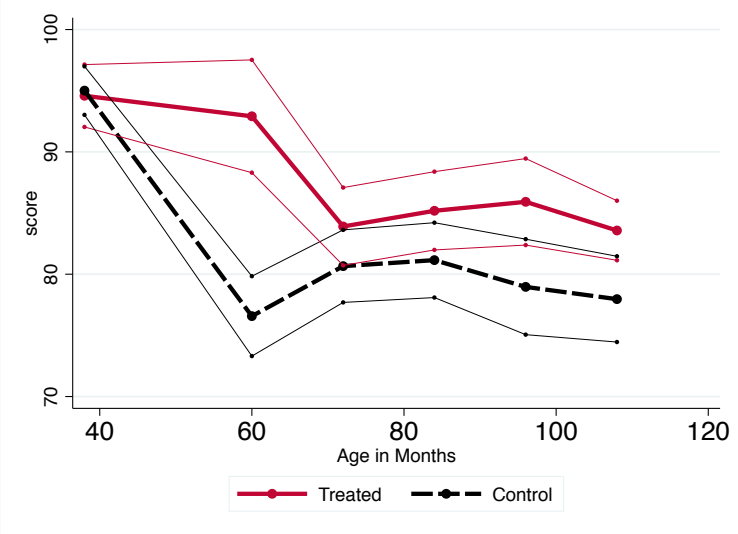
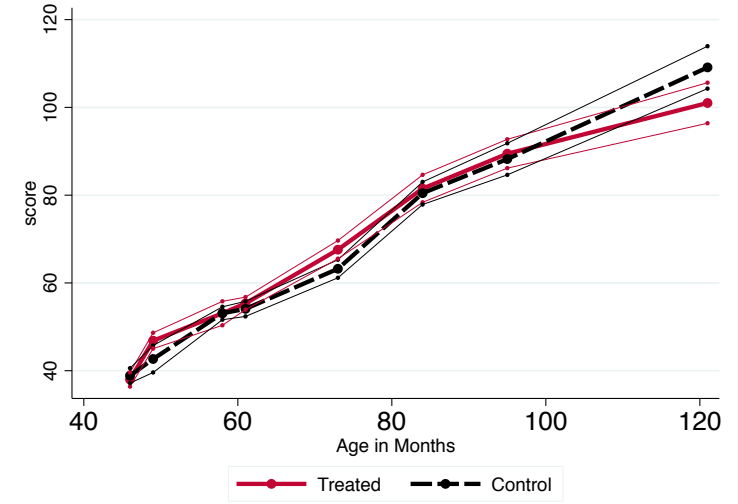
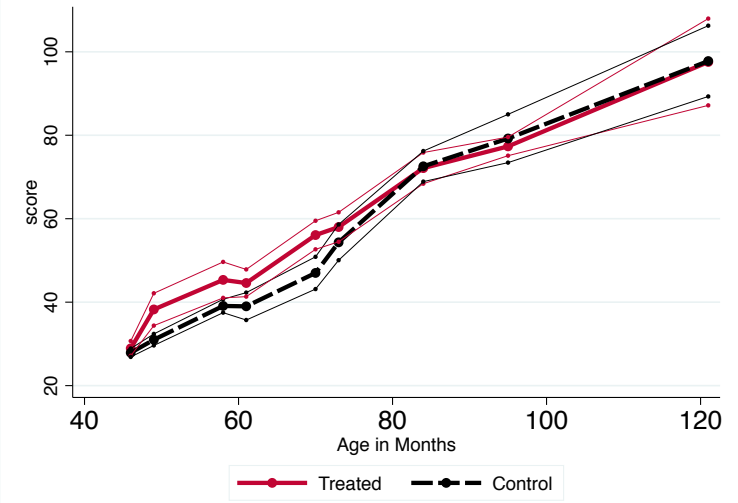


Figure 30: ETP, Mental age scores

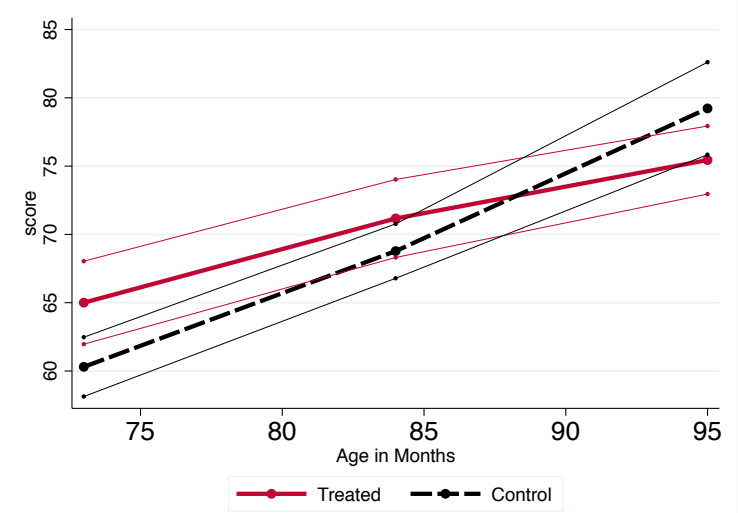
(a) Stanford Binet



(b) PPVT



(c) ITPA



(d) WISC

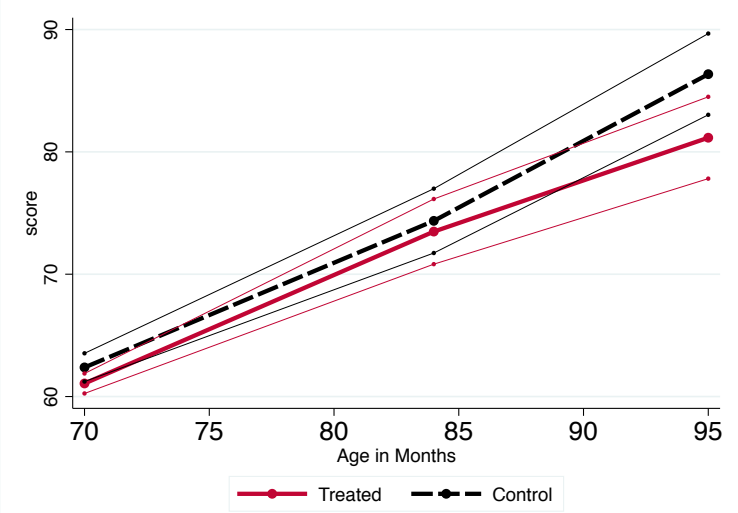
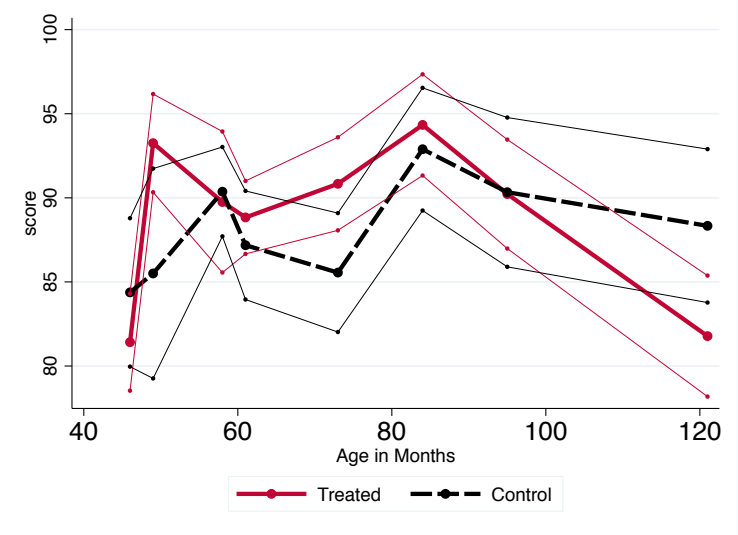
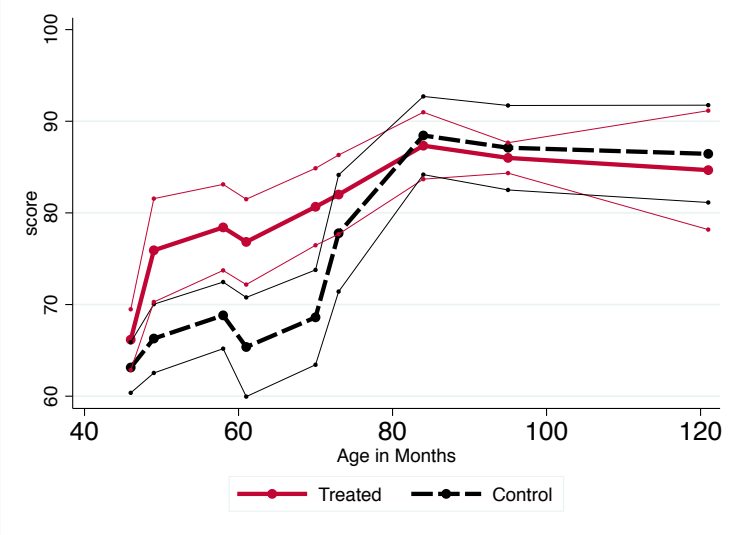


Figure 31: ETP, Standardized scores

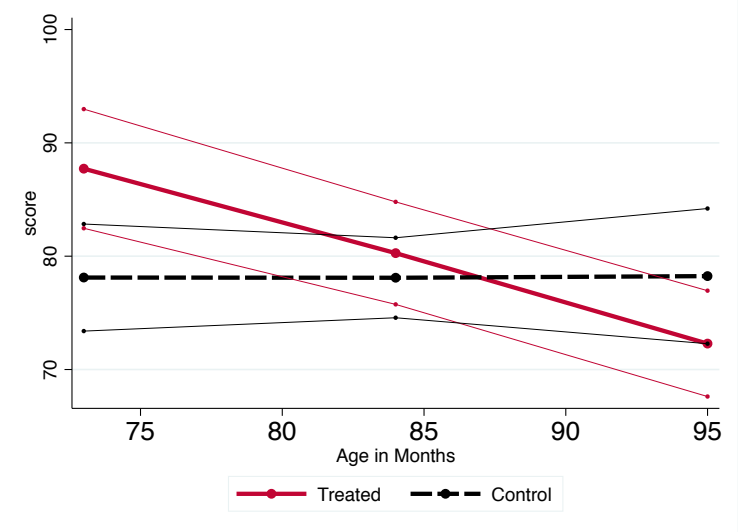
(a) Stanford Binet



(b) PPVT



(c) ITPA



(d) WISC

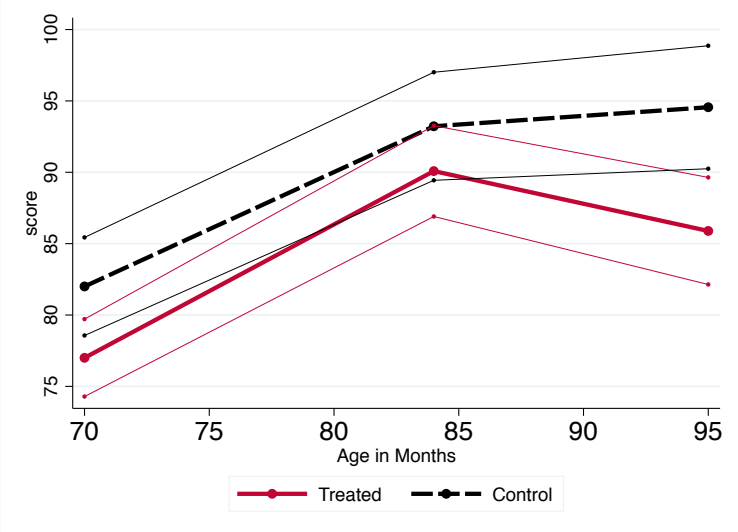
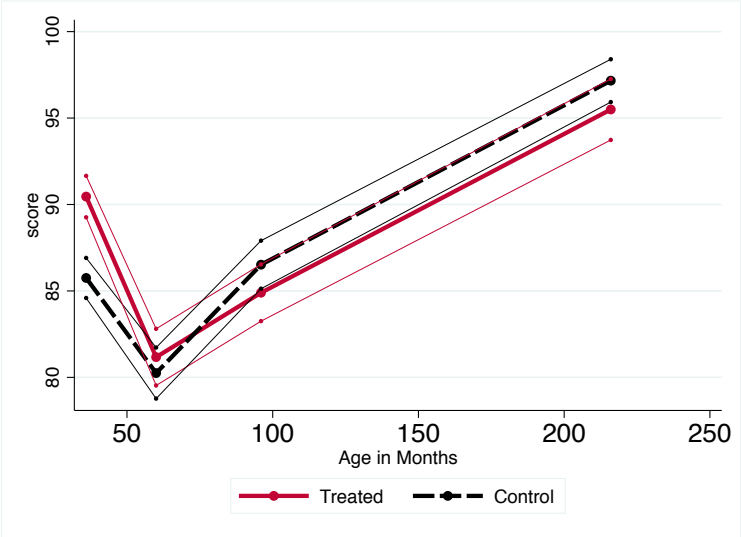


Figure 32: IHDP, Standardized scores

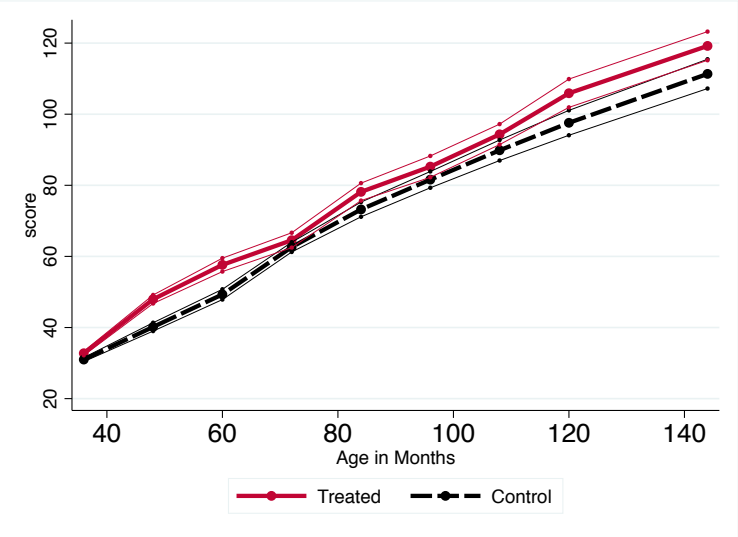
(a) PPVT



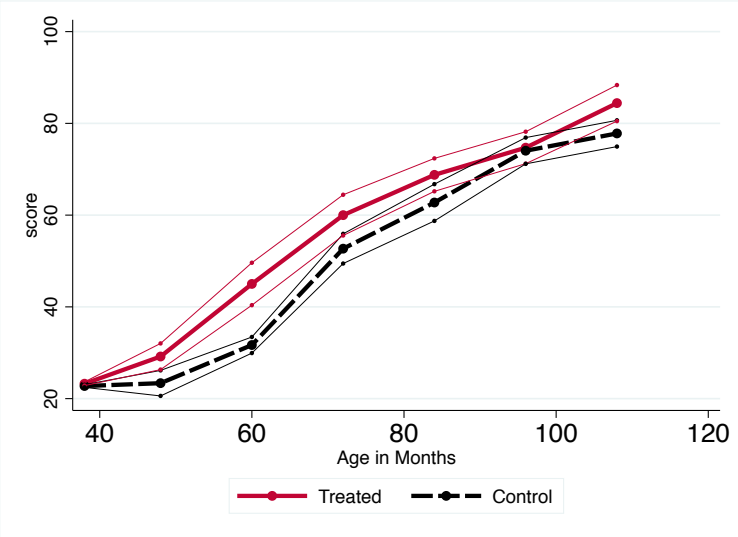
G.4 Trajectories of Cognitive Test Scores for Females

Figure 33: Perry, Mental age scores

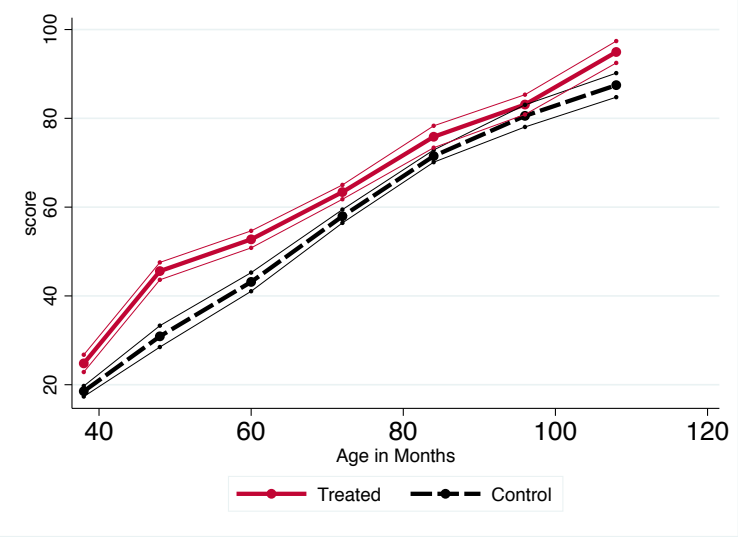
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

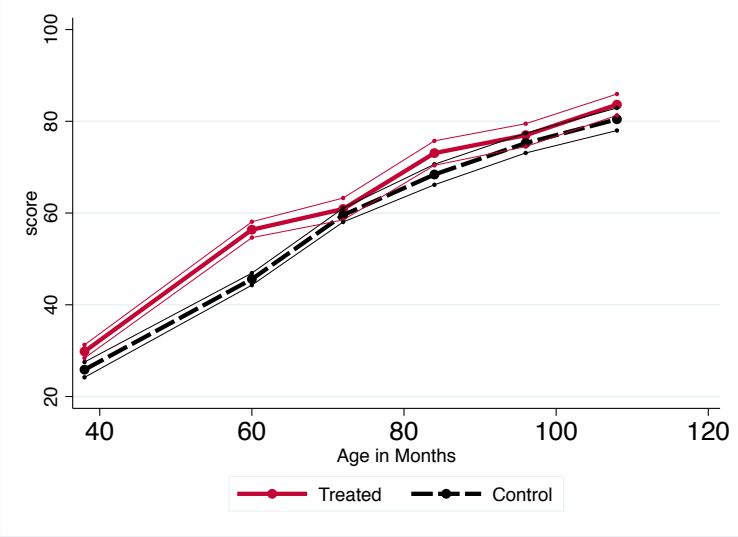
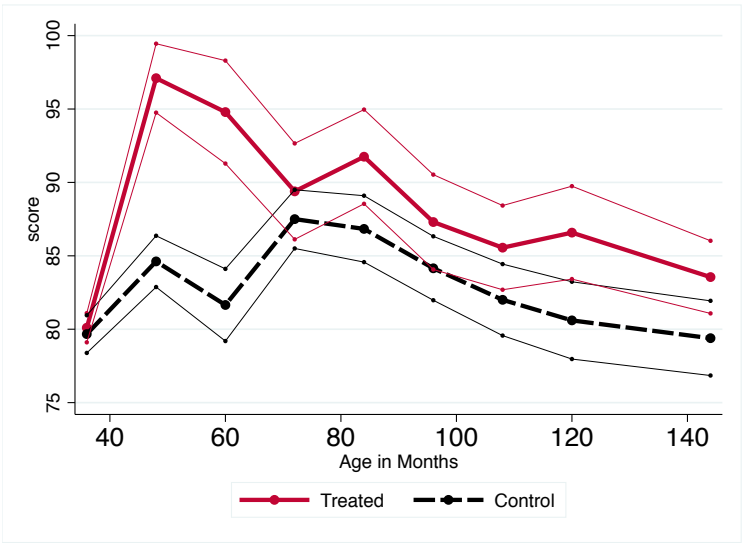
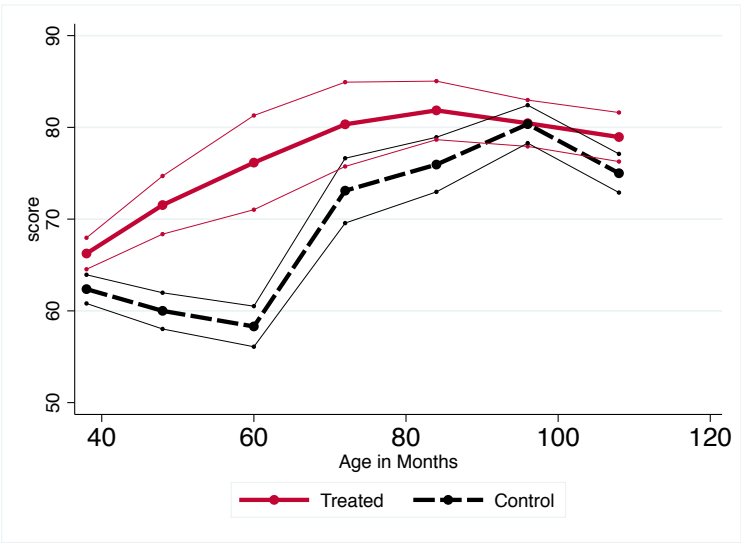


Figure 34: Perry, Standardized scores

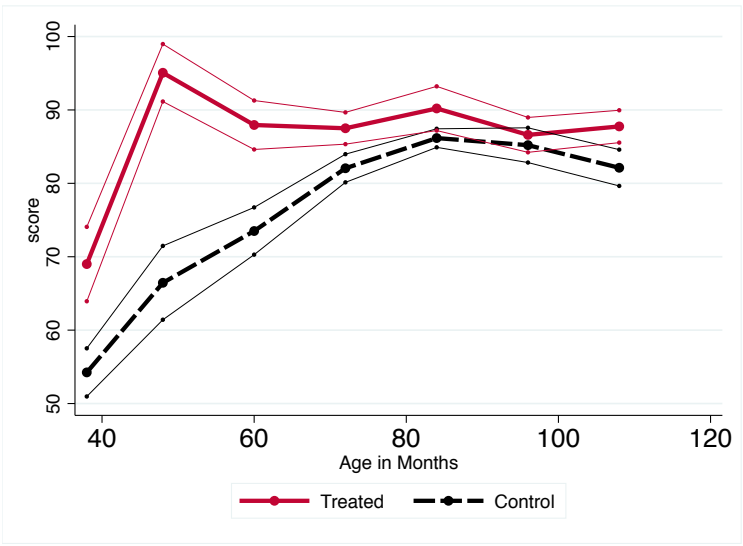
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

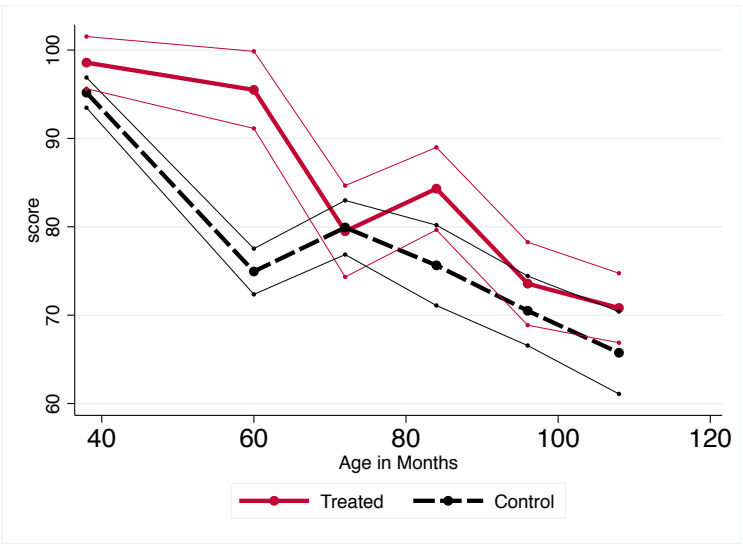
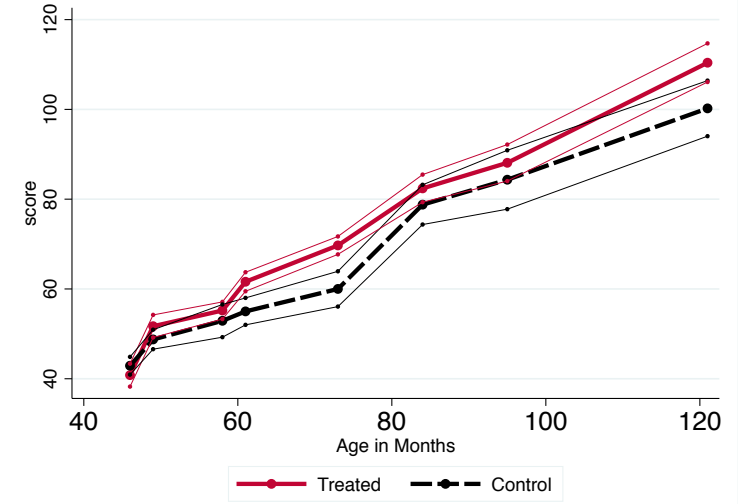
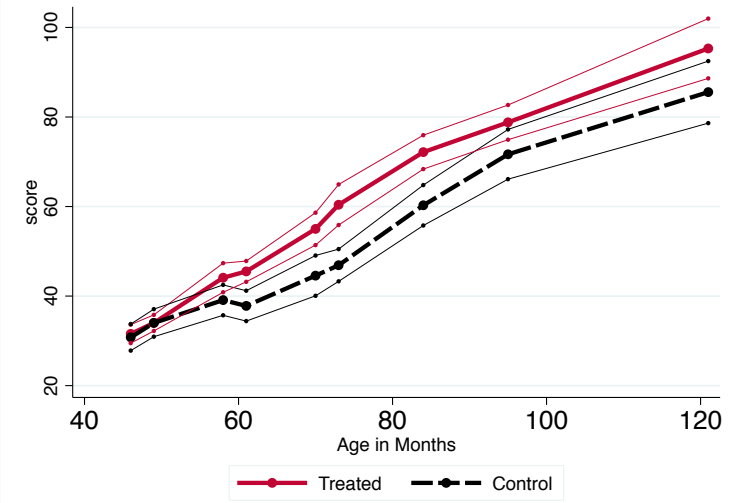


Figure 35: ETP, Mental age scores

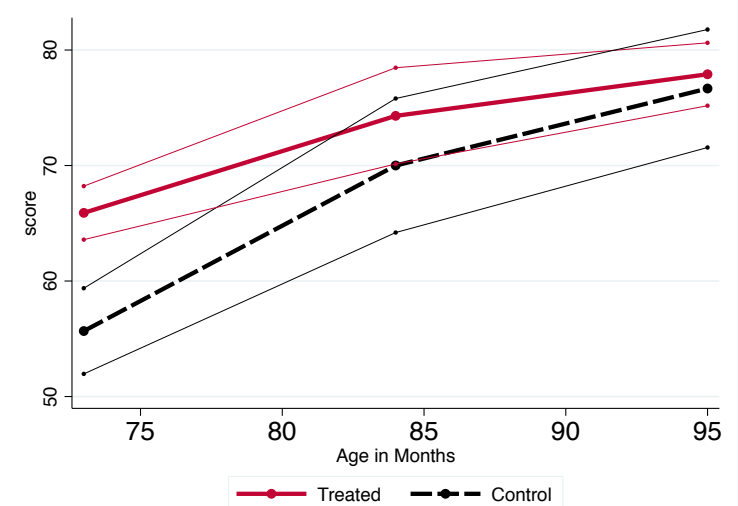
(a) Stanford Binet



(b) PPVT



(c) ITPA



(d) WISC

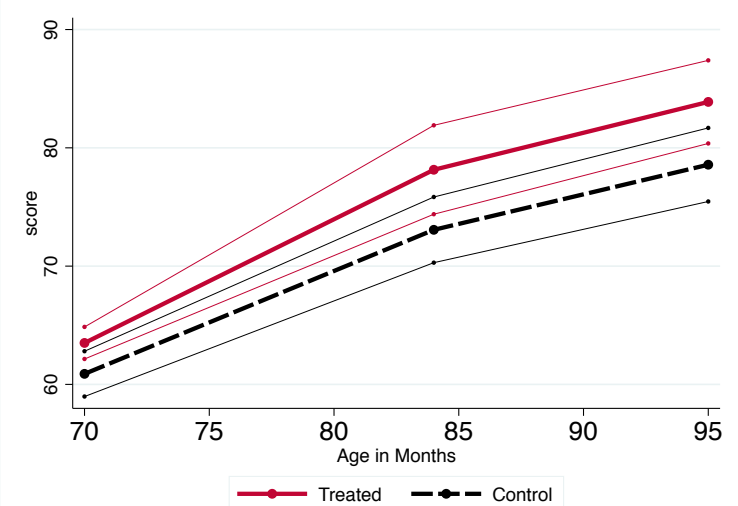
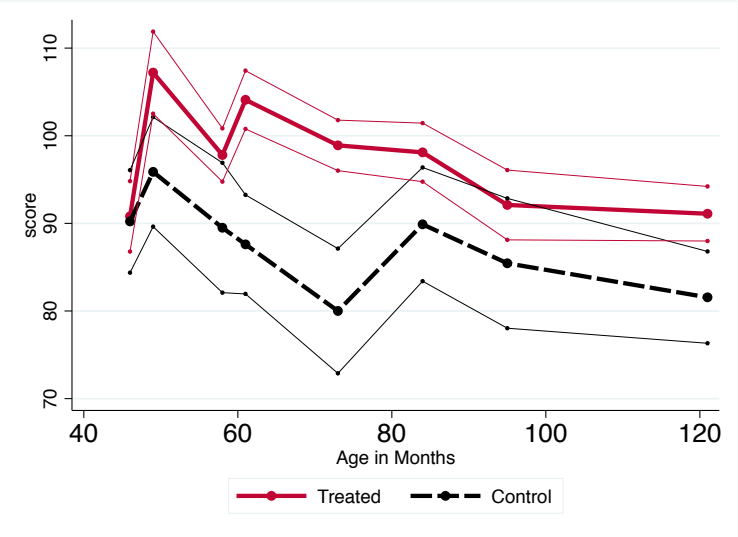
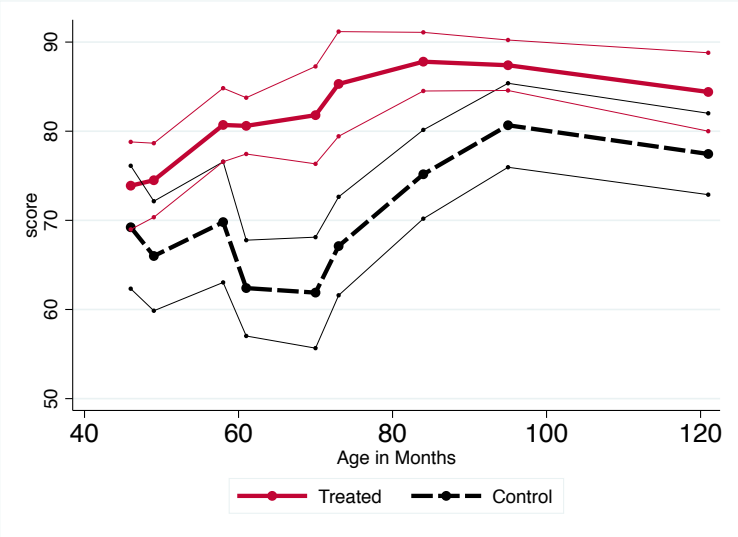


Figure 36: ETP, Standardized scores

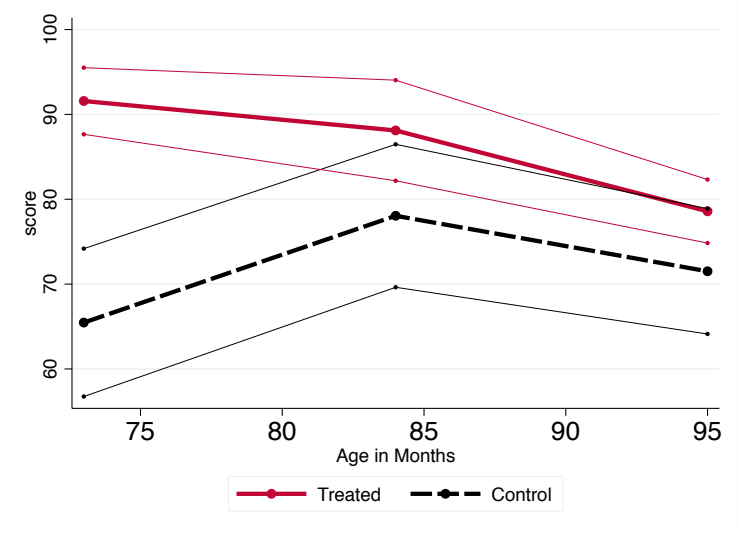
(a) Stanford Binet



(b) PPVT



(c) ITPA



(d) WISC

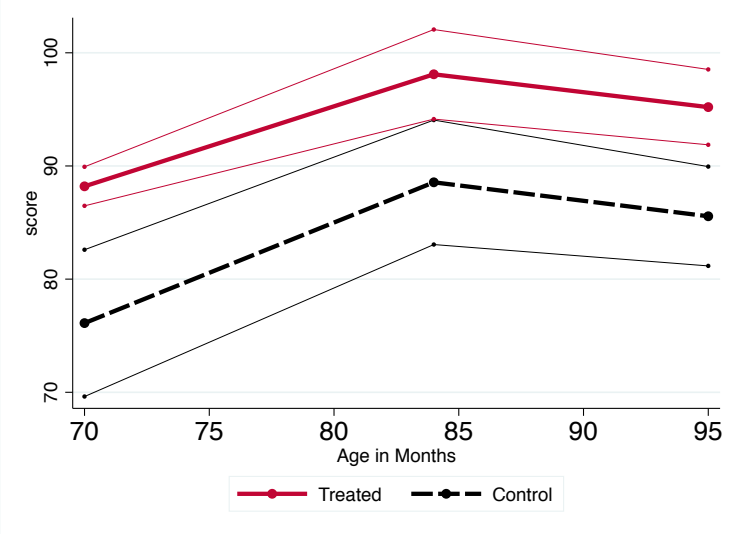
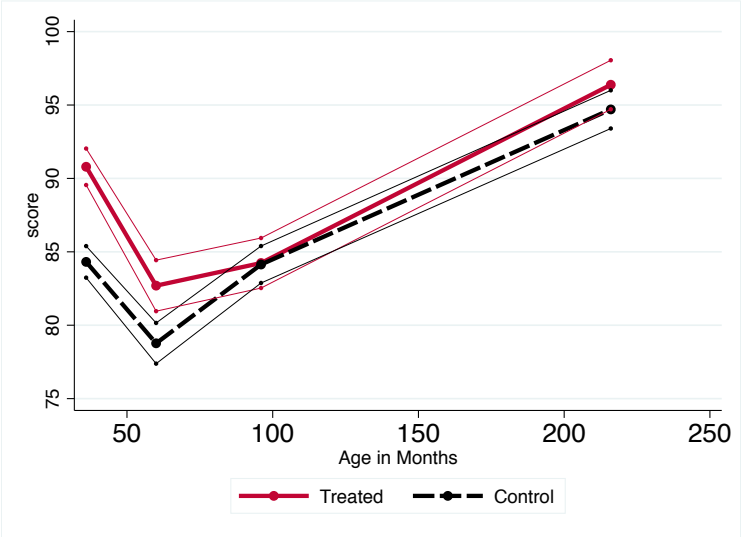


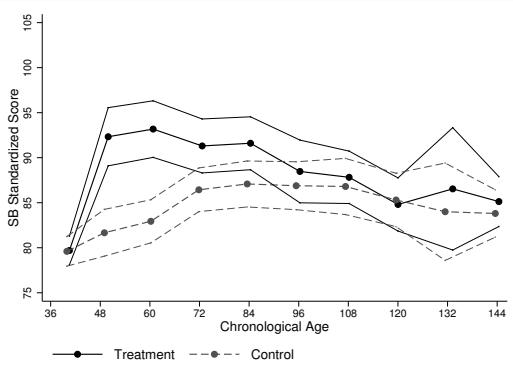
Figure 37: IHDP, Standardized scores

(a) PPVT



G.5 A Representative Graph from the Literature

Figure 38: Usual Graph of SB Standardized Scores for Perry, Mixing Cohorts with Different Timing



Note: These graphs show the average standardized IQ scores for the treatment and control groups for both programs. The IQ scores are measured using the Stanford-Binet IQ test and one measure of the WISC test. The program started at 48 months (60 months for one of the treatment groups) in ETP and 36 months (48 months for the first wave) in Perry. For both programs, there is an evident treatment effect right after the start of the program. The graphs show how the IQ scores of the treatment and control groups converge after the end of the programs.

H Impacts on Different Transformations of the Test Scores

I confirm the trends in Section 4 and the results in section 6 by showing that even using many different possible transformations of the test scores, fadeout is always present in our estimations. Notice that while the magnitudes are not comparable across the different types of measurements, because of the arbitrary in the scales, the t-tests are invariant to linear transformations of the data.

	36 months	48 months	60 months	72 months	84 months	96 months
Absolute Measures						
Raw Scores, T	81.16	93.52	94.72	92.09	92.82	89.92
Raw Scores, C	79.72	82.89	84.75	87.73	88.58	88.75
Difference	1.44	10.63**	9.97**	4.36**	4.24**	1.17
Percentiles, T	7.36	24.47	39.40	52.39	68.12	77.80
Percentiles, C	6.08	17.12	30.48	47.09	62.97	76.37
Difference	1.28	7.34**	8.92**	5.30**	5.15**	1.43
Relative Measures						
Age-Standardized, T	79.69	92.33	93.18	91.30	91.60	88.47
Age-Standardized, C	79.60	81.66	82.94	86.44	87.08	86.89
Difference	0.09	10.67**	10.24**	4.87**	4.52**	1.59
Percentiles by Age, T	51.25	62.80	63.20	55.64	54.45	50.86
Percentiles by Age, C	42.72	35.40	35.54	41.18	41.96	46.14
Difference	8.54	27.40**	27.66**	14.45**	12.49*	4.72
Anchoring, T	-0.04	-0.14	-0.14	0.25	0.19	0.08
Anchoring, C	0.04	0.12	0.13	-0.22	-0.18	-0.07
Difference	-0.08	-0.26**	-0.27**	0.47**	0.37**	0.16

This table uses Perry data with multiple transformations of the Standford-Binet test. Absolute tests are constructed to be always increasing in skills. Relative tests are constructed to be increasing in skills given an age. Raw scores are the simple sum of correct questions. "Percentiles" are constructed based on the raw scores in the whole sample. Age-standardized are raw scores less the national mean for the age and divided by the standard deviation for the age. "Percentiles by Age" are constructed based on the raw scores at each age. Anchoring scores are based on measurement error-corrected regressions of earnings in the scores, following Cunha et al. (2010)

I Estimation of the Autocorrelation Parameter

In this section I discuss how I can estimate autocorrelation parameters in my sample, rather than assuming a value for them. As I previously discussed, this sample is too small to be able to estimate an autoregressive parameter on it assuming individual fixed effects in a robust way. Then, I assume the following real model:

$$\theta_{it+1} = \rho\theta_{it} + \psi_t R_i + \omega_t + \eta_{it}$$

Estimating a meaningful parameter requires using tests that could plausibly have the same slopes, as discussed in Section 4:

$$M_{it}^m = a^m + b^m\theta_{it} + \varepsilon_{it}^m. \tag{13}$$

We can use this equation to express the above relationship in terms of the observed measurement instead of the unobserved skills, getting:

$$M_{it+1}^m = (1-\rho)a^m + \rho M_{it}^m + b^m\psi_t R_i + (b^m\omega_t + b^m\eta_{it} + \varepsilon_{it+1}^m - \rho\varepsilon_{it}^m)$$

There is measurement error in this equation, because ε_{it}^m is part of M_{it}^m. After taking that into account, this equation can be directly estimated, for example, using instrumental variables. We will obtain the ρ parameter. Tables 40 and 39 present my estimations of this parameter from the data used in this paper. The estimation for ETP is done period by period, because the spacing between the different periods varies between three months and around two years. For this reason, the estimated coefficients are highly instable. On the other hand, the estimation for Perry is done for all periods simultaneously, given that in the data the time between tests is always a year. The parameters are reasonably similar across specifications and, except for one case, cannot be statistically distinguished from 1, providing supporting

Figure 39: Estimation of Autoregressive Coefficients in Perry

	IQ	SB	PPVT	Leiter	ITPA
Coefficient	1.05	1.07	0.92	0.91	0.82
p value	0.30	0.21	0.67	0.21	0.02

The first row shows the autoregressive coefficient estimated in Perry of an Instrumental Variables regression of the measurements on the lagged measurements. The second row present the p-values of a test of the coefficients being equal to 1. The estimations presented in column 1 shows the results when all the tests are used, with the skill-standardization procedure explained in Section 3. The other columns show results using only one of the test types.

evidence towards the use of the assumption that $\rho = 1$.

Figure 40: Estimation of Autoregressive Coefficients in ETP

Age	IQ	SB	PPVT
49		1.11	1.20
58	1.38	1.28	1.02
61	0.75	0.87	0.83
70	1.16		1.11
73	1.39		0.80
84	1.08	0.98	1.10
95	1.76	1.33	0.97
121	0.56	0.98	1.55

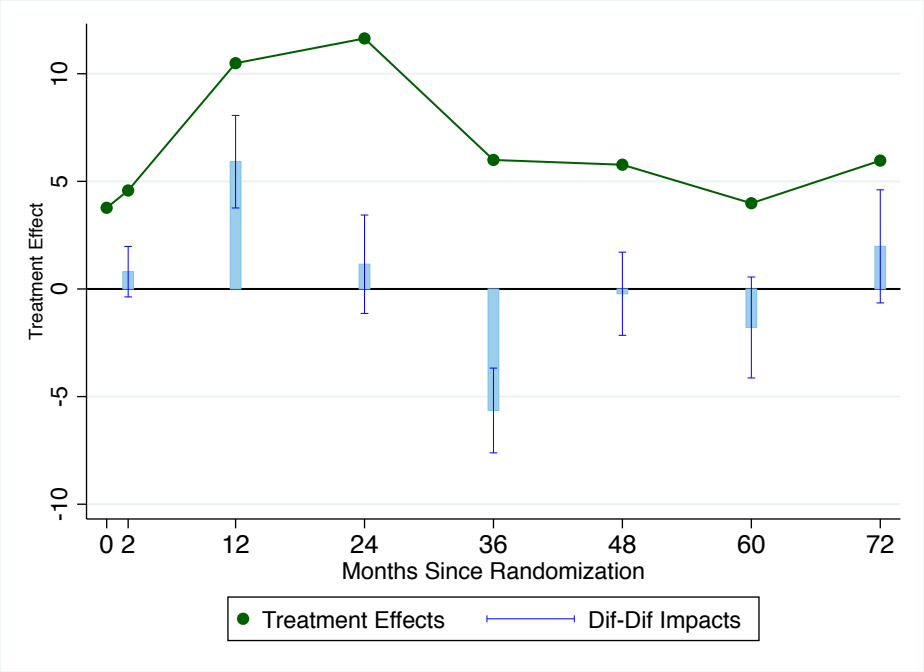
The coefficients in this table are estimated autoregressive coefficients in a regression of the skills at the age given in the row names on the lagged skills. I use previous lags or alternative measurements as instruments. Column IQ uses all available measurements, transformed using the procedure explained in Section 3. Column SB uses only measurements of the Stanford-Binet IQ test. Column PPVT uses only measurements of the PPVT test.

J Main Estimates Using Different Transformations of Test Scores

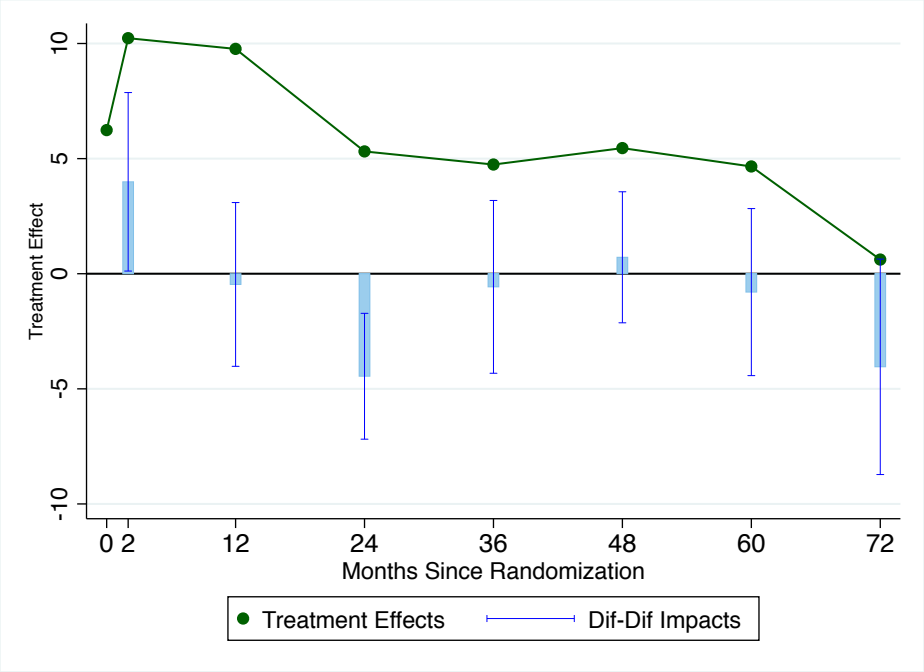
J.1 Estimates using All Mental Age Scores for Perry

Figure 41: Differences in Differences in Perry

(a) Two-Years-Treatment Cohorts



(b) One-Year-Treatment Cohort

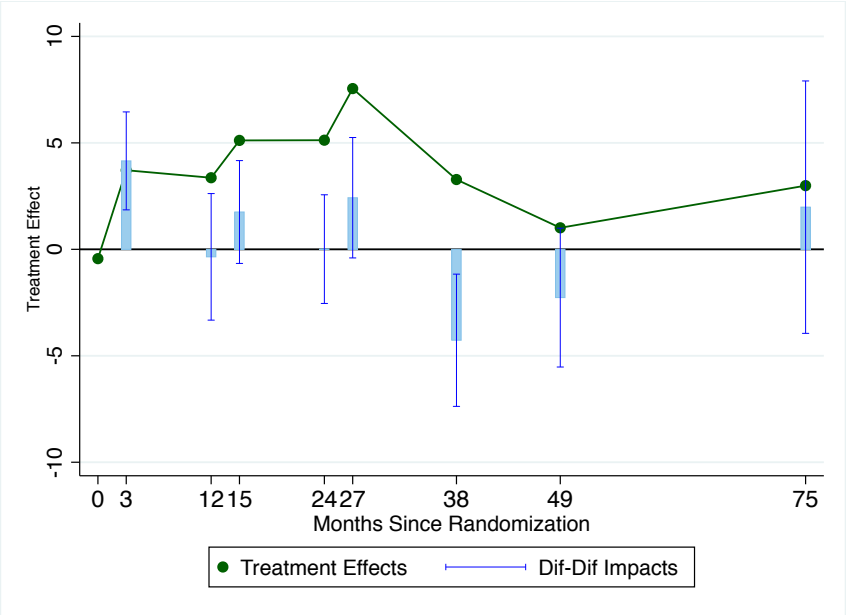


Note: The Y-axis represents the impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. Children in Perry take a baseline measurement at 36 months, are tested after the first year of preschool at 48 months, are tested after the second year of preschool at 60 months, are tested after the first year of school at 72 months and up to age 12. See Appendices C and G for an explanation of the construction of the different measurements.

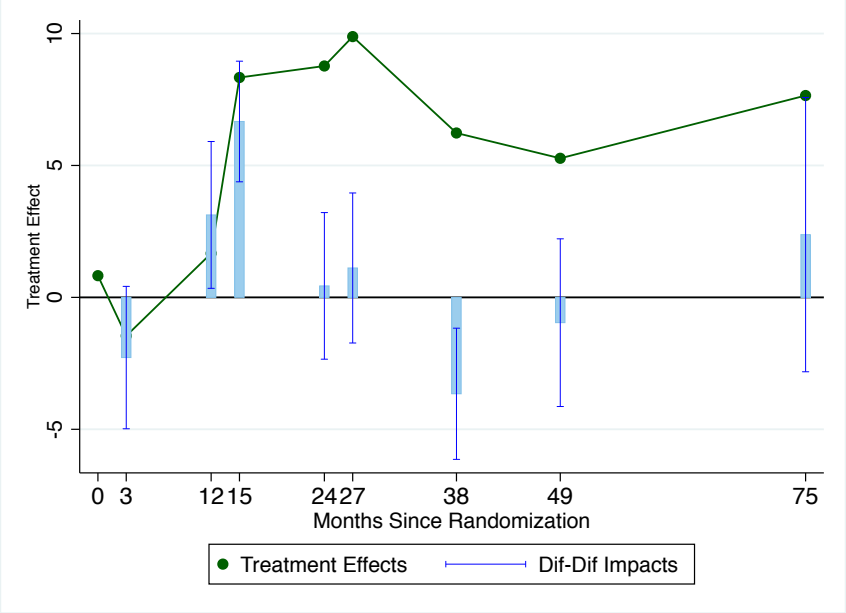
J.2 Estimates using All Mental Age Scores for ETP

Figure 42: Differences in Differences in ETP

(a) Two-Years-Treatment Cohorts



(b) One-Year-Treatment Cohort

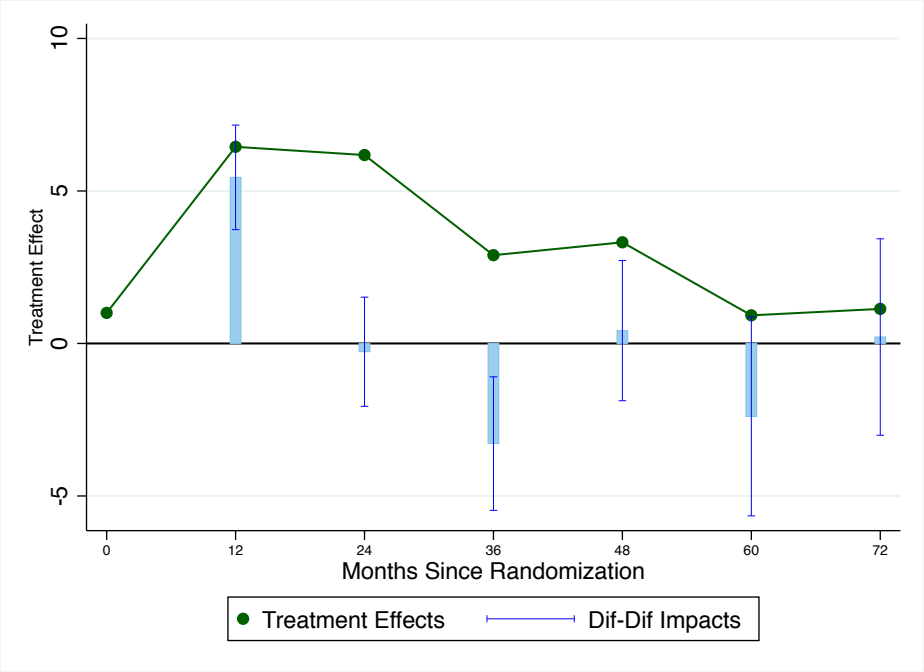


Note: The Y-axis represents the impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. Treated children in ETP take a baseline measurement at 46 months, are tested after the first summer school at 49 months, receive home visits during the second year, are tested before the second summer school at 58 months, then after it at 61 months, then they receive home visits during an additional year, and they are tested after the third summer school at 73 months, when they enter school. They are tested after the first year of school at 85 months, then they are tested again at 95 and 121 months. Control children experience the same testing schedule but no summer schools or home visits. See Appendices C and G for an explanation of the construction of the different measurements.

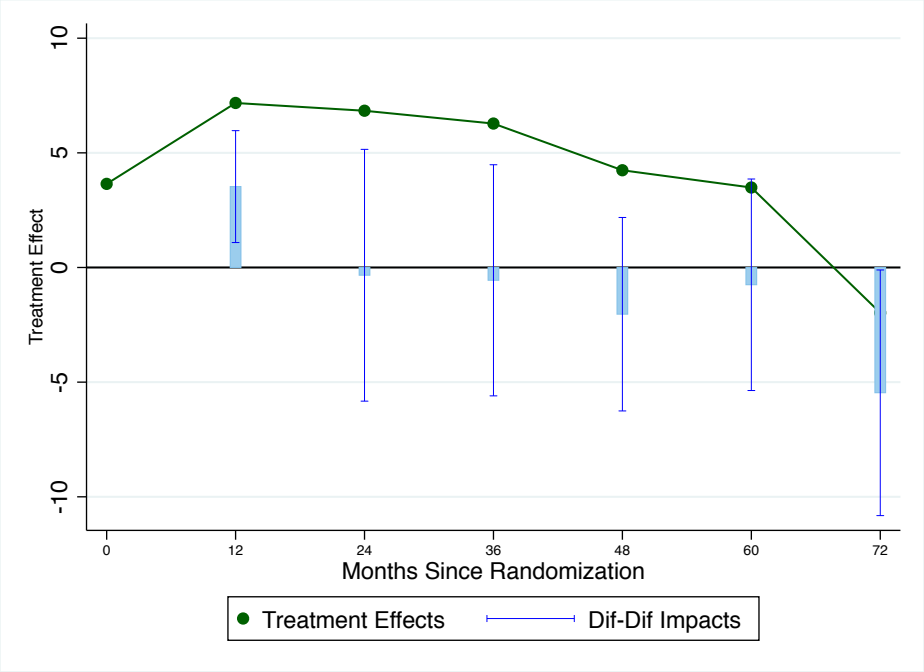
J.3 Estimates using Stanford-Binet Mental Age Scores for Perry

Figure 43: Differences in Differences in Perry

(a) Two-Years-Treatment Cohorts



(b) One-Year-Treatment Cohort

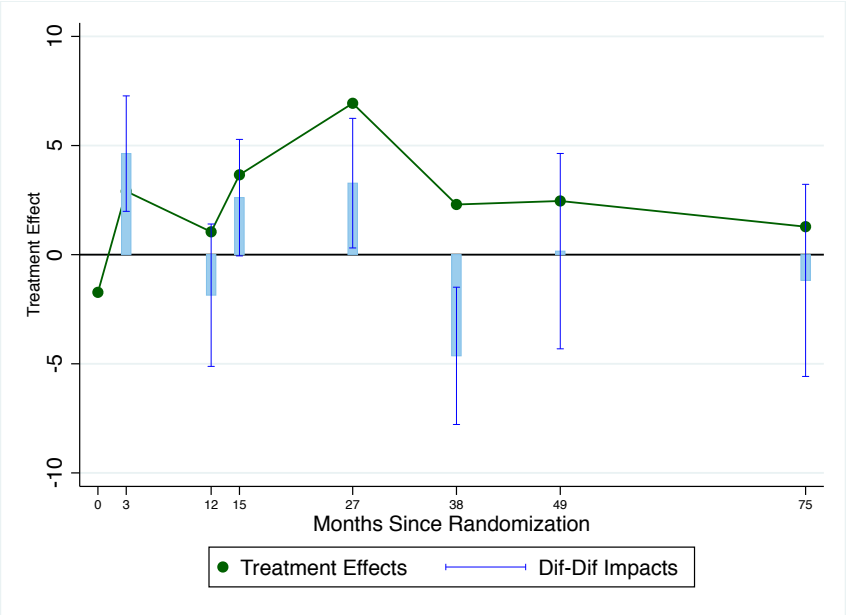


Note: The Y-axis represents the impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. Children in Perry take a baseline measurement at 36 months, are tested after the first year of preschool at 48 months, are tested after the second year of preschool at 60 months, are tested after the first year of school at 72 months and up to age 12. See Appendices C and G for an explanation of the construction of the different measurements.

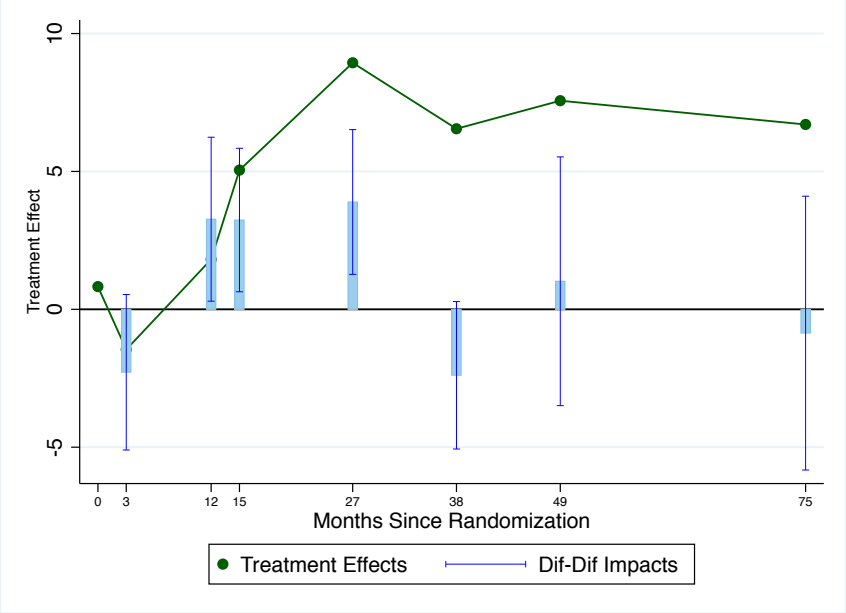
J.4 Estimates using Stanford-Binet Mental Age Scores for ETP

Figure 44: Differences in Differences in ETP

(a) Two-Years-Treatment Cohorts



(b) One-Year-Treatment Cohort

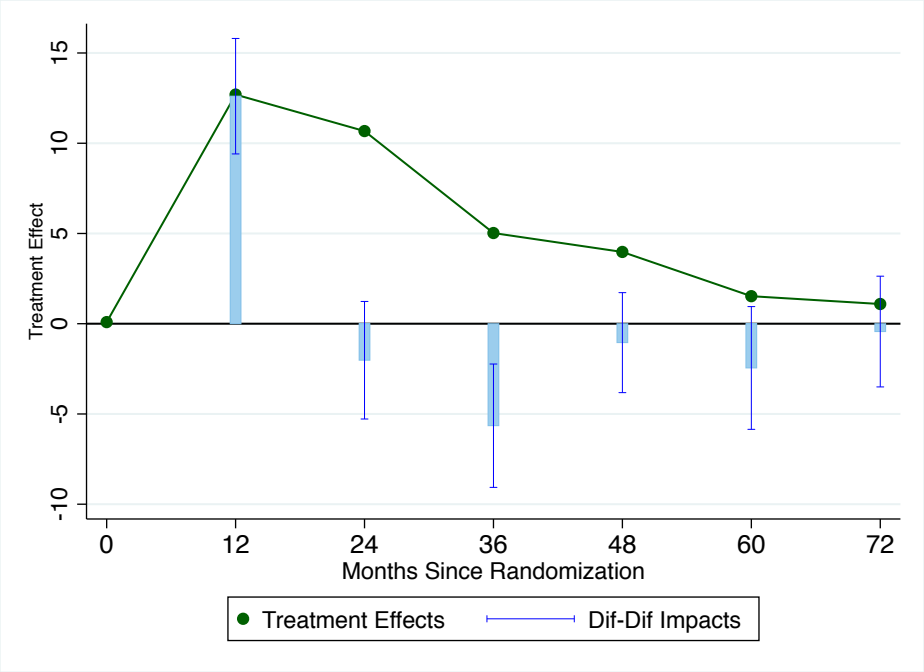


Note: The Y-axis represents the impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. Treated children in ETP take a baseline measurement at 46 months, are tested after the first summer school at 49 months, receive home visits during the second year, are tested before the second summer school at 58 months, then after it at 61 months, then they receive home visits during an additional year, and they are tested after the third summer school at 73 months, when they enter school. They are tested after the first year of school at 85 months, then they are tested again at 95 and 121 months. Control children experience the same testing schedule but no summer schools or home visits. See Appendices C and G for an explanation of the construction of the different measurements.

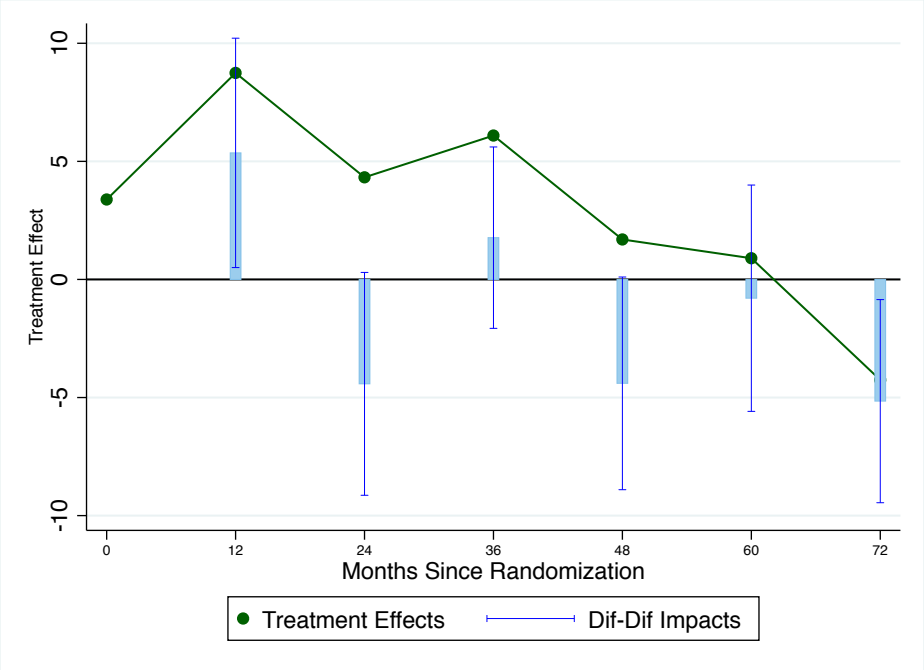
J.5 Estimates using Raw PPVT Scores for Perry

Figure 45: Differences in Differences in Perry

(a) Two-Years-Treatment Cohorts



(b) One-Year-Treatment Cohort

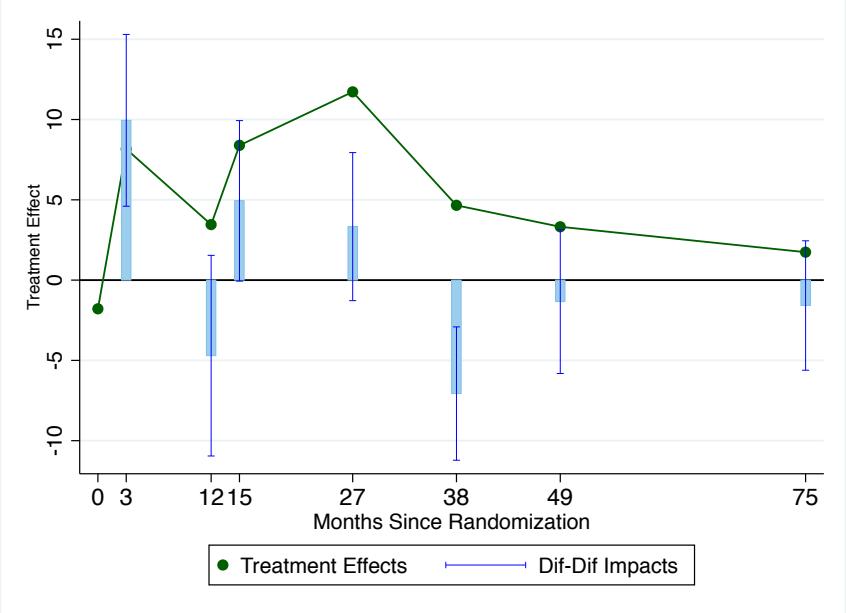


Note: The Y-axis represents the impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. Children in Perry take a baseline measurement at 36 months, are tested after the first year of preschool at 48 months, are tested after the second year of preschool at 60 months, are tested after the first year of school at 72 months and up to age 12. See Appendices C and G for an explanation of the construction of the different measurements.

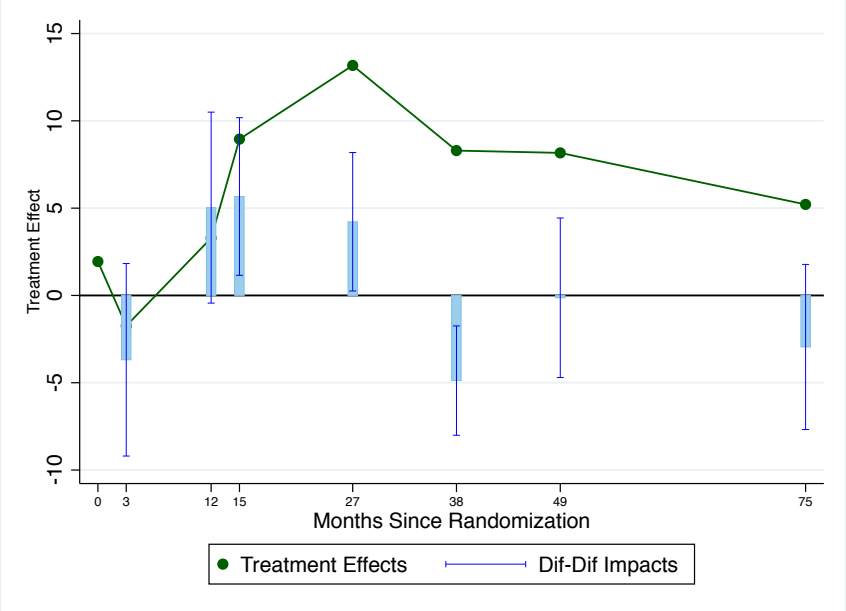
J.6 Estimates using Raw PPVT Scores for ETP

Figure 46: Differences in Differences in ETP

(a) Two-Years-Treatment Cohorts



(b) One-Year-Treatment Cohort

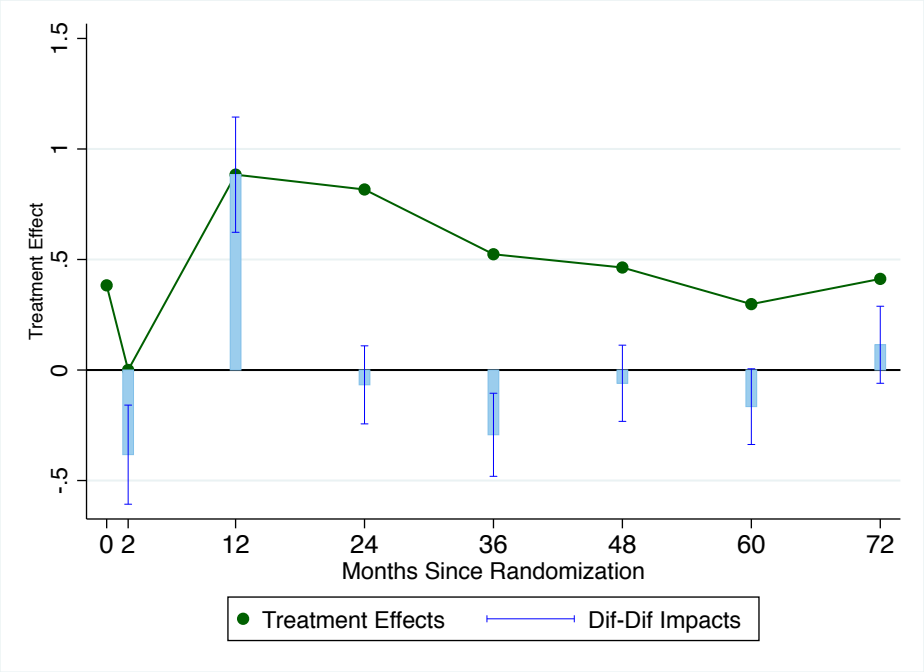


Note: The Y-axis represents the impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. Treated children in ETP take a baseline measurement at 46 months, are tested after the first summer school at 49 months, receive home visits during the second year, are tested before the second summer school at 58 months, then after it at 61 months, then they receive home visits during an additional year, and they are tested after the third summer school at 73 months, when they enter school. They are tested after the first year of school at 85 months, then they are tested again at 95 and 121 months. Control children experience the same testing schedule but no summer schools or home visits. See Appendices C and G for an explanation of the construction of the different measurements.

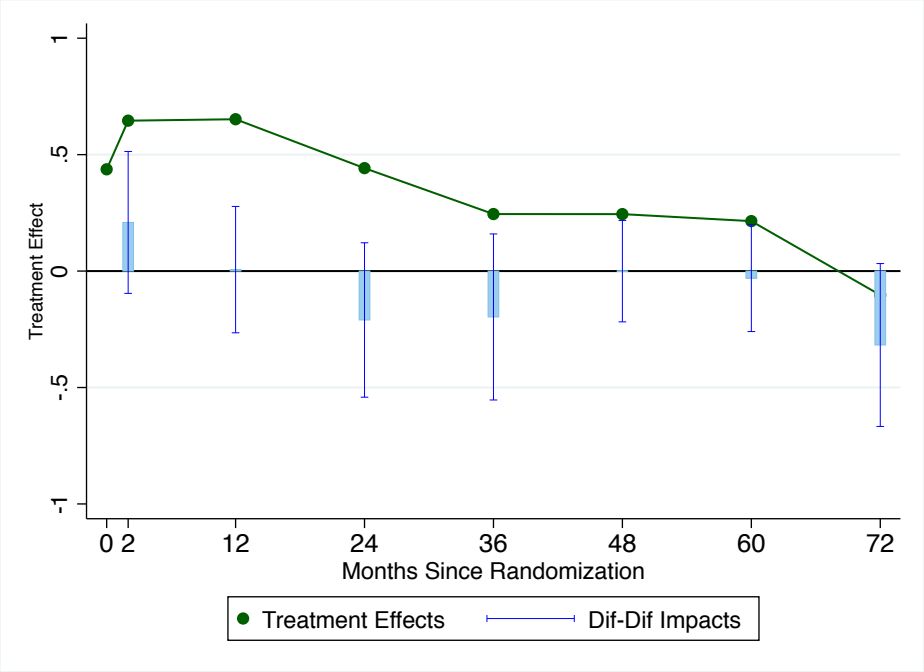
J.7 Estimates using Anchored Scores for Perry

Figure 47: Differences in Differences in Perry

(a) Two-Years-Treatment Cohorts



(b) One-Year-Treatment Cohort



Note: The Y-axis represents the impacts in terms of standard deviations. The X-Axis represents months after randomization. Asymptotic confidence intervals calculated at 90% level. Children in Perry take a baseline measurement at 36 months, are tested after the first year of preschool at 48 months, are tested after the second year of preschool at 60 months, are tested after the first year of school at 72 months and up to age 12. See Appendices C and G for an explanation of the construction of the different measurements.

K A Depreciation Model with Two Mutually Exclusive Types of Skills

In this appendix I present a formulation of the Depreciation model that can give rise to the model in the main paper. This formulation uses more notation, but it might be clearer.

Let the more trainable type of cognitive skills, which we can call K_{it}, An let the type of skills not affected by the programs be L_{it}. Let ρ_t be the persistence of L_{it}. Define δ_t as the difference in persistence between both types of skills. Finally, I assume that the relevant cognitive skills for this paper (the skills measured by the tests, as discussed in Section 3) are the sum of both types: $\theta_{it} = K_{it} + L_{it}$. The two skill production functions are:

$$K_{it} = (\rho_t - \delta_t) K_{it-1} + \beta_t F_{it} + \omega_t^K + \alpha_i^K + \eta_{it}^K \quad (14)$$

$$L_{it} = \rho_t L_{it-1} + \omega_t^K + \alpha_i^K + \eta_{it}^L \quad (15)$$

Where ω_t^K, ω_t^L are time fixed effects, α_i^K, α_i^L, are individual fixed effects, and η_{it}^K, η_{it}^L are the errors in the production functions. Then, in each period, the total cognitive skills can then be expressed, as in the main paper, as:

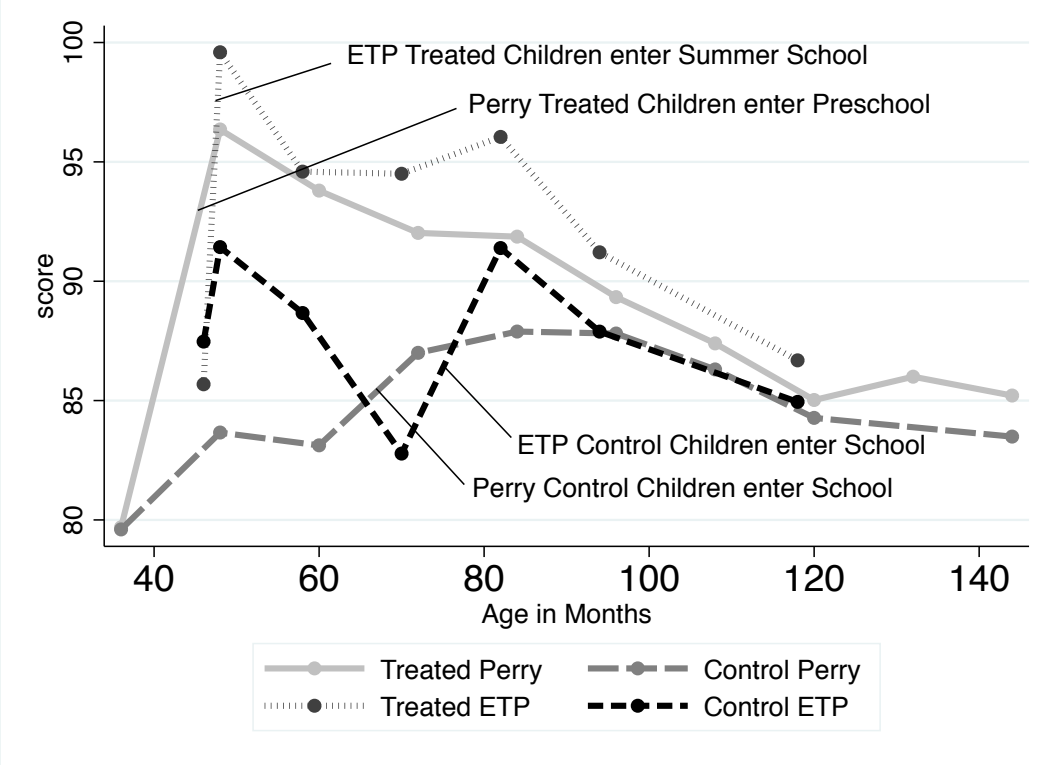
$$\theta_{it} = \rho_t \theta_{it-1} - \delta_t K_{it-1} + \beta_t F_{it} + \omega_t + \alpha_i + \eta_{it} \quad (16)$$

L Level of Standardized Test Scores Increase at School Age for the Control Group

L.1 Dynamics of Standardized Scores for Stanford-Binet

The following figure is analogous to the one presented for the PPVT test in Section 4.

Figure 48: Dynamics of the Standardized Scores in the SB Test: Perry and ETP



Note: Standardized Scores are constructed by substracting the mean for a representative sample of the age of the child to her raw score, and then dividing by the standard deviation for a representative sample for that age. The result is then multiplied by 15 and added 100, for simplicity and homogeneity across tests. Both programs used the Third Edition of the Stanford-Binet Test. Children in Perry take a baseline measurement at 36 months, are tested after the first year of preschool at 48 months, are tested after the second year of preschool at 60 months, are tested after the first year of school at 72 months and up to age 12. Treated children in ETP take a baseline measurement at 46 months, are tested after the first summer school at 49 months, receive home visits during the second year, are tested before the second summer school at 58 months, then after it at 61 months, then they receive home visits during an additional year, and they are tested after the third summer school at 73 months, when they enter school. They are tested after the first year of school at 85 months, then they are tested again at 95 and 121 months. Control children experience the same testing schedule but no summer schools or home visits.

The decline in Figure 48 for the control group in ETP around ages 60–70 months could be puzzling. One possible explanation is that most US children entered Kindergarten at that age (around 69.3% in 1970 (Snyder et al., 2016)), while Tennessee did not have KG at the time. The strong impacts at school entry for the control group are very clear.

L.2 A Formal Test of the Impacts of School Controlling for Age

To test the impact of school formally, it is necessary to aggregate the two datasets I use. I estimate a OLS regression to examine the determinants of the test scores in the control children, and in particular to quantify the effect of schooling. I include a set of dummy variables for a child being older than a given age.[11] Given that the ages are all different across the programs and that schooling is fully determined by age in each program, I average the test scores across groups of 2-3 months. Averaging makes the results more easily interpretable and less noisy. This regression can have a causal interpretation: entry into school is a deterministic function of age and study (Perry or ETP) and that no treatment substitutes were available for control children in these studies. I present the results of this exercise. The main takeaway from it, which is consistent across different specifications, is that even controlling for age non-parametrically, there is a large positive impact on controls at school entry, equivalent to 3/5 of a standard deviation of that test in the population.

I now show level plots with significance of the year-by-year changes for Perry and ETP. The results for Perry are easily interpretable, because there is little movement besides a positive shock at school entry age for controls. The results for ETP are more noisy, and there are trends in the level of the controls relative to the average US children that are not obvious to interpret. Those results can partially be explained by the relatively low number of individuals in this group (21) However, the impact at school entry is still significant and largest than any other impact.

[11] This is consistent with the model discussed in Section 5

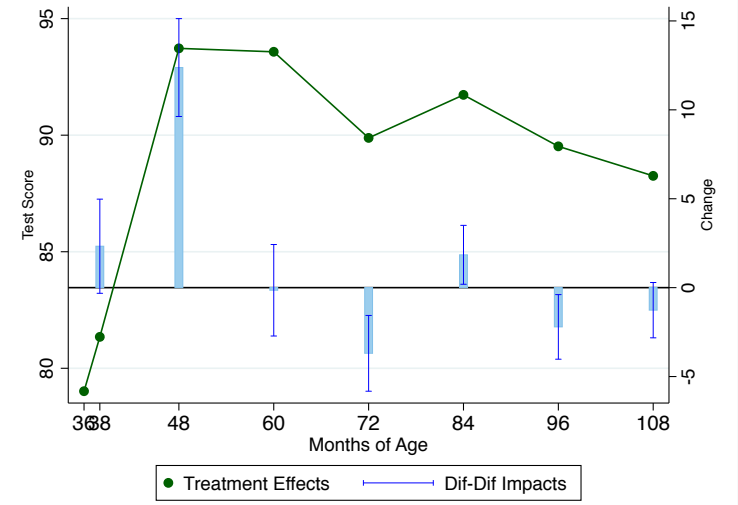
Table 9: Determinants of the Standardized Scores for Control Individuals

	All Tests	Stanford-Binet	PPVT
Age 36 Months	80.2**	82.9**	65.6**
Age 48 Months	0.26	4.93**	0.57
Age 60 Months	1.15	-0.74	-0.38
Age 72 Months	-2.84*	-4.34**	2.85
Age 84 Months	0.00036	1.06	3.70
Age 96 Months	-1.08	-0.98	2.86*
Age 108 Months	-1.47	-0.61	-1.94
Age 120 Months	-2.26*	-2.80**	-0.98
First Year of School	9.72**	7.55**	9.42**
Second Year of School	2.03	-0.17	0.20
Mother Works	3.83		
Test: PPVT	-12.8**		
Test: ITPA	-7.05**		
Test: Leiter	-6.59**		
Test: Weschler	-1.12		
Observation from Perry Study	-1.12	-3.32	-2.58

Note: all three specifications in this table use both Perry and ETP data together. Only the control group individuals are used. All the ages are in months. To calculate the impacts of age, grouped ages 46 and 49 in ETP are grouped with age 48 in Perry; ages 58 and 61 in ETP are grouped with age 60 in Perry; ages 70 and 73 in ETP are grouped with age 72 in Perry; ages 83 and 85 in ETP are grouped with age 84 in Perry; age 95 in ETP is grouped with age 96 in Perry; age 121 in ETP is grouped with age 120 in Perry. In all cases, the age dummies are constructed as "Individual has that age or more", for consistency with Section 5. Age 36 takes the value of 1 for all individuals and tests, so no additional constant is used in the estimation. First year of school is defined as tests taken at ages 72 or more for Perry and at ages 81 or more for ETP. Second year of school is defined as tests taken at ages 84 or more in Perry and at ages 95 or more for ETP. Standard errors clustered at the individual level.

Figure 49: Perry, Standardized Scores Trajectory, All Tests

(a) Treatment



(b) Control

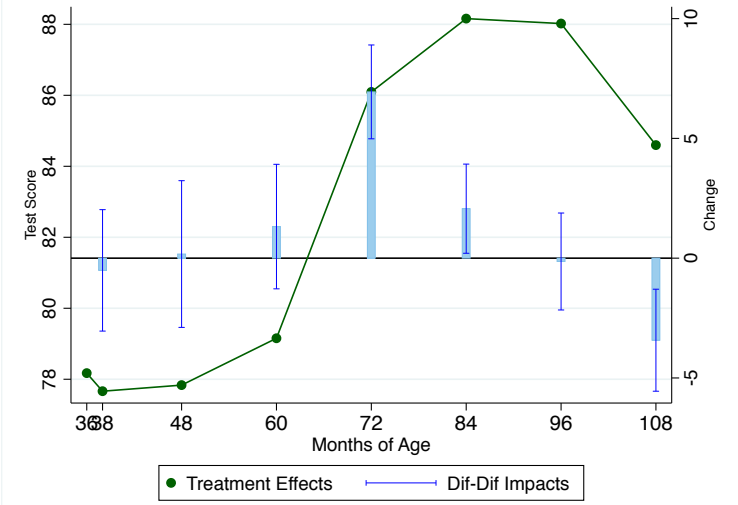
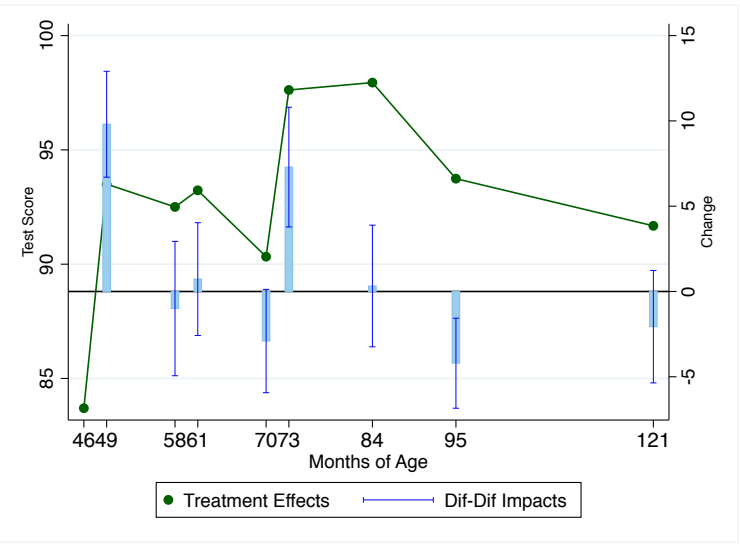
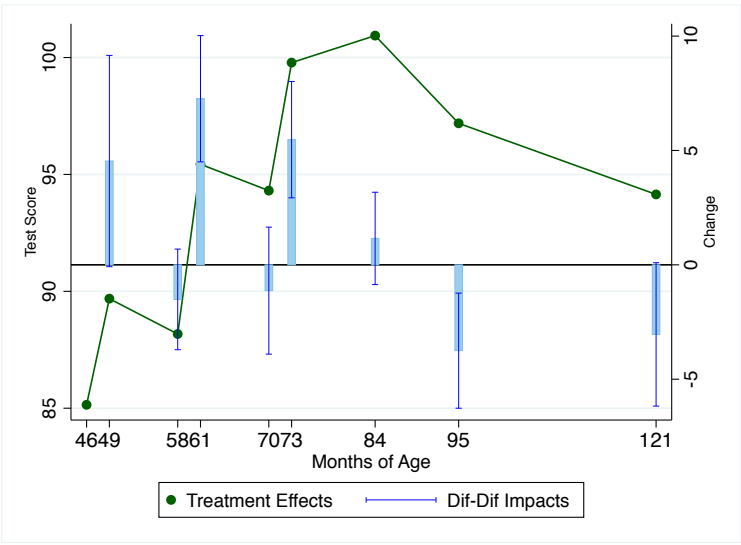


Figure 50: ETP, Standardized Scores Trajectory, All Tests

(a) 2-yrs. Treatment



(b) 1-yr. Treatment



(c) Control

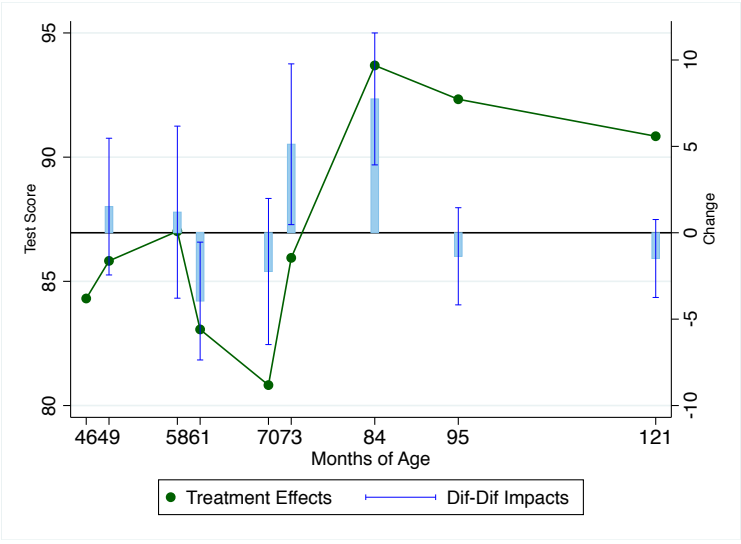
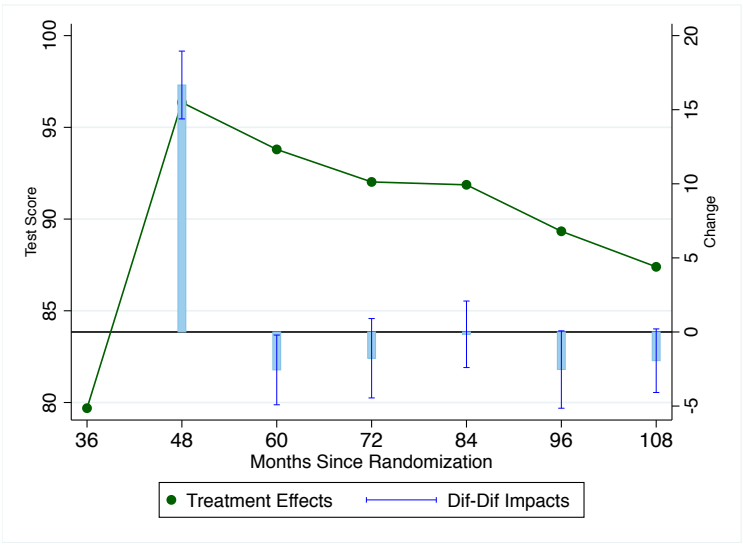
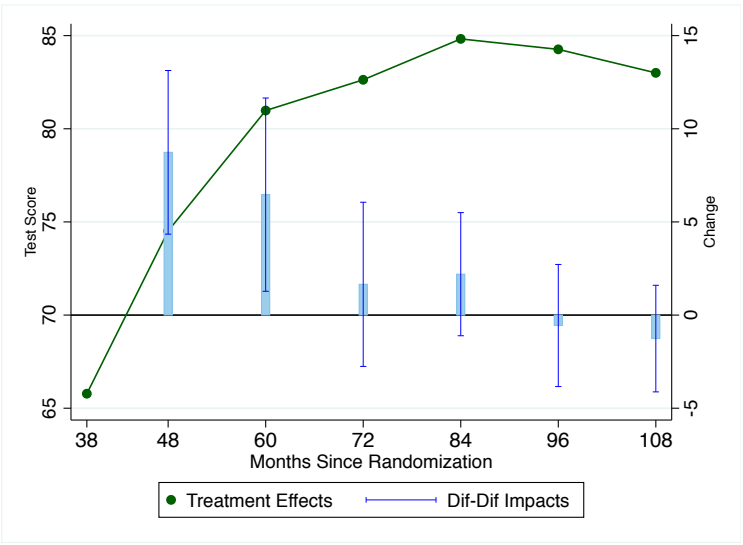


Figure 51: Perry, Standardized Scores Trajectory for Treatment Group

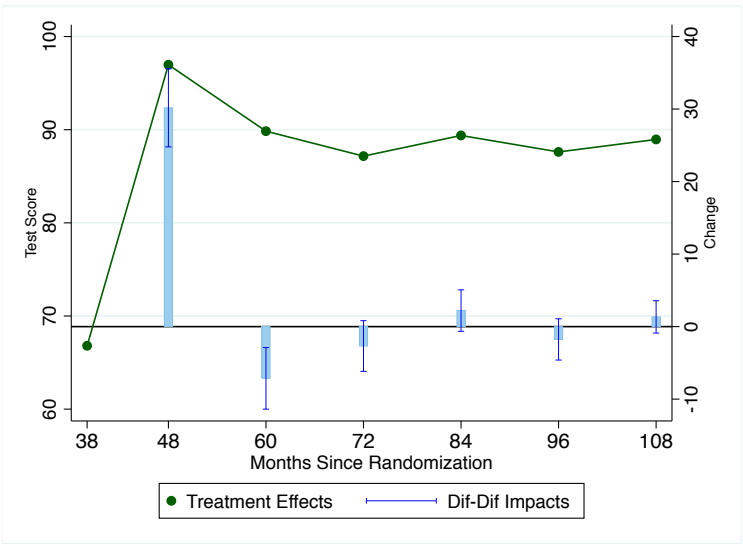
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

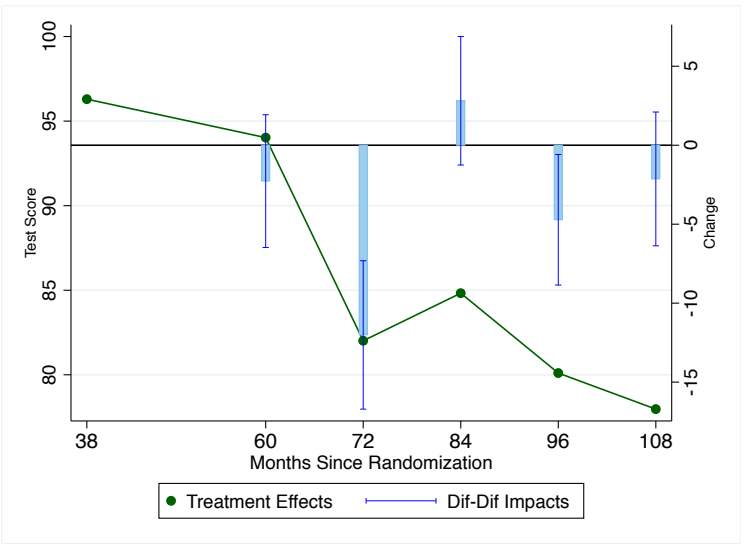
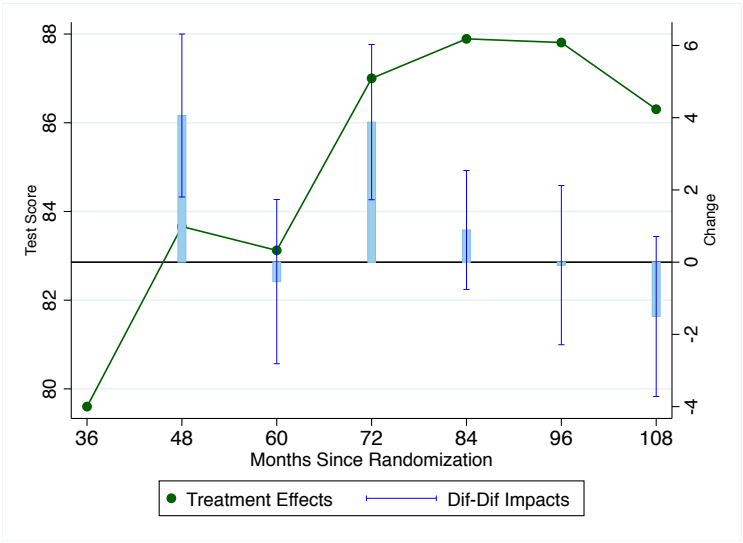
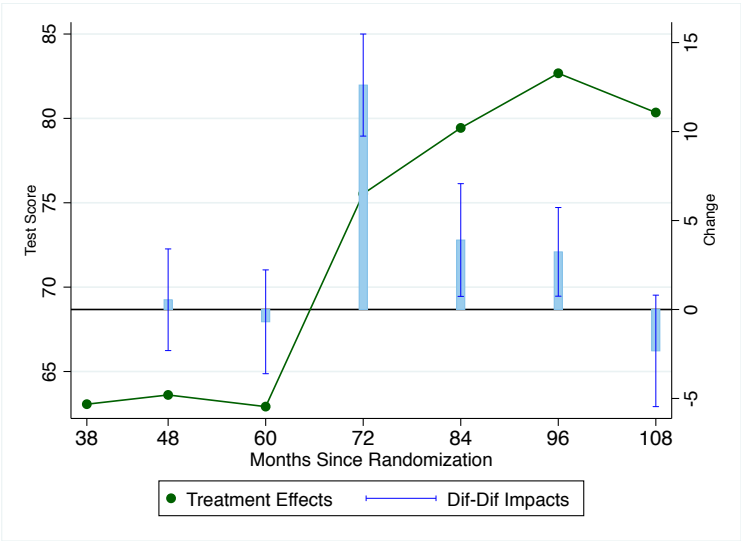


Figure 52: Perry, Standardized Scores Trajectory for Control Group

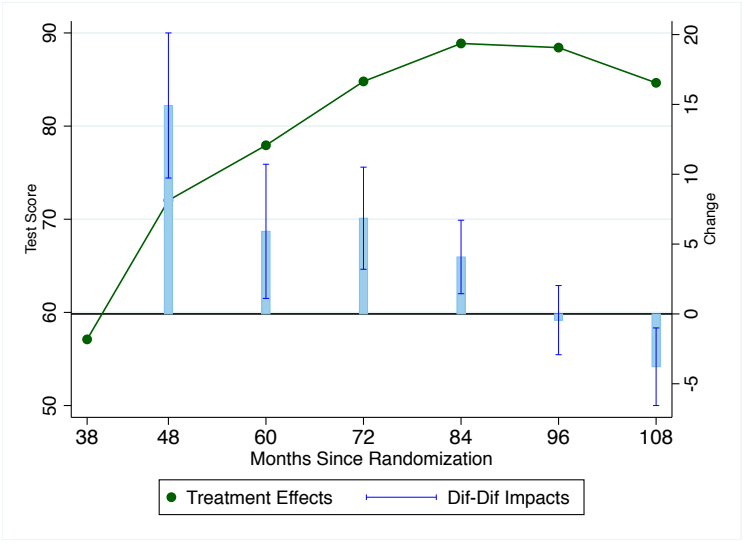
(a) Stanford Binet



(b) PPVT



(c) Leiter



(d) ITPA

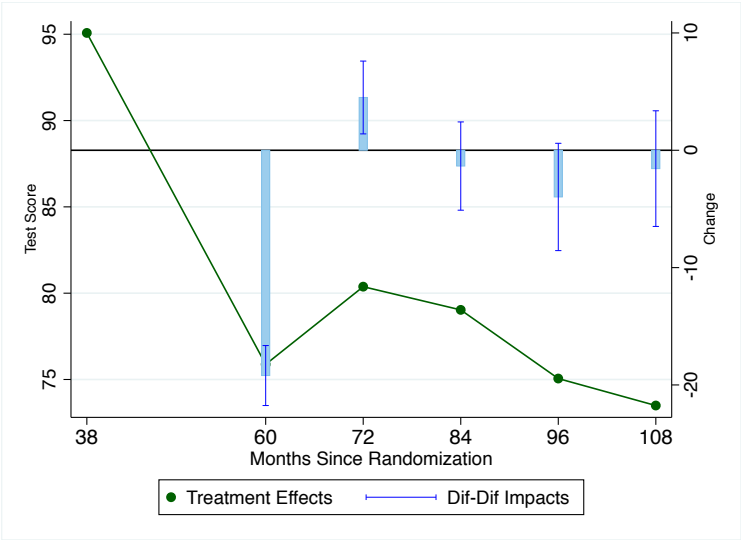
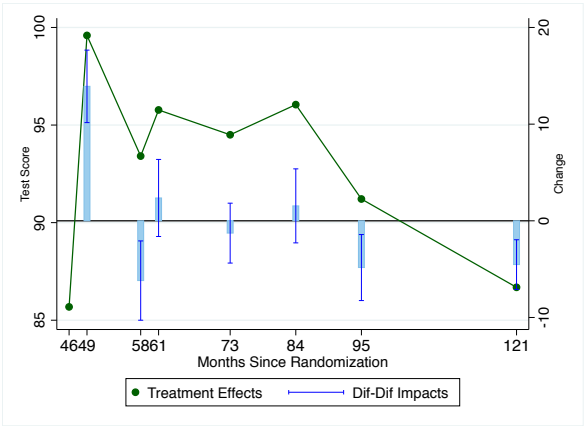
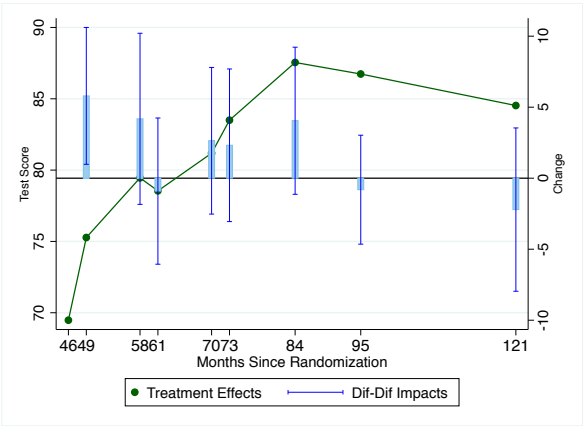


Figure 53: ETP, Standardized Scores Trajectory for All Groups

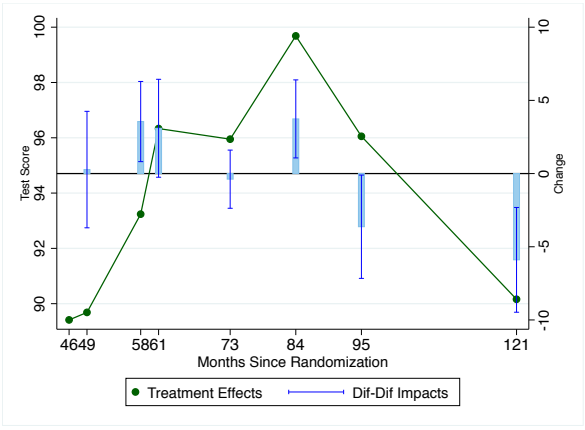
(a) Stanford Binet, 2-yrs Treatment



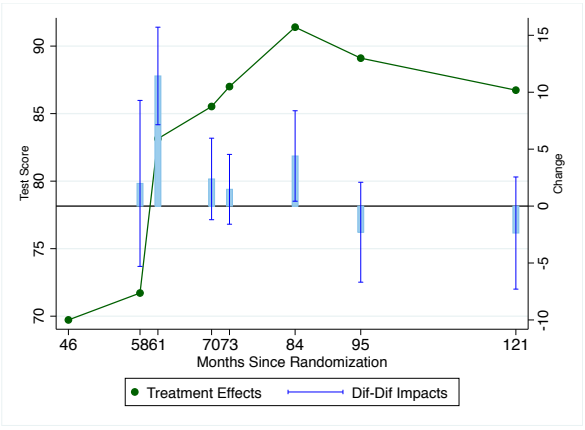
(b) PPVT, 2-yrs Treatment



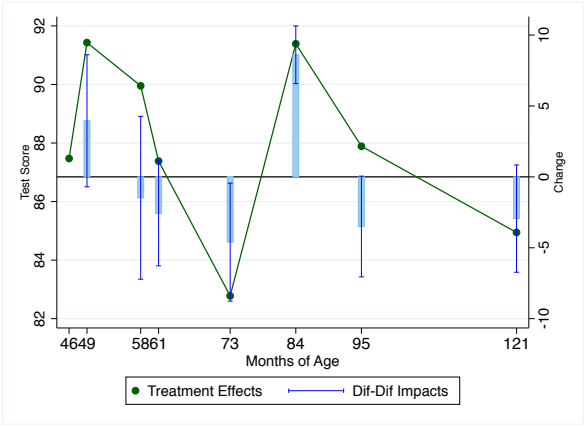
(c) Stanford Binet, 1-yr Treatment



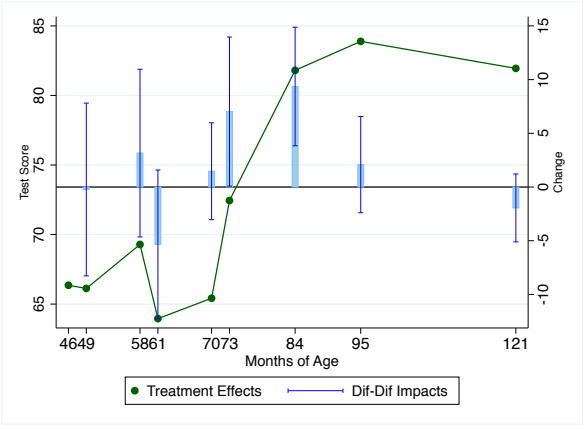
(d) PPVT, 1-yr Treatment



(e) Stanford Binet, Control



(f) PPVT, Control



M Evolution of Standardized Tests Using Differential Gains and Depreciation Models

In this section I assume that the raw tests can be represented by the measurement model in Equation (2). Age-standardized tests cannot be represented by it, because they are constructed to measure relative, not absolute levels of skills.[12] The age-standardized tests substact the mean for the age and divide by the standard deviation. We obtain:

$$M_{it}^{m,STD} = \frac{b^m}{\sigma_t^m}\left(\theta_{it} - E[\theta_t]\right) + \frac{\varepsilon_{it}^m}{\sigma_t^m} \tag{17}$$

Where $\sigma_t^m = \sqrt{(b^m)^2(\sigma_t^\theta)^2 + (\sigma_t^\epsilon)^2}$. Section 4 presents the evolution of the standardized scores across different ages for the control group. Using the assumptions discussed in Section 3, a measurement of the test for the control group reflects:

$$E\left[M_{it}^{m,STD}|R=0\right] = \frac{b^m}{\sigma_t^m}\left(E\left[\theta_{it}|R=0\right] - E[\theta_t]\right). \tag{18}$$

So the difference across two consecutive periods is:

$$E\left[M_{it}^{m,STD}|R=0\right] - E\left[M_{it-1}^{m,STD}|R=0\right] = \frac{b^m}{\sigma_t^m}\left(E\left[\theta_{it}|R=0\right] - E[\theta_t]\right) - \frac{b^m}{\sigma_{t-1}^m}\left(E\left[\theta_{it-1}|R=0\right] - E[\theta_{t-1}]\right). \tag{19}$$

We are interested in particular on the change in the level of the control group at school entry. To be able to interpret the estimates clearly, and for simplicity, I impose the approximation $\sigma_t^m = \sigma_{t-1}^m = \sigma^m$. The SD of each test is observable in the data, so we can know how far-fetched these approximations are. For Perry, $t-1 = 60$ months and $t = 72$ months. For ETP, $t-1 \sim 72$ months and $t \sim 84$ months. In the national norms for PPVT, the respective SDs are 8.17, 7.52 and 7.77. Then, these approximations have an error of 3–10%.

[12] Equation (2) has a fixed intercept for all ages, so higher skills will always translate into higher scores, regardless of the age. For age-standardized tests to be increasing in time more is required: they have to be each time larger than a mean that is also increasing in time. Moreover, we know how age-standardized tests are constructed, so we should be able to derive them from primitives.

Let $\bar{\Delta}V \equiv E[V|R=0] - E[V]$ for any variable V. The result is that the equations become:

$$E\left[M_{it}^{m,STD}|R=0\right] - E\left[M_{it-1}^{m,STD}|R=0\right] = \frac{b^m}{\sigma^m}\left(\bar{\Delta}\theta_t - \bar{\Delta}\theta_{t-1}\right). \tag{20}$$

The impacts using standardized scores are given in terms of the slope parameter over the standard deviations of the tests, $\frac{b^m}{\sigma^m}$, which is a magnitude that might be hard to interpret. A solution for that is to give the estimates as a proportion of the fadeout observed in the data:[13]

$$\frac{E\left[M_{it}^{m,STD}|R=0\right] - E\left[M_{it-1}^{m,STD}|R=0\right]}{\Delta M_{iP}^{m,STD} - \Delta M_{iT}^{m,STD}} = \frac{\bar{\Delta}\theta_t - \bar{\Delta}\theta_{t-1}}{\Delta\theta_P - \Delta\theta_T}. \tag{21}$$

The right-side hand estimate does not suffer from the interpretation difficulties, as it does not depend on the parameters of the specific test scores.

Interpreting the equations in the main paper for control-average US children difference instead than for the treated-control children difference, we can interpret the numerator in the light of both models we consider.

Taking the expectation of the difference between the control group and the average children in the Differential Gains Model in Equation (1), we obtain:

$$\bar{\Delta}\theta_t - \bar{\Delta}\theta_{t-1} = \beta_t\bar{\Delta}F_t + \beta_t\tau_t\bar{\Delta}F^t. \tag{22}$$

It is useful to remember that $\bar{\Delta}F_t = 0$ for the year of school entry, while $\bar{\Delta}F^t$ represents the difference between the control group and the average US children in previous participation in formal education. This is zero for control children in both programs. In the case of the Perry study, children entered school at age five, so $E[F^t]$ represents the percentage of children attending education at ages three and four in the US around 1967.[14] National statistics of preschool take-up for children of these ages in 1970 are available, and show a

[13] In the paper I define fadeout as the difference between the maximum and the minimum gap in skills: $\Delta\theta_P - \Delta\theta_T$

[14] The first cohort of Perry entered the program in 1962, but it is excluded from this graph. The next three cohorts entered in years 1963-1965. They entered school in years 1965-1967.

take-up rate of around 30% (Snyder et al., 2016). Then, $\bar{\Delta}F^t \sim 30\%$, implying that control children will benefit more than the average children from school entry because 30% of the US children had already experimented formal education. In the case of the ETP study, children entered school at age 6, and roughly 70% of US children entered Kindergarten at age 5 in 1970. This might be why the effect of school entry for ETP children seems stronger than the effect for Perry children.[15] It is possible to conclude that under this model, differences in previous learning opportunities, including a 30–70% of preschool attendance, imply that control children gain much more from entering schools than the average US children, even if those schools could have been of lower quality. The magnitudes of the impacts seem plausible and consistent with the rest of the evidence in the paper.

In the case of the Depreciation Model in Equation (3), we obtain:

$$\Delta\theta_t - \Delta\theta_{t-1} \quad = \quad -\delta_t \Delta K_{t-1} + \beta_t \Delta F_t. \tag{23}$$

It is hard to argue that a model in which depreciation depends on quality is reasonable here. Depreciation based on quality makes the most sense when children go from a high-quality program to a low-quality subsequent environment. Some of the literature assumes that this is an accurate description of the experiences of the children attending experimental programs. However, it does not have to be the case for the average US children. In general, the quality of the early childhood educational experiences should be positively related to the quality of schools. In that case, previous gains should generally not depreciate.

If depreciation is not based on quality, another possibility is that there is depreciation

[15] In practice, there are more factors in play. This simple model is fine for the comparison between treated and control children, because the quality of the schools and the quality of the home environments they experimented are similar. However, none of those factors are true in this case, and they deserve a further discussion. First, the fact that schools attended by children in the study might be of lower quality than schools attended by average children imply that if the gap narrows between the two groups at school entry, the effect of differential gains is being underestimated: children in the control group catch-up even though they participated in worse schools. Then, the effect that is generating the catch-up has to be of an even higher magnitude. On the contrary, the fact that the environments of control children were probably of lower quality than the environments implies that the 30% difference in previous schooling (70% for ETP) is only a lower bound for the difference in previous learning opportunities.

from previous gains for early childhood education programs in all cases. However, as mentioned before, relatively few (30%) of children experienced this type of education around the time the children in ETP and Perry entered school. Thus, it is hard to believe that the very large catch-up of control children compared to america children happened for this reason.

None of the two explanations based on depreciation seem plausible in explaining the fadeout observed at school entry: It is hard to believe that the average US children at the time suffered a strong decrease in quality from their preschool experiences to their school experiences. That implies that depreciation based on quality is not a convincing explanation. Given that the percentage of children attending preschool education was relatively low at the time, an explanation based on depreciation of any gains from preschool programs is also hard to sustain.

References

Arthur, G. (1952). *The Arthur Adaptation of the Leiter International Performance Scale*. Washington, D.C.: The Psychological Service Center Press.

Barlevy, G. and D. Neal (2012). Pay for percentile. *The American Economic Review 102*(5), 1805–1831.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology 54*(1), 1.

Cunha, F. and J. J. Heckman (2008). Formulating, Identifying, and Estimating the Technology of Cognitive and Non-cognitive Skill Formation. *Journal of Human Resources 43*(4), 738–782.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Dunn, L. M. (1965). *Expanded Manual Peabody Picture Vocabulary Test*. Minneapolis, MN: American Guidance Service, inc.

Leiter, R. G. (1940). *The Leiter International Performance Scale*. Santa Barbara, CA: Santa Barbara State College Press.

Leiter, R. G. (1952). *The Leiter International Performance Scale*. Washington, D.C.: The Psychological Service Center Press.

McCarthy, J. J. and S. A. Kirk (1961). *Examiners Manual Illinois Test of Psycholinguistic Abilities Experimental Edition*. Urbana, IL: Institute for Research on Exceptional Children.

Snyder, T., C. de Brey, and S. Dillow (2016). *Digest of education statistics 2014*. National Center for Education Statistics.

Terman, L. M. and M. A. Merrill (1960). *Stanford-Binet Intelligence Scale*. Cambridge MA: The Riverside Press.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children*. New York, NY: The Psychological Corporation.

Wechsler, D. (1951). Equivalent test and mental ages for the wisc. *Early Publication*.