

# Child Development Production Functions

Fali Huang

School of Economics and Social Sciences

Singapore Management University

90 Stamford Road

Singapore 178903

Email: fhuang@smu.edu.sg

Tel.: 65-68280859

Fax: 65-68280833

February 28, 2006

## Abstract

This paper estimates production functions of child cognitive and social development using NLSY79 child data. A sample of eight- and nine-year old children is constructed, which includes over two hundred home and school inputs starting from mother's prenatal care period as well as family background variables. A tree structured regression method is used to conduct estimation under various specifications, explicitly allowing for unobserved heterogeneity and non-linear structures. A small subset of earlier and current home inputs are found consistently important in explaining variances of child development results, while family backgrounds variables are seldom selected as primary predictors. The effect of maternal employment is negligible when detailed inputs are controlled, and the score gaps across race can be almost completely accounted for by home and school inputs.

*Key words: tree regression method, CART, child development.*

## 1 Introduction

An old Chinese saying claims that a person's lifetime achievements can be well predicted by his performance at age seven. Recent evidence suggests this may still be true in our modern times. For example, Currie and Thomas (1999) find that in Britain a person's test scores at age seven are significantly associated with his education level and earnings in thirties. Based on NLSY data, Keane and Wolpin (1997) show that skill endowment heterogeneity at age sixteen may account for ninety percent of the total variance of individual lifetime earnings. One possible reason for the vital importance of skill formation in early childhood is that "success or failure at this stage feeds

into success or failure in school which in turn leads to success or failure in post-school learning” (Heckman 1999).

The main goal of this paper is to estimate production functions of child cognitive and social development results. Our estimation strategy is guided by a conceptual framework where child development process is an integral part of a multi-period optimization problem: In each period starting at least from a mother’s prenatal care until the child is independent, parents choose various home and school inputs to maximize their remaining life-time utility, taking into consideration child endowments, family resource constraints, and the production functions of child development results, etc. Since home and school inputs are often endogenous choices of parents and hence are correlated with each other and across periods, any omitted inputs would necessarily cause bias in estimated effects of included ones. To mitigate this problem, many studies use family background variables as proxy. But as we will show in the paper, this would also lead to biased estimation if there is unobserved heterogeneity in family or child, or if there are non-linear interactions among inputs. Since different input choices may come out from identical family backgrounds, it is not surprising that the standard measures of family backgrounds cannot explain much about skill endowment heterogeneity (Keane and Wolpin, 1997).

To minimize the omitted variable problem, this paper uses a comprehensive set of detailed inputs, and adopts various within-child difference and value-added specifications to deal with unobserved heterogeneity. The data requirement of our estimation is best met by NLSY79 child data which contains rich information on age-appropriate home and school inputs. We construct a sample of 4726 eight- and nine-year-old children where each child has over two hundred inputs starting from mother prenatal care until current period. Even with this rich data set, one may still worry about omitted variable problems. To check this, we include in the regressions many family background variables, which should be correlated with omitted inputs. Since they never show up as primary explaining predictors, it suggests that the influence of omitted variables, if any, is very modest.

Current scientific knowledge, however, does not tell us which inputs among the two hundred plus available ones in the data affect child development and how they may interact with each other. Furthermore, most inputs are measured by categorical variables with multiple items, and many contain missing values. Researchers faced with these problems are often forced to choose, quite arbitrarily, which variables to include in and what structures to impose on production functions, how to combine different categories, and how to handle missing variables. Important information

may be lost during this subjective data reduction process (Harvey 1999), and different treatments per se may give rise to discrepant estimation results even for the same data (Haveman and Wolfe 1995). This motivates us to adopt a non-parametric method, namely tree structured regression, in estimation. The optimization mechanism underlying the tree regression is similar to the conventional linear regression; its non-parametric features are designed to select important explanation variables, detect non-linear structures, and treat missing values in a systematic way (Breiman, Friedman, Olshen, and Stone 1984).<sup>1</sup>

Our estimation results in various specifications yield a consistent picture: A reasonably small set of earlier and current inputs are important predictors for child development results at age eight or nine, while family background variables are never selected as primary predictors. The number of books a child has at various ages and how often a mother reads to her child before age 5 are the most important predictors of child math and reading scores from age five onwards; they are also primary predictors of child behavior problem scores at age 5. Spanking a child aged 8-9 seems to be mostly driven by some family or child fixed factors and have no effects on current result, while spanking a younger child may reduce his future behavior problems. When detailed home and school inputs are controlled, a mother's working hours in the first five years of a child's life have little effects on any child development results at age eight or nine; so does a child's race or sex.

Our results also suggest that, though earlier scores are very important in predicting future results, they may not be sufficient statistics for missing earlier inputs. Indeed, some earlier inputs have direct effects on current child performance beyond what is already captured by earlier scores. For example, the number of books a child has at age seven and how often his mother reads to him around age three are primary indicators of his math and reading scores at age nine, even when corresponding scores at age seven are controlled. We also find that production functions of cognitive scores are more stable over time than that of behavior problem scores; the latter is prone to the influence of home environment mostly at younger ages.<sup>2</sup>

There is a large body of literature studying the contributing factors to child cognitive and social-emotional development. A lot of work examines the effects of maternal employment on early child development results (Blau and Grossberg 1992, Parcel and Menahan 1994, Harvey 1999, Ruhm

---

<sup>1</sup>This is the first time regression tree analysis is used extensively in child development area, though it has been applied in medical, biological, weather forecasting and many other areas. In economics and finance literature, it was used occasionally to detect regime shifting patterns (e.g. Durlauf and Johnson, 1995).

<sup>2</sup>In a related study, Sacerdote (2002) finds that one's cognitive skills have more to do with nature while social ones with nurture.

2000, Neidell 2000, Waldfogel, Han, and Brooks-Gunn 2002, Baum 2003). Most of these studies use linear regression models taking control of various family background variables. Their findings are mixed and no consensus has been reached, probably due to ad hoc selection of control variables and omitted variable problems mentioned above (Haveman and Wolfe 1995, Todd and Wolpin 2003). The distinct features of the current paper are that we use a comprehensive set of detailed inputs, adopt a non-parametric tree regression to systematically detect important inputs and non-linear structures, and use within-child difference and value-added specifications to treat child-specific unobserved heterogeneity. We find no effects of mother working on child development once detailed inputs are included. This is consistent with the finding of Waldfogel et al. (2002) that the effects of maternal employment are much reduced after a couple of home quality indicators are controlled. A possible explanation is that a mother's time and attention to children are not much affected by her labor force participation (Bianchi 2000). A related interpretation is offered by our theoretic model: Maternal employment and other family backgrounds act as state variables and hence are less informative than the direct detailed inputs in predicting child development results.

Another important strand of literature examines the role of child care in child development (NICHD ECCRN 1998, 2000, 2001, 2003). These studies find that the overall quality of child care is consistently but modestly related to cognitive and social-emotional outcomes during the first three years of life; while family factors are "more consistent predictors of children's outcomes than any aspect of early non-maternal care experiences" (NICHD ECCRN 2001). This finding is also confirmed by our estimation results for children up to nine years old, where the type of child care, the duration and intensity of child care experiences in the first three years never show up as primary explaining variables.

The paper shows that the racial gaps of math and reading scores among eight and nine year-old children can be completely accounted for by home and school inputs. This is a striking result since a substantial Black-White test score gap persists even after controlling for a wide range of characteristics (Fryer and Levitt 2003, Todd and Wolpin 2004). We achieve this by using a more comprehensive set of detailed inputs and a non-parametric method to systematically select important explanatory variables. Similar to our results, the number of children books is found to be most important in explaining math and reading score gaps among Black and White children by Fryer and Levitt (2003), while the importance of earlier home environments for child cognitive development is demonstrated by Todd and Wolpin (2004).

The paper is organized as follows. In the next section the conceptual framework and its implications for estimation are discussed. The child history sample is described in section 3. Regression tree analysis and various estimation methods are discussed in the subsequent two sections. The results of regression trees are presented in section 6. The final section concludes.

## 2 The Conceptual Framework

The process of child development is a very complicated one. The following two period optimization model serves as the general conceptual framework in which our estimation is embedded. In order to focus on parent-child interaction, a mother is assumed to be the planner of a family, and bargaining among adult family members is omitted. To ease notation, ‘he’ is used for child and ‘she’ for mother. A mother’s time constraint in each period  $t \in \{1, 2\}$  is

$$h_t + H_t + L_t = 1, \tag{1}$$

where her total time is normalized to one and divided among working hours  $h_t$ , quality time with child  $H_t$ , and leisure time  $L_t$ .<sup>3</sup> Her budget constraint is

$$c_t + p_k k_t = w_t h_t + I_t, \tag{2}$$

where  $c_t$  is household consumption,  $k_t$  denotes market goods and services relevant in child development (such as books, toys, child care, and enriching activities),  $p_k$  the average price of  $k_t$ ,  $w_t$  mother’s wage,  $I_t$  household income net of mother’s wage.

For simplicity we ignore possible interactions between cognitive and social development. Let  $g_t$  denote a development result at period  $t$ . A child’s performance in the first period,  $g_1$ , is affected by  $k_1, H_1, h_1$ , in addition to his innate ability  $A$  that is stable across the two periods.<sup>4</sup> The second period score  $g_2$  is not only affected by current inputs  $k_2, H_2, h_2$ , but also by historical inputs  $k_1, H_1, h_1$ . Specifically, we assume

$$g_1 = G^1(A, k_1, H_1, h_1), \tag{3}$$

$$g_2 = G^2(A, k_2, H_2, h_2; k_1, H_1, h_1). \tag{4}$$

---

<sup>3</sup>This means that a mother could spend her non-working time either pursuing her own interests (say, watching TV programs for adults) or actively interacting with child to improve his development (such as reading to him). Bianchi (2000) shows empirical evidence in this matter.

<sup>4</sup>Note that  $h_t$  also represents a mother’s absent time from home, which may have its own effects on child development (Schore 1996).

Estimating these two production functions is the main goal of this paper.

A mother’s instant utility function is  $u(c_t, L_t, H_t, g_t)$ . To explicitly solve the optimal input choices, we adopt a simple functional form

$$u(c_t, L_t, H_t, g_t) = \log c_t + \alpha_1 L_t + a(\alpha_2 g_t + \alpha_3 \log H_t), \quad (5)$$

where  $a \in R_+$  is a mother’s altruism level, and  $\alpha_1, \alpha_2, \alpha_3, \in R_+$  are her preferences for leisure, child performance, and quality time with child, respectively. At each period, she will choose optimal inputs to maximize her remained life time utility discounted by  $\beta$ , subject to time, financial, and technical constraints listed above. The following proposition summarizes this conceptual framework’s empirical implications on omitted variable problem. The proof is in the appendix.

**Proposition 1** *Any omitted inputs will cause bias in the estimated effects of included ones since all inputs are correlated with each other. Using family background variables as a proxy for detailed inputs will also lead to biased estimation if there are non-linear interactions among inputs or unobserved heterogeneity in either family or child.*

The intuition is as follows. The home inputs are functions of the same set of state variables which are also inter-temporally linked, so any omitted inputs must be correlated with included ones and thus cause bias in the estimated effects of the latter. When there is unobserved heterogeneity or non-linear interactions among inputs, the mapping from family background variables to home input choices is no longer one-to-one. In other words, identical family backgrounds now lead to different inputs, so controlling the former can no longer effectively control the latter. This may explain why no consensus is reached on the effects of mother working despite so many studies. Similarly, the same family backgrounds now yield quite different child development results, which may be the fundamental reason why they are not very useful in explaining skill endowment heterogeneity at age 16 (Keane and Wolpin, 1997). So a comprehensive set of detailed home and school inputs should be used in order to get unbiased estimates.

Note family backgrounds act as state variables which affect inputs in child development production functions. For example, family incomes affect child development through purchasing relevant goods and services, and parent education levels matter when they are reflected in parent-child interactions. They would become irrelevant once relevant detailed inputs such as how many books parents buy for the child, how often they read to him, and how they treat him in various situations,

are included. This is indeed supported by our estimation results, in which no family background variables are selected as primary predictors when detailed inputs are controlled. A similar result in Feinstein and Symons (1999) shows that family background variables become non-significant once parents' interests in child education are included.

### **3 Data: Childhood History Samples**

Starting from 1986, children born to mothers of National Longitudinal Surveys 1979 Youth (NLSY79) are surveyed every two years. The set of child development results and inputs from birth up to 10 years is different across three age-groups, namely 0-2 years (A group), 3-5 years (B group), and 6-9 years (C group). We construct a sample of 4726 children each of whom has a single pair of comparable scores and inputs at ages 6-7 and 8-9, plus historical inputs at ages 0-5, mother prenatal care, mother working history up to the fifth year after child birth, and family backgrounds. This childhood history sample is called CCBA sample on which most of our regressions are based. We also assemble a BA sample of 7075 children around age five.

For children five years old and above, the cognitive development is measured by PIAT (Peabody Individual Achievement Test) math and reading recognition scores, while social and behavioral development is measured by BPITS (Behavior Problem Index Total Scores). For younger children, PPVT (Peabody Picture Vocabulary Test) is another widely-used measure of cognitive development.

A brief introduction of home and school inputs as well as maternal working and family backgrounds is in order (details in appendix). For children age three and above, home environment variables include how many books a child has, how often a mother reads to child, how often a father plays with him outdoors, whether there are music instruments and newspapers at home, whether parents encourage hobbies and bring a child to enriching activities, how often the family gets together with relatives and friends, how often a child watches TV and attends religious service etc. For younger children, home inputs also include the number of various toys a child has, how often he is talked to and taken to grocery shopping, whether he was breast-fed and how long the breast-feeding lasts, whether a mother teaches a child letters, numbers, and shapes, etc. The child care experiences in a child's first three years are measured by whether a child is in regular child care, the type of child care, the months per year and hours per week in child care. Variables about parenting styles include how often a child is expected to make bed, to clean room, etc.; and how mother responds

to low grades, tantrums, and hits. Mother prenatal care variables include mother usage of alcohol, cigarettes, vitamins, and sonograms during pregnancy, whether a child has low birth weight. Maternal employment history covers a mother's working hours one year before child birth up to the fifth year afterwards. Family backgrounds include a mother's highest grades, AFQT scores, age at child birth, her marital status, her wages, the salary income and total family incomes, the region the family lives, whether the mother ever lived in the south in her youth, whether there is a father figure at home, whether he is the biological father, etc. The school inputs include school types, number of hours a child working on homework, mother's rating of school quality including teacher skill and caring of students, safety of school, moral teaching, etc. Since school inputs for children below 10 years old are not available before 1996, there are quite a lot missing values in school inputs. There are also various degrees of missing values in other variables, where in general earlier inputs have more missing values. The treatment of missing values is introduced below.

## 4 Regression Tree Analysis

There are several reasons that motivate us to choose the regression tree analysis instead of the standard linear regression method. First, the current scientific knowledge does not provide any ex ante parametric model for child development production functions. We simply do not know which inputs are relevant among the two hundred plus inputs in the data. Secondly, most of these input variables are categorical with multiple items. Given our sample size, it is infeasible for the linear regression method to keep all the categories, even if we assume away all interaction terms and non-linear structures. Furthermore, as mentioned above, many variables contain missing values. Compared with the unavoidable randomness in choosing which variables to include, how to reduce data, and how to fill in the missing values, the non-parametric tree regression offers a more systematic way to handle these problems.

A regression tree is a piecewise constant or piecewise linear estimate of a regression function constructed by recursively partitioning the data (Breiman et al. 1984, Loh 2002). At each stage the binary partition that minimizes the sum of the squared errors is selected. It gets its name from the practice of displaying the partitions as a decision tree. By optimally splitting the sample into different subsets, the tree regression automatically performs internal feature selection and detects non-linear structures in the data. It is thereby resistant to the inclusion of many irrelevant predictor

variables and helps to solve model/variable selection problem. The binary splitting process naturally incorporates mixtures of numerical and categorical predictor variables, makes it immune to the effects of outliers, and localizes the influence of missing values (Hastie et al., 2001).

There are various ways in conducting tree regressions. The one used in our estimation is the CART (Classification And Regression Tree) method established by Breiman et al. (1984), which is also the most widely used tree-building methodology according to Hastie et al. (2001). A brief description of the CART methodology is presented below, while a more technical note is in the appendix.<sup>5</sup> A regression tree is grown by sequentially splitting the sample into binary nodes/subsets. At each node, every explaining variable competes in its ability to reduce the node variance. The variable with the best improvement score is selected to be the *primary splitter*. Unlike the linear regression, here not all available predictors show up as primary splitters: The primary splitters must be the best predictive variable at some node. The tree growing process continues until the variances in all nodes are less than some threshold. The nodes where splitting stops are called *terminal nodes*, the number of which measures the scale of the tree. In a terminal node the predicted value for the dependent variable is the sample mean of the node.

The mother tree thus grown has the most terminal nodes and accordingly the minimum (within-sample) resubstitution errors one can get from the sample data. But it is generally much larger than data warranted and the estimated error is overly optimistic.<sup>6</sup> It is then pruned backwards to get a sequence of trees with different number of terminal nodes. The pruning rule balances the trade-off between the out-of-sample resubstitution errors and the complexity of the trees. From this whole sequence the optimal tree is selected using *test sample* or *cross-validation* error estimate.

Take the test sample estimate as an example. A subset is selected randomly from the whole sample as the *learning sample* to grow the mother tree, while the remaining one is the independent test sample. Using each subtree on the test sample to predict the dependent variable and calculate the resulted sum of squared errors. The subtree with the minimum relative test error (compared to mean squared error) is selected as the optimal tree. The *importance* of a variable measures its overall ability as a primary splitter in an optimal tree to explain the variance of dependent variable in all nodes. It corresponds to the variable's total effect on the dependent variable.

---

<sup>5</sup>The tree regression can be conducted using the commercialized software named after CART, in a similar manner as implementing linear regressions using a standardized statistical software say STATA.

<sup>6</sup>Similarly, the  $R^2$  in standard linear regression methods also measures the within-sample relative substitution error and its explaining power is necessarily over-optimistic for out of sample observations.

CART’s missing value algorithm is designed to minimize the influence of missing values and make maximal use of the available data. At each node when all explaining variables compete in reducing node variance, those with missing values have their improvement scores calculated using the subset of complete data. If an observation has missing values so that the primary splitter is not defined for that case, CART uses the best surrogate to split it. That is, among all non-missing variables in the case, the one that best mimics the splitting result of the primary splitter is used.<sup>7</sup> So the missing values are not filled in; instead, their predicting roles are carried out by non-missing ones of other related variables. Thus, the efficiency of missing value algorithm mostly depends on whether surrogates are available in the data and how they are correlated with the missing values. Both conditions are met in our data since there are many home inputs which correlated with each other, and the missing school inputs are correlated with family backgrounds and home inputs.

The regression tree is consistent in the sense that as the sample size goes to infinity, the finite-node tree converges to the true structure of the data. Overall, the accuracy of regression tree has been generally competitive with linear regression. It can be much more accurate on nonlinear problems, though it tends to be somewhat less accurate on problems with good linear structure. This is not a problem for our case since most variables are categorical. A main drawback of the tree regression, determined by its non-parametric characteristics, is that it does not provide the traditional statistical significant tests. However, variables selected as primary splitters tend to be also significant in the conventional sense since they are the best predictive variables among many available ones and their predictive abilities are further tested on a random test sample.<sup>8</sup>

## 5 Estimation Methods

We run tree regressions according to various specifications used in the literature so that we can tell, among other things, whether earlier inputs affect child development in addition to current ones; whether former scores are sufficient statistics for missing earlier inputs; and whether home inputs

---

<sup>7</sup>This procedure is analogous to replacing a missing value in a linear model by regressing on the non-missing value most highly correlated with it. However, tree regression is more robust since the number of cases affected at any given node is generally small in a high-dimensional problem.

<sup>8</sup>Though it is technically infeasible to replicate the tree analysis using linear regression for the current project, I do try to use the linear regression method by reducing the categorical variables to binary versions, ignoring any interactive terms, and omitting missing values. The main results are similar: the family background and child care variables are rarely significant, while the home inputs selected as primary explaining variables in the tree regressions are often significant in the OLS as well.

have their own effects on child outcome other than just reflecting family- and child-specific fixed effects.<sup>9</sup>

For any child  $i$ , suppose each input  $X_i$  can be either high or low. Let  $X_h \equiv I\{X_i \text{ is high}\}$  denote the index function for any input  $X_i$  being high, and similarly  $X_l \equiv I\{X_i \text{ is low}\} = 1 - X_h$ . Child development functions (3) and (4) in our conceptual framework are assumed to be

$$g_{1i} = \alpha_{1a}A_h + \theta_{1a}A_l + \alpha_{1k}k_{1h} + \theta_{1k}k_{1l} + \alpha_{1ak}k_{1h} * A_h + \varepsilon_{1i}, \quad (6)$$

$$g_{2i} = \alpha_{2a}A_h + \theta_{2a}A_l + \alpha_{2k}k_{2h} + \theta_{2k}k_{2l} + \alpha_{2ak}k_{2h} * A_h \\ + \alpha_{21k}k_{1h} + \theta_{21k}k_{1l} + \alpha_{21ak}k_{1h} * A_h + \varepsilon_{2i}, \quad (7)$$

where subscript 2 indicates the second period and 1 the first period,  $\alpha$ 's and  $\theta$ 's are coefficients with  $\alpha > \theta$  for all subscripts,  $A$  denotes child ability and  $k$  a home input, and  $\varepsilon_{2i}$  and  $\varepsilon_{1i}$  are random noises with mean zero and cdf  $F(\cdot)$ .

These two functions are quite general in that child ability has different effects across periods and there are non-linear interactions between ability and measured inputs. For example, the marginal return of input  $k_{1h}$  on score  $g_{1i}$  varies with child ability: it is  $\alpha_{1k} + \alpha_{1ak}$  for a high ability child and  $\alpha_{1k}$  for a low ability one. When there are no omitted variables or when they are orthogonal to included ones, tree regression will yield unbiased estimates of all these parameters.<sup>10</sup> Since there is bound to be unobserved heterogeneity in family or child, we need within-child difference and value-added treatments to mitigate the problem.

## 5.1 Regressions with Current and Historical Inputs

The simplest estimation method is to use only contemporaneous home and school inputs in explaining child development results. The optimal regression tree under this specification is *current input regression*, which yields unbiased estimates when only current inputs matter and they are unrelated to unobserved child ability. That is,  $\alpha_{21k} = \theta_{21k} = \alpha_{21ak} = 0$  must hold. The *historical input regression* is the optimal tree when both current and historical inputs are included, which is unbiased when all inputs are unrelated to unobserved child ability. If some historical inputs act as primary splitters, then current input trees are biased.

<sup>9</sup>Following Todd and Wolpin (2003), appropriate assumptions are listed for each estimation method to yield unbiased estimates.

<sup>10</sup>See proposition 3 in Appendix B.

## 5.2 Within-Child Difference Regression

The *within-child difference regression* is the optimal tree with  $g_{2i} - g_{1i}$  as the regressant and all inputs as regressors. The corresponding regression equation is

$$\begin{aligned} g_{2i} - g_{1i} = & (\alpha_{2a} - \alpha_{1a})A_h + (\theta_{2a} - \theta_{1a})A_l + \alpha_{2ak}k_{2h} * A_h + (\alpha_{21ak} - \alpha_{1ak})k_{1h} * A_h \\ & + \alpha_{2k}k_{2h} + \theta_{2k}k_{2l} + (\alpha_{21k} - \alpha_{1k})k_{1h} + (\theta_{21k} - \theta_{1k})k_{1l} + \varepsilon_{2i} - \varepsilon_{1i} \end{aligned}$$

It eliminates the effects of unobserved child ability and gets unbiased estimates when (1) child ability has the same effect on scores  $g_{2i}$  and  $g_{1i}$ , that is, when  $\alpha_{2a} = \alpha_{1a}$  and  $\theta_{2a} = \theta_{1a}$  hold; (2) it does not interact with inputs in any non-linear way, that is,  $\alpha_{2ak} = 0$  and  $\alpha_{21ak} = \alpha_{1ak}$  hold; and finally (3) input choices in the second period do not depend on first period's random shock  $\varepsilon_{1i}$ . The third condition can be explicitly checked since we have variables about how a mother responds to low grades and behavioral problems. It holds for cognitive development production functions but not so well for social and behavioral development. To further check the degree of endogeneity problem for disciplines, a regression is run using the behavior scores at age 7 as the dependent variable, while including future inputs at age 9 as well as current and earlier inputs. Age 9 inputs would appear as primary splitters when the endogeneity problem is severe. But the opposite is true, where no age 9 inputs appear as primary splitters. Actually the optimal regression tree does no change before and after including these future inputs. This evidence suggests that the endogeneity problem is too weak to affect the main results.

Note that for within-child difference trees, we can only recover the differenced effects on  $g_{2i}$  and  $g_{1i}$  for historical inputs, but not absolute effects. For example, if the estimate for  $\alpha_{21k} - \alpha_{1k}$  is zero, it could be either that  $\alpha_{21k} = \alpha_{1k} > 0$  or  $\alpha_{21k} = \alpha_{1k} = 0$ . Another problem is that the variance of the new error term  $\varepsilon_{2i} - \varepsilon_{1i}$  goes up in general, hence the errors of within-child difference trees. This may cause non-existence of optimal test-sample validated tree. Furthermore, we cannot get any information about the effects of unobserved child ability. Since it is possible that child ability or its effects on scores change over time during the early years, within-child difference may not completely eliminate the effects of unobserved ability.

## 5.3 Value-Added Regressions

Another way to allow for unobserved heterogeneity is using an earlier score as a proxy. The *value-added historical regression* uses an earlier score in addition to current and historical inputs. The

widely used value-added estimation method, however, uses only current inputs as regressors, plus an earlier score as a proxy for both child ability and historical inputs. If earlier inputs act as primary splitters in a value-added historical tree, it suggests that they have direct effects on future child development and hence the earlier score is not a sufficient statistic for them.

The basic logic of finding appropriate assumptions for these two value-added trees is the same as in Todd and Wolpin (2003). The technical details are, however, different in that we use *subsets* instead of exact values of an earlier score. Let the support of the first period noise term  $\varepsilon_{1i}$  be  $[-\delta_1, \delta_2]$ , where  $\delta_1, \delta_2, \in R_{++}$ . The following condition

$$\delta_1 + \delta_2 \leq (\alpha_{1a} - \theta_{1a}) - (\alpha_{1k} - \theta_{1k}) \tag{8}$$

means that the difference between two children’s earlier scores caused by random noise and inputs is not bigger than that caused by child ability. The huge effects of earlier scores seem to confirm the plausibility of this assumption.

**Proposition 2** *Under condition (8) the earlier score  $g_{1i}$  is a perfect proxy for child innate ability so that the value-added historical tree of  $g_{2i}$  yields unbiased estimates.*

**Proof.** See the appendix. ■

In value-added regression trees, the effects of ability are allowed to change over time and non-linear interactions between ability and inputs are also accommodated. All of these effects are identifiable. Suppose another earlier score, say  $g_0$  at period zero, is available. If  $g_0$  has a smaller effect than  $g_1$  in their respective value-added historical trees, one explanation is that child ability or its effects change over time. Furthermore, if some inputs have large effects when  $g_0$  is included instead of  $g_1$ , then it is possible that child ability is affected by these inputs. In other words, some historical inputs, though not having direct effects on  $g_2$ , may still be important to  $g_2$  by affecting  $g_0$  or  $g_1$ . Thus we also grow historical input trees for  $g_0$  to see which inputs affect  $g_0$ .

## 6 Estimation Results

A series of regression trees described above are grown using the CCBA sample for PIAT math, reading, and BPI scores for children of age 9.<sup>11</sup> When a child’s race matters in current input trees, trees excluding race are grown for comparison. Historical input trees are also grown on the BA

---

<sup>11</sup>Detailed regression trees are available upon request.

sample for various scores at age 5. The main regression results are summarized in tables 1-4. In each table, the importance measures of primary splitters for the included regression trees are listed. An input's importance in a specific tree is equal to the aggregate variance it reduces as primary splitters at various nodes, which corresponds to its total effect on the dependent variable.

## 6.1 Math Scores for Children at Age 9

Table 1 summarizes regression results for six regressions of PIAT math scores for nine years old children. In the current input regression (Current), race is the most important predictor. When it is excluded (in C/race), however, the relative error increases by only 0.02, while the effects of books and TV watching hours at weekdays are almost doubled. The degree of physical affection shown by mother to child as well as several other home inputs become primary splitters. When earlier inputs are included (in History), the importance of race is halved; when race is excluded (in H/race), the errors on the test sample decrease. In the mean time, many other inputs become primary splitters including special activities and TV watching hours at weekdays, while the importance of existing primary splitters has little change. This suggests that most effects of race on math scores can be accounted for by earlier and current inputs a child receives; race is a good proxy for detailed home inputs when they are not available, but becomes less so when home inputs are available.

Comparing regressions with and without historical inputs, it is clear that earlier inputs are important to PIAT math scores.<sup>12</sup> In general, the importance of current inputs goes down in the historical input tree, which suggests that their estimated effects in current input tree are biased upward. For example, the importance of the number of books at age 9 is greatly reduced (from 15.3 to 1.29) while the number of books at age 7 becomes the most important predictor (with importance 17.4).

In the value-added historical regression where PIAT math score at age 7 is used (VAY), earlier inputs such as books at age 7 and mother reading to child at age 3-5 still have positive effects, though their importance is reduced. This implies that the previous PIAT math score, though with very high importance (67.7), is not a sufficient statistic for earlier inputs. The only current input selected is a mother's physical affection for child.

The extreme importance of the math score at age 7 in predicting the score two years later may

---

<sup>12</sup>Note that the missing values in earlier inputs increase the test sample error for history tree (0.844) compared with current tree (0.833), notwithstanding the reduction of resubstitution error due to more explaining variables.

suggest that child ability has large effects on math scores. But how does child ability in math evolve over time? To shed some light on this question, the math score at age 5 is used as a proxy for the earlier child ability in the value-added historical tree (VAB). The three primary splitters are the number of books at age 7 and 9, and special activities at age 9. The importance of math score at age 5 (39.8) is much lower than that at age 7, while the total importance of inputs increases. One implication is that child ability in math changes as a child grows, which is affected by home inputs especially books and enriching activities.<sup>13</sup> The explaining power of inputs alone, however, is quite low since no optimal within-child tree exists for the math score.

Among all inputs, the number of books a child has at age 7 and 9, mother’s physical affection for child, special activities, and mother reading to child at age 3-5 appear as primary splitters in at least one of the three value-added regressions.

## 6.2 Reading Scores for Children at Age 9

Table 2 summarizes six regression trees of PIAT reading scores. For both current and historical input trees, regressions with and without race are strikingly similar: most inputs have the same effects, and the difference in errors is negligible. This suggests that race does not have much explaining power beyond what is already exhibited by current (and earlier) inputs. When historical inputs are included, books at age 7 becomes the most important input (with importance 18.7) and the importance of books at age 9 is greatly reduced (from 19.8 to 3.52), which is exactly the same scenario as in the math score regressions.

In the value-added historical tree with reading score at age 7 (VAY), the top two inputs are the number of books at age 7 and child reading habit at age 9. This implies that the earlier reading score is not a sufficient statistic for earlier inputs. In the value-added historical tree with reading score at age 5 (VAB), the importance of this earlier reading score is 55, much lower than that at age 7 (97.83 in VAY); while the aggregate importance of inputs is higher. It suggests that child ability in reading also changes and is affected by home inputs. In within-child tree using age 7 reading score, the most important inputs are the number of books and child reading habit at age 9.<sup>14</sup>

Overall, the number of books at age 7 and 9 and child reading habit at age 9 are the most

---

<sup>13</sup> An alternative explanation is PIAT math measures different things at age 5, 6-7, and 8-9. Since it is difficult to tell the difference between ability and its manifestation (i.e. the measurable part), these two interpretations are actually quite similar.

<sup>14</sup> The without-race version is almost the same and thus not shown in the table.

important inputs predicting a child’s reading scores at age 9. The important inputs and their rankings are quite similar in both reading and math score regressions; a difference is that reading scores are affected by more inputs in value-added and within-child specifications.

### 6.3 Behavior Problem Scores for Children at Age 9

Table 3 summarizes five BPI regression trees. The difference between current and historical input trees is quite small; the only earlier input that matters is spanking frequency at age 7, while the importance levels of spanking and grounding a child at age 9 do not change. The number of grounding and sending a child to room are still primary splitters in the value-added historical tree using BPI score at age 7 (VAY), while the importance of grounding is much reduced. In contrast, the large effects spanking frequency at age 9 disappear once an earlier behavior score is controlled. Since no earlier inputs show up, it seems that the BPI score at age 7 can be used as a sufficient statistic for both ability and earlier inputs.

When BPI score at age five is included instead (VAB), the number of grounding and room-sending at age 9 are still the most important inputs, whose importance levels are almost the same as in the historical input tree. Many other inputs also become primary splitters including some inputs at age 3-5, which suggests BPI score at age five is not a sufficient statistic for earlier inputs. Again, the importance of age 5 BPI score (51) is smaller than that at age 7 (79.6 in VAY). However, the relative error in VAB is only 0.015 points higher than VAY, which suggests the difference between BPI scores at age 5 and 7 can be mostly accounted for by home inputs. In the within-child tree using age 5 BPI score, spanking a child at age 3-5 is negatively (though modestly) associated with his behavior problem at age 9, while the opposite is true for sending a child to room at age 9. The errors are very high, and no within-child tree exists using age 7 BPI score.

In these behavior problem regressions, the number of times sending a child to room, grounding him, and how often he reads for self-enjoyment at age 9 are selected as primary splitters in most specifications even when an earlier score is controlled. In sharp contrast, the spanking frequency at age 9 is the most important predictor (with importance 19.6) in both current and historical input trees, but it never appears in any value-added or within-child difference specifications. This suggests that spanking a child of age 9 merely reflects some family- or child-specific fixed factors. Spanking a child at age 3-5, however, is negatively associated with his behavior problem at age 9 as shown in the within-child regression. Though the evidence mentioned in section 5.2 suggests that

the endogeneity of disciplining inputs is not severe after an earlier score is controlled,<sup>15</sup> one should still be cautious in interpreting the positive correlation between disciplining (grounding and room-sending) and child behavior problems as causal. A possible explanation is that these disciplinary methods merely reflect some current random factors affecting a child's behavior problems, which are not completely controlled by the earlier scores.

## 6.4 Child Development Results at Age 5

A common feature of the various regressions above is that child development results at age 5 are very important for future outcomes. So it is interesting to know how the age 5 results are affected by home inputs. Table 4 summarizes historical input regression trees for PIAT math, reading, behavior problem index (BPI) total scores, and PPVT scores at age 5. The number of books and mother reading to child at age 3-5 are important in all of these scores, where mother reading is the most important input for both math (without race version) and reading scores, and the number of books is most important for PPVT scores (without race version). The number of books before age 3 and the number of magazines in home at age 3-5 are important for most scores. The spanking frequency at age 3-5 is the best predictor of BPI scores at age 5, and it's also associated with math scores. The first two borns seem to enjoy some advantages in cognitive development results, which may also reflect the positive effect of a small family size. Race matters only for math and PPVT scores, but its effects are partially accounted for by books, mother reading, and birth order. Note for PPVT score there are huge jumps in importance of books at age 0-2 (from 4.75 to 21.1) and books at age 3-5 (from 41.3 to 89.6) when race is taken out. Overall, books and how often a mother reads to her child appear to matter most for child cognitive development at age 5. They are also the primary predictors for behavior problems at age 5, though spanking is most predictive. The errors are generally higher than corresponding regressions of child development results at age 9.

## 7 Conclusion

Early child development is a crucial part of human capital formation. The paper estimates production functions of child cognitive and social development at age 8-9 using NLSY(79) child data, where over two hundred home and school inputs starting from mother prenatal care periods are included as well as many family background variables. A tree structured regression method is used to conduct

---

<sup>15</sup>This is also partially confirmed by the changes in the importance of spanking.

estimation, and the unobserved family or child heterogeneity is handled by within-child difference and value-added specifications. The omitted variable problem is greatly mitigated by using a very rich set of detailed inputs and further checked by including family backgrounds in the regressions. The endogeneity problem of disciplinary inputs is checked by putting future inputs at age 9 into the regressions of child development results at age 7. The evidence shows that the influence of these problems, if any, is very weak.

Detailed home and school inputs are the most important predictors of child development results. Family backgrounds are never selected as primary explaining variables for any child development results when detailed inputs are controlled. Production functions of child math and reading scores are more similar to each other than to behavior problem scores. The number of books a child has at various ages, how often a child reads for self-enjoyment, and how often a mother reads to her child before age 5 are among the most important inputs predicting both child math and reading scores from age five onwards. Child behavior problem scores at age 8-9 are mostly correlated with parental disciplining such as how often a mother grounds a child or sends a child to room, while the scores at age 5 are also associated with how often the mother read to him and the number of books the child had. Spanking a child aged 8-9 have no effects on current result, while spanking a younger child may reduce his future behavior problems.

The test score gaps across race are more likely to be caused by different inputs rather than race itself per se. The race of a child has little effects on his reading and behavior problem scores in various specifications. Though it appears to have large effects on child math and especially PPVT scores, its effects can be almost completely accounted for using current and historical home and school inputs for math, and partially for PPVT. The sex of child is never selected as a primary explaining variable, which implies that its effects on child development results are weak, if any. A mother's working time does not have any effects on child development once detailed inputs and child ability are controlled. A possible reason is that working mothers manage to get appropriate home and school inputs (including quality time) for their children as non-working mothers (Bianchi 2000).

There are fifteen school inputs for children older than six. A mother's rating of teachers' caring for students and the interaction between parents and school are associated with math and reading scores, while teacher's caring affects reading scores when child ability is controlled. The school inputs in general have less explanation power than home inputs. This is consistent with the finding of Parcel and Durfur (2001) that the effects of school inputs on child behavior are modest in size,

while home inputs are stronger. Similarly, the standard school inputs such as class size, pupil-teacher ratio, and teacher experiences are usually found to have little effects on children attainment (Feinstein and Symons 1999). In our case, the prevalence of missing values in school variables may also play some role. The effects of child care experiences in the first three years on a child's future development results are also modest compared with home inputs, since they are never selected as primary explaining variables.

A reasonably small set of inputs from over two hundred available in the data are selected as primary predictors for child cognitive and social development results at age eight and nine, which may be used as a rough guide for variable selection in relevant research. Though there is some evidence suggesting the estimated effects of home and school inputs in the paper are more than correlations, further research is needed to establish the causal links.

## References

- [1] Baharudin, R. and T. Luster (1998), "Factors Related to the Quality of the Home Environment and Children's Achievement," *Journal of Family Issues*, vol. 19 No. 4, 375-403.
- [2] Baum, Charles L. (2003), "Does Early Maternal Employment Harm Child Development? An Analysis of the Potential Benefits of Leave-Taking." *Journal of Labor Economics* 21 (2): 409-448.
- [3] Bianchi, Suzanne M. (2000), "Maternal Employment and Time with Children: Dramatic Change or Surprising Continuity?" *Demography*, Vol. 37, No. 4. (Nov., 2000), pp. 401-414.
- [4] Blau, F.D. and A. J. Grossberg (1992), "Maternal Labor Supply and Children's Cognitive Development," *The Review of Economics And Statistics*, pp 474-481.
- [5] Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. (1984), *Classification and Regression Trees*, Chapman and Hall, New York.
- [6] Currie, J. and D. Thomas (1999), "Early Test Scores, Socioeconomic Status and Future Outcomes," *NBER working paper #6943*.
- [7] Durlauf, S.N. and Paul A. Johnson (1995), "Multiple Regimes and Cross-Country Growth Behavior," *Journal of Applied Econometrics*, 10:365-384

- [8] Feinstein, L. and J. Symons (1999), "Attainment in Secondary School," *Oxford Economic Papers* 51 (1999), 300-321.
- [9] Fryer, R.G. Jr. and S. Levitt (2003), "Understanding the Black-White Test Score Gap in the First Two Years of School," forthcoming in the *Review of Economics and Statistics*.
- [10] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, New York : Springer, c2001.
- [11] Harvey, E. (1999), "Short-Term and Long-Term Effects of Early Parental Employment on Children of the NLSY," *Developmental Psychology*, vol. 35, No. 2, 445-459.
- [12] Haverman, R. and B. Wolfe (1995), "The Determinants of Children's Attainments: A Review of Methods and Findings," *Journal of Economic Literature*, vol. XXXIII, pp. 1829-78.
- [13] Heckman, J. (1999), "Policies to Foster Human Capital," *NBER working paper #7288*.
- [14] Leibowitz, A. (1974), "Home Investment in Children," *Journal of Political Economics*, 82 (2, II), pp. S111-31.
- [15] Loh, W.-Y. (2002), "Regression trees with unbiased variable selection and interaction detection," *Statistica Sinica*, 12, 361-386.
- [16] Neidell, Matthew J. (2000), "Early Parental Time Investment in Children's Human Capital Development: Effects of Time in the First Year on Cognitive and Non-cognitive Outcomes," mimeo, UCLA.
- [17] NICHD Early Child Care Research Network. (1998), "Early child care and self-control, compliance and problem behavior at twenty-four and thirty-six months," *Child Development*, 69, 1145-1170.
- [18] NICHD Early Child Care Research Network. (2000), "The relation of child care to cognitive and language development," *Child Development*, 71, 960-980.
- [19] NICHD Early Child Care Research Network (2001), "Nonmaternal care and family factors in early development: An overview of the NICHD Study of Early Child Care," *Journal of Applied Developmental Psychology*, 22, 457-492.

- [20] NICHD Early Child Care Research Network (2003), “Does amount of time spent in child care predict socioemotional adjustment during the transition to kindergarten?” *Child Development*, 74. 976-1005.
- [21] NICHD Early Child Care Research Network (2003), “Families matter—even for kids in child care,” *Journal of Developmental and Behavioral Pediatrics*, 24, 58-62.
- [22] Parcel, T.L. and E.g. Menaghan (1994), “Early Parental Work, Family Social Capital, and Early Childhood Outcomes,” *American Journal of Sociology*, Vol. 99 No. 4: 972-1009.
- [23] Parcel, Toby L. and Dufur, Mikael J., “Capital at Home and at School: Effects on Child Social Adjustment,” *Journal of Marriage and the Family* 63,1 (February 2001): 32-47.
- [24] Ruhm, Christopher J. (2000) “Parental Employment and Child Cognitive Development,” *Journal of Human Resources*, forthcoming.
- [25] Sacerdote, B. (2002), “The Nature and Nurture of Economics Outcomes,” *American Economic Review Papers and Proceedings*, Vol. 92 (May 2002).
- [26] Schore, A.N. (1996), “The Experience-Dependent Maturation of a regulatory System in the Orbital Prefrontal Cortex and the Origin of Developmental Psychopathology,” *Development and Psychopathology*, 8, 1996.
- [27] Todd, Petra, and Kenneth I. Wolpin (2003), “Towards a Unified Approach for Modeling the Production Function for Cognitive Achievement,” *Economic Journal*, Feb. 2003.
- [28] Todd, Petra, and Kenneth I. Wolpin (2004), “The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps,” working paper, University of Pennsylvania.
- [29] Waldfogel, Jane, Wen-Jui Han, and Jeanne Brooks-Gunn (2002), “The Effects of Early Maternal Employment on Child Cognitive Development,” *Demography*, Vol. 39 No. 2, May 2002: 369-392.

## Appendix A: Proofs

**Proof for Proposition 1:** We study the simplest case where partial derivatives of  $G(\cdot)$  are constant, and there is no intertemporal relation between mother’s two period wages. We use it as a benchmark situation for our proofs.

At the second and last period,  $u(c_2, g_2)$  is maximized by optimal choices  $k_2^*, H_2^*, h_2^*$  as functions of state variables  $p_k, w_2, I_2, A, k_1, h_1, H_1$ . Given  $(k_2^*, H_2^*, h_2^*)$  and the prices, her objective function in the first period is

$$\max_{k_1, H_1, h_1} U = u(c_1, g_1) + \beta u(c_2, g_2),$$

subject to constraints of time (1), budget (2) at  $t = 1$ , and technique (3).

**Lemma 1** *When the marginal returns  $G_{h_2}^2, G_{k_2}^2, G_{H_2}^2$  are constant, the optimal solutions in the second period are*

$$\begin{aligned} H_2^* &= \frac{a\alpha_3}{\alpha_1 - a\alpha_2 G_{H_2}^2}, \\ k_2^* &= \frac{I_2 + w_2 h_2^*}{p_k} - \frac{1}{a\alpha_2 G_{k_2}^2}, \forall h_2^* \in [0, 1 - H_2^*], \\ h_2^* &= \begin{cases} 0 & \text{if } w_2 < w_2^* \\ 1 - H_2^* & \text{if } w_2 > w_2^* \end{cases}, \end{aligned}$$

where  $w_2^* = \frac{(\alpha_1 - a\alpha_2 G_{h_2}^2)p_k}{a\alpha_2 G_{k_2}^2}$ .

**Proof.** The optimal choices  $k_2^*, H_2^*, h_2^*$  are determined by the following first order conditions, where interior solutions are assumed for  $k_2^*, H_2^*$ .

$$\frac{\partial u}{\partial k_2} = \frac{-p_k}{c_2} + a\alpha_2 G_{k_2}^2 = 0, \quad (9)$$

$$\frac{\partial u}{\partial H_2} = -\alpha_1 + a\alpha_2 G_{H_2}^2 + \frac{a\alpha_3}{H_2} = 0, \quad (10)$$

$$\frac{\partial u}{\partial h_2} = \frac{w_2}{c_2} - \alpha_1 + a\alpha_2 G_{h_2}^2 \begin{cases} \geq 0 & \text{if } h_2^* > 0 \\ < 0 & \text{if } h_2^* = 0 \end{cases}. \quad (11)$$

Then from (10) we get

$$H_2^* = \frac{a\alpha_3}{\alpha_1 - a\alpha_2 G_{H_2}^2},$$

which says a mother would spend more quality time with child if she is more altruistic, enjoys more with child, and concerns more about his performance.

Condition (9) implies  $c_2^* = \frac{p_k}{a\alpha_2 G_{k_2}^2}$ . By condition (11) we get that  $h_2^* > 0$  if  $w_2 \geq c_2^*(\alpha_1 - a\alpha_2 G_{h_2}^2) \equiv w_2^*$ , where  $w_2^*$  is mother's reservation wage. A mother with a lower wage than  $w_2^*$  would not work at all. If her wage is higher than  $w_2^*$ , she would work to the upper limit  $1 - H_2^*$ , where  $H_2^*$  is already optimally chosen. If her wage is  $w_2^*$ , she is indifferent about the exact working time. That is

$$h_2^* \begin{cases} 0 & \text{if } w_2 < w_2^* \\ \in [0, 1 - H_2^*] & \text{if } w_2 = w_2^* \\ 1 - H_2^* & \text{if } w_2 > w_2^*, \end{cases}$$

Given  $h_2^*$ , mother chooses  $k_2^*$  accordingly  $k_2^* = \frac{I_2 + w_2 h_2^*}{p_k} - \frac{1}{a\alpha_2 G_{k_2}^2}$ . ■

**Lemma 2** *When all the marginal returns  $G_{h_2}^2, G_{k_2}^2, G_{H_2}^2, G_{h_1}^2, G_{k_1}^2, G_{H_1}^2$  are constant, the optimal solutions in the first period are*

$$\begin{aligned} H_1^* &= \frac{a\alpha_3}{\alpha_1 - a\alpha_2(G_{H_1}^1 + \beta G_{H_1}^2)}, \\ k_1^* &= \frac{I_1 + w_1 h_1^*}{p_k} - \frac{1}{a\alpha_2(G_{k_1}^1 + \beta G_{k_1}^2)}, \forall h_1^* \in [0, 1 - H_1^*], \\ h_1^* &= \begin{cases} 0 & \text{if } w_1 < w_1^* \\ 1 - H_1^* & \text{if } w_1 > w_1^* \end{cases}, \end{aligned}$$

where  $w_1^* = \frac{(\alpha_1 - a\alpha_2(G_{H_1}^1 + \beta G_{H_1}^2))p_k}{a\alpha_2(G_{k_1}^1 + \beta G_{k_1}^2)}$ .

**Proof.** The first order conditions for interior solutions are

$$\frac{\partial u}{\partial k_1} = \frac{-p_k}{c_1} + a\alpha_2 G_k^1 + \beta a\alpha_2 G_{k_1}^2 = 0, \quad (12)$$

$$\frac{\partial u}{\partial H_1} = -\alpha_1 + a\alpha_2 G_{H_1}^1 + \frac{a\alpha_3}{H_1} + \beta a\alpha_2 G_{H_1}^2 = 0, \quad (13)$$

$$\frac{\partial u}{\partial h_1} = \frac{w_1}{c_1} - \alpha_1 + a\alpha_2 G_{h_1}^1 + \beta a\alpha_2 G_{h_1}^2 = 0. \quad (14)$$

Following similar procedure above we would get the optimal solutions. ■

All the optimal solutions are functions of mother characteristics  $(a, \alpha_1, \alpha_2, \alpha_3, \beta)$ , child ability  $A$ , market price  $p_k$ , mother wages  $w_1$  or  $w_2$ , non-wage family income  $I_1$  or  $I_2$ . Since these state variables are linked with each other and intertemporally related, all inputs are correlated with each other and across periods. So any omitted inputs must be correlated with included ones and thus cause bias in estimation.

In the above benchmark case, when all mothers have the same characteristics, face the same market prices, and the marginal returns of any inputs are constant for all children, then optimal inputs are completely determined by mother wage, working hours, and the family non-wage incomes. In other words, the detailed home inputs could be omitted when these three background variables are included. But this result no longer holds once we allow for heterogeneity among mothers, or there are non-linear interactions between different home inputs so that their marginal returns are not constant, or depend on a child innate ability that is known to be heterogenous.

If mothers are heterogenous in terms of  $(a, \alpha_1, \alpha_2, \alpha_3, \beta)$ , then for example  $k_1^*$  would be different among mothers with identical wages and working hours. Since the characteristics  $(a, \alpha_1, \alpha_2, \alpha_3, \beta)$  are generally not observable, omitted variable problem is sure to arise in the estimation.

If the marginal return  $G_{k_2}^2$  for child  $i$  with unobservable innate ability  $A_i$  is  $G_{k_2}^{2i} = \theta_1 A_i, k_{2i}^*$  becomes

$$k_{2i}^* = \frac{I_2 + w_2 h_2^*}{p_k} - \frac{1}{a\alpha_2 \theta A_i},$$

which implies that a mother would buy more goods and services for a high ability child than a lower one. Thus the inputs would differ across children even all mothers are identical. Again we cannot recover  $k_{2i}^*$  using family background variables as long as child ability is not completely observed or measured.

If non-linear interaction exists, for example  $G_{H_2}^2 = \theta_2 H_1^* + \theta_3 k_2^*$  and  $G_{H_1}^1 = \theta_4 k_1^*$ , then from conditions (10) and (13) we get (for interior solutions)

$$\begin{aligned} H_2^* &= \frac{a\alpha_3}{\alpha_1 - a\alpha_2(\theta_2 H_1^* + \theta_3 k_2^*)}, \\ H_1^* &= \frac{a\alpha_3}{\alpha_1 - a\alpha_2(\theta_4 k_1^* + \beta\theta_2 H_2^*)}. \end{aligned}$$

In this case, if we want to proxy  $H_2^*$ , former inputs like  $H_1^*$  has to be known. If we want to proxy  $H_1^*$ , future inputs like  $H_2^*$  must be included. Since child development is a multi-period process, each period is likely to have preceding and following periods. As a result, related inputs or background variables over all relevant periods have to be used in order to proxy such an input. If there is any non-linear interaction between working hours and other inputs, say  $G_{H_1}^1 = \theta_5 + \theta_6 k_1^*$ , then without knowing the exact value of  $k_1^*$ , we cannot get the unbiased estimates for  $\theta_5$  and  $\theta_6$ .

### Proof for Proposition 2.

**Proof.** To get unbiased estimate in value-added specification, there must exist a value  $g_1^*$  such that  $I\{g_{1i} \leq g_1^*\} = A_l$  and  $I\{g_{1i} > g_1^*\} = A_h$ . According to production function (6)  $g_{1i} = \alpha_{1a} A_h + \theta_{1a} A_l + \alpha_{1k} k_{1h} + \theta_{1k} k_{1l} + \alpha_{1ak} k_{1h} * A_h + \varepsilon_{1i}$ , the possible values of  $g_{1i}$  are

$$g_{1i} = \begin{cases} \alpha_{1a} + \alpha_{1k} + \alpha_{1ak} + \varepsilon_{1i} \equiv g_{1hh} + \varepsilon_{1i} & \text{iff } (A_h = 1, k_h = 1), & \text{with prob. } Pp_1 \\ \alpha_{1a} + \theta_{1k} + \varepsilon_{1i} \equiv g_{1hl} + \varepsilon_{1i} & \text{iff } (A_h = 1, k_l = 1), & \text{with prob. } P(1 - p_1) \\ \theta_{1a} + \alpha_{1k} + \varepsilon_{1i} \equiv g_{1lh} + \varepsilon_{1i} & \text{iff } (A_l = 1, k_h = 1), & \text{with prob. } Pq_1 \\ \theta_{1a} + \theta_{1k} + \varepsilon_{1i} \equiv g_{1ll} + \varepsilon_{1i} & \text{iff } (A_l = 1, k_l = 1), & \text{with prob. } P(1 - q_1) \end{cases}$$

Suppose child innate ability has much larger direct effect than other inputs such that  $\alpha_{1a} - \theta_{1a} > \alpha_{1k} - \theta_{1k}$  holds, which implies  $\alpha_{1a} + \theta_{1k} > \theta_{1a} + \alpha_{1k} \iff g_{1hl} > g_{1lh}$  (If not,  $g_{2i}$  would split first on inputs instead of  $g_{1i}$ , which is not the case). To get  $I\{g_{1i} > g_1^*\} = A_h$ , it must be true that

$$\Pr(g_{1hl} - g_{1lh} \geq \varepsilon_{1i} - \varepsilon_{1j}) = 0 \tag{15}$$

for all  $i, j$  in the sample, and  $g_1^* \in [g_{1lh} + \max \varepsilon_{1i}, g_{1hl} + \min \varepsilon_{1j}]$ . Let the support of the noise term  $\varepsilon_{1i}$  be  $[-\delta_1, \delta_2]$ , where  $\delta_1, \delta_2 \in R_{++}$ . Condition (15) is true if  $\delta_1 + \delta_2 \leq g_{1hl} - g_{1lh}$ , which is equivalent to (8). In this case it is certain that a child with scores higher than or equal to  $g_{1hl} - \delta_1$  must also have higher innate ability  $A_h$ . In other words, there is a one-to-one mapping between the set of scores and children's unobserved abilities:  $\Pr(A = A_h | g_{1i} > g_{1lh} + \delta_2) = 1$ . So the estimated total effects of unobserved child ability are unbiased as in the ideal tree where ability is observed. As a result, estimated total and direct effects of a splitting variable correlated with child ability are no longer inflated by the unobserved ability, while the effects of those independent of child ability would not change when  $g_1$  is included.

$$\begin{aligned}
g_{2i} = & \alpha_{2a} A_h + \theta_{2a} A_l + \alpha_{2k} k_{2h} + \theta_{2k} k_{2l} + \alpha_{2ak} k_{2h} * A_h \\
& + \alpha_{21k} k_{1h} + \theta_{21k} k_{1l} + \alpha_{21ak} k_{1h} * A_h + \varepsilon_{2i},
\end{aligned}$$

Suppose without loss of generality that the next optimal splitting rule is on  $k_1$ . Then in the value-added historical tree which variable would be used:  $g_1$  or  $k_1$ ? Since  $g_1$  contains random errors and is only an imperfect proxy for  $k_1$ , obviously the optimal rule would use  $k_1$  itself. Thus including the earlier score  $g_1$  would not mess up the effects of measured earlier inputs in  $G^2$ . Similarly, the estimated effects of current inputs are also not affected. ■

## Appendix B: An Example of Regression Tree

In this part we work out optimal splitting rules for a simple child development production function and use it to illustrate how to interpret tree-structured results in general. In the child development functions (6) and (7),  $\alpha_{1ak} > 0$  implies that the amount of input  $k_1$  is positively correlated with child ability  $A$  according to our behavior model.<sup>16</sup> Let  $P = \Pr(A = A_h) \in (0, 1)$  denote the exogenously determined proportion of high ability children in the population. Define  $p_1 \in (0, 1]$  as the probability that a high ability child has  $k_h$  and  $q_1 \in (0, 1]$  for a low ability child with  $k_h$ , i.e.

$$\begin{aligned}
\Pr(k_1^* = k_h | A = A_h) &= p_1, \\
\Pr(k_1^* = k_h | A = A_l) &= q_1.
\end{aligned}$$

Then  $p_1 \geq q_1$  holds.  $p_2$  and  $q_2$  are similarly defined for  $k_2^*$  where  $p_2 \geq q_2$  holds.

<sup>16</sup>The inequality may be reversed when a different model is used, for example, when parents compensate children of lower ability with more inputs. As long as inputs are correlated with each other, the results in the paper hold.

The optimal splitting rule  $\gamma$  in the mother node  $t$  allocates cases into two child nodes  $t_L$  (left) and  $t_R$  (right). It maximizes the variance decrease

$$\Delta R^r(\gamma|t) = p(t_L)p(t_R)[\bar{g}_{2i}(t_L) - \bar{g}_{2i}(t_R)]^2,$$

where  $p(t_L)$  and  $p(t_R)$  denote the frequency of a case falling into the left child node  $t_L$  and the right one  $t_R$  respectively, and  $\bar{g}_{2i}(t_L)$  and  $\bar{g}_{2i}(t_R)$  denote the average  $g_{2i}$  scores in these two child nodes.

Suppose for now the child innate ability is available in the data, and there are either no omitted variables or they are orthogonal to included ones. If child innate ability has much bigger influence on child development than any other inputs, the first primary splitter in the regression tree of  $g_{2i}$  must be that ‘all high ability children go to left node  $t_L$ , all low ability ones  $t_R$ .’ The means of nodes  $t_L, t_R$  are

$$\begin{aligned} E_\varepsilon[\bar{g}_{2i}(t_R)] &\equiv \mu(g_{2i}|t_R) = \mu(g_{2i}|A_h) \\ &= \alpha_{2a} + [(\alpha_{2k} + \alpha_{2ak})p_2 + \theta_{2k}(1 - p_2)] + [(\alpha_{21k} + \alpha_{21ak})p_1 + \theta_{21k}(1 - p_1)], \\ E_\varepsilon[\bar{g}_{2i}(t_L)] &\equiv \mu(g_{2i}|A_l) = \theta_{2a} + [\alpha_{2k}q_2 + \theta_{2k}(1 - q_2)] + [\alpha_{21k}q_1 + \theta_{21k}(1 - q_1)]. \end{aligned}$$

The mean difference between these two child nodes is

$$\underbrace{\mu(g_{2i}|A_h) - \mu(g_{2i}|A_l)}_{\text{total effect of ability change}} = \underbrace{(\alpha_{2a} - \theta_{2a} + \alpha_{2ak}p_2 + \alpha_{21ak}p_1)}_{\text{direct effect of ability change}} + \underbrace{(\alpha_{2k} - \theta_{2k})(p_2 - q_2) + (\alpha_{1k} - \theta_{1k})(p_1 - q_1)}_{\text{indirect effect through changes in } k_1^* \text{ and } k_2^*}, \quad (16)$$

which measures the *total effect* of child ability on  $g_{2i}$  in node  $t$ , including both direct and indirect effects.<sup>17</sup> The *direct effect* measures the *partial equilibrium effect* of ability change when other inputs remain the same. But due to the underlying optimization process, ability change would simultaneously cause corresponding changes in other inputs which also affect child development. In other words, ability change also has some *indirect effect* through changes of correlated inputs  $k_1^*$  and  $k_2^*$ . In this sense, the total effect actually measures *general equilibrium effect* of a variable’s marginal change.<sup>18</sup>

<sup>17</sup>If child ability is randomly given then the mean difference, by measuring the total effect, corresponds to the policy effect in program evaluation literature (see Todd and Wolpin 2003). A crucial difference, however, is that the direct effect or the production parameters can be uncovered in tree regression.

<sup>18</sup>In the continuous case where  $g_{2i} = G^2(A, k_2^*, k_1^*)$ , the mean difference is similar to the total derivative of  $g_{2i}$  with respect to  $A$ .

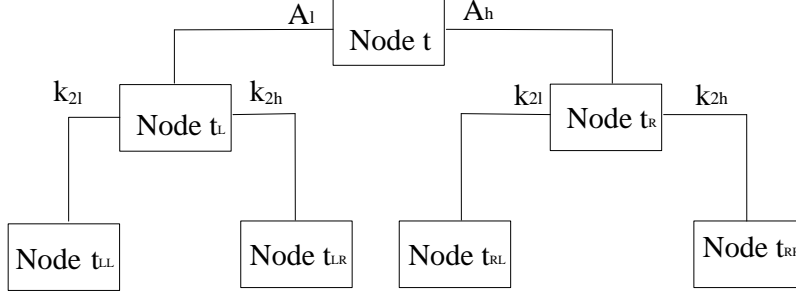


Figure 1: Regression Tree for  $g_{2i}$

For simplicity we assume  $p_1 = q_1 = 1$  so that all children have  $k_1 = k_h$ .<sup>19</sup> Then node  $t_R$  must be split on  $k_2$  into two terminal nodes  $t_{RL}$  and  $t_{RR}$ , and similarly  $t_L$  into  $t_{LL}$  and  $t_{LR}$ . See figure 1 for illustration. The means of nodes  $t_{RL}$  and  $t_{RR}$  are

$$\begin{aligned}\mu(g_{2i}|t_{RR}) &= \mu(g_{2i}|A_h, k_h, k_h) = \alpha_{2a} + (\alpha_{2k} + \alpha_{2ak}) + (\alpha_{21k} + \alpha_{21ak}), \\ \mu(g_{2i}|t_{RL}) &= \mu(g_{2i}|A_h, k_l, k_h) = \alpha_{2a} + \theta_{2k} + (\alpha_{21k} + \alpha_{21ak}).\end{aligned}$$

The mean difference between these two terminal nodes is

$$\mu(g_{2i}|t_{RR}) - \mu(g_{2i}|t_{RL}) = \alpha_{2k} + \alpha_{2ak} - \theta_{2k}, \quad (17)$$

which is exactly the partial effect of  $k_2$  on  $g_{2i}$  for high ability children.

The mean difference between  $t_{RR}$  and  $t_R$  is

$$\mu(g_{2i}|t_{RR}) - \mu(g_{2i}|t_R) = (\alpha_{2k} + \alpha_{2ak} - \theta_{2k})(1 - p_2).$$

This condition combined with (17) would solve  $p_2$ . Similarly from node  $t_L$  we can get  $\alpha_{2k} - \theta_{2k}$  and  $q_2$ . Plugging these parameters in (16), we can recover  $\alpha_{2a} - \theta_{2a} + \alpha_{2ak}p_2 + \alpha_{21ak}$ . Using the difference between  $\mu(g_{2i}|t_R)$  and  $\mu(g_{2i}|t)$  we can also find  $P$ . This means that all relevant direct effects of inputs and the correlations among them in child development production function would be identified by the regression tree. In other words, the structure of the data would be estimated unbiased. These results are summarized in the following proposition.

**Proposition 3** *The mean difference of two terminal nodes measures partial equilibrium effect of the splitting variable. In contrast, the mean difference of two non-terminal nodes measures general*

<sup>19</sup>Without this simplification, the analysis is similar but longer.

*equilibrium effect of the splitting variable in their mother node. Both effects can be estimated unbiased by the regression tree.*

When child innate ability is not measured in the data but is still perceivable to parents, the estimated effects of home inputs would be biased. For example, if we still assume  $p_1 = q_1 = 1$  as above, then the only available splitting variable is  $k_2$ . The mean difference of two child nodes thus split is the total effect of both child ability and  $k_2$ . In no way can we distinguish between them in this tree. However, within-child difference and value-added trees can be used to get unbiased estimates under certain assumptions, which are discussed in the paper.

## Appendix C: List of Variables

1. Child Development Results: PIAT math standard score at age 9, 7 and 5; PIAT reading recognition standard score at age 9, 7 and 5; Behavior Problem Index (BPI) total standard scores at age 9, 7 and 5; PPVT total standard scores at age 3-5.

2. Child Demographic Information: the race, sex, birth order of child, child ages (in months) at child assessment dates.

3. Home and school Inputs at age 6-9: how many books child has (1={none}, 2={1 or 2 books}, 3={3 to 9 books}, 4={10 or more books}). how often mom reads to child (1=none, 2=several times a year, 3=several times a month, 4=once a week, 5=at least 3 times a week, 6=everyday). how often child expected to make bed, clean room, clean spill, bathe self, pickup after self (1={almost never}, 2={less than 2 times}, 3={2 times}, 4={more than 2 times}, 5={almost always}). Is there music instrument at home (1=yes, 0=no). family gets newspaper daily (1=yes, 0=no). how often child reads for enjoyment (1=everyday, 2=several times a week, 3=several times a month, 4=several times a year, 5=never). family encourages hobbies (1=yes, 0=no). child gets special lessons/activities (1=yes, 0=no). how often child taken to museum, to performance (1=never, 2=once or twice, 3=several times, 4=about once a month, 5=once a week or more often). how often family get with relatives and friends (1=once a year or less, 2=a few times a year, 3=about once a month, 4=two or three times a month, 5=once a week or more). # of hours/weekday child sees TV, # hours/weekend day child sees TV.

child ever sees father(-figure) (1=yes, 0=no), is father bio/step/figure (1=biological father, 2=stepfather, 3=father-figure, 4=none of the above), how often child spends time w/dad (1=once a

day or more, 2=at least 4 times a week, 3=once a week, 4=once a month, 5=a few times a year, 6=never, 7=no father figure), how often child w/ dad outdoors, how often child eats w/mom & dad (1=more than once a day, 2=once a day, 3=several times a week, 4=once a week, 5=once a month, 6=never, 7=no father figure), parents discuss TV programs w/child (1=yes, 0=no).

Mom responds to tantrum-ground, -spanking, -talk w/child, -give chore, -ignore it, -send to room, -no allowance, -no TV, -time out (1=yes, 0=no). child close to mother, to bio father, to stepmother, to stepfather (1=extremely close, 2=quite close, 3=fairly close, 4=not at all close, 5=does not have this parent).

Mom responds to low grades: contact teacher, lecture child, keep closer eye on activities, punish child, talk w/child, see if child improves on own, make child spend more time on schoolwork, help more w/schoolwork, limit non-school activity, other responses (5=very likely, 4=somewhat likely, 3=not sure how likely, 2=somewhat unlikely, 1=not at all likely).

#times past week spanked child, grounded child, sent child to room, took away TV, took allowance, praised child, showed child physical affection, said positive things.

Is school public or private, is school gifted/handicapped/regular, reason child does not attend school. how many hours/week on math homework, on writing homework. child gets special help w/remedial work, gets special assignment for advanced work. Mom rating of teacher caring, of principal as leader, of teacher skill, of safety of school, of school communicating with parents, of parents participating with school, school teaches right and wrong, school maintains order.

# child close friends mom knows well, how often Mom knows who child is with, how often child attends religious service per year, how important to provide religious training.

#### 4. Home Inputs at age 3-5

how often mother reads to child (1={none}, 2={several times a year}, 3={several times a month}, 4={once a week}, 5={at least 3 times a week}, 6={everyday}). how many books does child have (1={none}, 2={1 or 2 books}, 3={3 to 9 books}, 4={10 or more books}). how many magazines does family get (1={none}, 2={1}, 3={2}, 4={3}, 5={4 or more}). child has record/tape player (1=yes, 0=no). amount of choice child has in food (1=a lot, 2=some, 3=little, 4=no). # of hours TV is on per day (0=less than 1 hour, 95=no TV at home). how often child taken on outing (1=a few times a year or less, 2=about once a month, 3=about 2 or 3 times a month, 4=several times a week, 5=about once a day). how often child taken to museum (1=never, 2=once or twice, 3=several times, 4=about once a month, 5=about once a week or more often). # of hours/weekday child sees

TV, # hours/weekend day child sees TV. how often child eats w/ mom & dad (1=more than once a day, 2=once a day, 3=several times a week, 4=once a week, 5=once a month, 6=never, 7=no father figure), is father bio/step/figure, child see father(-figure) daily (1=yes, 0=no), child ever sees father(-figure) (yes, 0=no). Mom helps child learn numbers, alphabet, colors, shapes (1=yes, 0=no).

Mom responds to hit-hit child back, -send to room, -spank child, -talk to child, -ignore it, -give chores, -take allowance, -hold child hands, -short time-out (1=yes, 0=no). how often spanked child in past week, did mother have to spank child per week? (1=yes, 0=no). child close to mother, to bio father, to stepmother, to stepfather (1=extremely close, 2=quite close, 3=fairly close, 4=not at all close, 5=does not have this parent).

#### 5. Home and school Inputs at age 0-2

how often child gets out of the house (1=never, 2=once a month or less, 3=a few times a month, 4=about once a week, 5=a few times a week, 6=four or more times a week, 7=everyday). how many children books child has, how often mother reads to child, how often mother takes child to grocery (1=twice a week or more, 2=once a week, 3=once a month, 4=hardly ever), how many cuddly or role-playing toys, how many push or pull toys child has. mothers attitude how child learns best (parents should: 1=always teach, 2=usually teach, 3=usually allow children to learn on their own, 4=always allow children to learn on their own). does child see father(-figure) daily? how often child eats with both mom and dad, how often mother talks to child while working (1=always, 2=often, 3=sometimes, 4=rarely, 5=never). mother had to spank child the past week (1=yes, 0=no), how often was child spanked past week. time child watches TV on typical weekday, time watching TV on typical weekend day, how many hours is the TV on in the home. does child ever see father(-figure)? (1=yes, 0=no) is person biological,step, father-figure.

#### 6. Child care, prenatal care, maternal employment, and family backgrounds.

In the first three years, whether child in regular child care, the number of child care arrangements, types of child care (relatives, non-relatives, center-based), # of hours per week in main child care, total months used in main child care with 10 hours+/week. Mother made prenatal visit, took vitamins and sonogram during pregnancy? Frequency of alcohol use, # cigarettes smoked, # sonograms done during pregnancy. Whether child was breast-fed, how long the breast-feeding lasts. Hours worked per week on main job 4th, 3rd, 2nd, and 1st quarter before and after child birth, in the 1st, 2nd, 3rd, 4th and 5th year after birth. Mother wages, family salary incomes, total family incomes, the region lived, and marital status in 1988. The highest grades of mom and whether mom

in school in 1988 and 2002. Mom's AFQT score in 1981, whether she lived in south at birth and age 14. Mom's age at child birth.

Table 1: Effects of Home and School Inputs on PIAT Math Scores at Age 8-9

Inputs	Current	C/race	History	H/race	VAY	VAB
<u>Current inputs at age 8-9</u>						
# of books a child has	6.04	15.3	1.29	1.29		2.53
# of hours/weekday child sees TV	3.07	6.07		2.82		
# showing child physical affection	3.12		0.75	1.02		
how often child reads for enjoyment	2.39	2.62	2.79	2.48		
child gets special help w/remedial work	3.42	1.98		1.10		
# past week mom spanked child	1.46	1.24	1.66			
child gets special lessons/activities	1.39			4.46		2.74
Mom keeps closer eye if low grades		1.09	0.96			
Mom responds to tantrum-time out	1.48					
Mom responds to tantrum-talk with child	1.24	1.23				
rating of parents participate with school		1.22				
<u>Earlier inputs</u>						
# of books a child has at age 6-7	--	--	17.4	17.4	1.40	2.14
teacher caring of child at age 6-7	--	--		1.26		
# of books a child has at age 3-5	--	--	1.91	1.45		
Mom helps child learn shapes at age 3-5	--	--	1.71	1.91		
child has record/tape player at age 3-5	--	--	1.27			
how often mother reads to child at age 3-5	--	--	1.05		0.59	
race (White versus Black/Hispanic)	15.95	--	7.72	--		
<u>Earlier Scores</u>						
PIAT math score at age 7	--	--	--	--	67.7	--
PIAT math score at age 5	--	--	--	--	--	39.8
Sample Variance	190	190	190	190	190	194
Sum of Squared Errors on test sample	0.83	0.85	0.844	0.838	0.65	0.79
Sample Size	2992	2992	2992	2992	2992	2008

Notes: The importance measures of primary splitters for the included regression trees are listed in the table. An input's importance in a specific tree is equal to the aggregate variance it reduces as primary splitters at various nodes, which corresponds to its total effect on the dependent variable.

Current -- Only current inputs at age 8-9 are included; C/race if race is excluded. History -- Historical inputs are included; H/race if race is excluded. VAY -- Value-Added regression with age 7 score and historical inputs. VAB -- Value-Added regression with age 5 score and historical Inputs.

Some inputs with importance less than one are not listed in the table. In the 'C/race' regression with current inputs at age 8-9, they are: how often mom reads to child, how often child eats w/ parents, is there music instrument at home, and #times past week mom says positive things. In the 'H/race' regression with historical inputs, they are: how often child taken to performance, whether child gets special assignment for adv. work, limits non-school activity if low grades, # past week mom grounded child, how often child is expected to bathe self, child ever sees father-figure at age 6-7, birth order, # child reads for enjoyment at age 6-7, mom responds to hit-send to room at age 3-5, #hours mom worked/week 4th quarter and 5th year after child birth, and how often child eats with parents at age 0-2.

Table 2: Effects of Home and School Inputs on PIAT Reading Scores of at Age 8-9

<u>Inputs</u>	<u>Current</u>	<u>C/race</u>	<u>History</u>	<u>VAY</u>	<u>Within</u>	<u>VAB</u>
<u>Current inputs at age 8-9</u>						
# of books a child has	19.8	19.8	3.52		2.99	3.18
how often child reads for enjoyment	8.9	8.9	9.28	2.51	2.37	
child gets special help w/remedial work	10.3	10.3	10.7			3.97
how often child eats with parents	1.67	1.67				0.6
#times past week mom spanked child	1.60	1.60		1.59		
parents participate with school	1.40					
child gets special lessons/activities	1.21	2.40				
how often child w/ dad outdoors					1.01	
Mom rating of teacher caring					0.92	
Mom punishes child for low grades					0.68	
school communicates with parents		1.15				
# hours/weekend day child sees TV		1.08				
how often child expected to clean room						1.34
<u>Earlier inputs</u>						
# of books a child has at age 6-7	--		18.7	3.57		3.91
# past week child spanked at age 6-7	--				1.66	
# mom shows child affection at age 6-7	--				0.68	
# child reads for enjoyment at age 6-7	--					1.95
# past week child grounded at age 6-7	--					1.59
child has record/tape player at age 3-5	--		2.7			
mom's attitude on child learning by age 3	--		1.55			
how often mother reads to child by age 3	--				0.70	
race (White vs. Black/Hispanic)	2.68	--	2.13		0.96	
<u>Earlier scores</u>						
PIAT reading score at age 7	--	--	--	97.8	--	--
PIAT reading score at age 5	--	--	--	--	--	55
Sample Variance	221	221	221	221	120	224
Sum of Squared Errors on test sample	0.825	0.827	0.835	0.57	0.966	0.74
Sample Size	2982	2982	2982	2982	2982	1935

Notes: The importance measures of primary splitters for the included regression trees are listed in the table. An input's importance in a specific tree is equal to the aggregate variance it reduces as primary splitters at various nodes, which corresponds to its total effect on the dependent variable.

Current -- Only current inputs at age 8-9 are included; C/race if race is excluded. History -- Historical inputs are included; H/race if race is excluded. VAY -- Value-Added regression with age 7 score and historical inputs. VAB -- Value-Added regression with age 5 score and historical Inputs. Within -- Within-child regression using age 7 score and all inputs.

Table 3: Effects of Home and School Inputs on BPI Total Scores at Age 8-9

<u>Input</u>	<u>Current</u>	<u>History</u>	<u>VAY</u>	<u>VAB</u>	<u>Within</u>
<u>Current inputs at age 8-9</u>					
#times past week mom spanked child	19.6	19.6			
#times past week mom grounded child	9.9	9.9	3.11	8.24	
#times past week mom sent child to room	4.65	4.0	3.98	4.14	3.92
how often child reads for enjoyment	3.93	1.84		1.09	1.53
how often family gets with relatives and friends	2.45				
how often child picks up after self				1.27	
Mom sees if child improves on own for low grades				0.92	
#times past week mom said positive things				0.61	
<u>Earlier inputs</u>					
#times past week mom spanked child at age 6-7	--	4.49		1.67	
#times past week mom took away TV at age 6-7	--				1.81
how often child w/ dad outdoors at age 6-7	--			1.35	
#times past week mom spanked child at age 3-5	--				1.31
Mom responds to hit--send to room at age 3-5	--			1.05	
Mom helps child learn alphabets at age 3-5	--			0.96	
<u>Earlier scores</u>					
BPI total score at age 7	--	--	79.6	--	--
BPI total score at age 5	--	--	--	51	--
Sample Variance	221	221	221	221	195
Sum of Squared Errors on test sample	0.87	0.86	0.68	0.69	0.98
Sample Size	3021	3021	3021	3021	2512

Notes: The importance measures of primary splitters for the included regression trees are listed in the table. An input's importance in a specific tree is equal to the aggregate variance it reduces as primary splitters at various nodes, which corresponds to its total effect on the dependent variable.

Current -- Only current inputs at age 8-9 are included. History -- Historical inputs are included. VAY -- Value-Added regression with age 7 score and historical inputs. VAB -- Value-Added regression with age 5 score and historical Inputs. Within -- Within-child regression using age 5 score and all inputs.

Table 4: Effects of Home Inputs on PIAT Math and Reading, BPI, and PPVT Scores at Age 5

<u>Inputs</u>	<u>math</u>	<u>math*</u>	<u>reading</u>	<u>BPI</u>	<u>PPVT</u>	<u>PPVT*</u>
<u>Inputs at age 3-5</u>						
# books child has	4.78	2.15	7.23	1.16	41.5	89.6
how often mother reads to child	1.88	12.1	12.1	1.92	5.27	9.77
# magazines family gets	4.10	2.85	2.01	1.39		
# mom spanked child in past week	1.79	2.91		21.7		
birth order		1.74	7.04			3.77
Mom responds to hit - short time-out			1.78			
child age at assessment date			2.44		7.89	
mother spanked child last week				1.19		
Mom responds to hit - spank child				1.18		
<u>Earlier inputs</u>						
# children's books child has by age 3		2.99		1.34	4.75	21.1
race (White vs. Black/Hispanic)	17.1	--			98.99	--
Sample Variance	220	220	227	222	479	479
Sum of Squared Errors on test sample	0.917	0.944	0.893	0.917	0.716	0.759
Sample Size	2236	2236	2180	4047	3023	3023

Note: The importance measures of primary splitters for the included regression trees are listed in the table. An input's importance in a specific tree is equal to the aggregate variance it reduces as primary splitters at various nodes, which corresponds to its total effect on the dependent variable.

\* -- race is excluded.