The Long-Term Effect of the Home Instruction for Parents of Preschool Youngsters (HIPPY) Program on Academic Achievement: Evidence from a School District in Texas

May 2014

Sherlene Chatterji[1]

Department of Economics
Stanford University
Stanford, CA 94305
sherlene@stanford.edu

Under the Direction of
Professor Caroline Hoxby

ABSTRACT

In this paper, I evaluate the effect of the Home Instruction for Parents of Preschool Youngsters (HIPPY) program, a home visitation program targeted at parents of children between the ages of 3-5, on long-term academic achievement using administrative data from a school district in Texas. To control for selection bias of program participants, I compare differences in scores between siblings where one of them was engaged in the program, and siblings where neither of them were. I find that HIPPY students see between 0.4 and 0.6 standard deviations of improvement relative to their non-HIPPY sibling in Mathematics test scores when I match sibling pairs directly on covariates. I do not find statistically significant results for Reading test scores. I also find that students who have undergone HIPPY are less likely to be held back in the same grade relative to their peers.

Keywords: Home instruction program for parents of preschool youngsters, parental intervention, sibling comparison

[1] I would like to thank Professor Hoxby for her tremendous insight, guidance and encouragement throughout this process. She has inspired me to think about the implications of education policy and research in new ways and always challenged me to go further. I thank Stanford's UAR Small Grant program and its donors for supporting my project and giving me the chance to visit a HIPPY site and gather my data. I am deeply grateful to Amber McEnturff, Carla Mowell, Lia Lent, Rachel Woodall, Olayinka Ojo, and Dr. Johnstone at HIPPY who not only gave me the opportunity to study the program but also their valuable time in making sure I had what I needed. My gratitude extends to Alejandro Molnar, Benjamin Shear, Michael Dinerstein and Wichsinee Wilbulpolprasert for taking the time to guide me through any challenges I was facing. Finally, my heartfelt thanks goes out to Rebecca Sachs and Yu Chen Xi for taking the time to give me valuable feedback and who constantly encouraged me.

Contents

I. Introduction

> "What's missing in the current debate over economic inequality is enough serious discussion about investing in effective early childhood development from birth to age 5." – *James J. Heckman, Professor of Economics at the University of Chicago, the New York Times, 2013*

Early childhood programs have been receiving increasing levels of federal funding, from an estimated $4.7 billion in 1999 to nearly $20 billion in 2011 (Barnett & Hustedt, 2011). Understanding what is the most cost-effective way to invest in early childhood, particularly in a period of rising concern surrounding national debt and budget deficits, is therefore paramount. It is evident that home background, largely driven by the parents, is a major input into academic performance; the 1966 Coleman Report explained that the major source of both between-school and within-school variation in academic achievement of children is the variation in home environments, not other commonly cited reasons that pertain to school factors, such as expenditure per student or class size. Yet, some of the largest policy initiatives continue to emphasize targeting students in classrooms, for example the Class Size Reduction program in California which was established in 1996, or the more recent Class Size Amendment in Florida which was enacted in 2003. Additionally, as Heckman alludes to in the quote above, the general debate in the media surrounding early childhood has not moved significantly beyond the question of what should be done in schools.

As Heckman and Masterov (2007) put it, "Schools work with what parents give them." There is a need to understand how we might better prepare children at home, even before they enter school. Many children, particularly poor children, enter kindergarten already academically behind their peers (Jacobson-Chernoff et al., 2007; Cunha et al., 2006). Using results from the first follow-up of the Early Childhood Longitudinal Study-Birth Cohort, Jacobson-Chernoff et al. (2007) show that the average literacy score for children between 4 and 5 years old below the 20th percentile in socioeconomic status is roughly half a standard deviation below the middle 60th percentile, and slightly over one standard deviation below the highest 20th percent. Similar results are seen for average Math scores. These differences early on widen as children progress through grades (Brooks-Gunn & Markman, 2005; Baydar, Brooks-Gunn & Furstenberg, 1993).

Voluntary parental involvement, i.e. highly motivated parents, in a child's life early on improves educational outcomes, but there is little consensus on the effectiveness of parental intervention programs based on home visitations (Jeynes, 2012). In this thesis, I present an empirical evaluation of the Home Instruction for Parents of Preschool Youngsters (henceforth "HIPPY") program. HIPPY works directly with parents by teaching them a structured curriculum. Parents in turn engage their child using the HIPPY-provided educational material.

I counter the non-random selection into HIPPY by first comparing children who were enrolled, with their siblings who were not enrolled, so as to control for common confounding family background effects. I then compare this sibling contrast to a pair of siblings where both siblings were not enrolled in HIPPY. This controls for any common trends in sibling scores that may arise, for example, that the older child systematically does better or worse than the younger child.

I find positive and statistically significant effects of HIPPY on Math test scores although these are not consistent across matching models. I find positive, but insignificant effects of HIPPY on Reading test scores. In addition, there is suggestion of spillover effects from the older child to the younger child if the older child is the sibling who attends HIPPY. I also find a significant negative effect of HIPPY on the probability of a student being "held back" in a grade relative to his or her peers.

I begin by providing a review of the current evidence on the link between parental-based early childhood programs and human capital theory. I then present the background of HIPPY and existing research on HIPPY's effectiveness. Next, I examine the empirical methods of sibling comparisons and matching. I go on to present my empirical methodology, data and results. I conclude by comparing my results to other early childhood intervention programs and commenting on steps to take in the future.

II. Literature Review

A. Early Childhood Parental Intervention Programs

Theoretical research in the field of education seeks to overcome many myths in human capital theory. One myth is that the return on investment at all stages of a child's development is the same. Using a dynamic model where skills in one period foster skill attainment in the next period, Cunha and Heckman (2007) show that not all investments are made equal – investment earlier, rather than later, reaps greater returns. Empirical work provides evidence for this theory. For example, Reynolds et al. (2001) finds that compared to intervention in elementary school, intervention in preschool via the Chicago Child-Parent Center produced bigger reductions.

It is also clear that supporting family environments for children at an early stage are more effective investments in human capital (Cunha et al., 2006). However it is not clear which types of parental-based intervention programs are more effective. Parental intervention programs vary greatly in terms of whether or not they are home visitation programs, what topics they focus on (e.g. health, parenting attitudes) and who the ultimate service recipient is (parent or child).

Brooks-Gunn, Berlin and Fuligni (2012) list the general four service types of parental based intervention programs – (i) parent-focused home visiting programs (programs that focus on parent and family functioning e.g. John Hopkins Children and Youth Program), (ii) parent-focused combined center- and home-based programs (programs that provide treatment to both children and their parents in multiple settings, e.g. the New Orleans Parent Child Development Center (PCDC), Infant Health and Development Program (IHDP), the Milwaukee Project, Project CARE), (iii) intergenerational literacy programs (programs that seek to provide comprehensive services for the parent and the child by also trying to reduce adult illiteracy, e.g. Parent and Child Education (PACE)) and (iv) parent-focused literacy programs (programs that focus solely on the parent-child interaction e.g. Parents as Teachers (PAT) Program).

HIPPY is a type of parent-focused literacy program. Thus, it is important to note here that the results from this evaluation are not indicative of parental based intervention programs in general.

B. Background on HIPPY

HIPPY originated in Israel and was brought to the U.S. in 1984. It is a three year (ninety weeks) long parent involvement and school readiness program for 3, 4 and 5 year olds. HIPPY started in Dallas, Texas in 1998 as a pilot project serving 14 migrant families, and has since grown to serve ten cities in Texas. My analysis is based on data from one of the ten cities.

HIPPY is one of the few programs that specifically looks at improving parenting skills through a narrowly defined method – role play. HIPPY staff visit parents at home, and by taking turns with parents to play the role of the child, the HIPPY staff walks parents through a weekly curriculum that parents are expected to deliver to their child daily. The HIPPY curriculum encourages the development of cognitive, literacy, and social/emotional skills that help children succeed in school, such as the ability to recognize shapes and colors, tell stories, follow directions and solve logical problems.[2] In addition, the set of materials that complement this curriculum typically require parent-child interaction on the activity for only fifteen minutes a day. At the next home visit, HIPPY staff collect activity sheets that parents are supposed to complete with their child during the week. As there is almost no interaction between HIPPY staff and children, this allows one to be more confident of isolating the effect of programs that target parenting behavior. However, one drawback is that no evidence is collected on how frequently parents work with their children and how the parent carries out the activities with the child.

One issue that most researchers find difficult to deal with is the selection bias inherent in such programs: parents who sign up for programs like HIPPY are also more likely to possess characteristics that can help their child perform better later in life. My personal on-site visit to a HIPPY location in Texas confirmed that self-selection is indeed a serious problem – parents in the program often hear about HIPPY through outreach events and word-of-mouth, and were not randomly enrolled into the program. While HIPPY is implemented only in neighborhoods identified by school districts as having more children who enter school unprepared, any parent living in HIPPY neighborhoods may enroll their child in HIPPY. In other words, HIPPY enrollment is not limited to families from a particular socioeconomic background within the HIPPY neighborhood. This further blurs the criteria for who are "treated" and who are not. Further, a general rule is that only one child per family is allowed to attend HIPPY, the rationale being that the parent can apply what he/she has learned with that child to his/her other children. However exceptions are frequently made e.g. if the children are close in age or if the parent is

[2] The curriculum is translated into both Spanish and English and parents may choose which medium they are more comfortable working with. While there are slight differences in the content, the learning objectives are identical.

persistent with his/her request as long as the child is between three and five years old. I thus exclude sibling pairs where both siblings underwent HIPPY from my dataset.

Methods used by others studying HIPPY to counter this selection bias include randomized controlled trials (RCT) (e.g. Baker, Piotrkowski & Brooks-Gunn, 1999; Bradley & Gilkey, 2002) and matching on demographic variables to produce a similar control group to the HIPPY group in non-experimental settings (e.g. Niever et al., 2011). While many studies of both experimental and non-experimental natures have been conducted on the HIPPY program, most involve a very small sample size of less than fifty HIPPY students, and few look at long-term outcomes. In a review of the literature on HIPPY, Paulsell et al. (2010) found that many papers failed to establish "a baseline equivalence for program and comparison groups" (p. 6). Thus, more rigorous studies that properly control for selection bias is needed. Of the current small pool of rigorous studies, the results are rather mixed.

One of the more rigorous RCTs was conducted by Baker et al. (1999). They randomly assigned children in New York to HIPPY and no-treatment control conditions using a lottery. They conducted this for two cohorts of children with Cohort I beginning HIPPY in the winter of 1990 and Cohort II beginning in the fall of 1991. The study conducted in New York found that Cohort I of HIPPY children performed 0.56 and 0.75 standard deviations above the mean in measures of classroom adaptation and standardized reading respectively, but this result could not be replicated in Cohort II. One possible reason might be the small sample size in Cohort I – only 69 families were in Cohort I while there where were 113 families in Cohort II. This led the authors to caution against concluding the effectiveness of the program.

The same authors also conducted a non-experimental study in Arkansas, where they separately compared two cohorts of HIPPY children to children who participated in a high quality prekindergarten program. In Arkansas, HIPPY students performed 0.59 standard deviations above the mean for classroom adaptation, but the results were insignificant for standardized achievement. In Cohort I of the Arkansas study, the control group performed better than the HIPPY group on school readiness and standardized achievement tests, with effect sizes of 0.47 and 0.63 standard deviations above the mean respectively. The authors note that variation in parental involvement may have influenced outcomes as qualitative research showed that parental involvement varied substantially. However, there were no measures to quantify the intensity of parental involvement.

Utilizing data from 21 HIPPY programs in Arkansas in a non-experimental setting, Bradley and Gilkey (2002) study the effects of HIPPY on outcomes in 3^{rd} and 6^{th} grade by performing a within-classroom matching of HIPPY children with children who are demographically similar, did not participate in HIPPY but either had no or some other preschool experience. As with Baker et al. (1999), the authors find mixed results with regards to academic performance. HIPPY students outperformed both groups of children in both the standardized tests and the school exams for English and Language Arts. They performed significantly better in standardized Math tests only when compared to students with other preschool experience. While this is a positive result, it does not resolve the discrepancy in the finding that the difference in effect size between HIPPY students and those with other preschool experience was generally *larger* than the difference in effect size between HIPPY students and students with no preschool experience. One hypothesis is family compensation for students with no preschool experience which the study was unable to measure and control for. For example, the lack of preschool could have motivated parents to prepare their children for school more intensively than they would have otherwise.

Niever et al. (2011) specifically examines the effect of HIPPY on Latino immigrant families. The authors compare a randomly selected group of HIPPY mothers with HIPPY-waitlisted mothers to examine the effect of HIPPY on the home environment and parenting behavior using the Home Observation for Measurement of the Environment (HOME) survey. They find that HIPPY families tended to have significantly higher HOME scores compared to the waitlisted families. The authors then compare a group of former HIPPY program participants in 3^{rd} grade to a control group of non-HIPPY 3^{rd} graders who had similar characteristics. They find that HIPPY students performed 0.43 standard deviations higher than the non-HIPPY students on math achievement. There was no significant difference on reading achievement.

As seen from the above studies in quasi-experimental settings, researchers commonly assume that matching on demographics such as income, gender, ethnicity and whether or not the child has attended some other preschool program, eliminates the selection bias that occurs when parents select into the program. However, this assumption may not hold if the parent that enrolls in HIPPY is different in other ways that do not show up in the demographic characteristics. These unobserved characteristics can, in turn, have an effect on how the child performs academically.

C. The Use of Sibling Comparisons

Sibling comparisons are one way to eliminate confounding from unobservable, shared family background characteristics (Almond & Currie, 2011; Griliches, 1979). Depending on the age difference, siblings are more likely to be alike than a randomly selected pair of individuals in terms of family environment. Thus, the use of sibling comparisons can provide a more credible counterfactual to a HIPPY child. In this section, I explain possible cases where the non-random selection of participation in HIPPY could bias program effects.

First, if HIPPY parents are more motivated than non-HIPPY parents, we might overestimate program effects. The converse however, is more likely to be true. A study by Baker and Roth (1997) reported how HIPPY staff were concerned about the "lack of enthusiasm" many HIPPY parents brought to the program even though they would finish the program.

On another front, we may underestimate the effect of the program because the child already starts from a poorer academic foundation than non-participants, such that post-program, the child still performs worse than non-participants. For example, Currie and Thomas (1995) find that children who attended Head Start did worse than children who did not attend Head Start. However, when compared to their own siblings, they tended to do better.

Another reason for underestimation is if there are spillover effects, particularly if the HIPPY child is the older child and he/she teaches his/her younger sibling. Garces et al. (2002) examines the longer-term effects of Head Start by comparing siblings who have been through the program with siblings who have not. They find that siblings who attended Head Start earned significantly more and were less likely to be booked for or charged with a crime, than siblings who attended another preschool program. To assess spillover effects, the authors estimated models that allow the effects of Head Start to differ if an older sibling participated in the program. They find that adults with an older sibling who attended Head Start are considerably less likely to have been booked or charged.

However, the sibling comparison model is not without its limitations. For example, it assumes that parents use the same parenting style with both children. If parents change their parenting habits from the first child to the next child, e.g. to "compensate" for lacking qualities in the second child, the sibling fixed effects estimate can underestimate the total effect of the intervention (Almond & Currie, 2011). Information on individual child-level characteristics at the time that an intervention occurs can suggest if parents are compensating for any early childhood event. This would help inform the interpretation of the estimates. Unfortunately, information on the child when he/she was undergoing HIPPY was not available for this thesis. That being said, the model can be adjusted to control for *systematic* changes in parental habits from the older child to the younger child. This is done by comparing a sibling pair where only one sibling attends HIPPY and the other does not (henceforth "the treated pair"), to a sibling pair where *both* siblings do not attend HIPPY (henceforth "the control pair").

The next section reviews two matching methods that may be used to match the treated pair with the control pair.

D. Matching

Matching enables one to provide a better counterfactual to the treated pair. Define Y_{1i} as the outcome of unit i if it is treated, and Y_{0i} as the outcome of unit i if it is not treated. In this case, I clearly cannot observe both outcomes for the same unit, that is, I cannot observe the difference in test scores in the treated pair, if the pair had not been treated. This is important because if the *potential* untreated outcome, Y_{0i}, of the treated pair systematically differs from the actual untreated outcome of the control pair, then, in observing only the outcome from the untreated and differencing that from the treated outcome, we will incorrectly estimate the effect of the treatment. For example, suppose the parents that enroll one child in HIPPY are different from those who do not in such a way that if the former had *not* enrolled in HIPPY, the potential untreated outcome would still have been different from the control parents. Then, simply comparing parents that did enroll with parents that did not enroll will incorrectly estimate the effect of the treatment on the treated.

The non-experimental studies that have been conducted on HIPPY match treated individuals to control individuals directly on covariates. This means that a control student that looks similar to a HIPPY student based on a set of demographic characteristics, is matched to that HIPPY student as the counterfactual, rendering an estimate of the outcome of the HIPPY student had he not been through HIPPY. Many matching algorithms exist to do this, for example the nearest neighbor matching algorithm. The nearest neighbor matching algorithm selects a control unit that has the smallest Mahalanobis distance ("the nearest neighbor") to each treated unit, in the specified set of covariates, and matches that to the treated unit (Abadie et al., 2004).

Formally, two conditions are assumed when implementing matching methods:

(a) $Y_{0i} \perp D_i | X_i$, and

(b) $P(D_i = 1 | X_i) < 1$

where Y_{1i}, Y_{0i} are the potential treatment and control outcomes for unit i respectively, D_i is the treatment assignment which takes a value of either 1 or 0, and X_i is the set of all covariates.

Condition (a) is the conditional independence assumption, which requires that the treatment assignment be independent of the potential outcomes conditional on the observed baseline covariates. Condition (b) indicates that the covariates do not perfectly predict participation in the treatment so that there will be similar units in the control group to match to the units in the treated group (Angrist & Pischke, 2010).[3]

[3] In this case, I estimate the average treatment on the treated. It is arguable that this is the estimate of interest since HIPPY is a targeted program and we would like to understand the impact of HIPPY on the people that attend HIPPY. Nonetheless, if one would like to estimate the average treatment effect, the conditional independence assumption will have to be strengthened to $Y_{1i}, Y_{0i} \perp D_i | X_i$. Condition (b) will also have to be strengthened to $0 < P(D_i = 1 | X_i) < 1$.

Another method is propensity score matching, a methodology first introduced by Rosenbaum and Rubin (1983). A propensity score, that is, the probability that a unit is assigned treatment, is first estimated based on some observed covariates (see equation (1)), and a control unit is then matched to a treated unit with a similar propensity score.

Mathematically, a propensity score, $p(X_i)$, is defined as follows:

$$p(X_i) = p[D_i = 1|X_i] \tag{1}$$

where $p(X_i)$ is the propensity score gives the likelihood of treatment, D_i, based on the covariates, X_i.

By matching treated and control sibling units that have a similar propensity score, one removes the systematic bias that is associated with differences in the observables in the treated and control groups.[4] If there are multiple control units with a propensity score that is close to the treated unit's score, the potential outcome of the treated unit is imputed using an average of the outcomes of the similar control subjects.

I employ both types of matching in this paper to obtain a set of controls that better mimic the treated pairs and to test for robustness. One limitation of matching directly on covariates is the dimensionality – as the number of covariates to match on increases, it becomes more difficult to find the right matches. Consequently, propensity score matching is helpful as we are now matching on an (estimated) index $p(X_i)$ that is constructed from the observables. Further, with

[4] There is a substantial volume of theoretical and empirical research using propensity scores (Dehejia & Wahba, 1999; Abadie & Imbens, 2012; Angrist, 1998). Dehejia and Wahba (1999) compare estimates obtained with propensity score matching, with the experimental estimates of a randomized control trial studying the National Supported Work program. By matching on demographic controls – age, years of schooling, ethnicity, marital status and historical earnings – they find that the non-experimental control group match more closely with the treated group than the full sample of observational control groups taken from the Panel Survey of Income Dynamics and Current Population Survey. By enforcing common support, the non-experimental estimates are consistent with the experimental estimates. This suggests that propensity score matching is a viable way of estimating average treatment effects in observational data.

the right specification for the propensity score, we essentially randomize treatment based on a set of covariates that predict likelihood of treatment. We can then compare sibling pairs in families that are likely to attend HIPPY and do end up attending it, to sibling pairs in families that are just as likely to attend HIPPY but do not. On the flipside, a large set of observables is needed to estimate the propensity score in order to better capture what drives the decision to participate in a program. I find that my data lacks sufficient information to accurately predict propensity scores for each individual, resulting in low propensity score estimates. It is important to note that for either form of matching, we can only match on the observables and not on the unobservables e.g. behavior. However, observable characteristics can be correlated to unobservable characteristics e.g. low income parents are more or less likely to read to their child. If they are correlated, matching on observables can control for the corresponding unobservable behavior.

III. Empirical Methodology

The aim of this study is to investigate whether participation in HIPPY has long-term academic effects. I look at three outcomes: standardized Reading scores, standardized Math scores and the probability of being held back a grade relative to one's peers. I employ the empirical strategy of sibling comparisons, as was used in Currie and Thomas (1995). An RCT would likely produce the most credible and robust results, but my sibling-based differences-in-differences method eliminates selection biases that plague most of the aforementioned non-RCT studies. Furthermore, no study on HIPPY thus far has used sibling differences to investigate the effect of HIPPY.

A. Differences-in-Differences Model

The differences-in-differences approach requires that we have control and treatment groups over a period of time. I exploit the sibling pairs where for one sibling, the parent has undergone HIPPY but for the other sibling in the same family, the parent has yet to undergo HIPPY. An instance in which this could happen is when a parent has not heard about HIPPY when the older child is three years old, but hears about it when the younger child turns three years old.

Figure 1 explains the procedure. Note that the time periods are interpreted slightly differently in this context than from regular differences-in-differences analyses. The earlier time period, T=0, represents the time period in which the older child is, say, a 3rd grader. The later time period, T=1, represents the time period in which it is the younger child that is now the 3rd grader. As an illustration of the first difference, consider Bob who is a 6th grader in 2011. He has a younger brother Larry who is currently a 3rd grader. Assuming that the parents treated them the same and controlling for different socioeconomic environments when they attended 3rd grade, Bob is likely to look very similar as a 3rd grader in 2008 as Larry does as a 3rd grader in 2011. The only difference is that Larry is taking the same test three years later.

Figure 1: Illustration of differences-in-differences estimate

	T=0	**T=1**
Control pair: Sibling pair where none of the siblings have undergone HIPPY	Older/non-HIPPY child's 3rd Grade performance	Younger/HIPPY child's 3rd Grade performance
Treated pair: Sibling pair where only one sibling underwent HIPPY	Non-HIPPY child's 3rd Grade performance	HIPPY child's 3rd Grade performance

Furthermore, the empirical strategy does not simply stop at comparing the younger child's academic performance (treated) with the older child's academic performance (control). This is because there may be common factors affecting parenting trends over time. For example, during the period of time between the first and second sibling, schools in this school district may promote a certain style of parenting that generally either improve or worsen their parenting behavior. Parents could also simply mature between the two children regardless of what the school has been doing. Moreover, it is worth noting that earlier-born children generally have higher achievement (Black, Devereux & Salvanes, 2005; Black, Devereux & Salvanes, 2011). Because of this, it is highly likely that the first difference taken in my empirical methodology is made smaller and potentially negative. Thus, by next comparing a treated sibling pair to a control sibling pair, I control for such systematic parenting changes and birth order effects.

B. Nearest Neighbor Matching on Covariates and Propensity Score Matching

To match control sibling pairs to treated sibling pairs, I employ two types of matching. I match treated pairs to control pairs based on a set of observed individual-level and family-level characteristics. The *teffects nnmatch* command in Stata allows one to implement nearest neighbor matching directly on covariates.

To implement propensity score matching, I estimate a propensity score for each pair with a logit model using family-level and individual sibling characteristics as the covariates. I then match sibling pairs on the estimated propensity score using the nearest neighbor matching method. The *teffects psmatch* command in Stata allows one to implement propensity score matching (Abadie & Imbens, 2012).

Formally, the propensity score is estimated with the following logit specification:

$$Pr(D_i = 1|\boldsymbol{X_i}) = logit^{-1}(\beta\boldsymbol{X_i}) \qquad (2)$$

where D_i is the treatment indicator and $\boldsymbol{X_i}$ is the set of all covariates that affect treatment status.

Next, I match sibling pairs with similar propensity scores using the nearest neighbor algorithm.

For both the model that matches directly on covariates and the propensity score matching model, family-level characteristics include ethnicity, whether or not both siblings attended the same school[5], immigrant status, and difference in age of siblings, unless otherwise noted. The demographic characteristics of the HIPPY sibling and the non-HIPPY sibling include gender and the economic disadvantage status (measured by his/her eligibility for the free or reduced price lunch program). All these individual-level characteristics are measured at the time the child took the exam. For example if the HIPPY sibling took the exam in 2008, I use his economic disadvantage status in 2008. At the same time, his corresponding non-HIPPY sibling may have taken the same exam in 2007. I control for the non-HIPPY sibling's economic disadvantage status by using his economic disadvantage status in 2007.

One concern is the reduced size of my dataset after conducting a match, making it more difficult to identify the effect of HIPPY. However, I find that not a substantial number of treated units are dropped due to the inability of finding a matching control unit. Thus, this is not a big concern.

[5] Due to the small sample size, I do not use fixed effects for the school each individual attends as there are insufficient data points for each school.

With the matched dataset, the specification tested is as follows:

$$Y_I = \alpha_I + \beta_I HIPPY_I + \theta \boldsymbol{X}_I + \varepsilon_I \tag{3}$$

where Y_i is the difference in standardized test scores of sibling pair i, $HIPPY_i$ is the variable of interest, that is, whether or not the sibling pair is a treated sibling pair, and $\boldsymbol{X}_I$ is a vector of family-level controls.

A key assumption is control families change in the same way that treated families change and that these changes affect the child similarly, for example, parents maturing, earning higher incomes, learning more about community programs, except that only the treated families enroll their (usually younger) child in HIPPY. With my differences-in-differences model that compares treated pairs to control pairs, I control for these systematic changes in families over time.

IV. Data

A. Building Sibling Dataset

I obtained data from a school district in Texas. This dataset contained 19,305 cross-sectional records of students who were officially enrolled in 2011. It contained information such as whether the child had been through HIPPY, and raw test scores for each grade between 2007 to 2011.

I first matched siblings by checking if they shared the same family name, parents' name, contact numbers and addresses. From this, 3,149 families were identified, 213 of which had at least one child that had been through HIPPY.

Next, I obtained these students' test scores from 2007 to 2011. The Texas Assessment of Knowledge and Skills (TAKS) has been the state's standardized mode of academic measurement since 2003. It assesses reading at Grades 3-9, English Language Arts from Grades 10-11 and Math from Grades 3-11, among other subjects. In 2012, Texas replaced TAKS with STAAR from Grades 3-9. In 2013 and 2014, this was extended to Grades 10 and 11 respectively. Full conversion will be completed in 2015 (Texas Education Agency, 2008). This affords me a window period of nine years of TAKS data from Grades 3-9 and ten years for Grade 10.

One challenge with the data is that the student population in the school district is known to be highly transient – student records are seen to "disappear" for a while as if the child leaves for a while and comes back again later. Thus, a student who was in Grade 8 in 2011 might not necessarily have his test scores recorded for Grades 4 in 2007, 5 in 2008, 6 in 2009 and 7 in 2010 simply because he did not take those tests in this particular school district. I would then only have his Grade 8 score.

One implication regards how many treatment and control sibling pairs I can use. For example, I may have a sibling pair where the older child's Grade 8 test score is recorded in 2007 and the records for the following years are unavailable. The younger child's Grade 6 test score is recorded in 2007, but again, no test scores are recorded between 2008-2012.[6] As a result, I cannot find a common grade between the two children and by discarding this sibling pair, further reduce the size of my dataset.

Another implication is a bias in the results as a student could do worse or better than others, not because of the treatment per se but because of his/her mobility. It is arguable that a student who is highly transient (moves in and out of the school district) is also more likely to do

[6] Another reason for this, besides being in a physically mobile familial environment, may be that both children happened to drop out of school at the same time.

worse academically and be at higher risk of retention (Burkam, Lee & Dwyer, 2009; Engec, 2006; Rumberger & Larson, 1998). In that case, my estimates will more likely underestimate the effect of HIPPY. Although I do not have information on the student's yearly location, the data suggests that the treated families may be more transient than control families because the proportion of treated pairs with siblings that attend the same school (approximately 68%) is smaller than that of control pairs (approximately 83%). I attempt to partially control for mobility by including a variable that indicates if both siblings attended the same school at the time they took their tests. Since students usually attend school within their neighborhood, if both siblings were enrolled in the same school, it suggests that the family did not move.

The dataset also includes an indicator variable on whether a student is in the Limited English Proficient (LEP) program, however I do not use this variable in my estimations. This is because LEP status is identified in part by how well a student reads and his/her familiarity with the English language (Texas Education Agency, 2012). A student's language ability is very likely to be influence the decision of the parent to participate in HIPPY. This indicator could thus be endogenous to the treatment on HIPPY.

B. Sibling Spillover Effects

To test for spillovers from older to younger siblings, I create a variable that indicates if it is the older sibling that underwent HIPPY (Garces et al., 2002). In the treated group, such sibling pairs account for approximately 16% of all sibling pairs.

C. Outcome Variables

As mentioned earlier, I aim to study the effect of HIPPY on three outcomes: standardized Reading scores, standardized Math scores and the probability of being held back a grade relative to one's peers.

Given that I am comparing the difference in test scores between one sibling pair, to the difference between another sibling pair, the outcome variable is then the *difference* in standardized Reading scores and standardized Math scores. To create these two variables, I first standardize the individual sibling's test score, and then take the differences in test scores within a sibling pair.

Standardizing test scores

Within each sibling pair, the test score compared is always of the same grade. However, between control and treated sibling pairs, the grades compared may be different. For example, treated pair A may have sibling differences in test scores corresponding to Grade 5. However, control pair B may have sibling differences in test scores corresponding to Grade 7. Thus, we need to standardize test scores. This is a common method used to "stack" grades such that we might measure the impact of an intervention on the same scale even if the test taken is different.

The state-level mean and standard deviations of test scores from Grade 3 to Grade 10 during the period that TAKS was implemented (see Table A in Appendix) is used to implement the following score standardization formula:

$$Z_{jgt} = \frac{TestScore_j - Mean_{gt}}{SD_{gt}} \quad (4)$$

where Z_{jgt} is the standardized score of student j at grade g and test year t, $TestScore_j$ is the raw score of student j, $Mean_{gt}$ and SD_{gt} are the mean score and standard deviation of the scores of all students who took the test at grade g and test year t, respectively.

Taking the difference within siblings

After standardizing individual scores, I take the difference between the standardized scores of siblings. For treated pairs, I subtract the non-HIPPY sibling's score from the HIPPY sibling's score to obtain the difference in test scores.

It is not as simple for control pair siblings, both of whom have not been through HIPPY. This raises the question of which direction scores should be subtracted in. Thus, to derive the differences between them, I randomly select 16% of the control pairs and create differences that assume the following direction:

$$Difference_i = Score_{OlderSibling_i} - Score_{YoungerSibling_i} \tag{5}$$

where $Difference_i$ is the difference in test scores between siblings in pair i. This proportion matches the roughly 16% of treated pairs where the HIPPY sibling is the older sibling.

I switch the direction of differencing for the remaining 84% of control pairs because I find that the HIPPY sibling is the younger sibling in roughly 84% of the treated pairs. I implement this to construct outcome variables for the difference in standardized Reading scores, and the difference in standardized Math scores.

I randomly assign which of the control pairs is more likely to be an older-sibling-is-HIPPY-sibling variant because there is no logical variable to match on. Given the data, it is unclear what the selection process is behind which child in birth order goes through HIPPY.

Tables 1 and 2 show the descriptive statistics for the full sample of data.

In both cases, the majority of students in both control and treatment pairs are Hispanic and are on the free lunch program. Treated pairs have larger differences (in either direction) in standardized test scores than control pairs, on average. There are two possible reasons for the difference in magnitude. One is that HIPPY may have had an additional effect on the HIPPY child such that he/she does better above and beyond the difference that accrues from birth order and time lapse (this is accounted for by the control pair's "HIPPY" sibling). Another possible reason is that parents are treating each sibling within a pair differently in the control group, in a way that is not the same within other pairs in the treated group. These differences in parenting style in turn are not distributed in a similar proportion in the treated group as they are in the control group. Matching to yield more suitable control pairs to each treated pair should help alleviate this.

Figure 2 shows how the HIPPY sibling does in comparison to his counterpart HIPPY sibling in the control pair (where HIPPY sibling of the control pair is the younger sibling for 84% of the control pairs and the older sibling for 16% of the control pairs). Within pairs, the HIPPY sibling tends to do better than the non-HIPPY sibling on average, scoring 0.235 standard deviations below the mean as compared to the control pair. Between pairs, the HIPPY sibling of a treated pair does better in both Reading and Math than the chosen HIPPY sibling of the control pair. In contrast, the non-HIPPY sibling of the treated pair does poorer than the chosen non-HIPPY sibling of the control pair. This suggests that the siblings in the treatment pair may be starting with a lower academic foundation than the siblings in the control pair.

Figure 2: Simple comparison of standardized scores of HIPPY and non-HIPPY siblings

	Reading		**Math**	
	Treatment pair	**Control pair**	**Treatment pair**	**Control pair**
HIPPY sibling	-0.235 (1.11)	-0.406 (1.27)	-0.0175 (1.07)	-0.220 (1.28)
Non-HIPPY sibling	-0.458 (1.32)	-0.407 (1.27)	-0.274 (1.366)	-0.242 (1.34)

Standard deviations are in parentheses.

Table 1: Descriptive statistics for sample used to estimate effect on TAKS Reading score

		Treatment Pairs	Control Pairs
Number of Pairs		151	2,772
HIPPY sibling is the older sibling		25 (16%)	450* (16%)
Average difference in standardized test scores (% pairs)**			
	Grade 3[a]	0.413 (0.7%)	-0.669 (1.0%)
	Grade 4	-0.160 (21.9%)	-0.0256 (19.7%)
	Grade 5	0.452 (33.8%)	-0.103 (20.0%)
	Grade 6	-0.0880 (18.5%)	0.0968 (22.2%)
	Grade 7	0.136 (13.9%)	0.0131 (17.2%)
	Grade 8	1.19 (7.3%)	0.151 (11.8%)
	Grade 9	-0.144 (3.3%)	0.242 (6.5%)
	Grade 10[b]	1.79 (0.7%)	0.244 (1.6%)
Income (% pairs)			
	Free Lunch	129 (85.4%)	2,112 (76.2%)
	Reduced Lunch	7 (4.6%)	133 (4.4%)
	NA	15 (9.9%)	537 (19.4%)
School (% pairs)			
	Same School	104 (68.9%)	2,294 (82.8%)
Ethnicity (% pairs)			
	American Indian/Native	0 (0%)	19 (0.7%)
	Asian/Pacific Islander	1 (0.7%)	85 (3.1%)
	Black	2 (1.3%)	202 (7.3%)
	Hispanic	140 (92.7%)	2,071 (74.7%)
	Multiple	4 (2.7%)	35 (13%)
	White	3 (2.0%)	323 (11.7%)
Age Difference (days)***		881	779

**Control pairs were randomly chosen whereby the difference in test scores were calculated such that it was of the older sibling-younger sibling variant. Means and standard deviations of each test by grade by year are in Table A of the Appendix.*
***Difference in test scores = Score of HIPPY sibling - Score of non-HIPPY sibling*
****Age difference = Age of oldest sibling - Age of youngest sibling*
[ab] There is only one sibling pair in the treated sample where this was the grade of comparison.

Table 2: Descriptive statistics for sample used to estimate effect on TAKS Math score

	Treatment Pairs	Control Pairs
Number of Pairs	154	2774
HIPPY sibling is the older sibling (% pairs)	25 (16%)	450* (16%)
Average difference in standardized test scores (% pairs)**		
Grade 3[a]	-0.725 (0.65%)	0.225 (0.79%)
Grade 4	0.233 (21.4%)	0.0455 (19.7%)
Grade 5	0.302 (34.4%)	-0.0446 (20.2%)
Grade 6	-0.224 (18.2%)	-0.0523 (22.3%)
Grade 7	0.217 (14.29%)	-0.0536 (17.2%)
Grade 8	0.952 (7.14%)	0.141 (11.9%)
Grade 9	0.0742 (3.25%)	0.347 (6.49%)
Grade 10[b]	3.693 (0.65%)	0.122 (1.48%)
Economic Disadvantaged Status (% pairs)		
Free Lunch	128 (83.1%)	2,091 (75.4%)
Reduced Lunch	6 (3.9%)	122 (4.4%)
NA	20 (13.0%)	561 (20.2%)
School (% pairs)		
Same School	104 (67.5%)	2,296 (82.8%)
Ethnicity (% pairs)		
Asian/Pacific Islander	1 (0.6%)	85 (3.1%)
American Indian/Native	0 (0.0%)	19 (0.7%)
Black	3 (2.0%)	202 (7.3%)
Hispanic	141 (91.6%)	2,073 (74.7%)
Multiple	5 (3.3%)	36 (1.3%)
White	3 (2.0%)	323 (11.6%)
Age Difference (days)**	894	780

**Control pairs were randomly chosen whereby the difference in test scores were calculated such that it was of the older sibling-younger sibling variant. Means and standard deviations of each test by grade by year are in Section A of the Appendix.*
***Difference in test scores = Score of HIPPY sibling - Score of non-HIPPY sibling*
****Age difference = Age of oldest sibling - Age of youngest sibling*
[ab] *There is only one sibling pair in the treated sample where this was the grade of comparison.*

Held Back A Grade

The third outcome I seek to understand is the probability of being held back a grade relative to one's peers. The data does not specifically report whether or not a student was held back. However, I am able to use information on grades students enroll in, in two consecutive years 2011 and 2012, their age and birth month to generate two types of indicator variables to capture whether the child is in the "right" grade for his age. This helps me determine if a child is progressing through the grades at the same time relative to others his/her age.

The first indicator variable I create is $HELDBACK_{DIFF}$, which uses the difference in enrolled grades between academic years 2011 and 2012.[7] One would predict that a student moves up a grade with each successive year. $HELDBACK_{DIFF}$ takes the value of 1 if the difference in enrolled grades between academic years 2011 and 2012 is 0. If the student enrolls two grades above the grade they were previously at, they are "ahead" of their peers, and if they stay at the same grade, they are "held back" relative to their peers.

The second indicator, $HELDBACK_{PREDICT}$, uses the age and birth month to determine the predicted grade a student should be in, in the academic year 2011, given that the school cut-off date for age calculation is 1st September of every year. $HELDBACK_{PREDICT}$ takes the value of 1 if the student is enrolled in a grade lower than the predicted grade.

One potential limitation of the indicator variable $HELDBACK_{PREDICT}$ as shown by Cascio (2005) is the presence of red-shirting, that is, when parents start their child in school late. Due to red-shirting, there can be a higher number of false positives. Approximately one-fifth of students are not necessarily repeaters, but instead are simply old for their class. Of those that are thought to be repeaters, about one-tenth are not. This means that the result is likely to be

[7] The years 2011 and 2012 were chosen as they had the least number of missing observations in demographic measures and test scores.

overestimated. In my case, I find that HIPPY and non-HIPPY students are largely similar in their patterns of red-shirting. Of the non-HIPPY students, only 1% are enrolled in a grade that is the same level as the predicted grade based on their age. No HIPPY students are enrolled in a grade that is the same level as their predicted grade. Since the two groups do not differ much in the extent that red-shirting occurs, this concern is not as important.

V. Results and Discussion

A. On Academic Achievement

I first present differences in standardized Reading and Math scores within the treated pairs, that is, the 'first-differences', in Table 3.

Three sets of estimates are presented for Reading and Math scores. Specification (1) does not include any observed covariates – I am thus only testing the means of the test scores. Specification (2) controls for individual characteristics i.e. gender and age of sibling at the time the test is taken find. Specification (3) includes a family-level characteristic i.e. the economic disadvantage status of sibling at the time the test is taken.

The variable of interest is $TREATED$, which is a dummy variable for whether or not an individual is the HIPPY sibling in a sibling pair. I find that compared to their sibling, being in HIPPY is associated with a positive increase in test scores for Reading of 0.217 standard deviations (with no controls), 0.168 standard deviations (with only individual-level controls) and 0.180 standard deviations (with individual and family-level controls). However, this is not statistically significant. For Math, HIPPY siblings see a statistically significant increase by 0.234 standard deviations over their non-HIPPY siblings. However, as I control for individual-level and family-level characteristics, this statistical significance goes away. Controlling for individual and family-level characteristics, HIPPY is associated with a positive 0.167 standard deviations increase over one's non-HIPPY sibling.

These results suggest that HIPPY could potentially be ineffective in improving the child's test scores. Another possible reason is that parents in the treatment group are compensating in different ways for their non-HIPPY and HIPPY child. Matching may help to reduce this problem if the matched control pair's parents treat their "HIPPY" child and "non-HIPPY" child in a similar way as the treated pair's parents.

First differences do not control for the systematic changes that happen in the family as time passes between two children. For example, parents may mature and change their parenting style accordingly. To control for these systematic changes, I compare treated pairs with control pairs of siblings.

I present these results in Table 4, where the dependent variables are the difference in standardized Reading scores between siblings in Panel A, and the difference in standardized math scores between siblings in Panel B. The specifications are run on the entire sample, and thus the treated pairs are matched to a control group that may not resemble the treated pairs as well.

Most of the coefficient estimates are statistically insignificant. However, controls for whether or not a sibling pair is from an immigrant family is seen to be consistently statistically significant across Panel A and B. Being Black, Hispanic is associated with a wider gap between children for Math scores, but not for Reading scores, suggesting that HIPPY may have different effects based on race, positively impacting students of Black and Hispanic ethnicity.

Table 3: First differences in standardized Reading and Math scores

	Panel A			Panel B		
	Reading			Math		
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Treated	0.217	0.168	0.180	0.234**	0.160	0.167
	(0.135)	(0.178)	(0.179)	(0.114)	(0.150)	(0.152)
Gender		-0.241	-0.196		-0.355**	-0.343**
		(0.192)	(0.192)		(0.161)	(0.162)
Age		-0.0220	0.0128		-0.0301	-0.0260
		(0.0669)	(0.0684)		(0.0558)	(0.0575)
Economic Disadvantage Status						
Free Lunch			0.251			-0.0340
			(0.502)			(0.366)
Reduced Lunch			-0.0898			-0.209
			(0.549)			(0.412)
Constant	-0.459***	-0.0113	-0.692	-0.252***	0.377	0.365
	(0.0953)	(0.973)	(1.126)	(0.0808)	(0.813)	(0.947)
Observations	302	302	298	308	308	308
R-squared	0.554	0.559	0.575	0.670	0.68	0.682

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4: Results for differences-in-differences (whole sample)

	Panel A			**Panel B**		
	Reading			Math		
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Treated	0.183	0.157	0.155	0.204	0.197	0.196
	(0.175)	(0.173)	(0.173)	(0.147)	(0.144)	(0.144)
	2.86e-06	-1.76e-05	-1.96e-05	9.75e-05	0.000100	9.99e-05
Difference in age (days)	(6.37e-05)	(7.89e-05)	(7.90e-05)	(6.39e-05)	(7.83e-05)	(7.84e-05)
Immigrant status	0.290**	0.285**	0.279**	0.465***	0.457***	0.458***
	(0.131)	(0.139)	(0.139)	(0.121)	(0.126)	(0.126)
HIPPY student is the older of the sibling	-0.168**	-0.176**	-0.174**	-0.213***	-0.206***	-0.206***
	(0.0819)	(0.0824)	(0.0824)	(0.0759)	(0.0764)	(0.0765)
Ethnicity						
American Indian/Native	-0.0533	-0.0553	-0.0632	-0.526	-0.491	-0.499
	(0.229)	(0.232)	(0.228)	(0.369)	(0.375)	(0.378)
Asian/Pacific Islander	0.0637	0.0700	0.0529	0.176	0.247	0.236
	(0.152)	(0.175)	(0.176)	(0.181)	(0.209)	(0.213)
Black	0.0414	0.0303	0.0325	0.292**	0.328**	0.321**
	(0.118)	(0.134)	(0.130)	(0.139)	(0.162)	(0.162)
Hispanic	0.0995	0.106	0.0933	0.167**	0.210**	0.202*
	(0.0794)	(0.0943)	(0.0920)	(0.0795)	(0.102)	(0.103)
Missing	-0.183	-0.232	-0.259	0.362	0.364	0.354
	(0.269)	(0.288)	(0.292)	(0.278)	(0.287)	(0.289)
Multiple	-0.242	-0.284	-0.346	-0.0200	-0.00707	-0.0139
	(0.365)	(0.403)	(0.407)	(0.527)	(0.597)	(0.596)
Same School[a]		-0.0724	-0.0778		-0.126	-0.129*
		(0.0851)	(0.0852)		(0.0782)	(0.0782)
Characteristics of HIPPY sibling						
Economic disadvantage status						
Free Lunch		-0.0173	-0.0476		-0.0143	-0.0282
		(0.0895)	(0.130)		(0.0983)	(0.134)
Reduced Lunch		-0.0880	-0.146		-0.292**	-0.326**
		(0.116)	(0.138)		(0.123)	(0.143)
Gender		-0.131**	-0.136**		0.102*	0.102*
		(0.0567)	(0.0568)		(0.0607)	(0.0610)

			Characteristics of non-HIPPY sibling			
Economic disadvantage status						
Free Lunch			0.0549			0.0278
			(0.118)			(0.124)
Reduced Lunch			0.165			0.0932
			(0.129)			(0.135)
Gender			0.173***			0.00629
			(0.0561)			(0.0602)
Constant	-0.0422	0.121	0.0202	-0.209**	-0.163	-0.174
	(0.0836)	(0.146)	(0.144)	(0.0816)	(0.136)	(0.136)
Observations	2,923	2,773	2,773	2,928	2,781	2,781
R-squared	0.008	0.011	0.016	0.016	0.021	0.021

Standard errors are clustered on the family and are in parentheses.

[a] Same school is an indicator variable for whether or not the HIPPY sibling and the non-HIPPY sibling attended the same school for the grade which they are compared with. This is used instead of categorical variables for each school because I find that individual school fixed effects difference away much of the variation in test scores. This is especially so because in some schools, I only have data on a few pairs of siblings.

*** p<0.01, ** p<0.05, * p<0.1

The variable of interest is again, $Treated$, which is an indicator variable that tells us whether the pair is treated or not i.e. one of the siblings has been through HIPPY. If the pair is a treated pair, the difference in standardized Reading test scores between siblings in a treated pair is on average 0.155 to 0.183 standard deviations higher than between siblings in a control pair. This difference is also positive, which means that the HIPPY sibling is scoring higher relative to his/her non-HIPPY sibling. The estimated difference in standardized Math test scores between siblings in a treated pair is slightly higher – the difference between siblings in a treated pair is on average 0.196 to 0.204 standard deviations higher than between siblings in a control pair. However, these results are not statistically significant.

It is interesting to note that there is a strong suggestion of spillover from the older sibling to the younger sibling when it is the older sibling that undergoes HIPPY. To interpret the coefficient on $HIPPY\ Student\ is\ the\ Older\ of\ the\ Sibling,$ I add the coefficient estimate to the coefficient on $Treated$. For standardized Reading scores, I find that if the pair is a treated pair and the HIPPY sibling is also an older sibling, the difference becomes negligible. Comparatively, when the HIPPY sibling is a younger sibling, there is a positive but statistically insignificant difference in achievement. The same results are seen with standardized Math scores. This result is robust across all specifications. This suggests that there is no sibling difference in achievement if the older child is in HIPPY, but that there is a sibling difference if it is the younger child that attended HIPPY, particularly for Math.

I next use nearest neighbor matching on demographic variables directly and propensity score matching to narrow the control pairs down to pairs which more closely resemble the treated pairs. For easier comparison, these results are presented side by side in Table 5. The results from the matching on demographic variables directly are presented in columns (i) and (ii), and the results from propensity score matching are presented in columns (iii) and (iv).[8]

By directly matching on demographic covariates, I find that there is no change in the significance of the effects on Reading scores, but that the effect on Math scores is highly significant. The effects on Math range from 0.416 to 0.588.

Apart from the reason that perhaps HIPPY indeed has no effect on a child's reading achievement, another reason for these results may be the possibility that either of the siblings have attended a program outside of HIPPY that affected Reading or Math scores and this is not captured by the demographic characteristics used in the matching process. For example, parents may have decided that the two siblings should attend different programs: the non-HIPPY child attends pre-kindergarten while the HIPPY sibling attends none. This is a plausible scenario given that the first few public pre-kindergarten programs in this school district were established in 1998, overlapping with the birthdays when siblings turn of age for siblings of approximately half of all sibling pairs. There would also have been high publicity to create awareness surrounding these new programs. It is likely that such a program could affect Reading scores because one of the criteria to attend pre-kindergarten in this school district is that English must not be the child's primary language.[9] This is more likely to happen for a non-HIPPY child who, in turn, is often the older child. Such a criterion suggests that pre-kindergarten aims in part to be a place where the child interacts in English and improves their fluency at the language. Additionally, the curriculum for pre-kindergarten focuses strongly on reading. It then becomes

[8] See Appendix Section B for a snapshot of the statistics pertaining to the matched samples.

[9] In terms of the impact this has on the validity of the estimates here, the only scenario in which we may have grossly overestimated the effect sizes is if the HIPPY child simultaneously attended pre-K and HIPPY, while his/her sibling did not attend either and both the control pair's HIPPY and non-HIPPY sibling did the same thing (i.e. either attended or did not attend pre-K or HIPPY). Otherwise, for the most part, our estimates are likely to either be almost unchanged or biased downwards.

important to control for this intervention. However, such information was not available for this paper.

Table 5: Estimates from matching models

SPECIFICATION	Matching on Demographic Variables: Reading	Matching on Demographic Variables: Math	Propensity Score Matching: Reading	Propensity Score Matching: Math
	(i)	(ii)	(iii)	(iv)
Family-level demographic controls[a]	0.376	0.588**	0.387	0.739*
	(0.306)	(0.255)	(0.418)	(0.404)
Family-level demographic controls, HIPPY sibling characteristics[b]	0.194	0.434***	0.247	0.322
	(0.178)	(0.149)	(0.180)	(0.271)
Family-level demographic controls, HIPPY and non-HIPPY sibling characteristics[c]	0.199	0.416***	0.133	0.268
	(0.182)	(0.154)	(0.209)	(0.249)

Standard errors are clustered on the pair and are in parentheses.
[a] Family demographic controls include: age difference between siblings, whether or not the siblings went to the same school, immigration status, and ethnicity
[b] HIPPY sibling characteristics include: the economic disadvantage status of family at the time the HIPPY sibling sat for the exam, and gender
[c] Non-HIPPY sibling characteristics include: the economic disadvantage status of family at the time the non-HIPPY sibling sat for the exam, gender
*** p<0.01, ** p<0.05, * p<0.1

Propensity score matching yields slightly different results. The propensity score matching estimates indicate that relative to their non-HIPPY siblings, HIPPY siblings were between 0.133 and 0.387 standard deviations ahead in Reading scores. However, none of the estimates are statistically significant. Relative to their non-HIPPY siblings, HIPPY siblings were between 0.268 and 0.739 standard deviations ahead in Math. Only the specification which accounts for only family-level demographic controls is statistically significant at the 10% level. This suggests that the gap between the effect of HIPPY versus the effect of “no HIPPY”, that is the “difference in differences,” is approximately 0.7 when you take into consideration the demographic characteristics of the family.

One plausible reason for the difference in results obtained from the two matching methods could be that the propensity scores are not able to identify good controls since the covariates used to estimate the propensity scores of each pair appear to lack predictive capability. Propensity scores estimates ranged from 0.000141 to 0.251, depending on which covariates were used to predict propensity scores. Further, the sample of control pairs matched directly on covariates to treated pairs look marginally more similar to the treated pairs than do those matched using propensity scores.

B. On Being Held Back A Grade

Table 6 presents the estimates from matching directly on demographic covariates for the effect of HIPPY on a student, relative to others his/her age.[10] I find that there is a small but statistically significant effect of HIPPY on the likelihood of being held back in school. HIPPY students are between 3.0% and 11.7% less likely to be held back relative to their peers. This is a wide range that could be because students are more likely to be held back in certain grades than others. Thus, I include a variable that indicates which Grade the student is in, in 2011, and present the results in columns (ii) and (iv) of Table 6. When matched on the grade the student is enrolled in, in 2011, one finds the statistical significance disappear when $HELDBACK_{DIFF}$ is the dependent variable. One reason for this may be that the progression from Grade 3 to 4 is likely to be different from the progression from Grade 8 to 9. For example, it may be easier to progress from Grade 3 to 4 than it is to progress from Grade 8 to 9.

A logical way of segmenting difficulty in progression might be to look at progression through Grades 3-6 differently from progression through Grades 7-10. Thus, as a robustness check, I restrict my sample to students who are below the age of 13 in 2011. I find that the

[10] I also employ propensity score matching and yield the same results.

results hold for all four model specifications (see Table 7). HIPPY students are on average between 0.4% to 8.4% less likely to be held back relative to their peers.

Table 6: Estimates on probability of being held back

Indicator variable	**HELDBACK_DIFF (i)**	**HELDBACK_DIFF (ii)**	**HELDBACK_PREDICT (iii)**	**HELDBACK_PREDICT (iv)**
Average treatment effect	-0.0299***	-0.0163	-0.104***	-0.117***
S.E. (Robust)	0.00440	0.0128	0.104	0.0404
Matched on	Economic disadvantage, Immigrant status, Ethnicity, Gender	Economic disadvantage, Immigrant status, Ethnicity, Gender, Grade in 2011	Economic disadvantage, Immigrant status, Ethnicity, Gender	Economic disadvantage, Immigrant status, Ethnicity, Gender, Grade in 2011

Standard errors are clustered on the pair and are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 7: Estimates on probability of being held back between Grades 3 and 6

Indicator variable	**HELDBACK_DIFF (i)**	**HELDBACK_DIFF (ii)**	**HELDBACK_PREDICT (iii)**	**HELDBACK_PREDICT (iv)**
Average treatment effect	-0.00402***	-0.00415***	-0.087**	-0.084**
S.E. (Robust)	0.000831	0.000829	0.011	0.03
Matched on	Economic disadvantage, Immigrant status, Ethnicity, Gender	Economic disadvantage, Immigrant status, Ethnicity, Gender, Grade in 2011	Economic disadvantage, Immigrant status, Ethnicity, Gender	Economic disadvantage, Immigrant status, Ethnicity, Gender, Grade in 2011

Standard errors are clustered on the pair and are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

VI. Conclusion

In this paper, I compare siblings to one another to obtain the first difference, and then compare this difference to control sibling pairs in an attempt to control for selection bias in HIPPY participation. I find that when comparing to their siblings, HIPPY siblings do not achieve significantly better when additional controls are added. I match treated sibling pairs to control sibling pairs directly on covariates and find statistically significant results for the effect of HIPPY on Math scores. This may be because when the treated pair is matched to a control pair that closely resembles it, it controls for the unobservable nature of differences in parental contribution to each sibling's development. However, I caution against simply concluding that HIPPY has been effective because a robustness check with propensity score matching finds that similar results (in terms of magnitude and sign) are insignificant. There also appears to be a strong spillover effect from the treated sibling to the non-treated sibling within a pair. I find a statistically significant negative association between HIPPY participation and the likelihood of being held back.

To put in perspective my estimates of 0.42 to 0.59 in improvements of a HIPPY sibling compared to a non-HIPPY sibling for Math, I compare my estimates with results of other HIPPY studies. Niever et al. (2011) find HIPPY effects of 0.43 standard deviations above the mean in Math achievement, and similarly no significant effect in reading achievement. Bradley and Gilkey (2002) find effect sizes of 0.45 to 0.50 standard deviations in Math achievement but only when comparing HIPPY students to students with some other preschool experience.

There has been much hype in recent years about expanding universal pre-K access. Using a differences-in-differences model, Fitzpatrick (2008) finds that universal Pre-K availability is associated with an increase in 4th Grade test scores of economically disadvantaged children in urban fringes ranging from -0.002 to 0.09 standard deviations, depending on race. Universal Pre-K availability, at a cost of $16,000 to $41,000 per student, however is far more expensive than implementing a targeted parental intervention program such as HIPPY, which costs approximately $1,600 per student (Fitzpatrick, 2008; Kilburn & Karoly, 2008).

As a policymaker deciding which options to put taxpayer money, it is the relative effect of this program versus others that is also valuable to know. Gomby (2005) conducts a meta-analysis of home visitation programs and finds that while they generally do have a statistically significant effect on the outcomes they hope to measure, the effect is by and large, rather small (<0.20). The effect sizes of all estimates we have here are in the moderate range.

Finally, more research needs to be done to understand if programs like HIPPY are useful. This study is limited because of the small sample size, which may lack the statistical power to pick up any statistically significant and positive effects of HIPPY, if any. The lack of knowledge on how intensively parents participate e.g. a parent could conceivably drop out of the program half-way but this is not recorded in the data, and whether the parent enrolled the non-HIPPY child in some other program, suggest that the results here may be underestimations of the true effect of the program. My results may thus be more indicative of the effects of intent to treat rather than the effects of treatment on the treated. More information also needs to be known about how parents treat each child, for example the other early childhood programs children attend. In addition, it is important to note that cognitive development is only one part of a child's development. My estimates do not capture the non-cognitive effects of a program like HIPPY, which in turn are associated with many other positive outcomes. Thus, any cost-benefit analysis should take this into consideration.

VII. Appendix

Section A: Mean and Standard Deviations of TAKS tests

Exhibit 1: Mean and standard deviations of TAKS tests by year administered and grade (Texas Education Agency, 2007-2012[11])

Year Administered	Grade	TAKS Math		TAKS Reading	
		Mean	SD	Mean	SD
2003	3	30.28	6.63	28.78	6.32
2004	3	31.964	5.955	30.51	5.41
2005	3	31.755	6.09	30.195	5.944
2006	3	31.95	6.39	30.88	5.63
2007	3	31.973	6.356	30.079	5.944
2008	3	32.358	6.385	30.404	5.797
2009	3	32.843	6.549	30.249	5.713
2010	3	32.43	6.37	30.18	5.77
2011	3	32.705	6.251	29.689	5.938
2003	4	30.92	7.30	30.84	7.29
2004	4	32.52	6.876	31.122	6.875
2005	4	33.424	7.099	32.257	6.553
2006	4	34.02	6.78	31.53	5.88
2007	4	34.525	6.538	32.369	6.157
2008	4	34.064	6.773	32.779	6.298
2009	4	34.961	6.881	33.326	6.073
2010	4	34.81	6.62	33.13	6.14
2011	4	34.858	6.588	32.981	6.234
2003	5	31.95	7.35	31.07	7.68
2004	5	33.646	7.436	32.871	7.287
2005	5	34.757	6.843	33.101	6.596
2006	5	35.56	7.28	33.22	6.56
2007	5	36.213	6.667	33.783	6.283
2008	5	35.911	7.071	34.111	6.467
2009	5	36.354	7.224	33.781	6.783
2010	5	35.91	7.41	35.10	5.82
2011	5	35.84	7.425	35.525	5.749
2003	6	30.65	9.01	30.65	8.23
2004	6	31.584	8.685	32.163	7.181
2005	6	33.487	8.985	33.039	7.046
2006	6	35.06	7.96	34.58	6.09

[11] The statistics in this table were derived from annual state reports on TAKS performance.

2007	6	35.278	8.264	35.18	5.741
2008	6	35.548	8.652	35.222	6.201
2009	6	35.543	8.366	35.115	5.817
2010	6	35.60	8.35	34.72	6.21
2011	6	35.735	8.294	35.04	6.159
2003	7	28.33	9.48	36.29	7.90
2004	7	30.15	9.251	37.579	8.079
2005	7	31.178	9.82	38.208	7.704
2006	7	32.59	9.33	38.79	7.30
2007	7	34.067	9.162	38.416	6.941
2008	7	33.807	9.805	39.426	7.877
2009	7	34.927	9.101	39.882	7.09
2010	7	34.77	8.93	38.49	7.34
2011	7	34.756	8.944	38.532	7.344
2003	8	29.97	9.31	38.27	8.23
2004	8	32.141	10.071	39.192	6.84
2005	8	32.088	10.046	40.06	7.974
2006	8	33.644	9.75	40.008	7.77
2007	8	34.836	9.262	40.61	6.66
2008	8	35.781	9.52	41.696	6.306
2009	8	36.964	9.114	42.322	6.192
2010	8	36.54	9.15	42.23	5.91
2011	8	36.462	9.112	41.967	6.284
2003	9	29.64	10.37	29.33	6.38
2004	9	30.989	10.831	26.679	6.929
2005	9	32.306	10.355	30.676	5.84
2006	9	32.13	10.49	31.13	5.81
2007	9	33.058	10.528	31.935	5.2
2008	9	33.327	11.376	30.802	6.276
2009	9	34.946	10.808	32.645	5.18
2010	9	33.99	10.93	32.34	4.92
2011	9	34.038	10.855	33.077	5.601
2003	10	32.75	11.07	50.51	10.38
2004	10	34.089	11.262	46.674	10.887
2005	10	35.047	11.393	50.78	9.633
2006	10	35.821	11.432	55.048	9.007
2007	10	35.791	11.563	54.551	9.171
2008	10	37.179	12.166	55.128	9.027
2009	10	37.266	11.551	57.199	8.408
2010	10	38.59	10.66	56.49	8.44
2011	10	38.553	10.577	56.367	8.411
2012	10	38.524	10.71	57.046	8.118

Section B: Statistics from Matching

Tables 8 and 9 are snapshots of the descriptive statistics for the matched pairs when matched directly on covariates and when matched on propensity scores, using the full range of covariates (family-level and sibling-level).

Table 8: Descriptive statistics of treated pairs and matched control pairs in sample matched directly on covariates, used to estimate effect on TAKS Math score

	Treatment Pairs	Control Pairs
Number of Pairs	154	165
HIPPY sibling is the older sibling	25 (16%)	27 (16%)
Average difference in standardized test scores* (% pairs)		
Grade 3[a]	-0.726 (0.7%)	-1.04 (0.6%)
Grade 4	0.233 (21.4%)	-0.228 (19.4%)
Grade 5	0.302 (34.4%)	-0.117 (16.4%)
Grade 6	-0.224 (18.2%)	0.0759 (20.0%)
Grade 7	0.217 (14.3%)	-0.0557 (23.6%)
Grade 8	0.952 (7.1%)	0.136 (15.2%)
Grade 9	0.0742 (3.3%)	0.598 (4.2%)
Grade 10[b]	3.69 (0.7%)	-1.78 (0.6%)
Income (% pairs)		
Free Lunch	128 (83.1%)	140 (84.9%)
Reduced Lunch	6 (3.9%)	6 (3.6%)
NA	20 (13.0%)	19 (11.5%)
School (% pairs)		
Same School	104 (67.5%)	120 (72.7%)
Ethnicity (% pairs)		
American Indian/Native	0 (0.0%)	0 (0.0%)
Asian/Pacific Islander	1 (0.7%)	1 (0.6%)
Black	3 (2.0%)	4 (2.4%)
Hispanic	141 (91.6%)	154 (93.3%)
Multiple	5 (3.3%)	2 (1.2%)
White	3 (2.0%)	3 (1.8%)
Age Difference (days)**	894	879

**Difference in test scores = Score of HIPPY sibling - Score of non-HIPPY sibling*
***Age difference = Age of oldest sibling - Age of youngest sibling*
[ab] There is only one sibling pair in the treated sample where this was the grade of comparison.

Table 9: Descriptive statistics of treated pairs and matched control pairs in sample matched with propensity scores, used to estimate effect on TAKS Math score

	Treatment Pairs	Control Pairs
Number of Pairs	154	136
HIPPY sibling is the older sibling	25 (16%)	27 (20%)
Average difference in standardized test scores* (% pairs)		
Grade 3[a]	-0.726 (0.7%)	- (0.0%)
Grade 4	0.233 (21.4%)	0.510 (18.4%)
Grade 5	0.302 (34.4%)	0.277 (18.4%)
Grade 6	-0.224 (18.2%)	-0.602 (26.5%)
Grade 7	0.217 (14.3%)	-0.254 (20.6%)
Grade 8	0.952 (7.1%)	0.0716 (12.5%)
Grade 9	0.0742 (3.3%)	1.07 (3.7%)
Grade 10[b]	3.69 (0.7%)	- (0.0%)
Income (% pairs)		
Free Lunch	128 (83.1%)	114 (83.8%)
Reduced Lunch	6 (3.9%)	12 (8.8%)
NA	20 (13.0%)	10 (7.4%)
School (% pairs)		
Same School	104 (67.5%)	104 (76.5%)
Ethnicity (% pairs)		
American Indian/Native	0 (0.0%)	0 (0.0%)
Asian/Pacific Islander	1 (0.7%)	0 (0.0%)
Black	3 (2.0%)	1 (0.7%)
Hispanic	141 (91.6%)	128 (94.1%)
Multiple	5 (3.3%)	4 (2.9%)
White	3 (2.0%)	2 (1.5%)
Age Difference (days)**	894	923

**Difference in test scores = Score of HIPPY sibling - Score of non-HIPPY sibling*
***Age difference = Age of oldest sibling - Age of youngest sibling*
[ab] *There is only one sibling pair in the treated sample where this was the grade of comparison.*

VIII. Reference List

Abadie, Alberto, Drukker, David, Herr, Jane Leber and Imbens, Guido. 2004. "Implementing matching estimators for average treatment effects in Stata," *The Stata Journal* 4(3): 290-311.

Abadie, Alberto and Imbens, Guido. 2012. "Matching on the estimated propensity score," Harvard University and National Bureau of Economic Research, December.

Almond, Douglas and Currie, Janet. 2011. "Human capital development before age five," National Bureau of Economic Research, Working Paper 15827, January.

Angrist, Joshua D. 1998. "Estimating the labor market impact of voluntary military service using social security data on military applicants," *Econometrica* 66(2): 249-288.

Angrist, Joshua D. and Pischke, Jörn-Steffen. 2009. *Mostly harmless econometrics: An empiricist's companion.* New York, NY: Princeton University Press.

Baker, Amy JL, Piotrkowski, Chaya S. and Brooks-Gunn, Jeanne. 1999. "The home instruction program for preschool youngsters (HIPPY)," *The Future of Children* 9(1): 116-133.

Baker, Amy J. L. and Roth, Jodie, L. 1997. "Predictors of parent involvement in an early intervention program: Comparing sites, cohorts, and type of involvement," *Applied Behavioral Science Review* 5(2): 199-217.

Barnett, W. Steven and Hustedt, Jason, T. 2011. "Improving public financing for early learning programs," National Institute for Early Education Research, Issue 23, April.

Baydar, Nazli, Brooks-Gunn, Jeanne and Furstenberg, Frank F. 1993. "Early warning signs of functional illiteracy: Predictors in childhood and adolescence," *Child Development* 64(3): 815-829.

Black, Sandra E., Devereux, Paul G. and Salvanes, Kjell G. 2005. "The more the merrier? The effect of family composition on children's education," *Quarterly Journal of Economics* 120(2): 669-700.

Black, Sandra E., Devereux, Paul G. and Salvanes, Kjell G. 2011. "Older and wiser? Birth order and IQ of young men," *CESifo Economic Studies* 57(1): 103-120.

Bradley, Robert H. and Gilkey, Barbara. 2002. "The impact of the home instruction program for preschool youngsters (HIPPY) on school performance in 3rd and 6th grades," *Early Education and Development* 13(3): 301-312.

Brooks-Gunn, Jeanne and Markman, Lisa, B. 2005. "The contribution of parenting to ethnic and racial gaps in school readiness," *The Future of Children* 15(1): 139-168.

Brooks-Gunn, Jeanne, Berlin, Lisa J. and Fuligni, Allison S. Early childhood intervention programs: What about the family? Cambridge University Press, New York: Handbook of Early Childhood Intervention (2nd ed.), pp. 549-588.

Burkam, David T., Lee, Valerie E. and Dwyer, Julie. 2009. "School mobility in the early elementary grades: Frequency and impact from nationally-representative data," Workshop on the Impact of Mobility and Change on the Lives of Young Children, Schools, and Neighborhoods, June.

Cascio, Elizabeth U. 2005. "School progression and the grade distribution of students: Evidence from the Current Population Survey," Institute for the Study of Labor, Discussion Paper 1747.

Coleman, James S., Campbell, Earnest Q., Hobson, Carol J., McPartland, James, Mood, Alexander M., Weinfeld, Frederic D. and York, Robert L. 1966. "Equality of educational opportunity," U.S. Department of Health, Education, and Welfare.

Cunha, Flavio and Heckman, James J. 2007. "The technology of skill formation," National Bureau of Economic Research, Working Paper 12840, January.

Cunha, Flavio, James J. Heckman, Lance J. Lochner and Dimitriy V. Masterov. 2006. "Interpreting the evidence on life cycle skill formation," In Handbook of the Economics of Education, edited by Eric A. Hanushek and Frank Welch. Amsterdam: North-Holland, 697–812.

Currie, Janet and Thomas, Duncan. 1995. "Does Head Start make a difference?" *American Economic Review* 85(3): 341-364.

Dehejia, Rajeev H. and Wahba, Sadek. 1999. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American Statistical Association*, 94(448): 1053-1062.

Engec, Necati. 2006. "Relationship between mobility and student performance and behavior," *The Journal of Educational Research* 99(3): 167-178.

Fitzpatrick, Maria. 2008. "Starting school at four: The effect of universal pre-kindergarten on children's academic achievement," *The B.E. Journal of Economic Analysis and Policy*, 8(1): 1935-1682.

Garces, Eliana, Thomas, Duncan and Currie, Janet. 2002. "Longer-term effects of Head Start," *American Economic Review* 92(4): 999-1012.

Gomby, Deanna S. 2005. "Home visitation in 2005: Outcomes for children and parents," Committee for Economic Development: Invest in Kids Working Group, Working Paper No. 7.

Griliches, Zvi. 1979. “Sibling models and data in economics: Beginnings of a survey,” *Journal of Political Economy* 87(5): S37-S64.

Heckman, James J. 2013. “Lifelines for Poor Children,” *The New York Times* (New York Edition), September 14: p. SR5.

Heckman, James J. and Masterov, Dimitriy V. 2007. “The productivity argument for investing in young children,” *Review of Agricultural Economics* 29(3):446–493.

Jacobson-Chernoff, Jodi, Flanagan Kristin, D., McPhee, Cameron and Park, Jennifer. 2007. “Preschool: First findings from the third follow-up of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B),” U.S. Department of Education: National Center for Education Statistics.

Jeynes, William. 2012. “A meta-analysis of the efficacy of different types of parental involvement programs for urban students,” *Urban Education* 47(4): 706-742.

Kilburn, Rebecca M. and Karoly, Lynn A. 2008. “The economics of early childhood policy: What the dismal science has to say about investing in children,” RAND Corporation.

Nievar, M. Angela, Jacobson, Arminta, Chen, Qi, Johnson, Ursula and Dier, Shannon. 2011. “Impact of HIPPY on home learning environments of Latino families,” *Early Childhood Research Quarterly* 26:268-277.

Paulsell, Diane, Avellar, Sarah, Martin, Emily Sama and Del Grosso, Patricia. 2010. “Home visiting evidence of effectiveness review: Executive summary,” Office of Planning, Research and Evaluation.

Reynolds, Arthur J., Temple, Judy A., Robertson, Dylan L. and Mann, Emily A. 2001. “Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools,” *Journal of the American Medical Association* 285(18): 2339-2346.

Rosenbaum, Paul R. and Rubin, Donald B. 1983. “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70(1): 41-55.

Rumberger, Russell W. and Larson, Katherine A. 1998. “Student mobility and the increased risk of high school dropout,” *American Journal of Education* 107.

Texas Education Agency. 2008. “2008 Comprehensive Annual Repot on Texas Public Schools: A Report to the 81st Legislature from the Texas Education Agency,” Document No. GE09 601 03, Austin, TX: Author.

Texas Education Agency. 2012. “Limited English Proficient (LEP) Student Success Initiative (SSI).” Last updated 10/2012. http://www.tea.state.tx.us/index2.aspx?id=4537, accessed May 5, 2014.

Texas Education Agency. 2007-2012. "Technical Digests and Reports," Austin, TX: Author.