Tests of Exchangeability for Siblings and Twins*

[FIRST VERSION: PRELIMINARY AND INCOMPLETE]

Please do not cite or circulate this draft.

Andrew Elzinga† Bruno Gasperini‡

February 15, 2015

Abstract

Studies using twins and siblings as natural experiments to identify causal effects have become widespread in economics. In this paper, we test the validity of these strategies using existing panel data techniques; particularly Chamberlain's correlated random effects (CRE) model. The CRE framework nests the conventional fixed effects specification by allowing for additional heterogeneity that can be empirically tested and provide insight into the credibility of strategies using twins and siblings. We take advantage of this model by focusing on one of the most common relationships of interest involving within-family variation: the effects of birth weight on children's development. Using a rich, longitudinal dataset - Collaborative Perinatal Project (CPP) - our results cast doubt on research designs based on siblings and twins. Future empirical work employing fixed effects could benefit from the generality of the CRE framework.

Keywords: correlated random effects, exchangeability, twins, siblings, birth weight
JEL: C12, C33, C51, J13, I14

*We have benefited from comments and discussions with Anna Aizer, Kenneth Chay, Philipp Ketz, and Blaise Melly. We are responsible for any errors that may remain.

†Brown University, Department of Economics, 64 Waterman Street, Providence, RI 02912. E-mail: andrew_elzinga@brown.edu

‡Brown University, Department of Economics, 64 Waterman Street, Providence, RI 02912. E-mail: bruno_gasperini@brown.edu

1 Introduction

A common empirical strategy in economics is to use fixed-effects and within-variation to alleviate the omitted variables problem. Consider the problem of estimating the economic returns to education. Ashenfelter and Zimmerman (1997) use within-sibling variation to estimate the returns to schooling, while Ashenfelter and Krueger (1994) use twins to estimate the same relationship. Both research designs are based on the idea that individuals within a family will face the same environment and have similar genetic characteristics. Therefore, siblings and twins will act as natural experiments, that is, for some exogenous reason, there exists differences in educational levels between them.

In the economics literature, the use of siblings and twins to control for omitted variables has become widespread and they have been used as natural experiments for many other relationships of interest. For example, twins have been used to identify the effects of birth weight on a myriad of outcomes (Almond, Chay and Lee, 2005; Behrman and Rosenzweig, 2004; Black, Devereux and Salvanes, 2007; Royer, 2009) and the effects of education on health (Amin, Behrman and Spector, 2013; Lundborg, 2013).

On the other hand, siblings have been used to identify the effects of birth weight (Oreopoulos, Stabile, Walld and Roos, 2008; Fletcher, 2011b), the effects of education on health (Behrman and Wolfe, 1989), the effects of breastfeeding on cognitive development (Rothstein, forthcoming), the effects of maternal education on children's outcomes (Behrman and Wolfe, 1989), the effects of neighborhoods on children's outcomes (Aaronson, 1998), the effect of child health on test scores and schooling attainment (Currie and Stabile, 2006) and on adult outcomes (Currie, Stabile, Manivong and Roos, 2010; Fletcher, Green and Neidell, 2010); and the effects of teenage childbearing on mother's behavior (Fletcher, 2011a). Outside of the economics literature, siblings have been used to analyze the effect of maternal smoking on cognitive and physical development (Gilman et al., 2008) and the effect of maternal BMI and gestational weight gain on children's BMI (Branum et al., 2011). Further inquiries into the literature would results in many other examples.

In spite of this, recall that the intuition behind a natural experiment is that we have exogenous variation in a treatment of interest while everything else is constant, mimicking random assignment. But as more and more predetermined differences and outcomes are studied between family pairs, the validity of their use as a natural experiment becomes debatable. Nevertheless, to the best of our knowledge, there does not exist any thorough analysis justifying the use of siblings and twins in these circumstances.

The objective of this paper is to fill this gap and investigate the exchangeability of siblings and twins. By exchangeability of a pair we mean that individuals are indeed valid counterfactuals for one another, implying they have identical observable and unobservable characteristics. Exchangeable pairs provide the ideal setting for a natural experiment because, by definition, they are perfect substitutes for one another. Rejecting the exchangeability of twins or siblings would bring doubts on research designs based on within-family variation.

Our strategy focuses on one of the most commonly used treatment variables involving within-family variation: birth weight. Using twins and siblings, the common strategy is to use fixed effects (FE). Instead, we will analyze the effects of birth weight using the correlated random effects (CRE) framework (Chamberlain, 1982, 1984; Mundlak, 1978), which is a simple extension of the FE specification that adds room for heterogeneity. It is more general in the sense that is allows for dependence between the group heterogeneity and the regressors, which provides two overidentifying restrictions that can be tested in the data. First, it allows for heterogeneity in the returns to the treatment across siblings. Second, it decomposes the OLS bias for each individual separately. That is, for each person the CRE will give a parameter that shows the partial correlation between the family effect and the treatment. If twins and siblings are truly exchangeable, then we should expect no type of heterogeneity in either of these parameters.

In order to achieve our objectives, we need data that satisfy three conditions: (i) contain a substantial number of pairs of twins and siblings, (ii) provide detailed information for a large number of variables, and (iii) follow individuals over time. We will take advantage of the Collaborative Perinatal Project (CPP) dataset which was carried out from 1959-1974 and collected data on pregnant women and children's developmental outcomes until the age of eight.

Our empirical results point to a lack of exchangeability for twins and siblings. We find numerous differences within pairs and the tests based on the CRE estimates reject exchangeability. For example, measurements of IQ suggest the first born is better off at 8 months, worse off at 4 years, and no different from their sibling by age 7. Relative to siblings, differences within twin pairs are less evident. However, by exploiting the time dimension of our data, the exchangeability of twins comes into question. We have also expanded the model to include nonparametric estimation and our results remain unchanged. Finally, we perform simulations of the extent to which measurement error could be driving our results and our conclusions are unchanged.

The CRE framework has been more frequently applied to non-linear models in the context of panel data with a time invariant, unobserved component (Wooldridge, 1999, 2005, 2010a). The CRE and its structure has been used to circumvent the incidental parameter problem, which leads to biased marginal effects in these classes of models. In addition to this use, we feel that the CRE framework can be implemented in a linear, non-panel data setting to provide valuable testing. In this regard, the aim of this paper is to show how a useful framework can be implemented at low cost to validate research designs using within-family variation. Our results suggest that these strategies should be taken cautiously.

2 Correlated Random Effects

Our test of exchangeability will be based on panel data techniques (Hsiao, 2003; Wooldridge, 2010b). More specifically, we will work within the CRE device framework (Chamberlain, 1982, 1984). By nesting the popular fixed-effects (FE) specification, this more general model will provide two overidentifying restrictions that will be useful in testing for exchangeability. The simplest framework is as follows:

$$Y_{1j} = \beta_1 X_{1j} + F_j + v_{1j} \tag{1}$$

$$Y_{2j} = \beta_2 X_{2j} + F_j + v_{2j} \tag{2}$$

$$F_j = \lambda_0 + \lambda_1 X_{1j} + \lambda_2 X_{2j} + e_j \tag{3}$$

where X_{ij} is the treatment variable, F_j is a group-specific unobserved heterogeneity, and v_{ij} is the idiosyncratic error term. In a panel data setting, i would represent time period and j would be the individual. In our context, i indicates a sibling/twin and j a family.

As is well known, pooled OLS is biased if $cov(X_{ij}, F_j) \neq 0$. To account for this potential bias, the typical FE analysis explicitly assumes that $E[v_{ij}|X_{ij}, F_j] = 0$ and then differences equations (1) and (2) to remove the unobserved heterogeneity. In contrast, the CRE supplements the FE model with equation (3), which is a linear projection of the unobserved heterogeneity onto the regressors. A linear projection always exists and is unique if all the variables have finite second moments. The idea of relating F_j to X_{ij} dates back to Mundlak (1978) who projected the unobserved heterogeneity onto the average of the regressors. Plugging equation (3) into equations (1) and (2) gives:

$$Y_{1j} = \lambda_0 + (\beta_1 + \lambda_1)X_{1j} + \lambda_2 X_{2j} + v_{1j} + e_j$$
$$Y_{2j} = \lambda_0 + \lambda_1 X_{1j} + (\beta_2 + \lambda_2)X_{2j} + v_{2j} + e_j$$

Although equation (3) makes no assumption about the conditional distribution of F_j given X_{ij}, in order to obtain consistent estimates of the parameters, we need $E[v_{ij} + e_j] = 0$ and $E[v_{ij} + e_j|X_{ij}] = 0$. In order to satisfy these, we assume $E[F_j|X_{ij}]$ is linear, which is a sufficient condition.

To estimate this model, we need a criteria that allows us to sort siblings and twins within a family. For example, in panel data models there is a natural progression from one individual to the next, e.g. time periods. However, when dealing with family members, one could propose numerous ordering criteria, each of which may lead to different results. In the analysis that follows we will sort pairs according to their order of birth. In our view, ordering by age is the most natural criteria and does the best job of emulating the panel data structure.

2.1 β_1 and β_2

Two major differences between this model and the traditional FE framework come up. First, the model allows for differential treatment effects within group, that is, β_1 may differ from β_2. In the family context, this allows the return to the treatment to vary according to the birth order of the sibling; in usual panel context this allows the return to treatment to vary across time.

A benefit of this generalization is that we can test for the equality of treatment effects, which, if rejected, implies the FE specification is inappropriate. Estimates from the CRE under the restriction that $\beta_1 = \beta_2$ are equivalent to that from FE. Therefore, our first test utilizing the CRE framework will be whether $\beta_1 = \beta_2$. Differential returns to treatment might be due to inherent differences between siblings, casting doubts on the validity of them as counterfactuals and their exchangeability.[1]

[1] Although we relax the assumption of a constant treatment effect, we will assume that the returns to

Allowing heterogeneous treatment effects is the first departure from the canonical fixed effects model. To understand the importance of this generalization, we derived the consequences of falsely restricting the treatment effect to be constant within group ($\beta_1 = \beta_2 = \beta$). Consider the following fixed effects setup:

$$Y_{ij} - \overline{Y}_j = \beta(X_{ij} - \overline{X}_j) + v_{ij} - \overline{v}_j$$

After some algebra contained in Appendix A.1, we find:

$$plim\ \hat{\beta}_{FE} = \theta\beta_1 + (1-\theta)\beta_2$$

Define $E[X_{1j}] = \mu_1, E[X_{2j}] = \mu_2, Var(X_{1j}) = \sigma_1^2, Var(X_{2j}) = \sigma_2^2$, and $Corr(X_{1j}, X_{2j}) = \rho$. θ can be expressed as:

$$\theta = \left(\frac{\sigma_1^2 + \mu_1^2 - \rho\sigma_1\sigma_2 - \mu_1\mu_2}{\sigma_1^2 + \sigma_2^2 + \mu_1^2 + \mu_2^2 - 2\rho\sigma_1\sigma_2 - 2\mu_1\mu_2}\right)$$

Table 1 enumerates an exhaustive list of all possible combinations of the first two moments of the distribution of X_1 and X_2. Panel A includes the cases which can arise when the group distributions have different means. It is important to note that in addition to the estimates being biased, $\hat{\beta}_{FE}$ does not necessarily lie in the convex hull, $co(\beta_1, \beta_2)$, in cases 1–4.

Furthermore, Panel B lists the potential scenarios when the group distributions have the same mean, which occurs whenever a group fixed effect is included. In our case, this refers to including a first born dummy variable. Although the estimates are still biased, only in 1 of the 3 cases can the estimate lie outside of $co(\beta_1, \beta_2)$. This occurs in case 7 when $\sigma_1 < \rho\sigma_2$.

The likelihood of the potential scenarios from Table 1 is fundamentally an empirical question dependent upon the data generating process in question. However, the importance of including a group dummy variable is highlighted as it reduces the number of cases for which the estimate, even though biased, can lie outside the convex hull. Overall, these results show that falsely assuming homogeneous treatment effects can lead to misleading estimates

treatment are the same for individual 1 and individual 2, across all pairs.

2.2 λ_1 and λ_2

The second major differences is the role played by the group specific unobserved heterogeneity. Within the CRE framework, the second test of exchangeability will examine if $\lambda_1 = \lambda_2$. Although a rejection of this test does not imply the FE specification is incorrect, it suggests the pair is not perfectly comparable. Recall that the idea behind using family members to remove any omitted variable bias is that they are counterfactuals for one another and therefore should be similar along every dimension. If this is truly the case, then we should expect $\lambda_1 = \lambda_2$ because it says that the family effect is correlated with the treatment equally for each individual. Finding that $\lambda_1 \neq \lambda_2$ suggests that the unobservables for individual 1 are different than for individual 2; a direct contradiction of exchangeability.

Additional intuition for λ comes from the fact that they are related to the bias from pooled OLS using (1) and (2); as λ increases, the omitted variables bias gets larger because it implies the partial correlation between F_j and X_{ij} increases. Using the model introduced above, it can be shown that the bias is related to the coefficients from the linear projection in the following way:[2]

$$Bias(\hat{\beta}_{1,OLS}) = \lambda_1 + \lambda_2 \frac{cov(X_{1j}, X_{2j})}{var(X_{1j})}$$
$$Bias(\hat{\beta}_{2,OLS}) = \lambda_2 + \lambda_1 \frac{cov(X_{1j}, X_{2j})}{var(X_{2j})}$$

In addition to the general test of $\lambda_1 = \lambda_2$, the above equations suggest testing for any OLS bias; $\lambda_1 = \lambda_2 = 0$. The equations also present a mechanical link between the OLS estimates and the CRE estimates. For example, if the regressors are positively correlated and both λ_1 and λ_2 are positive, then OLS will overestimate the treatment effect and CRE will be smaller in magnitude.

[2]The biases derived are those from running OLS for individual 1 and individual 2 separately. They are not the bias coming from the pooled OLS estimate.

2.3 Identification and Estimation

Identification and estimation of the CRE device is straightforward. Plugging equation (3) into equations (1) and (2) we obtain the following reduced-form system of equations:

$$\begin{aligned} Y_{1j} &= \lambda_0 + (\beta_1 + \lambda_1)X_{1j} + \lambda_2 X_{2j} + v_{1j} + e_j \\ &= \Pi_{10} + \Pi_{11}X_{1j} + \Pi_{12}X_{2j} + v_{1j}^* \end{aligned} \tag{4}$$

$$\begin{aligned} Y_{2j} &= \lambda_0 + \lambda_1 X_{1j} + (\beta_2 + \lambda_2)X_{2j} + v_{2j} + e_j \\ &= \Pi_{20} + \Pi_{21}X_{1j} + \Pi_{22}X_{2j} + v_{2j}^* \end{aligned} \tag{5}$$

where Π_{10}, Π_{11}, Π_{12}, Π_{20}, Π_{21}, Π_{22} are reduced-form parameters. The structural parameters can be backed out as follows:[3]

$$\lambda_0 = \Pi_{10} = \Pi_{20}$$

$$\lambda_1 = \Pi_{21} \qquad \lambda_2 = \Pi_{12}$$

$$\beta_1 = \Pi_{11} - \Pi_{21} \qquad \beta_2 = \Pi_{22} - \Pi_{12}$$

Estimation of the reduced-form system of equations will be performed via system OLS (SOLS). It is well known that the estimates from SOLS are numerically equivalent to those from OLS equation-by-equation. However, the latter method does not estimate the full-blown variance-covariance matrix that takes into account correlations between cross-equation error terms. Since we are ultimately interested in testing cross-equation restrictions, SOLS is the preferred method.[4] Finally, standard errors of the structural parameters are estimated using the delta method.

With the estimates of the structural parameters we will perform three tests in order to verify the exchangeability of siblings and twins: (1) $\beta_1 = \beta_2$, (2) $\lambda_1 = \lambda_2$, and (3) $\beta_1 = \beta_2$ and $\lambda_1 = \lambda_2$ jointly. In doing so, we will say that pairs are exchangeable if we cannot reject both that $\beta_1 = \beta_2$ and $\lambda_1 = \lambda_2$.

[3] As can be seen, this model is overidentified with 6 reduced form parameters and 5 structural parameters. However, the only overidentified parameter is λ_0 which has little importance in our context.

[4] Conventionally, system of equations have been estimated using SURE. However, because all of our equations contain the same set of regressors, SURE and SOLS are numerically equivalent.

2.4 Nonparametric Estimation

Our discussion thus far has assumed a linear impact of the treatment. However, this might not be the case, especially when estimating the returns to birth weight, which has been shown to exhibit nonlinear effects on health outcomes and test scores (Almond, Chay and Lee, 2005; Figlio, Guryan, Karbownik and Roth, 2013).

For example, consider a nonlinear (e.g strictly concave) birth weight "production function" that maps birth weight into outcomes (e.g health measures or IQ tests). Rejecting exchangeability implies that siblings have different production functions. However, under nonlinearities, our previous testing results using CRE are subject to potential issues. Although the estimated treatment effects are not the focus of this paper, rather we are concerned with the testing, misspecification could well affect our conclusions from the tests proposed.

First, we could face size problems (type I error), that is, rejecting exchangeability when pairs are actually valid counterfactuals. We could well be rejecting exchangeability under a nonlinear production function if both siblings have markedly different birth weights. In this case, the estimated derivatives of the shared production function would mechanically be different, even though the pair is, in truth, perfectly exchangeable.

Second, our tests could also have low power (type II error), that is, failing to reject exchangeability when pairs are not valid counterfactuals. This could occur if siblings have different birth weight production functions, but due to differences in birth weight, by chance, they have similar slopes at those points.

Our solution to these problems is to estimate the CRE device nonparametrically via a piecewise linear spline specification. Using knot points $p_0, p_1, p_2, \ldots, p_K$ (where $p_0 = 0$), the splines are generated as follows:

$$Y_{1j} = \sum_{k=0}^{K} \beta_{1k} X_{1j} \cdot D_{1j}^k + F_j + v_{1j} \tag{6}$$

$$Y_{2j} = \sum_{k=0}^{K} \beta_{2k} X_{2j} \cdot D_{2j}^k + F_j + v_{2j} \tag{7}$$

$$F_j = \sum_{k=0}^{K} \lambda_{1k} X_{1j} \cdot D_{1j}^k + \lambda_{2k} X_{2j} \cdot D_{2j}^k + e_j \tag{8}$$

where $D_{ij}^k = \mathbb{1}\{X_{ij} \geq p_k\}$. Plugging (8) into equations (6) and (7) results in the following reduced-form equations

$$Y_{1j} = \sum_{k=0}^{K} \left((\beta_{1k} + \lambda_{1k})X_{1j} \cdot D_{1j}^k + \lambda_{2k}X_{2j} \cdot D_{2j}^k\right) + v_{1j}$$
$$Y_{2j} = \sum_{k=0}^{K} \left(\lambda_{1k}X_{1j} \cdot D_{1j}^k + (\beta_{2k} + \lambda_{2k})X_{2j} \cdot D_{2j}^k\right) + v_{2j}$$

Identification and estimation of $\{\beta_{1k}, \beta_{2k}, \lambda_{1k}, \lambda_{2k}\}_{k=1}^K$ is equivalent to the linear case. In addition to avoiding the testing problems discussed above, the crucial advantage of using splines is that it allows us to test for exchangeability across the entire distribution of the regressor.

In spite of this, we believe that the testing results from the linear case may still be informative. If birth weight differences are not too large[5] and the production function is not highly nonlinear, the tests are unlikely to have size distortions. In the same spirit, a small difference in birth weight suggests that if first and second borns have different production functions, and thus are not exchangeable, then it is unlikely that we will estimate the same treatment effects, limiting the incidence of type II error.

2.5 Measurement Error

Griliches and Hausman (1986) and Bound and Solon (1999) show that measurement error problems are exacerbated in first-differencing or FE specifications. By nesting these, the CRE device is also prone to this problem. However, the CRE can be easily modified to allow for the possibility of measurement error. Suppose that X_{ij} are measured as:

$$X_{1j} = X_{1j}^* + m_1$$
$$X_{2j} = X_{2j}^* + m_2$$

where X_{1j}^* and X_{2j}^* are the true levels of the regressors and m_1 and m_2 are the measurement

[5]For the CPP, birth weight differences between first-borns and second-borns are 0.1 and 0.06 standard deviations for twins and siblings, respectively

error terms. Also, assume that m_1 and m_2 are classical in the following sense:

$$\begin{array}{llllll} i) & m_1 \perp X_{1j}^* & ii) & m_2 \perp X_{1j}^* & iii) & m_1 \perp m_2 \\ & m_1 \perp X_{2j}^* & & m_2 \perp X_{2j}^* & & \end{array}$$

that is, the measurement error terms are orthogonal to the true level of the regressors and are also also orthogonal to each other. Under these assumptions, the probability limit of the reduced form OLS estimators can be derived. With these estimates, it can be shown that the probability limit of the estimators for the structural parameters are:

$$plim\ \hat{\beta}_1 = \beta_1 - \frac{\beta_1\psi_1 + \frac{cov(X_1,X_2)}{var(X_1)}\beta_2\psi_2}{1-\rho^2} \tag{9}$$

$$plim\ \hat{\beta}_2 = \beta_2 - \frac{\beta_2\psi_2 + \frac{cov(X_1,X_2)}{var(X_2)}\beta_1\psi_1}{1-\rho^2} \tag{10}$$

$$plim\ \hat{\lambda}_1 = \lambda_1 - \frac{\lambda_1\psi_1 - \frac{cov(X_1,X_2)}{var(X_1)}(\beta_2+\lambda_2)\psi_2}{1-\rho^2} \tag{11}$$

$$plim\ \hat{\lambda}_2 = \lambda_2 - \frac{\lambda_2\psi_2 - \frac{cov(X_1,X_2)}{var(X_2)}(\beta_1+\lambda_1)\psi_1}{1-\rho^2} \tag{12}$$

where $\psi_i \equiv \frac{var(m_i)}{var(X_{ij}^*)+var(m_i)}$ for $i = 1, 2$ are the ratio of noise to total variance and ρ is the correlation between the observed regressors. Note that if $cov(X_1, X_2) = 0$ or $\psi_j = 0$ for $j \neq i$, then the traditional result for classical measurement error holds.

Let $plim\ \hat{\beta}_1 = \Delta_1$ and $plim\ \hat{\beta}_2 = \Delta_2$. With observed levels of $cov(X_1, X_2)$, $var(X_1)$, $var(X_2)$ and assumed levels of ψ_1 and ψ_2, we can back out the measurement error-corrected structural parameters by solving the below system of linear equations,

$$\left(1 - \frac{\psi_1}{1-\rho^2}\right)\beta_1 + \left(-\frac{cov(X_1,X_2)\psi_2}{var(X_1)(1-\rho^2)}\right)\beta_2 = \Delta_1 \tag{13}$$

$$\left(-\frac{cov(X_1,X_2)\psi_1}{var(X_2)(1-\rho^2)}\right)\beta_1 + \left(1 - \frac{\psi_2}{1-\rho^2}\right)\beta_2 = \Delta_2 \tag{14}$$

Similarly, after solving equations (13) and (14) for β_1 and β_2, we can back out the measurement error-corrected values of λ_1 and λ_2 from equations (11) and (12).

3 Data - Collaborative Perinatal Project (CPP)

To compare the exchangeability of siblings and twins, we need datasets that satisfy three conditions. First, in order to produce estimates with any precision, the data will need to have a substantial number of pairs of siblings and twins. Second, assessments of exchangeability will be much more convincing if our data provides detailed and accurate information for a large number of variables. Finally, in order to track exchangeability over time, the data will need to follow the same individuals and ideally provide measurements of the same indices.

With this in mind, our results will be based on the Collaborative Perinatal Project (CPP) produced by the National Institute of Neurological Disorders and Stroke (NINDS) and made publicly and electronically available by Lawlor et al. (2005). The CPP was a prospective study of neurological disorders and other conditions in children. Pregnant women were enrolled from 1959–1965 during their first prenatal care visit at any of 12 university hospital clinics located throughout the United States.[6] The dataset contains detailed information on 59,391 children and 48,197 families. Beginning at registration, pregnant women had information recorded on their medical history, pregnancy, labor and delivery. Sociodemographic characteristics and health behavior information were recorded; including maternal age, race, gestation length, smoking behavior during pregnancy, and SES index.

After birth, the children were followed for the first eight years of life. Children were systematically assessed at 4, 8, and 12 months, and 3, 4, 7, and 8 years. Most of the analysis that follows takes advantage of data from birth, 8 months, 4 years, and 7 years.[7] At each checkup, trained psychologists administered a battery of cognitive, sensory, and motor tests, in addition to measuring the child's behavior and development. For example, at birth, the dataset contains information regarding delivery type and short-run health outcomes, such as APGAR scores, number of days in hospital and incubator use. To measure mental skills, the CPP administered the Bayley Scales of Infant Development at 8 months, the Stanford-Binet Intelligence Scale at 4 years, and the Wechsler Intelligence Scale for Children at 7 years. Each of these three tests are aimed at measuring an individual's IQ. Motor skills were assessed using the Bayley Scales of Infant Development at 8 months, a general motor

[6] Although the CPP does not provide sampling weights, the results can be thought as nationally representative since our sample of pregnant women comes from major cities across the U.S.: Baltimore, Boston, Buffalo, Memphis, Minneapolis, New Orleans, New York, Philadelphia, Portland, Providence and Richmond.

[7] Although the study aimed at interviewing everyone at a fixed age, there exist small differences in the age at which participants were interviewed. We control for these age differentials in our analysis.

test at age 4, and at age 7 via the Bender Gestalt Visual Motor Test. Also, at age 7, the CPP administered the Wide Range Achievement Test (WRAT) aimed at evaluating spelling, reading and arithmetic abilities of children. Because each of these tests vary in the range of their scores and in order to make scores comparable over time, we normalize all scores into percentile ranks across the whole population in our data (with the exception of age 4 motor skills, which ranges discretely from 0–13). The appendix contains more information regarding these tests.

The current study includes only liveborn individuals from families that had multiple pregnancies in the study. Under these restrictions we are able to identify 13,568 sibling pairs and 640 twin pairs. We are able to determine the order of birth directly because the CPP records the pregnancy number for each mother. In the case of multiple birth pregnancies, the CPP also records the birth order of each individual. Importantly, infant birth weight is measured at the time of delivery by the medical staff present. Follow-up rates for survivors in the full CPP sample were 88% at 1 year, 75% at 4 years, and 79% at 7 years. For more information on the CPP, see Niswander and Gordon (1972); Klebanoff (2009).

4 Empirics

4.1 Mean Comparisons

In any research design involving random assignment or a natural experiment, the first analysis typically consists of comparing sample averages of the control and treatment (or encouraged) groups. If they are suitable counterfactuals for one another, then, on average, we should find no observable differences between them in controls or predetermined outcomes.

In principle, existent significant differences between groups can be controlled for. However, these differences will call into question the validity of the research design, since differences in observables could well be related to differences in unobservable (to the researcher) characteristics that are potentially correlated with outcomes. With this in mind, our first pass of exchangeability is based on mean comparisons when siblings are ordered by birth order.

Table 2 presents the mean comparison results for twins using CPP data. The first column represents the sample average among all first born twins and the second column shows the average for second borns. The numbers in parentheses in columns (1) and (2) are standard deviations for the given variable for first and second born twins, respectively. Column (3) presents the raw differences between the twins, without any controls. There are two potential problems with this raw difference. First, the differences are purely cross-sectional and it has been well documented that maternal characteristics impact the early life outcomes of children (Almond, Chay and Lee, 2005). Secondly, as the first row of Table 2 reports, there is a significant difference between the birth weight of the first born and the second born. With this in mind, it may be that any subsequent difference between the twins is being driven entirely by the difference in birth weight rather than the order of birth. To correct these problems, column (4) presents the differences between twins having controlled for mother fixed effects and differences in birth weight; when examining test outcomes we also control for age differences. In columns (3) and (4), the numbers in brackets are the standard deviation effects of the difference using both first and second born twins. Finally, column (5) shows the number of pairs used in each calculation.

Outside of expected discrepancies, the only time there are statistically significant differences is during the first moments of life. Overall, Table 2 provides evidence suggesting that, under the birth order criteria, twins look exchangeable, except for potentially very early on in life.

Turning to Table 3, the sample mean differences for siblings in the CPP, we see markedly different results. An important thing to note about going from twins to siblings is the significant increase in the number of pairs. As a result, given a fixed difference for twins and siblings, the sibling difference is mechanically more likely to be significantly different than zero. With this in mind, there are some unsurprising differences: the second born has slightly higher birth weight and is born to an older mother. Family characteristics and early life health outcomes appear to be better for the second born and do not seem to be driven solely by differences in birth weight, but the size of the impacts are not large. Interestingly, first borns tend to have higher mental and motor skills at 8 months old with magnitudes of 0.13 and 0.22 standard deviation effects, but by age 4 the gap seems to be reversed, implying that the second born is catching up and exceeding during that period. However, by age 7 the difference in motor and mental skills seem to have decreased significantly in terms of magnitude.

In general, Table 3 find numerous statistically significant differences across siblings, although the magnitude of the effects are less condemning of their exchangeability. In comparison to twins, the size of the differences are more pronounced for siblings and they can be found during the entire childhood, which provides preliminary evidence that they are less exchangeable. Corroborating this conclusion is the fact that the differences between siblings do not appear to be following any pattern nor are the differences consistent across multiple measures of the same attribute. The inability to reconcile these inconsistencies suggests siblings are not exchangeable under the birth order criteria.

From the mean comparison results discussed above, implications regarding twin's and sibling's exchangeability can be made. For twins, except for very early in life, there seem to be few meaningful differences between first and second borns, while siblings show pointed differences, which are also not consistent over time. Our conclusion from these tables is that twins appear to better satisfy the conditions of a valid natural experiment than siblings.

Finally, it is worth noting the difference between the tests of exchangeability using mean differences and the CRE framework. Both tests are focused on measuring the validity of pairs as natural experiments. However, mean comparisons, which are performed by regressing variables on a constant and a dummy for first-born, capture level differences between first born and second born. Differently, CRE aims to test for slope differences by interacting the order of birth with the treatment and also provides estimates of the partial correlation between X_{ij} and F_j. Therefore, we see the tests as complements to each other, neither being more complete than the other.

4.2 Correlated Random Effects

Complementing the mean comparison results, this section presents tests of exchangeability using the CRE with respect to the effects of birth weight. As previously mentioned, the goal of this paper is not to estimate the causal impacts of birth weight, rather, our decision to use birth weight was made mainly for three reasons. First, it is conceivably exogenous to the individual. Second, among studies using twins and siblings, it is one of the most common treatment variables. Third, and most importantly, by being positively correlated with initial health endowments, looking at how the effects of birth weight vary over time and how they differ across first and second borns could shed light on family behavior and the mechanisms behind exchangeability.

CRE tables are arranged as follows: for a given outcome, we first show the cross-sectional

OLS estimates using only first borns and also for only second borns. The next column shows the commonly used fixed effects estimate. Then, the estimated treatment effects (β_1 and β_2) and partial correlations (λ_1 and λ_2) using the CRE are displayed. Results include controls for zygosity and gender when analyzing twins and gender, maternal age, SES at registration, gestational length, and the maximum number of cigarettes smoked consistently during pregnancy for siblings. Accompanying each estimate are numbers in parenthesis giving the heteroskedasticity-robust standard errors and numbers in brackets presenting the standard deviation effects; measured as $\beta_{bw} \times \frac{\sigma_{bw}}{\sigma_y}$. The tables also give p-values for the various tests of exchangeability and finally some statistics of the sample that was used in the estimation.

Tables 4, 5 and 6 are the twin results using the CRE device for early in life outcomes. The OLS estimates of the birth weight effects in Tables 4 and 5 indicate that increasing birth weight increases APGAR scores and lowers incubator usage and hospital days, and are significant and large in magnitude. Estimates using the CRE device have the same signs as the OLS estimates, however the magnitude of the effects decrease for all outcomes. Similar results are seen in Table 6, which estimates the impact of birth weight on infant mortality; OLS finds a large and significant effect, whereas CRE essentially finds zero effect across all time frames. Compared to Almond, Chay and Lee (2005), we find larger APGAR results and similarly small and insignificant results for infant death. It appears that birth weight has an impact on early infant health, but not large enough to result in death.

More importantly for our context, in neither table are we able to firmly reject the hypothesis that twins are not exchangeable. Only for hospital days are we able to reject the joint test of exchangeability and this is clearly driven by differences in the partial correlations because the treatment effects are nearly identical. For every other outcome in Tables 4, 5 and 6 we do not have enough evidence to reject differences in the effects of birth weight and differences in the size of the partial correlations.

Table 7 shows estimates regarding the effects of birth weight on short run motor and mental skills. OLS finds large and significant effects for both motor and mental at all three ages, however the CRE estimates suggest there is an impact of birth weight on motor and mental skills at 8 months, but little effect afterwards. For mental skills, the estimated impact at 8 months is 1.002 and 0.996 for twin 1 and twin 2 (standard deviation effects are 0.198 and 0.197, respectively), while the estimates drop to 0.149 and 0.372 by age 4 (with standard deviation effects of 0.028 and 0.071). Regarding twin's exchangeability, we are again unable to reject exchangeability with respect to birth weight and mental skills, although the partial correlations appear to be substantially different at 8 months.

Also, twins do not appear to be exchangeable in regards to their early motor skills. Similar to mental skills, we see large differences in the partial correlations at 8 months, but now also find significant differences in the effects of birth weight at 8 months and 4 years. By age 7 the differences in motor skills have subsided and twins appear to have become more exchangeable. Overall the motor and mental skills results present evidence of early differences between twins and a convergence of these differences as the children age. In fact, for both mental and motor we see increases in p-values and a disappearance of the effects of birth weight after 8 months. These results are in contrast to Black, Devereux and Salvanes (2007) who, using Norwegian data, find positive and significant effects of birth weight on IQ at age 18.

Consistent with our IQ estimates in Table 7, OLS finds large effects of birth weight on WRAT scores, but insignificant impacts under CRE in Table 8. Interestingly, we find an effect of birth weight on the probability of having failed a grade by age 7. In contrast to our results, Figlio, Guryan, Karbownik and Roth (2013), using data from Florida, find significant effects of birth weight on the cognitive test of children. Meanwhile p-values from Table 8 suggest exchangeability for all four outcomes. These results corroborate the idea that twin differences have subsided by age 7, at least in terms of their mental skills.

Our main twin's findings, that the effects of birth weight disappear after 8 months and that as twins age it becomes harder to reject their exchangeability, is in fact evidence against their exchangeability. If twins truly face the same family background and levels of investment, then any initial differences between them should persist over time. But the estimates thus far contradict this idea, suggesting there may have been alternative experiences across twins.

Moving from twins to siblings, we expect to find more statistically significant effects due to the increased sample size and therefore a failure to reject exchangeability should be taken as an even stronger statement. The surprising fact from Table 9 is that the estimated results tend to get larger in magnitude when going from OLS to CRE. The estimated impact of birth weight on the 5 minute APGAR score is 0.052/0.041 (standard deviation effects of 0.237/0.188) using OLS and 0.083/0.80 (0.380/0.365) from CRE. Using the fact that sibling birth weights are positively correlated and the negative estimates of λ, the equations for the

OLS bias are consistent with these increases.

However, the size of the estimated effects of birth weight are similar to that of twins for all five outcomes. The APGAR test statistics imply that siblings are exchangeable, but we are able to reject exchangeability regarding incubators and hospital days. Even though the p-values are quite small for incubator days and use, the differences in β and λ are nearly identical to the differences found between twins. Thus the fact that we can reject exchangeability for siblings and not for twins appears to be driven largely by sample size. The infant mortality estimates for siblings in Table 11 are similar to twins in the regard that they fail to reject exchangeability at all time periods, but differ with respect to the magnitude of the effects. Also, the sibling estimates of λ are all positive, which, like Tables 9 and 10, is the opposite of twins.

Looking at the mental and motor characteristics of siblings in Table 12, the magnitude effects from the CRE device tend to be only somewhat similar to that of twins. Similar to the early life outcomes, the OLS estimates are typically smaller in absolute value than the CRE results. The conclusions from the p-values, for both mental and motor, is that siblings are not exchangeable. The fact that we can reject exchangeability for siblings, but were unable to for twins, is likely driven, in part, due to the large sample size. But it should be noted that the difference in the effects of birth weight on mental skills between sibling 1 and sibling 2 are, on average, larger than those for twins. Therefore, the frequent rejections of exchangeability for mental and motor skills appear to be a result of a large number of pairs, but also because siblings truly are less exchangeable. The estimated effects for WRAT and grade repetition in Table 13 provide mixed evidence of sibling's exchangeability. Much like mental skills at age 7, the estimated λ are quite similar, but different treatment effects lead to rejections of exchangeability for the spelling test and the likelihood of repeating a grade.

5 Robustness Checks

In this section we probe our results to further robustness checks. Looking at Table 7 we see that (i) the effects of birth weight on IQ tests for twins disappear after 8 months and (ii) twins are becoming more exchangeable over the course of their lives. In addition to a behavioral explanation involving family investments, this result could be driven by sample selection for two reasons. First, in pairs comprised of a low birth weight infant and a healthy infant, which are expected to be the least exchangeable, it may be that the low birth weight child is dying and thus preventing us from including the pair in subsequent regressions. Second, as an artifact of the data, it could be that less exchangeable pairs are missing non-randomly from our sample.

To address the infant mortality hypothesis, recall in Table 6 we find that the effects of birth weight on infant mortality are constant and very small across different time periods. Consistent with this, as shown in the same table, the average of infant mortality barely changes from 1 week to 1 year, suggesting that most of the mortality is occurring within the first week of life. Also, despite the reduction in sample size, we also do not believe that non-random missing is an important factor. Looking at Table 7, there is no evidence of changes in the mean or standard deviation of birth weight, pointing to the fact that the distribution of characteristics is being held constant across time periods. However, we should note that another robustness check that we have in mind is to perform the analysis with a fixed sample, including only pairs that we have data across all years.

Until this point, all of our CRE estimates have assumed a linear effect of birth weight. As discussed in the methodology section, nonlinearities can affect both the size and power of our linear tests. To be more flexible, use nonparametric methods in order to relax the linearity assumption. While these models allow for more generality, they also require more data and therefore, our nonparametric analysis will be based only on sibling pairs due to the low sample sizes for twins.

To address this, we estimate piecewise linear splines of 4 segments with knot points at 1600g, 2500g, and 3000g. Based the the chosen knot points, the first segment contains 4% of the pairs, the second has 9%, and the third and fourth contain 23% and 64%, respectively. Tables 14-16 give the spline estimates for the following outcomes: APGAR 5-minutes, infant mortality within 1 month and mental scores 8 months, 4 years, and 7 years. Entries in the tables report the slope coefficient for each segment of the spline for both cross-sectional OLS and CRE specifications.

Table 14 presents the spline estimates for early life outcomes. The results before the first knot point are rather unintuitive, but this may be driven by the fact that the infants used in estimation are all very low birth weight and likely have many health issues other than just low birth weight. The remaining segments all present positive results that are consistent with a concave birth weight production function. However, the linear CRE results are quite similar to these nonparametric results suggesting only a slight concavity. Interestingly, the returns to birth weight are always larger for the older individual even though the older child is not always the heavier. Also consistent with the linear estimates, we find negative λ for siblings when looking at APGAR and the opposite for infant mortality. The testing implications of the spline regression of infant mortality are that we can not reject the exchangeability of siblings, but looking at the 5 minute APGAR results, there appears to be a small, constant difference between the older and younger individuals suggesting they are not exchangeable.

Nonparametric estimates of the impact of birth weight on IQ are shown in Tables 15 and 16. We find small, but significant effects, which is consistent with the linear results with the exception of 8 months. However, there does not appear to be any pattern of exchangeability over time, in fact the p-values suggest they become less exchangeable as they age.

As previously discussed in Section 2.5, measurement error problems are exacerbated when using the CRE device. Even though birth weight records from the CPP come directly from medical records we test the robustness of our conclusions. However, we are less interested in the impact of measurement error on the actual coefficients and more concerned about its impact on our tests of exchangeability. To address this, we use the results derived in equations (13) and (14) to first back out the values of β_1, β_2, λ_1 and λ_2 corrected for an assumed value of the measurement error variance and after that, we test for the equality of these parameters.

In the following we assume that the measurement error variance in both regressors is the same. We perform this procedure using 0.1 increments from zero until the point where half of the variance of observed birth weight would be coming from measurement error. Figure 1 shows the p-values as function of the assumed measurement error variance for siblings using our measure of mental and motor skills over time. For all of the variables, the p-values are flat lines, which is evidence that our conclusions from the tests are not driven by measurement error.[8]

[8]Nonetheless, we should note that this conclusion rests on the classical structure imposed on the measurement error. Different assumptions of the measurement error could imply different conclusions. Future work includes relaxing some of these assumptions (for example, the orthogonality between the measurement errors) to see how our p-values would change.

6 Conclusion

In this paper we have examined the exchangeability of twins and siblings. Using mean comparisons and the CRE device, our results indicate that twins appear to better satisfy the conditions for natural experiments, relative to siblings. As shown in the empirical section, there are numerous differences between siblings, providing evidence against their suitability as a natural experiment. However, this is not to say that twins are perfectly exchangeable. The fact that twins become more alike over time and that health endowments do not have persistent effects are indicators that go against exchangeability and their validity as counterfactuals.

The sibling results are not surprising based on the existing literature regarding intra-household allocation and their implications of birth order. Theoretically, Behrman, Pollak and Taubman (1982) develop a model of household allocation that explains differential familial investments in siblings due to differences in initial endowments, resources and parental preferences. Empirically, Behrman and Taubman (1986), using U.S. data, show that birth order has an effect on schooling and earnings. In light of this literature and our evidence, our results point to a rejection of the validity of siblings and valid counterfactuals.

The correlated random effects model and our results could also have implications regarding parental behavior and family's allocation of resources. For instance, a possible channel would be that as a result of differences in health endowments, as measured by birth weight, these exchangeability patterns may suggest that families are making compensatory investments in their children. Initially, health endowments have positive effects on cognitive development and as children grow up, parents realize differences in their health endowments and make higher investments in the disadvantaged child, attenuating the effects of birth weight on later outcomes.

In analyzing the results, the most striking pattern amongst siblings was the consistency at which the OLS estimates were a fraction of the Chamberlain estimates. Although this pattern might be surprising, we are not the first to find it. Behrman and Rosenzweig (2004) use female monozygotic twins and find a FE estimate larger than OLS regarding the effects of birth weight on schooling and Conley and Bennett (2000) find the same pattern for siblings and high school graduation using PSID data. Behrman and Rosenzweig (2004) suggest that one potential explanation for this result is that maternal/family investments tend to be negatively correlated with birth weight. However, we are able to run the same analysis for both siblings and twins and find estimates of λ that suggest OLS biases of different directions and that the downward OLS bias for siblings shows up for outcomes that are only minutes after birth. Therefore this explanation needs to be augmented to understand why λ is positive for twins and negative for siblings and how differential investments occur even for very early life outcomes.

In conclusion, although we are not the first to implement the CRE model, we believe its value is not fully exploited in empirical studies. Our aim is to show that the CRE is a simple framework that encompasses the popular fixed effects approach, which is a more general and provides useful tests of the validity of strategies using within family variation. By highlighting the generality of the CRE, we hope that future applied research can benefit from it.

References

Aaronson, Daniel, "Using Sibling Data to Estimate the Impact of Neighborhoods on Children's Educational Outcomes," *Journal of Human Resources*, 1998, *33* (4), 915–946.

Almond, Douglas, Kenneth Y. Chay, and David S. Lee, "The Costs of Low Birth Weight," *The Quarterly Journal of Economics*, 2005, *120*, 1031–1083.

Amin, Vikesh, Jere R. Behrman, and Tim D. Spector, "Does More Schooling Improve Health Outcomes and Health Related Behaviors? Evidence from UK Twins," *Economics of Education Review*, 2013.

Ashenfelter, Orley and Alan B. Krueger, "Estimates of the Economic Returns to Schooling from a New Sample of Twins," *American Economic Review*, December 1994, *84*, 1157–73.

_ **and David J. Zimmerman**, "Estimates of the Returns to Schooling from Sibling Data: Fathers, Sons, and Brothers," *Review of Economics and Statistics*, 1997, *79*, 1–9.

Behrman, Jere R. and Barbara L. Wolfe, "Does More Schooling Make Women Better Nourished and Healthier? Adult Sibling Random and Fixed Effects Estimates for Nicaragua," *Journal of Human Resources*, 1989, *24* (4), 644–63.

_ **and Mark R. Rosenzweig**, "Returns to Birthweight," *Review of Economics and Statistics*, 2004, *86* (2), 586–601.

_ **and Paul Taubman**, "Birth Order, Schooling, and Earnings," *Journal of Labor Economics*, 1986, *4* (3), S121–45.

_ **, Robert A. Pollak, and Paul Taubman**, "Parental Preferences and Provision for Progeny," *Journal of Political Economy*, 1982, *90* (1), 52–73.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes, "From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes," *The Quarterly Journal of Economics*, 2007, *122* (1), 409–439.

Bound, John and Gary Solon, "Double Trouble: On the Value of Twins-Based Estimation of the Return to Schooling," *Economics of Education Review*, April 1999, *18* (2), 169–182.

Branum, Amy M., Jennifer D. Parker, Sarah A. Keim, and Ashley H. Schempf, "Prepregnancy Body Mass Index and Gestational Weight Gain in Relation to Child Body Mass Index Among Siblings," *American Journal of Epidemiology*, 2011, *174* (10), 1159–1165.

Chamberlain, Gary, "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 1982, *18*, 5–46.

_ , "Panel Data," *Handbook of Econometrics*, 1984, *2*, 1247–1318.

Conley, Dalton and Neil G. Bennett, "Is Biology Destiny? Birth Weight and Life Chances," *American Sociological Review*, 2000, *65*, 458–467.

Currie, Janet and Mark Stabile, "Child Mental Health and Human Capital Accumulation: The Case of ADHD," *Journal of Health Economics*, 2006, *25* (6), 1094–1118.

_ , _ , Phongsack Manivong, and Leslie L. Roos, "Child Health and Young Adult Outcomes," *Journal of Human Resources*, 2010, *45* (3), 517–548.

Figlio, David N., Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth, "The Effects of Poor Neonatal Health on Children's Cognitive Development," Technical Report, National Bureau of Economic Research 2013.

Fletcher, Jason M., "The Effects of Teenage Childbearing on the Short- and Long-Term Health Behaviors of Mothers," *Journal of Population Economics*, October 2011, *25* (1), 201–218.

_ , "The Medium Term Schooling and Health Effects of Low Birth Weight: Evidence from Siblings," *Economics of Education Review*, 2011, *30* (3), 517–527.

_ , Jeremy C. Green, and Matthew J. Neidell, "Long Term Effects of Childhood Asthma on Adult Health," *Journal of Health Economics*, 2010, *29*, 377–387.

Gilman, Stephen E., Hannah Gardener, and Stephen L. Buka, "Maternal Smoking during Pregnancy and Children's Cognitive and Physical Development: A Causal Risk Factor?," *American Journal of Epidemiology*, 2008, *168* (5), 522.

Griliches, Zvi and Jerry A. Hausman, "Errors in Variables in Panel Data," *Journal of Econometrics*, February 1986, *31* (1), 93–118.

Hsiao, Cheng, *Analysis of Panel Data*, Cambridge University Press, 2003.

Klebanoff, Mark A., "The Collaborative Perinatal Project: a 50-year Retrospective," *Paediatric and Perinatal Epidemiology*, 2009, *23* (1), 2–8.

Lawlor, J.P., E. Gladen, D. Dhavale, D. Tamagoglu, J.B. Hardy, A.K. Duggan, W.W. Eaton, and E.F. Torrey, "Modernization and Enhancement of the Collaborative Perinatal Project (1959-74)," 2005. Funded by the Stanley Medical Research Institute, Bethesda MD and NIMH grant 070333.

Lundborg, Petter, "The Health Returns to Schooling–What Can we Learn from Twins?," *Journal of Population Economics*, 2013, *26* (2), 673–701.

Mundlak, Yair, "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 1978, *46*, 69–85.

Niswander, Kenneth R. and Myron Gordon, *The Women and their Pregnancies: the Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*, National Institute of Health, 1972.

Oreopoulos, Philip, Mark Stabile, Randy Walld, and Leslie L. Roos, "Short-, Medium-, and Long-Term Consequences of Poor Infant Health An Analysis Using Siblings and Twins," *Journal of Human Resources*, 2008, *43* (1), 88–138.

Rothstein, Donna S., "Breastfeeding and Children's Early Cognitive Outcomes," *Review of Economics and Statistics*, forthcoming.

Royer, Heather, "Separated at Girth: U.S. Twin Estimates of the Effects of Birth Weight," *American Economic Journal: Applied Economics*, 2009, *1*, 49–85.

Wooldridge, Jeffrey M., "Distribution-Free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics*, 1999, *90* (1), 77–97.

_ , "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 2005, *20* (1), 39–54.

_ , "Correlated Random Effects Models with Unbalanced Panels," May 2010.

_ , *Econometric Analysis of Cross Section and Panel Data*, 2 ed., The MIT Press, 2010.

Appendix

A.1 Technical Appendix

Consider the general two sibling/twin model:

$$Y_{1j} = \beta_1 X_{1j} + F_j + v_{1j} \tag{15}$$

$$Y_{2j} = \beta_2 X_{2j} + F_j + v_{2j} \tag{16}$$

What is the probability limit of $\hat{\beta}_{FE}$ if we restrict the treatment effect to be constant within group? Let $\beta_i = \beta + \eta_i$ where $E[\eta_i] = 0$ for $i = 1, 2$. For ease of explanation, the following focuses on a first difference setting, however since we are using a 2 person model, the results are the same for the fixed effect setup.

Now differencing equations (15) and (16) and then differencing we get:

$$\Delta Y_j = \beta \Delta X_j + \Delta \tilde{v}_j$$

where $\tilde{v}_{ij} = v_{ij} + \eta_i X_{ij}$. From this setup we know the equation for $\hat{\beta}_{FE}$:

$$\begin{aligned}
\hat{\beta}_{FE} = \frac{\sum_j \Delta X_j \Delta Y_j}{\sum_j {\Delta X_j}^2} &= \frac{\sum_j \Delta X_j \left(\beta \Delta X_j + \Delta \tilde{v}_j\right)}{\sum_j {\Delta X_j}^2} \\
&= \beta + \frac{\sum_j \Delta X_j \Delta \tilde{v}_j}{\sum_j {\Delta X_j}^2} \\
&= \beta + \frac{\sum_j \Delta X_j \left[(v_{2j} - v_{1j}) + (\eta_2 X_{2j} - \eta_1 X_{1j})\right]}{\sum_j {\Delta X_j}^2}
\end{aligned}$$

$$
\begin{aligned}
plim\ \hat{\beta}_{FE} &= \beta + \frac{E\left[\Delta X_j \Delta v_j\right] + E\left[\Delta X_j (\eta_2 X_{2j} - \eta_1 X_{1j})\right]}{E\left[\Delta {X_j}^2\right]} \\
&= \beta - \eta_1 \frac{E[\Delta X_j X_{1j}]}{E\left[\Delta {X_j}^2\right]} + \eta_2 \frac{E[\Delta X_j X_{2j}]}{E\left[\Delta {X_j}^2\right]} && (E\left[\Delta X_j \Delta v_j\right] = 0) \\
&= \frac{\beta E\left[\Delta {X_j}^2\right] - \eta_1 E[\Delta X_j X_{1j}] + \eta_2 E[\Delta X_j X_{2j}]}{E\left[\Delta {X_j}^2\right]} \\
&= \frac{\beta E\left[\Delta X_j \left(X_{2j} - X_{1j}\right)\right] - \eta_1 E[\Delta X_j X_{1j}] + \eta_2 E[\Delta X_j X_{2j}]}{E\left[\Delta {X_j}^2\right]} \\
&= \frac{-(\beta + \eta_1) E[\Delta X_j X_{1j}] + (\beta + \eta_2) E\left[\Delta X_j X_{2j}\right]}{E\left[\Delta {X_j}^2\right]} \\
&= -\beta_1 \left(\frac{E[\Delta X_j X_{1j}]}{E\left[\Delta {X_j}^2\right]}\right) + \beta_2 \left(\frac{E\left[\Delta X_j X_{2j}\right]}{E\left[\Delta {X_j}^2\right]}\right) \\
&= \beta_1 \left(\frac{E\left[{X_{1j}}^2\right] - E\left[X_{1j} X_{2j}\right]}{E\left[{X_{1j}}^2\right] + E\left[{X_{2j}}^2\right] - 2E\left[X_{1j} X_{2j}\right]}\right) + \beta_2 \left(\frac{E\left[{X_{2j}}^2\right] - E\left[X_{1j} X_{2j}\right]}{E\left[{X_{1j}}^2\right] + E\left[{X_{2j}}^2\right] - 2E\left[X_{1j} X_{2j}\right]}\right)
\end{aligned}
$$

Let $E[X_{1j}] = \mu_1, E[X_{2j}] = \mu_2, Var(X_{1j}) = \sigma_1^2, Var(X_{2j}) = \sigma_2^2$, and $Corr(X_{1j}, X_{2j}) = \rho$,

$$
plim\ \hat{\beta}_{FE} = \beta_1 \underbrace{\left(\frac{\sigma_1^2 + \mu_1^2 - \rho\sigma_1\sigma_2 - \mu_1\mu_2}{\sigma_1^2 + \sigma_2^2 + \mu_1^2 + \mu_2^2 - 2\rho\sigma_1\sigma_2 - 2\mu_1\mu_2}\right)}_{\theta} + \beta_2 \underbrace{\left(\frac{\sigma_2^2 + \mu_2^2 - \rho\sigma_1\sigma_2 - \mu_1\mu_2}{\sigma_1^2 + \sigma_2^2 + \mu_1^2 + \mu_2^2 - 2\rho\sigma_1\sigma_2 - 2\mu_1\mu_2}\right)}_{1-\theta}
$$

Note that θ is not restricted to be between 0 and 1. $\hat{\beta}_{FE}$ converges in a probability to a weighted average of β_1 and β_2 (not necessarily a convex combination).

A.2 Data Appendix

This section describes in more details the standardized tests used in the paper. With the exception of APGAR and GENERAL MOTOR TEST, we standardized all tests to percentile rank to perform our analysis. For more information on these, see Niswander and Gordon (1972).

1. **APGAR** (at birth): The APGAR score was devised as a simple and repeatable method to quickly and summarily assess the health of newborn children immediately after birth. The APGAR score is determined by evaluating the newborn baby on five simple criteria on a scale from zero to two, then summing up the five values thus obtained. The resulting APGAR score ranges from 0 to 10. The five criteria are summarized using words chosen to form a Appearance (skin color), Pulse (heart rate), Grimace (reflex irritability), Activity (muscle tone), and Respiration. The test is generally done at one and five minutes after birth, and may be repeated later if the score remains low. Scores 7 and above are generally normal, 4 to 6 fairly low, and 3 and below are generally regarded as critically low.

2. **Bayley Scales of Infant (Mental and Motor) Development** (8 months): The Bayley Scales of Infant Development (BSID) is a standard series of measurements used primarily to assess the motor (fine and gross), language (receptive and expressive), and cognitive development of infants with ages one to 42 months.

   This measure consists of a series of developmental play tasks and takes between 45-60 minutes. Scores of successfully completed items are used to determine the child's performance compared with norms taken from typically developing children of their age (in months). The CPP used the BSID to children along two scales:

   Mental scale: This part of the evaluation, which yields a score called the mental development index, evaluates several types of abilities: sensory/perceptual acuities, discriminations, and response; acquisition of object constancy; memory learning and problem solving; vocalization and beginning of verbal communication; basis of abstract thinking; habituation; mental mapping; complex language; and mathematical concept formation.

   Motor scale: This part of the BSID assesses the degree of body control, large muscle coordination, finer manipulatory skills of the hands and fingers, dynamic movement, postural imitation, and the ability to recognize objects by sense of touch (stereognosis).

3. **General Motor Test** (4 years): This General motor test was constructed from data on 4 years from the following 13 specific tests: (1) line walk; (2) hopping right foot; (3) hopping left foot; (4) ball catch trial 1; (5) ball catch trial 2; (6) call catch trial 3; (7) wallin pegboard right; (8) wallin pegboard left; (9) copy circle; (10) copy square; (11) copy cross; (12) stringing beads; (13) maze IV. Scores go from 0 to 13, where 1 point is equivalent to passing a specific test.

4. **Stanford-Binet Intelligence Scales** (4 years): The Stanford-Binet Intelligence Scales attempts to measure cognitive ability of children. This test measures primarily visual-motor capabilities, nonverbal reasoning, social intelligence, and language functions

between the ages of 2 and 5 years, and abstract reasoning and memory skills at older ages. The CPP applied a short form of the third revision of the Stanford-Binet Test.

5. **Bender-Gestalt Visual Motor Test** (7 years): The Bender-Gestalt Visual Motor Test is used to evaluate "visual-motor maturity", to screen for developmental disorders, or to assess neurological function or brain damage. The original test consists of nine figures, each on its own 3 5 card, comprising dots, lines, angles, and curves. The subject is shown each figure and asked to copy it onto a piece of blank paper. The test typically takes 7-10 minutes, after which the results are scored based on accuracy and other characteristics.

6. **Weschler Intelligence Scales** (7 years): The Wechsler Intelligence Scale for Children (WISC) is an individually administered intelligence test for children between the ages of 6 and 16 that can be completed without reading or writing. The WISC takes 65-80 minutes to administer and generates an IQ score which represents a childs general cognitive ability. The WISC is divided into two summary scales of verbal (Verbal IQ) and nonverbal (Performance IQ) intelligence and also includes a total summary intelligence scale (Full Scale IQ). The CPP administered a short form of the WISC that comprised 4 verbal subtests and 3 of the performance subtests.

7. **Wide Range Achievement Test** (7 years): The Wide Range Achievement Test (WRAT) is an achievement test which measures an individual's ability to read, spell and solve simple arithmetical problems.The test takes approximately 20-30 minutes and is appropriate for individuals ages 5-94 years. The reading part measures letter and word decoding through letter identification and word recognition. The spelling test estimates an individual's ability to encode sounds into written form through the use of a dictated spelling format containing both letters and words. The math sections is aimed to evaluate an individual's ability to perform basic mathematics computations through counting, identifying numbers, solving simple oral problems, and calculating written mathematics problems. The current WRAT version 4 has an additional test on sentence comprehension

Table 1: Potential Scenarios

Case	Mean	Variance	Correlation	θ
Panel A: $\mu_1 \neq \mu_2$				
1	$\mu_1 \neq \mu_2$	$\sigma_1^2 = \sigma_2^2 = \sigma^2$	$corr(X_1, X_2) = 0$	$\frac{\sigma^2+\mu_1^2-\mu_1\mu_2}{\mu_1^2+\mu_2^2+2\sigma^2-2\mu_1\mu_2}$
2	$\mu_1 \neq \mu_2$	$\sigma_1^2 = \sigma_2^2 = \sigma^2$	$corr(X_1, X_2) \neq 0$	$\frac{\sigma^2(1-\rho)+\mu_1^2-\mu_1\mu_2}{\mu_1^2+\mu_2^2+2\sigma^2(1-\rho)-2\mu_1\mu_2}$
3	$\mu_1 \neq \mu_2$	$\sigma_1^2 \neq \sigma_2^2$	$corr(X_1, X_2) = 0$	$\frac{\sigma_1^2+\mu_1^2-\mu_1\mu_2}{\mu_1^2+\mu_2^2+\sigma_1^2+\sigma_2^2-2\mu_1\mu_2}$
4	$\mu_1 \neq \mu_2$	$\sigma_1^2 \neq \sigma_2^2$	$corr(X_1, X_2) \neq 0$	$\frac{\sigma_1^2+\mu_1^2-\rho\sigma_1\sigma_2-\mu_1\mu_2}{\mu_1^2+\mu_2^2+\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2-2\mu_1\mu_2}$
Panel B: $\mu_1 = \mu_2$				
5	$\mu_1 = \mu_2 = \mu$	$\sigma_1^2 = \sigma_2^2 = \sigma^2$	–	$\frac{1}{2}$
6	$\mu_1 = \mu_2 = \mu$	$\sigma_1^2 \neq \sigma_2^2$	$corr(X_1, X_2) = 0$	$\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}$
7	$\mu_1 = \mu_2 = \mu$	$\sigma_1^2 \neq \sigma_2^2$	$corr(X_1, X_2) \neq 0$	$\frac{\sigma_1^2-\rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}$

Table 2: Sample Means for Twins CPP

	(1) 1^{st}-Born	(2) 2^{nd}-Born	(3) Raw Diff.	(4) Adj. FE Diff.	(5) Number of Pairs
Birth Weight (in 100g)	21.969	21.241	0.728***	–	615
	(7.149)	(7.327)	[0.100]		
Health Outcomes					
APGAR 1-min	7.267	5.982	1.285***	1.249***	494
	(2.349)	(2.711)	[0.491]	[0.477]	
APGAR 5-min	8.509	8.069	0.440***	0.416***	491
	(1.934)	(2.118)	[0.216]	[0.204]	
Incubator Used	0.587	0.608	-0.021	-0.010	533
	(0.493)	(0.489)	[-0.042]	[-0.020]	
Incubator Days	2.026	2.041	-0.015	0.047	533
	(2.462)	(2.410)	[-0.006]	[0.019]	
Hospital Days	13.975	14.018	-0.044	0.318	597
	(14.988)	(14.906)	[-0.003]	[0.021]	
Infant Death within 1 Day	0.059	0.077	-0.018**	-0.018	558
	(0.236)	(0.267)	[-0.071]	[-0.072]	
Infant Death within 1 Month	0.084	0.095	-0.011	-0.011	558
	(0.278)	(0.293)	[-0.038]	[-0.038]	
Infant Death within 1 Year	0.106	0.106	0.000	0.001	558
	(0.308)	(0.308)	[0.000]	[0.004]	
Motor Skills					
8 Months Old	28.954	29.227	-0.273	-0.153	412
	(26.316)	(28.392)	[-0.010]	[-0.006]	
4 Years Old	9.071	9.095	-0.024	-0.042	252
	(2.154)	(2.263)	[-0.011]	[-0.019]	
7 Years Old	53.932	55.603	-1.670	-1.305	370
	(30.298)	(30.176)	[-0.055]	[-0.043]	
Mental Skills (IQ Tests)					
8 Months Old	30.185	29.106	1.079	1.119	411
	(26.964)	(26.009)	[0.041]	[0.042]	
4 Years Old	34.831	34.181	0.649	0.529	356
	(27.413)	(27.703)	[0.024]	[0.019]	
7 Years Old	38.361	39.332	-0.971	-1.037	376
	(28.293)	(28.163)	[-0.034]	[-0.037]	
WRAT (by Age 7)					
Reading	40.437	38.714	1.723*	1.707	371
	(27.522)	(27.985)	[0.062]	[0.062]	
Spelling	38.452	37.770	0.681	0.657	372
	(28.859)	(27.804)	[0.024]	[0.023]	
Arithmetic	37.034	36.525	0.508	0.338	371
	(27.920)	(27.455)	[0.018]	[0.012]	

Notes: Data are from the Collaborative Perinatal Project (CPP). Columns (1) and (2) are the sample means for first and second borns, respectively. Column (3) represents the raw difference within the pair and column (4) is the difference having linearly controlled for birth weight and mother fixed effects. Column (4) of motor skills, mental skills, and WRAT also controls for differences in age at the time of the test. Motor skills, mental skills, and WRAT are in percentile rank. Standard deviation are reported in parenthesis. Numbers in brackets report the standard deviation effect for being the first-born. Standard errors are heteroskedasticity-robust and clustered at the pair level; * denotes 90% significance, ** denotes 95% significance, and *** denotes 99% significance.

Table 3: Sample Means for Siblings CPP

	(1) 1^{st}-Born	(2) 2^{nd}-Born	(3) Raw Diff.	(4) Adj. FE Diff.	(5) Number of Pairs
Birth Weight (in 100g)	30.855	31.249	-0.394***	–	12,678
	(6.917)	(6.764)	[-0.058]		
Characteristics					
Female	0.487	0.495	-0.008	-0.011	12,678
	(0.500)	(0.500)	[-0.015]	[-0.023]	
Gestational length (in weeks)	38.868	38.806	0.062	0.213***	12,562
	(4.177)	(4.028)	[0.015]	[0.052]	
Mother Age	22.945	25.131	-2.186***	-2.187***	12,678
	(5.268)	(5.447)	[-0.400]	[-0.400]	
SES (family) at registration	45.073	47.904	-2.831***	-2.827***	11,767
	(20.901)	(20.515)	[-0.136]	[-0.136]	
SES (family) 7 Years	45.239	48.179	-2.940***	-2.932***	8,666
	(20.823)	(20.428)	[-0.142]	[-0.142]	
Mother Smoke pregnancy (0 - 1)	0.522	0.531	-0.009***	-0.010**	12,678
	(0.500)	(0.499)	[-0.018]	[-0.020]	
Mother Smoke pregnancy (# cigs)	7.443	8.372	-0.929***	-0.933***	12,115
	(10.241)	(11.236)	[-0.086]	[-0.087]	
Health Outcomes					
APGAR 1-min	7.676	7.897	-0.221***	-0.193***	10,181
	(2.039)	(1.837)	[-0.114]	[-0.099]	
APGAR 5-min	8.839	8.968	-0.129***	-0.106***	10,624
	(1.357)	(1.224)	[-0.100]	[-0.082]	
Incubator Used	0.166	0.183	-0.017***	-0.023***	11,518
	(0.372)	(0.386)	[-0.045]	[-0.060]	
Incubator Days	0.386	0.410	-0.023*	-0.046***	11,518
	(1.176)	(1.167)	[-0.020]	[-0.039]	
Hospital Days	6.019	6.346	-0.326***	-0.450***	12,324
	(7.540)	(10.565)	[-0.036]	[-0.049]	
Infant Death within 1 Day	0.018	0.014	0.003**	0.001	12,113
	(0.132)	(0.119)	[0.028]	[0.005]	
Infant Death within 1 Month	0.025	0.018	0.007***	0.003	12,113
	(0.155)	(0.133)	[0.046]	[0.022]	
Infant Death within 1 Year	0.035	0.029	0.006***	0.002	12,113
	(0.183)	(0.167)	[0.033]	[0.013]	
Motor Skills					
8 Months Old	45.250	42.790	2.461***	3.833***	8,905
	(28.663)	(28.264)	[0.086]	[0.135]	
4 Years Old	9.671	10.012	-0.341***	-0.323***	5,791
	(2.007)	(1.942)	[-0.172]	[-0.163]	
7 Years Old	47.478	47.561	-0.083	1.982***	8,818
	(29.926)	(28.724)	[-0.003]	[0.068]	
Mental Skills (IQ Tests)					
8 Months Old	48.713	43.472	5.242***	6.376***	8,891
	(29.312)	(27.925)	[0.182]	[0.222]	
4 Years Old	47.724	51.694	-3.970***	-3.929***	7,662
	(28.795)	(28.404)	[-0.138]	[-0.137]	
7 Years Old	46.529	47.930	-1.401***	-0.990**	8,865
	(29.108)	(28.187)	[-0.049]	[-0.035]	
WRAT (by Age 7)					
Reading	49.704	46.210	3.494***	3.555***	8,784
	(29.602)	(28.747)	[0.120]	[0.122]	
Spelling	49.244	45.303	3.940***	4.010***	8,786
	(30.160)	(29.173)	[0.133]	[0.135]	
Arithmetic	47.154	45.389	1.765***	1.856***	8,789
	(29.155)	(27.280)	[0.062]	[0.066]	

Notes: See notes to Table 2

Table 4: Short Run for Twins (CPP)

	APGAR 1-min			APGAR 5-min		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE
$\hat{\beta}_{FE}$		0.122**			0.077**	
		(0.054)			(0.039)	
$\hat{\beta}_1$	0.152***		0.107***	0.169***		0.079***
	(0.017)		(0.036)	(0.017)		(0.026)
	[0.359]		[0.253]	[0.511]		[0.239]
$\hat{\beta}_2$	0.136***		0.093**	0.150***		0.058**
	(0.018)		(0.037)	(0.017)		(0.027)
	[0.322]		[0.220]	[0.455]		[0.174]
$\hat{\lambda}_1$			0.033			0.055**
			(0.032)			(0.022)
$\hat{\lambda}_2$			0.016			0.047**
			(0.025)			(0.019)
p-value Joint Test	–	–	0.699	–	–	0.365
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.510	–	–	0.167
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.693	–	–	0.786
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0.414	–	–	0.00106
mean of outcome		6.624			8.289	
sd of outcome		2.615			2.039	
Number of Pairs		494			491	

Notes: Data are from the Collaborative Perinatal Project (CPP). Columns (1) and (4) show the OLS coefficient using only first-borns on the first row and only second-borns on the second row, columns (2) and (5) report the pooled fixed effect estimate, and columns (3) and (6) show the structural parameters from the CRE. Controls used include dummy for zygosity and gender. p-value Joint Test is the pvalue of testing jointly $H_0 : \beta_1 = \beta_2$ and $\lambda_1 = \lambda_2$. Heteroskedasticity-robust standard errors are reported in parenthesis. Numbers in brackets report the standard deviation effect of birth weight ($\beta_{bw} \times \frac{\sigma_{bw}}{\sigma_y}$); * denotes 90% significance, ** denotes 95% significance, and *** denotes 99% significance.

Table 5: Short Run for Twins (CPP)

	Incubator Used			Incubator Days			Hospital Days		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		-0.039***			-0.218***			-0.520**	
		(0.008)			(0.034)			(0.256)	
$\hat{\beta}_1$	-0.051***		-0.040***	-0.260***		-0.223***	-0.715***		-0.490***
	(0.002)		(0.005)	(0.019)		(0.022)	(0.129)		(0.182)
	[-0.622]		[-0.482]	[-0.634]		[-0.543]	[-0.330]		[-0.226]
$\hat{\beta}_2$	-0.049***		-0.038***	-0.244***		-0.218***	-0.598***		-0.487***
	(0.003)		(0.005)	(0.019)		(0.023)	(0.115)		(0.167)
	[-0.588]		[-0.459]	[-0.595]		[-0.532]	[-0.276]		[-0.225]
$\hat{\lambda}_1$			-0.009*			-0.038**			-0.363**
			(0.005)			(0.019)			(0.142)
$\hat{\lambda}_2$			-0.003			0.002			0.159
			(0.005)			(0.020)			(0.108)
p-value Joint Test	–	–	0.392	–	–	0.283	–	–	0.00242
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.340	–	–	0.735	–	–	0.976
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.425	–	–	0.146	–	–	0.000656
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0.0738	–	–	0.143	–	–	0.00300
mean of outcome		0.598			2.034			14.00	
sd of outcome		0.491			2.435			14.94	
Number of Pairs		533			533			597	

Notes: See notes to Table 4.

Table 6: Infant Mortality for Twins (CPP)

	Within 1 Day			Within 1 Month			Within 1 Year		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		0.001			0.001			-0.004	
		(0.004)			(0.003)			(0.005)	
$\hat{\beta}_1$	-0.018***		0.002	-0.023***		0.001	-0.024***		-0.003
	(0.003)		(0.003)	(0.003)		(0.003)	(0.003)		(0.004)
	[-0.454]		[0.058]	[-0.497]		[0.032]	[-0.488]		[-0.056]
$\hat{\beta}_2$	-0.021***		0.000	-0.025***		-0.001	-0.026***		-0.005
	(0.003)		(0.003)	(0.003)		(0.003)	(0.003)		(0.004)
	[-0.508]		[0.002]	[-0.548]		[-0.021]	[-0.526]		[-0.101]
$\hat{\lambda}_1$			-0.013***			-0.014***			-0.013***
			(0.003)			(0.003)			(0.003)
$\hat{\lambda}_2$			-0.010***			-0.013***			-0.011***
			(0.002)			(0.003)			(0.003)
p-value Joint Test	–	–	0.453	–	–	0.365	–	–	0.554
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.261	–	–	0.157	–	–	0.281
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.397	–	–	0.765	–	–	0.722
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	1.80e-09	–	–	0	–	–	2.85e-07
mean of mortality		0.0681			0.0896			0.106	
sd of mortality		0.252			0.286			0.308	
Number of Pairs		558			558			558	

Notes: See notes to Table 4.

Table 7: Mental Skills for Twins (CPP)

	8 Months			4 Years			7 Years		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		1.090**			0.287			0.321	
		(0.463)			(0.496)			(0.477)	
$\hat{\beta}_1$	1.727***		1.002***	1.269***		0.149	1.226***		0.478
	(0.246)		(0.336)	(0.255)		(0.339)	(0.273)		(0.361)
	[0.341]		[0.198]	[0.242]		[0.028]	[0.222]		[0.087]
$\hat{\beta}_2$	1.428***		0.996***	1.385***		0.372	1.244***		0.416
	(0.228)		(0.329)	(0.280)		(0.361)	(0.291)		(0.340)
	[0.282]		[0.197]	[0.264]		[0.071]	[0.225]		[0.075]
$\hat{\lambda}_1$			0.728**			0.815**			0.413
			(0.366)			(0.411)			(0.400)
$\hat{\lambda}_2$			-0.061			0.477			0.544
			(0.349)			(0.408)			(0.422)
p-value Joint Test	–	–	0.413	–	–	0.620	–	–	0.959
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.983	–	–	0.361	–	–	0.803
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.194	–	–	0.633	–	–	0.853
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0.101	–	–	0.00639	–	–	0.0778
mean of mental		29.65			34.51			38.85	
sd of mental		26.48			27.54			28.21	
Number of Pairs		411			356			376	

Table 7b: Motor Skills for Twins (CPP)

	8 Months			4 Years			7 Years		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		0.955**			0.035			-0.746	
		(0.422)			(0.046)			(0.649)	
$\hat{\beta}_1$	1.768***		0.739**	0.062***		0.006	-1.241***		-0.691
	(0.230)		(0.308)	(0.023)		(0.033)	(0.296)		(0.451)
	[0.338]		[0.141]	[0.153]		[0.015]	[-0.210]		[-0.117]
$\hat{\beta}_2$	1.906***		1.123***	0.096***		0.071**	-1.370***		-0.794
	(0.254)		(0.317)	(0.025)		(0.029)	(0.307)		(0.484)
	[0.365]		[0.215]	[0.236]		[0.175]	[-0.232]		[-0.134]
$\hat{\lambda}_1$			1.017***			0.081**			-0.448
			(0.381)			(0.035)			(0.414)
$\hat{\lambda}_2$			0.050			-0.031			-0.247
			(0.380)			(0.034)			(0.446)
p-value Joint Test	–	–	0.107	–	–	0.0276	–	–	0.931
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.0794	–	–	0.0191	–	–	0.764
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.145	–	–	0.0502	–	–	0.772
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0.00593	–	–	0.0701	–	–	0.367
mean of motor		29.09			9.083			54.77	
sd of motor		27.36			2.207			30.23	
Number of Pairs		412			252			370	

Notes: See notes to Table 4. Mental scores for 8 Months, 4 Years and 7 Years are from the Bayley Scales, Stanford-Binet and Weschler, respectively. Motor scores for 8 Months and 7 Years are from Bayley Scales and Bender-Gestalt, respectively. Motor test for 4 Years ranges from 0 to 13 and is a combination of 13 other specific motor tests. All tests scores, except Motor 4 Years, are in percentile rank. See Appendix for more information on these.

Table 8: Wide Range Achievement Tests (WRAT) 7-Years for Twins (CPP)

	Reading			Spelling			Arithmetic		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		0.199			0.305			0.577	
		(0.412)			(0.403)			(0.480)	
$\hat{\beta}_1$	0.872***		0.140	1.073***		0.389	1.020***		0.506
	(0.284)		(0.308)	(0.280)		(0.293)	(0.271)		(0.362)
	[0.161]		[0.026]	[0.194]		[0.070]	[0.189]		[0.094]
$\hat{\beta}_2$	0.991***		0.179	1.042***		0.243	1.046***		0.588*
	(0.285)		(0.275)	(0.283)		(0.277)	(0.271)		(0.338)
	[0.183]		[0.033]	[0.189]		[0.044]	[0.194]		[0.109]
$\hat{\lambda}_1$			0.491			0.457			0.494
			(0.409)			(0.380)			(0.385)
$\hat{\lambda}_2$			0.440			0.442			0.114
			(0.403)			(0.411)			(0.396)
p-value Joint Test	–	–	0.984	–	–	0.823	–	–	0.836
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.860	–	–	0.546	–	–	0.770
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.942	–	–	0.982	–	–	0.570
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0.0766	–	–	0.0769	–	–	0.265
mean of outcome		39.58			38.11			36.78	
sd of outcome		27.75			28.32			27.67	
Number of Pairs		371			372			371	

Notes: See notes to Table 4. WRAT scores are in percentile rank. See Appendix for more information on these.

Table 9: Short Run for Siblings (CPP)

	APGAR 1-min			APGAR 5-min		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE
$\hat{\beta}_{FE}$		0.069***			0.068***	
		(0.009)			(0.007)	
$\hat{\beta}_1$	0.058***		0.092***	0.052***		0.083***
	(0.005)		(0.007)	(0.004)		(0.005)
	[0.177]		[0.279]	[0.237]		[0.380]
$\hat{\beta}_2$	0.049***		0.090***	0.041***		0.080***
	(0.005)		(0.006)	(0.004)		(0.005)
	[0.148]		[0.273]	[0.188]		[0.365]
$\hat{\lambda}_1$			-0.022***			-0.023***
			(0.005)			(0.003)
$\hat{\lambda}_2$			-0.024***			-0.019***
			(0.004)			(0.003)
p-value Joint Test	–	–	0.895	–	–	0.563
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.751	–	–	0.485
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.812	–	–	0.382
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0	–	–	0
mean of outcome		7.781			8.903	
sd of outcome		1.938			1.283	
Number of Pairs		9237			9660	

Notes: Data are from the Collaborative Perinatal Project (CPP). Columns (1) and (4) show the OLS coefficient using only first-borns on the first row and only second-borns on the second row, columns (2) and (5) report the pooled fixed effect estimate, and columns (3) and (6) show the structural parameters from the CRE. Controls used include gender, maternal age, SES at registration, gestational length, and the maximum number of cigarettes smoked consistently during pregnancy. p-value Joint Test is the pvalue of testing jointly $H_0 : \beta_1 = \beta_2$ and $\lambda_1 = \lambda_2$. Heteroskedasticity-robust standard errors are reported in parenthesis. Numbers in brackets report the standard deviation effect of birth weight ($\beta_{bw} \times \frac{\sigma_{bw}}{\sigma_y}$); * denotes 90% significance, ** denotes 95% significance, and *** denotes 99% significance.

Table 10: Short Run for Siblings (CPP)

	Incubator Used			Incubator Days			Hospital Days		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		-0.023***			-0.094***			-0.330***	
		(0.001)			(0.006)			(-0.046)	
$\hat{\beta}_1$	-0.021***		-0.027***	-0.084***		-0.106***	-0.267***		-0.323***
	(0.001)		(0.001)	(0.004)		(0.004)	(0.022)		(0.036)
	[-0.317]		[-0.417]	[-0.414]		[-0.523]	[-0.183]		[-0.222]
$\hat{\beta}_2$	-0.023***		-0.030***	-0.090***		-0.117***	-0.383***		-0.410***
	(0.001)		(0.001)	(0.003)		(0.004)	(0.025)		(0.033)
	[-0.356]		[-0.458]	[-0.446]		[-0.578]	[-0.263]		[-0.282]
$\hat{\lambda}_1$			0.006***			0.018***			0.057**
			(0.001)			(0.003)			(0.029)
$\hat{\lambda}_2$			0.002***			0.010***			-0.020
			(0.001)			(0.002)			(0.017)
p-value Joint Test	–	–	4.82e-07	–	–	0.00195	–	–	3.68e-05
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.00283	–	–	0.00819	–	–	0.0125
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.000961	–	–	0.0251	–	–	0.0189
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0	–	–	0	–	–	0.0635
mean of outcome		0.174			0.393			6.179	
sd of outcome		0.379			1.163			9.288	
Number of Pairs		10425			10425			11038	

Notes: See notes to Table 9.

Table 11: Infant Mortality for Siblings (CPP)

	Within 1 Day			Within 1 Month			Within 1 Year		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		-0.008***			-0.010***			-0.010***	
		(0.001)			(0.001)			(0.001)	
$\hat{\beta}_1$	-0.007***		-0.010***	-0.008***		-0.012***	-0.009***		-0.013***
	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)
	[-0.321]		[-0.507]	[-0.338]		[-0.522]	[-0.309]		[-0.457]
$\hat{\beta}_2$	-0.005***		-0.010***	-0.006***		-0.012***	-0.007***		-0.013***
	(0.000)		(0.001)	(0.000)		(0.001)	(0.001)		(0.001)
	[-0.248]		[-0.493]	[-0.253]		[-0.511]	[-0.226]		[-0.445]
$\hat{\lambda}_1$			0.003***			0.003***			0.003***
			(0.000)			(0.000)			(0.000)
$\hat{\lambda}_2$			0.003***			0.003***			0.004***
			(0.000)			(0.000)			(0.000)
p-value Joint Test	–	–	0.748	–	–	0.715	–	–	0.240
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.613	–	–	0.676	–	–	0.584
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.646	–	–	0.448	–	–	0.113
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0	–	–	0	–	–	0
mean of mortality		0.0150			0.0202		0.0304		
sd of mortality		0.122			0.141		0.172		
Number of Pairs		10827			10827		10827		

Notes: See notes to Table 9.

Table 12: Mental Skills for Siblings (CPP)

	8 Months			4 Years			7 Years		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		0.901***			0.402***			0.461***	
		(0.136)			(0.110)			(0.104)	
$\hat{\beta}_1$	0.777***		0.948***	0.654***		0.341***	0.529***		0.390***
	(0.065)		(0.096)	(0.064)		(0.078)	(0.058)		(0.073)
	[0.145]		[0.177]	[0.122]		[0.064]	[0.099]		[0.073]
$\hat{\beta}_2$	0.854***		1.200***	0.699***		0.519***	0.620***		0.566***
	(0.058)		(0.084)	(0.060)		(0.071)	(0.054)		(0.065)
	[0.159]		[0.224]	[0.130]		[0.097]	[0.116]		[0.106]
$\hat{\lambda}_1$			-0.156**			0.164**			0.033
			(0.072)			(0.073)			(0.068)
$\hat{\lambda}_2$			-0.141**			0.054			0.024
			(0.069)			(0.066)			(0.062)
p-value Joint Test	–	–	0.00774	–	–	0.0652	–	–	0.0286
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.00593	–	–	0.0202	–	–	0.0114
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.885	–	–	0.316	–	–	0.933
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0.00765	–	–	0.0296	–	–	0.781
mean of mental		46.21			50.23			47.98	
sd of mental		28.62			28.68			28.67	
Number of Pairs		8094			6954			7986	

Table 12b: Motor Skills for Siblings (CPP)

	8 Months			4 Years			7 Years		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		0.916***			0.041***			-0.486***	
		(0.128)			(0.012)			(0.131)	
$\hat{\beta}_1$	0.844***		1.011***	0.042***		0.044***	-0.501***		-0.419***
	(0.061)		(0.090)	(0.006)		(0.009)	(0.064)		(0.092)
	[0.158]		[0.190]	[0.113]		[0.119]	[-0.092]		[-0.077]
$\hat{\beta}_2$	0.957***		1.219***	0.047***		0.054***	-0.543***		-0.645***
	(0.058)		(0.079)	(0.005)		(0.008)	(0.059)		(0.085)
	[0.180]		[0.229]	[0.126]		[0.145]	[-0.099]		[-0.118]
$\hat{\lambda}_1$			-0.177**			-0.002			-0.103
			(0.073)			(0.007)			(0.074)
$\hat{\lambda}_2$			-0.028			-0.005			0.123*
			(0.067)			(0.006)			(0.069)
p-value Joint Test	–	–	0.00115	–	–	0.463	–	–	0.0214
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.0155	–	–	0.218	–	–	0.0122
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.154	–	–	0.749	–	–	0.0333
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0.0420	–	–	0.589	–	–	0.0970
mean of motor		44.22			9.867			46.96	
sd of motor		28.41			1.972			29.30	
Number of Pairs		8108			5252			7956	

Notes: See notes to Table 9. Mental scores for 8 Months, 4 Years and 7 Years are from the Bayley Scales, Stanford-Binet and Weschler, respectively. Motor scores for 8 Months and 7 Years are from Bayley Scales and Bender-Gestalt, respectively. Motor test for 4 Years ranges from 0 to 13 and is a combination of 13 other specific motor tests. All tests scores, except Motor 4 Years, are in percentile rank. See Appendix for more information on these.

Table 13: Wide Range Achievement Tests (WRAT) 7-Years for Siblings (CPP)

	Reading			Spelling			Arithmetic		
	(1) OLS	(2) FE	(3) CRE	(4) OLS	(5) FE	(6) CRE	(7) OLS	(8) FE	(9) CRE
$\hat{\beta}_{FE}$		0.389***			0.376***			0.469***	
		(0.117)			(0.123)			(0.127)	
$\hat{\beta}_1$	0.353***		0.433***	0.389***		0.411***	0.311***		0.539***
	(0.062)		(0.081)	(0.063)		(0.084)	(0.063)		(0.088)
	[0.065]		[0.080]	[0.070]		[0.074]	[0.059]		[0.103]
$\hat{\beta}_2$	0.416***		0.509***	0.486***		0.587***	0.402***		0.614***
	(0.057)		(0.075)	(0.058)		(0.078)	(0.056)		(0.080)
	[0.077]		[0.094]	[0.088]		[0.106]	[0.077]		[0.117]
$\hat{\lambda}_1$			-0.051			0.009			-0.165**
			(0.071)			(0.073)			(0.071)
$\hat{\lambda}_2$			-0.121*			-0.132*			-0.194***
			(0.066)			(0.068)			(0.068)
p-value Joint Test	–	–	0.602	–	–	0.0922	–	–	0.669
p-value $H_0 : \beta_1 = \beta_2$	–	–	0.339	–	–	0.0334	–	–	0.372
p-value $H_0 : \lambda_1 = \lambda_2$	–	–	0.509	–	–	0.191	–	–	0.781
p-value $H_0 : \lambda_1 = \lambda_2 = 0$	–	–	0.106	–	–	0.153	–	–	0.000418
mean of outcome		48.32			47.64			46.62	
sd of outcome		29.21			29.66			28.16	
Number of Pairs		7930			7928			7930	

Notes: See notes to Table 9. WRAT scores are in percentile rank. See Appendix for more information on these.

Table 14: APGAR and Infant Death for Siblings (CPP)

	APGAR 5-min				Infant Death within 1 Month			
	OLS		CRE		OLS		CRE	
	(1) 1^{st}-Born	(2) 2^{nd}-Born	(3) 1^{st}-Born	(4) 2^{nd}-Born	(5) 1^{st}-Born	(6) 2^{nd}-Born	(7) 1^{st}-Born	(8) 2^{nd}-Born
slope$_1$	-0.139***	-0.174***	-0.081***	-0.121***	0.013***	0.015***	0.008**	0.006*
$[\beta_1]$	(0.026)	(0.023)	(0.029)	(0.027)	(0.004)	(0.003)	(0.004)	(0.004)
slope$_2$	0.065***	0.030***	0.093***	0.062***	-0.013***	-0.010***	-0.016***	-0.014***
$[\beta_1 + \beta_2]$	(0.009)	(0.009)	(0.011)	(0.011)	(0.001)	(0.001)	(0.001)	(0.001)
slope$_3$	0.064***	0.039***	0.090***	0.062***	-0.012***	-0.009***	-0.014***	-0.012***
$[\beta_1 + \beta_2 + \beta_3]$	(0.007)	(0.007)	(0.009)	(0.009)	(0.001)	(0.001)	(0.001)	(0.001)
slope$_4$	0.050***	0.030***	0.076***	0.054***	-0.009***	-0.007***	-0.012***	-0.010***
$[\beta_1 + \beta_2 + \beta_3 + \beta_4]$	(0.006)	(0.006)	(0.007)	(0.007)	(0.001)	(0.001)	(0.001)	(0.001)
linear projection$_1$			-0.043***	-0.037**			0.003	0.008***
$[\lambda_1]$			(0.016)	(0.016)			(0.002)	(0.002)
linear projection$_2$			-0.020***	-0.023***			0.001	0.003***
$[\lambda_1 + \lambda_2]$			(0.007)	(0.007)			(0.001)	(0.001)
linear projection$_3$			-0.019***	-0.015***			0.001**	0.003***
$[\lambda_1 + \lambda_2 + \lambda_3]$			(0.006)	(0.005)			(0.001)	(0.001)
linear projection$_4$			-0.019***	-0.016***			0.001***	0.003***
$[\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4]$			(0.005)	(0.004)			(0.000)	(0.000)
p-value Joint Test	–		0.0483		–		0.0282	
p-value Slope Test	–		0.0911		–		0.111	
p-value Lin. Proj. Test	–		0.184		–		0.0912	
p-value slope$_{11}$=slope$_{21}$	–		0.255		–		0.815	
p-value slope$_{12}$=slope$_{22}$	–		0.0168		–		0.187	
p-value slope$_{13}$=slope$_{23}$	–		0.00618		–		0.0837	
p-value slope$_{14}$=slope$_{24}$	–		0.00713		–		0.0659	
p-value Proj$_{11}$=Proj$_{21}$	–		0.787		–		0.0348	
p-value Proj$_{12}$=Proj$_{22}$	–		0.714		–		0.0172	
p-value Proj$_{13}$=Proj$_{23}$	–		0.628		–		0.0601	
p-value Proj$_{14}$=Proj$_{24}$	–		0.651		–		0.0363	
Number of Pairs	9660		9660		10827		10827	

Notes: Data are from the Collaborative Perinatal Project (CPP). Entries for slope and linear projection are the slope coefficients within each birth weight segment from the piecewise linear spline specification. Birth weight is measured in 100g. The segments are: [0,16); [16,25); [25,30) and ≥ 30. p-value slope$_{11}$=slope$_{21}$ shows the p-value of testing the equality of the slope entry between first and second born in the first segment. p-value Proj$_{11}$=Proj$_{21}$ shows the p-value of testing the equality of the partial correlation entry between first and second born in the first segment. p-value Slope Test shows the p-value of jointly testing for the equality of slopes at all segments. p-value Lin. Proj. Test shows the p-value of jointly testing for the equality of partial correlations at all segments. p-value Joint Test tests the equality of slopes and partial correlation at all segments jointly. Controls include gender, maternal age, SES at registration, gestational length, and the maximum number of cigarettes smoked consistently during pregnancy. Heteroskedasticity-robust standard errors are reported in parenthesis; * denotes 90% significance, ** denotes 95% significance, and *** denotes 99% significance.

Table 15: Mental Skills (CPP)

	8 Months (Bayley Scales IQ Test)				4 Years (Stanford-Binet IQ Test)			
	OLS		CRE		OLS		CRE	
	(1) 1^{st}-Born	(2) 2^{nd}-Born	(3) 1^{st}-Born	(4) 2^{nd}-Born	(5) 1^{st}-Born	(6) 2^{nd}-Born	(7) 1^{st}-Born	(8) 2^{nd}-Born
slope$_1$	-0.874***	-0.825***	-0.394	-0.427	1.168***	0.003	0.092	-0.163
[β_1]	(0.318)	(0.265)	(0.458)	(0.409)	(0.381)	(0.391)	(0.379)	(0.429)
slope$_2$	0.295*	0.304**	0.589***	0.420**	0.511***	0.353**	0.178	0.277
[$\beta_1 + \beta_2$]	(0.164)	(0.152)	(0.213)	(0.203)	(0.159)	(0.153)	(0.177)	(0.171)
slope$_3$	0.492***	0.550***	0.778***	0.725***	0.538***	0.467***	0.196	0.423***
[$\beta_1 + \beta_2 + \beta_3$]	(0.129)	(0.120)	(0.168)	(0.161)	(0.126)	(0.120)	(0.141)	(0.134)
slope$_4$	0.523***	0.592***	0.761***	0.741***	0.585***	0.531***	0.248**	0.412***
[$\beta_1 + \beta_2 + \beta_3 + \beta_4$]	(0.104)	(0.095)	(0.138)	(0.130)	(0.100)	(0.095)	(0.115)	(0.108)
linear projection$_1$			-0.472	-0.308			0.887**	-0.110
[λ_1]			(0.360)	(0.372)			(0.375)	(0.349)
linear projection$_2$			-0.268*	-0.048			0.245	-0.100
[$\lambda_1 + \lambda_2$]			(0.163)	(0.166)			(0.160)	(0.158)
linear projection$_3$			-0.257**	-0.118			0.246*	-0.101
[$\lambda_1 + \lambda_2 + \lambda_3$]			(0.130)	(0.131)			(0.127)	(0.124)
linear projection$_4$			-0.205*	-0.096			0.248**	-0.027
[$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$]			(0.106)	(0.107)			(0.104)	(0.101)
p-value Joint Test	–		0.915		–		0.0747	
p-value Slope Test	–		0.612		–		0.0688	
p-value Lin. Proj. Test	–		0.869		–		0.282	
p-value slope$_{11}$=slope$_{21}$	–		0.953		–		0.630	
p-value slope$_{12}$=slope$_{22}$	–		0.503		–		0.649	
p-value slope$_{13}$=slope$_{23}$	–		0.793		–		0.190	
p-value slope$_{14}$=slope$_{24}$	–		0.900		–		0.216	
p-value Proj$_{11}$=Proj$_{21}$	–		0.753		–		0.0572	
p-value Proj$_{12}$=Proj$_{22}$	–		0.348		–		0.143	
p-value Proj$_{13}$=Proj$_{23}$	–		0.457		–		0.0646	
p-value Proj$_{14}$=Proj$_{24}$	–		0.476		–		0.0775	
Number of Pairs	8094		8094		6954		6954	

Notes: See notes to Table 14.

Table 16: Mental Skills continued (CPP)

	7 Years (Weschler IQ Test)			
	OLS		CRE	
	(1) 1^{st}-Born	(2) 2^{nd}-Born	(3) 1^{st}-Born	(4) 2^{nd}-Born
slope$_1$	0.640*	-0.168	0.188	-0.073
$[\beta_1]$	(0.345)	(0.310)	(0.360)	(0.300)
slope$_2$	0.402***	0.272*	0.299*	0.305*
$[\beta_1 + \beta_2]$	(0.150)	(0.139)	(0.165)	(0.159)
slope$_3$	0.469***	0.312***	0.307**	0.334***
$[\beta_1 + \beta_2 + \beta_3]$	(0.119)	(0.110)	(0.131)	(0.125)
slope$_4$	0.483***	0.413***	0.333***	0.413***
$[\beta_1 + \beta_2 + \beta_3 + \beta_4]$	(0.096)	(0.087)	(0.108)	(0.102)
linear projection$_1$			0.375	-0.200
$[\lambda_1]$			(0.348)	(0.299)
linear projection$_2$			0.090	-0.112
$[\lambda_1 + \lambda_2]$			(0.149)	(0.148)
linear projection$_3$			0.123	-0.090
$[\lambda_1 + \lambda_2 + \lambda_3]$			(0.119)	(0.117)
linear projection$_4$			0.113	-0.071
$[\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4]$			(0.098)	(0.096)
p-value Joint Test	–		0.130	
p-value Slope Test	–		0.0675	
p-value Lin. Proj. Test	–		0.621	
p-value slope$_{11}$=slope$_{21}$	–		0.554	
p-value slope$_{12}$=slope$_{22}$	–		0.975	
p-value slope$_{13}$=slope$_{23}$	–		0.868	
p-value slope$_{14}$=slope$_{24}$	–		0.521	
p-value Proj$_{11}$=Proj$_{21}$	–		0.227	
p-value Proj$_{12}$=Proj$_{22}$	–		0.366	
p-value Proj$_{13}$=Proj$_{23}$	–		0.230	
p-value Proj$_{14}$=Proj$_{24}$	–		0.212	
Number of Pairs	7986		7986	

Notes: See notes to Table 14.

Figure 1: Simulated p-values