Stockholms universitet

Research Report

Department of Statistics

No. 2010:4

Adjusting for Selection Bias in the Relationship between Sibship Size and Cognitive Performance

Gebrenegus Ghilagaber
Linda Wänström

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Adjusting for Selection Bias in Assessing the Relationship between Sibship Size and Cognitive Performance

Gebrenegus Ghilagaber* and **Linda Wänström***

Abstract

Consistent negative correlations between sibship size and cognitive performance (as measured by IQ and other mental aptitude tests) have been observed in past empirical studies. However, the issue of potential selection process in the decision to have larger families (larger sibship size) has been partly neglected in past single- and multilevel investigations. The present work extends existing knowledge in three aspects: (1) as factors affecting decision to increase family size may vary across the number and composition of current family size, we propose a sequential probit model (as opposed to binary or ordered models) for the propensity to increase sibship size; (2) we investigate if families who choose to increase family size are a representative random sample of the population of families or there exists selection; (3) in order to disentangle selection and causality we propose a multilevel multiprocess modelling where a continuous model for performance is estimated jointly with a sequential probit model for family-size decisions. The issues are illustrated through analyses of scores on PIAT tests among children of the NLSY79. We find substiantial between-family heterogeneity in the propensity to increase family size - thereby providing empirical evidence in support of the admixture hypothesis. Ignoring such adverse selection led to overestimation of the negative effects of sibship size on cognitive performance but our multiprocess modelling could mitigate the biasing effects of selection.

*Department of Statistics, Stockholm University, SE-106 91 Stockholm - Sweden. (Gebre@stat.su.se or Linda.Wanstrom@stat.su.se). The work was financed partially through grants by the Faculty of Social Sciences, Stockholm University.

1 Introduction

Studies have consistently found negative correlations between children's sibship size and outcome variables such as their intelligence test scores, achievement test scores, or educational attainment (e.g. Anastasi, 1956; Higgins, Reed, & Reed, 1962; Belmont & Marolla, 1973; Jæger, 2008; 2009; Nisbet & Entwistle, 1967; Page & Grandon, 1979; Velandia, Grandon, & Page, 1978; Zajonc & Markus, 1975; Zajonc, Markus, Berbaum, Bargh & Moreland, 1991).

These findings have resulted in theories such as the confluence model (Zajonc & Markus, 1975; Zajonc, 2001) and the resource dilution theory (Blake, 1981; Downey, 1995; 2001; Armor, 2001). In brief, the confluence model states that the intellectual levels in our families influence our own intellectual levels, and this level will be lower in families with more children. In addition, those of us who have younger siblings benefit from a "tutoring" function, i.e. our intellectual levels are positively influenced from us teaching our younger siblings. Thus, we are both affected by the intellectual climate of the home, which in turn is affected by the number of siblings, and by our birth orders. The resource dilution theory explains the negative correlations by a dilution of parental resources that occurs in families with more children. Children in smaller families get more attention, more help with homework etc. than children in larger families.

However, parents that have large families may differ from parents that have small families (e.g. Page & Grandon, 1979; Velandia, Grandon, & Page, 1978; Rodgers et al., 2000; Rodgers, 2001), and the variables influencing decisions to have large families may also influence children's cognitive performance. For example, parents with a lower level of education may decide to have larger families whereas families with a higer level of education may decide to have smaller families because they need to focus more on work-related issues. On the other hand, parents with a higher level of education may decide to have larger families because they have better paying jobs and can afford to spend more money on children. Parental education, household income and related variables, may also affect childrens' cognitive performance. If researchers fail to account for these types of variables in looking at the relationship between children's sibship size and cognitive performance, the results may be invalid. Most previous studies of this relationships have acknowledged this and included what has appeared to be relevant variables into statistical models. All relevant variables may not be known to the researcher,

however, resulting in misspecified models and biased estimates of effects.

Most data in the social sciences have a hierarchical or clustered structure. For example, the analysis of child data involves a natural hierarchy where children are grouped within mothers or families, classes, etc., and the latter, in turn, are grouped into households, communities, or schools. Children in the same family tend to be more alike in their characteristics than children chosen at random from the population at large. Ignoring this grouping risks overlooking the importance of group effects, and may also render invalid many of the traditional statistical analysis techniques used for studying data relationships. This is relevant in studying the relationship between child sibship size and cognitive performance as data often consist of multiple children from the same family and this clustered structure should be accounted for in the models.

In the present study, we address the relationship between cognitive performance on the one hand, and child- and family-specific background variables on the other, using data on children of the NLSY79 (National Longitudinal Study of Youth, 1979). In contrast to traditional approaches where the collection of children is assumed to be an independent random sample, we treat children of the same mother as correlated cases (multilevels) within the same observation (mother). Our formulation also enables us to allow for unobserved family-specific heterogeneity (shared-frailty) in the models. This, in turn, enables us to address an important but partly neglected issue - **selection bias**. Since there are no randomized trials of couple's decisions on their family-size, it is difficult to assess the impact of such size on children's cognitive ability without accounting for selection processes in the decision on family size. Thus, we will examine the biasing effects of selection on estimates of sibship size by using multiprocess models in order to model the outcome variable and endogenous explanatory variable(s) simultaneously.

The paper is organized as follows. We begin, in Section 2, by describing the data used in the illustration together with some summary statistics across the variables used as correlates of congitive performance. In Section 3, we first fit standard models to get a preliminary idea about correlates of cognitive performance by treating children as independent observations. We then build on the standard models and treat children of the same mother as nested cases (multilevels) within one observation (mother) and fit a multilevel model of perfomance that allows for mother-specific unobserved heterogeneity. In Section 4, we explore for possible selection processes in the decision to have larger family size (sibship size) by fitting binary-, and sequential pro-

bit models of decision. Section 5 is devoted to joint modelling of cognitive performance and family size decisions and, thereby, adjusting for selection bias. Section 6 ties up the findings of the paper in the form of summary and concluding remarks.

2 Data and variables

The NLSY79 Child and Young Adult Data originated as a multi-stage stratified area probability sample of 12,686 males and females in 1979 in the U.S. Starting in 1986, the biological children of the 6283 females were also interviewed, and they have been interviewed every second year since then. Some of the assessments administered to the children include the PPVT, the PIAT tests, and the Digit Span test. These data are suitable for our analyses because they contain the outcome of cognitive assessment as well as vast background information about the 11,428 children and their families. The present study is based on 6430 children who were assessed as a 5- or 6 year old in 1986, 1988, 1990, ..., or 2006. Thus, a child who was assessed as a 5- or 6-year old in 1986 will contribute with his/her 1986 score, whereas his/her brother or sister who was assessed as a five- or six-year old in, say, 1992 will contribute with his/her 1992 score.

The response variable, cognitive performance, is a subset of the Peabody Individual Achievement Test (PIAT) assessment. The PIAT subtests were administered to children ages five to 14 biannually starting in 1986. They measure academic achievement and have high test-retest reliability and concurrent validity. They have also been found to be predicted by and to predict other assessment tests. The Mathematics subtest, used in our study, consists of 85 multiple-choice questions of increasing difficulty, assessing skills such as numerals recognition, geometry, and trigonometry. The child answers each problem by pointing to or naming one of four options.The response variable will from now on be referred to as Score

Five mother-specific and six child-specific variables were used as explanatory variables. The mother-specific variables are Mother's IQ (Standardized score of the Armed Forces Qualifications Test in 1979), Mother's Race with 3 levels (Black, Hispanic, Other), Mother's age at birth of first child, Household Income (standardized score), and Mother's Level of Education with 2 leves (No College, Some College). Household income and Educational level refer to the values attained in 2006 or, in the absence of a value in 2006, to

the closest year of measurement with available value. These family-specific variables take the same value across children of the same mother (siblings) even though siblings have measurements in different years.

The child-specific variables are Year of measurement (1 for 1986, 2 for 1988, ..., 11 for 2006), Birth order, Sibship size - the number of biological children born to the mother at time of measurement (it is also referred to as family size and is the main explanatory variable of interest), Mother's marital status at time of measurement with 3 levels (Married, Never married, Other), Father's presence in the household (Yes, No), Sex (Male, Female). These child-specific variables may differ between siblings because siblings have often been measured (when they are five or six years old) at different times. Summary statistics of these variables is shown in Table 1.

The above variables are among those considered to be correlated with cognitive performance in previous analyses of the same data set (Wänström, 2007; Wichman, Rodgers and MacCallum, 2006) or other data sets. Based on previous studies (Wänström, 2007; Rodgers and Wänström, 2006; Wichman, Rodgers and MacCallum, 2006), mother's IQ, mother's age at birth of first child, household income, and mother's education are expected to have positive effects on cognitive performance. On the other hand, we expect negative effects of sibship size and birth order. Further, children with married mothers as well as non-Black, non-Hispanic children, and children with present fathers are also expected to have higher cognitive performance scores than children with non-married mothers, or Black or Hispanic children. The children of the NLSY79 show increasing scores over time (Rodgers & Wanstrom, 2006), a phenomenon referred to as the Flynn effect (Flynn, 1984; 1987) and the year of measurement is thus expected to have positive effects on cognitive performance.

3 Modelling Correlates of Cognitive Performance

3.1 Standard (single-level) linear regression on log-Score

An extract from the data set is shown is Table 2 below. The mean score by sibship size (plotted in Figure 1) shows a linearly decreasing relationship.

Table 1: Summary statistics across mother- and child-specific variables

```
Number of observations:    3340
Maximum number of level 2 branches in any observation:      8

-------------------------------------------------------------

LEVEL 1 VARIABLES:
Variable     N       Mean    Std Dev        Min        Max
_id       3340     1670.5   964.3193        1.0     3340.0
IQ        3340   -.003457   .9849737    -1.3919   2.144257
Race      3340   2.346707   .7697706        1.0        3.0
Agefirst  3340   22.71437   5.230692       13.0       41.0
Income    3340   -0.02885   .9542863   -.873624    10.2515
Educ      3340   .4404192   .4965118        0.0        1.0

LEVEL 2 VARIABLES:
Variable     N       Mean    Std Dev        Min        Max
Score     6430   99.88927   14.00594       65.0      135.0
Year      6430   4.524261   2.604294        1.0       11.0
Order     6430   2.025039   1.133558        1.0       10.0
Sibsize   6430   2.639191   1.245113        1.0       11.0
Marstat   6430   2.056921   .5925297        1.0        3.0
Father    6430   .6430793   .4791284        0.0        1.0
Sex       6430   1.491602   .4999683        1.0        2.0

-------------------------------------------------------------
```

Suppose we fit a standard linear regression model on log-Score:

$$\log\left(Score_j\right) = \beta' X_j + \varepsilon_j \tag{1}$$

where j ($j = 1, 2, ..., 6430$) indicates index child, X_j is a vector of mother- and child-specific variables; and $\varepsilon_j \tilde{} N(0, \sigma_\varepsilon^2)$ represents child-specific residuals. Results of fitting such model are shown in Table 3:

Table 2: An extract from the data set (the first 11 and the last 2 children)

Child	Mother-specific variables					Child-specific variables						
	IQ	Race	Age	Inc	Ed	Score	Yr	Ord	Sib	Mar	Fath	Sex
1	-1.15	3	34	-0.64	0	90	7	1	2	2	1	2
2	-1.15	3	34	-0.64	0	103	8	2	2	2	1	2
3	0.36	3	19	-0.34	1	98	4	3	3	2	1	2
4	0.16	3	17	-0.40	1	125	1	2	3	2	1	1
5	0.16	3	17	-0.40	1	107	2	3	3	3	0	2
6	-0.33	3	22	-0.17	0	103	2	1	2	2	1	2
7	-0.33	3	22	-0.17	0	106	5	2	2	2	1	2
8	1.78	3	30	0.32	1	106	5	1	2	2	1	1
9	0.39	3	31	0.88	1	110	6	1	2	2	1	1
10	0.39	3	31	0.88	1	97	8	2	3	2	1	1
11	0.39	3	31	0.88	1	112	9	3	3	2	1	2
...	...	...	...	...	...	...	...	...	...	...	...	...
6429	-0.44	1	21	0.49	0	114	1	1	2	2	1	1
6430	-0.44	1	21	0.49	0	69	4	3	3	3	0	2

Figure 1: PIAT Mean Scores by SibSize

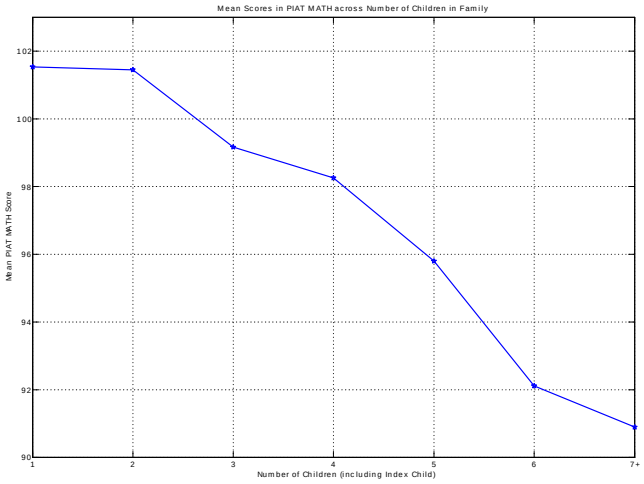

Table 3: Estimates of effects on log(Score) in a standard linear model

Parameter	**Estimate**	**z-value**
IQ	.033	15.769
Hisp	-.0276	-6.3645
Black	-.023	-5.6030
Age	.00168	3.2155
Inc	.009	5.6203
No Coll	-.0078	-2.2096
Year	.00226	2.3557
Order	-.00053	-0.1892
Sibsize	**-.0104**	**-5.2011**
Married	.0139	2.3041
Other	0.00625	1.2298
Father	.0079	1.4603
Girls	.0184	5.6160
σ_ϵ	.1302	118.0222

The effects are in the expected direction - positive and significant effects of Mother's IQ, Mother's age at first birth, Household Income, Year of measurement, Married women and Girls; and negative and significant effects of Sibship size, and children of Hispanics, Blacks, and those with lower education. Note that all estimates are based on the Maximum-Likelihood (ML) method and the z-values (including for the estimate of standard error) are results of asymptotic properties of ML-estimates.

3.2 A multilevel model on log(Score) with family-specific heterogeneity

Consider the data structure again - now with one more dimension (mother ID) shown in the first column of Table 4. We note that child 1 and 2 are siblings as are child 4 and 5. Child 6 and 7 also have the same mother while Mother 6 contributes 3 children (child 9, 10, and 11), etc...

Table 4: An extract from the data set (the first 11 and the last 2 children) with Mother ID

Moth	Child	Mother-specific variables					Child-specific variables						
		IQ	Race	Age	Inc	Ed	Score	Yr	Ord	Sib	Mar	Fath	Sex
1	1	-1.15	3	34	-0.64	0	90	7	1	2	2	1	2
1	2	-1.15	3	34	-0.64	0	103	8	2	2	2	1	2
2	3	0.36	3	19	-0.34	1	98	4	3	3	2	1	2
3	4	0.16	3	17	-0.40	1	125	1	2	3	2	1	1
3	5	0.16	3	17	-0.40	1	107	2	3	3	3	0	2
4	6	-0.33	3	22	-0.17	0	103	2	1	2	2	1	2
4	7	-0.33	3	22	-0.17	0	106	5	2	2	2	1	2
5	8	1.78	3	30	0.32	1	106	5	1	2	2	1	1
6	9	0.39	3	31	0.88	1	110	6	1	2	2	1	1
6	10	0.39	3	31	0.88	1	97	8	2	3	2	1	1
6	11	0.39	3	31	0.88	1	112	9	3	3	2	1	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3340	6429	-0.44	1	21	0.49	0	114	1	1	2	2	1	1
3340	6430	-0.44	1	21	0.49	0	69	4	3	3	3	0	2

A legitimate question that arises now is if our standard (single-level) model is appropriate. In other words, is it appropriate to ignore the hierarchical nature of the data and what is the price of ignoring it? We recall that the variance of the sum of two variables is the sum of their individual variances plus twice their covariance. Children of the same mother are expected to be positively correlated. Thus, if we treat children of the same mother as independent - as in the single-level model (1), we ignore their positive covariance and, thereby, underestimate the variance (standard error). Such underestimated standard error will, in turn, lead to spurious significance, as the z-values are often computed as ratios of estimated values to their standard errors. This problem and its alternative solution (multilevel modelling) is well known in the litterature.

For the jth child in the ith family a two-level model of performance with a family-specific heterogeneity-term is given by

$$\log\left(Score_{ij}\right) = \beta' X_{ij} + \tau_i + \varepsilon_{ij} \tag{2}$$

where i indicates mother (family), j indicates index child, X_{ij} is a vector of child- and mother-specific variables; and $\varepsilon_{ij} \tilde{} N(0, \sigma_\varepsilon^2)$ represents child-specific

residuals. This is a continuous two-level model with family/mothers as the experimental units and children as repeated outcomes (multilevels) within observations. Because children from the same family are likely to be more correlated than children from the population at large, τ_i captures family-level unobserved heterogeneity (shared-frailty) that may affect children's scores, τ_i ~ $N(0, \sigma_\tau^2)$. Models of this type (or their variants) have been used in a number of previous studies - among others Wichman, Rodgers and MacCallum (2006) and Wänström (2007). Results of fitting such model to our data set are shown in Table 5:

Table 5: Estimates of effects on log(Score) in a multilevel standard linear model

Parameter	**Estimate**	**z-value**
IQ	.0333	**13.0710**
Hisp	-0.0267	**-5.0405**
Black	-.0226	**-4.5052**
Age	.0015	**2.5925**
Inc	.009	**4.4520**
No coll	-.0074	-1.7205
Year	.0025	**2.4435**
Order	-.0014	-0.5262
Sibsize	**-.0096**	**-4.4118**
Married	.01334	1.9917
Other	0.0068	1.1668
Father	.008	1.5045
Girls	.0176	**5.4156**
σ_ε	**0.119457**	88.8254
σ_τ	**0.05179**	18.3733

We see the same pattern as in Table 3, but the size of estimated effects (specially the mother-specific variables) is now deflated - for some even changing the level of significance. Why? Consider the variances in the two tables. We note (where 1L and 2L stand for single level and multilevel models, respectively):

$$\begin{aligned} \sigma^2_{\varepsilon(1L)} &= (0.1302)^2 = 0.016952 \\ \sigma^2_{\varepsilon(2L)} &= (0.119457)^2 = 0.014270 \\ \sigma^2_{\tau(2L)} &= (0.05179)^2 = 0.0026822 \end{aligned}$$

Thus,

$$\begin{aligned} 0.014270 + 0.0026822 &= 0.016952 \\ \sigma^2_{\varepsilon(2L)} + \sigma^2_{\tau(2L)} &= \sigma^2_{\varepsilon(1L)} \end{aligned}$$

and, hence, the **Intra-class (Intra-mother) correlation coefficient (ICC)** is given by:

$$ICC = \frac{\sigma^2_{\tau(2L)}}{\sigma^2_{\varepsilon(2L)} + \sigma^2_{\tau(2L)}} = \frac{0.0026822}{0.014270 + 0.0026822} = 0.16$$

implying that 16% of the total variance (in the single-level model) comes from **variability between mothers**. If we igniore this between-mother correlation the result is that our effect-estimates will be inflated as shown in Table 3 (compared to Table 5).

4 Exploring for selection in the process of family-size decisions

Our inferences above about the possible effect of sibship size on cognitive performance are based on observational study where the assignment of children into small versus large families is outside the control of the investigator. Since there are no randomized trials on decisions to have small or large sibship size, it is difficult to assess its impact on cognitive ability without controlling for selection processes that can inhibit or promote decisions on larger sibship size. For more details on selection bias see, for instance, Heckman (1979), Heckman and Singer (1984), Winship and Mare (1992), Frick & Lantz (1996), Vella (1998), Ghilagaber (2004), Yoo and Frick (2006).

4.1 A binary probit model for sibship size: with family specific heterogeneity

In the log(Score) model above, one of the explanatory variables in the X-vector is sibship size. This is the number of children in the family (including the index child) and ranges between 1 and 11 in our data set. Let's begin

with dichotomizing sibship size into 'small' and 'large':

$$P_{ij} = \begin{cases} 0, & \text{if } j^{th} \text{ child in } i^{th} \text{ family belongs to a small family} \\ \\ 1, & \text{if } j^{th} \text{ child in } i^{th} \text{ family belongs to a large family} \end{cases}$$

The propensity (for family i of index child j) to have a large family may then be modelled as a binary probit model:

$$P_{ij}^{*} = \alpha' X_{ij}^{*} + \upsilon_i \tag{3}$$

where P_{ij}^{*} represents the propensity of having a large family size, X_{ij}^{*} are family specific explanatory variables (that have same value for all children within the same family); and υ_i captures family level unobserved heterogeneity (shared-frailty) that may affect the decision on family size, $\upsilon_i \tilde{} N(0, \sigma_{\upsilon}^2)$.

Table 6: Estimates of effects on propensity to have a large family

	large family is defined as:			
	$>$ 2 children	$>$ 3 children	$>$ 4 children	$>$ 5 children
IQ	0.050	-0.023	0.16**	-0.220
Hisp	-0.0014	0.21	0.67***	0.450**
Black	-0.34**	0.059	0.51***	0.017
Age	-0.14***	-0.20***	-0.25***	-0.28***
Inc	0.39***	0.25***	0.14**	-0.033
No Coll	-0.074	-0.10	0.056	0.0087
σ_{υ}	1.70***	1.88***	1.92***	2.17***

with *, **, and ***, indicating significance at 5%, 1%, and 0.1% levels, respectively.

While the effects of family-specific variables vary depending on how we define 'large family', we note that there are sociodemographic differentials in the decision to have a large family. As would be expected, women who get their first child late are less likely to have large families, households with higher incomes are more likely to have larger families, while Blacks and Hispanics tend to have larger families (when large family is defined as having more than 4 children including the index child).

4.2 A sequential probit model for the propensity to increase sibship size

We note that dichotomization into 'small' and 'large' family is too subjective and that effects vary depending on how it is defined. Since the actual family–size-decision process requires successful completion of the prior level (parity) for passage into the subsequent one, a sequential decision model accurately reflects the real decision process (Yamaguchi & Ferguson, 1995; Upchurch, Lillard, & Panis, 2002). The model of family-size-decision used in this paper specifies the propensity of progressing to successively higher parity levels, conditional on having completed the next lower parity – a discrete sequential choice model. Apart from measured covariates, the sequential probabilities may depend on individual and decision-varying covariates and unobserved heterogeneity in the propensity to continue to the next parity. We group the ten possible transitions into four:

- transition from 1 child to 2 children
- transition from 2 children to 3 or 4 children
- transition from 4 children to 5 or 6 children
- transition from 6 children to 7 children or more

As the reasons to have a 2^{nd} child may differ from those to have a 3rd and 4th child, etc..., we allow the effects on the transition propensities to vary between these four transitions.

Thus, there are up to four sequential choices of whether to continue to the next level ($s = 1, 2, \ldots, 4$), each conditional on having continued to the previous level. Here, $s = 1$ corresponds to transition from 1 to 2 children, $s = 2$ corresponds to transition from 2 to 3-4 children, $s = 3$ corresponds to transition from 4 to 5-6 children, and lastly $s = 4$ corresponds to transition from 6 to 7 or more children.

We use a multilevel sequential probit model of individual-family (mother) choice. Family j progresses from having completed parity s to complete the next parity $s + 1$ if its propensity to continue is positive, $I_s > 0$. The propensity of mother (family) i progressing is thus determined by the probit index function

$$I_{i_{(s)}} = \alpha_{0s} + \alpha'_{1s} X_{i(s)} + \upsilon_i + \theta_s \tag{4}$$

for $s = 1, 2, 3, 4$, where X_s is a vector of exogenous covariates affecting sibship-size decisions, α_{0s} and α_{1s} are decision-specific intercepts and coefficients, respectively, υ_i is a residual term capturing family level unobserved heterogeneity (shared-frailty) that may affect all levels of decision on family size, and θ_s is a decision specific stochastic element (normalized to $\theta_s = 1$, for all s). Each is assumed to be normally distributed:

$$\upsilon \tilde{} N\left(0, \sigma_\upsilon^2\right), \quad \theta_s \tilde{} N\left(0, 1\right)$$

The model also allows parameters to vary across decisions (hence the s subscript on the parameter vector α). In other words, we will estimate four intercepts and four sets of coefficient estimates, one set for transition to 2^{nd} child, another set for transition from 2^{nd} to 3^{rd} and 4^{th} child, a third set for transition from 4^{th} to 5^{th} and 6^{th} child, and a fourth set for transition from 6^{th} to 7^{th} child and above.

Table 7: Estimates of effects propensity of transition to the next group of family-size

	Constant effect		Varying effects on transition to:			
			$\mathbf{2^{nd}}$	$\mathbf{3^{rd} - 4^{th}}$	$\mathbf{5^{th} - 6^{th}}$	$>= 7$
IQ	.004	-.004	0.017	0.044	-.0101	-0.034
Hisp	.082**	.037	0.047	0.23***	.33**	-0.17
Black	.015	-.020	-0.21**	0.11	.066	-0.72*
Age	-.037***	-.035***	-0.085***	-0.12***	-.15***	-0.088
Inc	.034**	.040	0.26***	0.11***	.028	-0.42
No Coll	.0081	-.028	-.070	-.040	.15	.22
σ_υ	-	0.85***	.85***			

Columns 2 and 3 in Table 7 report results from models (without and with heterogeneity, respectively) assuming that the effects of variables are constant across all levels of decision. Accordingly, we note that Hispanics and households with higher income tend to have larger families while women who begin child birth late are less likely to end up with large families. We note, further, that there is an unignorable between-mother heterogeneity in the propensity to have a large family. In columns 4-7, we allow the effects to vary across decision levels and we observe that they do vary. For instance, differentials due to household incomes show only until transition to 3^{rd} and 4^{th} child while higher age at first birth continues, as would be expected, to be an inhibiting factor for larger families.

5 Adjusting for selection bias with a multiprocess model

We have now estimated models for cognitive performance (2) and for sibship-size decision (a sequential variant of Eq. (3)). In both cases, we found evidence of unobserved family specific characteristics that affect children's performance and family-size decisions. If families of children with predominantly below-average cognitive ability have higher propensity to increase their family size, the result will be that the large values of the regressor sibship size in (2) consists of a disproportionately high mix of low performing children. If ignored, this **adverse selection** will inflate (overestimate) the negative effect of sibship size on cognitive ability. Conversely, the regressor sibship size in (2) may consist of a disproportionately high mix of children with above average cognitive ability if selection is **favorable** in the sense that it is the families with well performing children that have a higher propensity of increasing their family size. These may include more wealthy families with extra resources to enhance children's learning and who can afford larger family sizes. In this later type of selection, ignoring the favorable selection will deflate (underestimate) the negative effect of sibship size on cognitive performance.

The above limitations prompt us to address the potential endogeneity of sibship size and estimate a joint model (multiprocess model) of cognitve ability and sibship size decisions. The joint model consists of two sets of equations:

- a linear model for cognitive performance,

$$Score_{ij} = \beta' X_{ij} + \tau_i + \varepsilon_{ij} \tag{5}$$

and

- a sequential probit model for sibship size (a sequential variant of Eq. (3)),

$$P^*_{ij} = \alpha' X^*_{ij} + \upsilon_i \tag{6}$$

If the correlation between τ and υ is nonzero, estimation of equation (5) without regard to equation (6) will be biased. Thus, the main issue addressed

here is that we wish to allow for the possibility that unobserved family specific characteristics affect both the child's cognitive performance and decisions on family size. In other words, we wish to allow for a correlation between τ and υ:

$$\begin{pmatrix} \tau \\ \upsilon \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\tau^2 & \\ \sigma_{\tau\upsilon} & \sigma_\upsilon^2 \end{pmatrix}\right] \tag{7}$$

The bias due to selection effects is mitigated by making the source of the bias (the correlation) part of the model. In our present case, the effect of sibship size on cognitive performance will be biased because of non random decisions on sibship-size.

The joint likelihood of the continuous model for log(Score) and sequential probit model outcomes may be separated into a continuous and a probit part, where the probit residual υ becomes conditional on the realized value of τ, and thus on the continuous outcome:

$$L^{(CP)} = L_1^{(C)} L_2^{(P)}, \tag{8}$$

where (with $y = \log(\text{Score})$),

$$L_1^{(C)} = \frac{1}{\sigma_\tau \sqrt{2\pi}} \exp\left\{-\frac{(y - \beta' X)^2}{2\sigma_\tau^2}\right\} \tag{9}$$

$$L_2^{(P)} = \begin{cases} 1 - \Phi\left(\frac{\alpha' X^* + \mu_{\upsilon|\tau}}{\sigma_{\upsilon|\tau}}\right), & \text{if } P = 0 \ (\Longrightarrow P_j^* < 0) \\ \Phi\left(\frac{\alpha' X^* + \mu_{\upsilon|\tau}}{\sigma_{\upsilon|\tau}}\right), & \text{if } P = 1 \ (\Longrightarrow P_j^* > 0) \end{cases}$$

where (τ, υ) is bivariate normal:

$$\begin{pmatrix} \tau \\ \upsilon \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\tau^2 & \\ \sigma_{\tau\upsilon} & \sigma_\upsilon^2 \end{pmatrix}\right]$$

so that

$$\upsilon|\tau \sim N\left(\frac{\sigma_{\tau\upsilon}}{\sigma_\tau^2}(y - \beta' X), \quad \sigma_\upsilon^2 - \frac{\sigma_{\tau\upsilon}^2}{\sigma_\tau^2}\right)$$

or, equivalently,

$$\upsilon|\tau \sim N\left[\rho_{\tau\upsilon}\frac{\sigma_\upsilon}{\sigma_\tau}(y - \beta' X), \quad \sigma_\upsilon^2\left(1 - \rho_{\tau\upsilon}^2\right)\right]. \tag{10}$$

We fit this multiprocess model to our data with results as shown in Table 8. Comparing the results in Table 8 with those in Table 5, we now note that the effect of sibship size on log-score is reduced from a highly significant effect -0.0096 (z-value = -4.41) in Table 5 to an insignificant (or marginally significant) effect $\mathbf{-0.0064}$ **(with z-value** $\mathbf{= -1.94}$**)** in Table 8**.** This is in accordance with the estimate of the correlation coefficient. We note that $\rho_{\tau\upsilon}$ (the correlation between family-specific unobserved heterogeneity terms that affect cognitive performance and family-size decisions) is estimated at $\rho_{\tau\upsilon} = -0.091$. This negative correlation, though statistically insignificant, indicates **adverse selection** in the sense that it is families to children with below-average performance that are more likley to increase family size. If ignored and the log(Score) model is estimated separately without due account of the probit model, this negative correlation pushes the effect of sibship negatively (to the left) from $\mathbf{-0.0064}$ **(with z-value** $\mathbf{= -1.94}$**)** to -0.0096 (with z-value = -4.41). In other words, the adverse effect of sibship size on cognitive ability will be inflated. To focus thoughts, consider the last row in Table 9 where we report results from a standard model and selection (multiprocess) model when we exclude the Ethnicity variable:

We recall from Table 5 that children of hispanics and blacks have lower performance than their baseline counterparts. We also recall from results in our probit model (Tables 6 and 7) that ethnicity is a potential source of selection as we found ethnic differentials in family-size decisions. Specifically, we found that hispanics are more likely to have larger families than their baseline counterparts. Thus, if ethnicity is excluded from the models and becomes part of the unobserved heterogeneity, the effect is that the negative correlation between the unobserved heterogeneity terms gets stronger $\rho_{\tau\delta} = -0.201$ (z-value = $\mathbf{-3.4306}$**).** More interestingly, the effect of ignoring such a strong negative correlation will be to overestimate the adverse effect of sibship size from an insignificant effect -0.00286 (z-value = $-$ 0.8912**)** to a highly significant effect -0.00998 (z-value = $-$ 4.5829**).**

Table 8: Estimates of effects on log(Score) in a multilevel multiprocess model

	est.	z-value
IQ	0.033	12.9953
Hisp	-0.027	-5.0522
Black	-0.023	-4.4607
Age	0.0019	2.8701
Inc	0.0090	4.2302
No Coll	-0.00745	-1.7183
Year	0.0022	2.0484
Order	-0.0017	-0.6309
Sibsize	**-0.0064**	-1.9463
Married	0.013	1.9821
Other	0.0068	1.1513
Father	0.0082	1.4517
Girls	0.018	5.3811
σ_ε	0.12	87.3452
σ_τ	0.052	18.0021
σ_υ	0.85	35.8924
$\rho_{\tau\upsilon}$	-0.091	-1.4561

Table 9: Changing effects of sibship size on log(Score) in three models

variables	Standard model		Selection model		Correlation ($\rho_{\tau\delta}$)	
	est.	z-value	est.	z-value	est.	z-value
All (no het)	-0.0104	-5.2011	$-$	$-$	$-$	$-$
All (het)	-0.00958	-4.4118	-0.0064	-1.9463	-0.0908	-1.4561
- Ethn	-0.00998	-4.5829	-0.00286	-0.8912	-0.201	-3.4306

6 Concluding Remarks

In the present paper we have used recently developed multilevel multiprocess modelling (Lillard and Panis, 2003) to address selection-bias in the relationship between sibship and size and cognitive performance. Our concern was that sibship-size decisions are not random but rather that there is a selection

in the process. If so, selection-bias can arise because the regressor "sibship size" will be correlated with the residual term in the equation.

Some previous theories in the area - like the **Confluence** and **Resource-dilution** theories focus within the family - invoking interactions of family members. Others have attempted to explain the patterns through the **admixture hypothesis** in the sense that the causal sources of the systematic patterns observed in their data were outside the family. For instance, Page & Grandon (1979) note:

> "The apparent effects of family size, far from explaining population differences, seem themselves to be better explained as the result of group **admixtures**".

Our empirical analysis based on data from the NLSY79 reveals an appreciable between-family variation in sibship-size decisions. In other words, our results provide empirical evidence in support of the **admixture hypothesis.** This, in turn, casts doubts on previous conlcusions that relied on within-family factors and processes in explaining sibship size - intelligence patterns. Overall, family-specific unobserved heterogeneity that affects sibship-size decision was negatively correlated to family-specific unobserved heterogeneity that affect cognitive ability. Ignoring such negative correlation (**adverse selection)** led to overestimation of the negative effects of sibship size on cognitive ability. Thus, we recommend mutilevel-multiprocess modelling to account for between-family variability, address selection bias and come up with adjusted estimates of sibship size effects on cognitive ability.

The multiprocess procedure we used estimates two equations (continuous and probit models) jointly and it allows to model that correlation (source of selection-bias) directly, thereby mitigating (if not eliminating) the bias. The tradeoff in the procedure we have used is that we have to assume the mother-specific heterogeneity terms in the two equations are jointly normally distributed. If the true model is known, then a proper model will **eliminate** selection bias. In real life, we don't know the precise nature of selection. We may theorize that there is unobserved mother-effect which is normally distributed. If the distribution of that effect is in fact something other than normal, the selection bias will be **reduced** but not necessarily eliminated. And if there is some other selection effect (say from siblings who share the same teacher), the resulting bias would remain. Thus, while the modeling approach we used is designed to certainly mitigate selection-biases, it may not be taken for granted that such biases are fully eliminated. A possible area for

future studies would, therefore, be a deeper examination on the validity of the distributional assumptions (and proposal of alternative distributions) as well as investigation of the robustness to violations of distributional assumptions of the procedure used here.

References

[1] Anastasi, A. (1956). Intelligence and family size. *Psychological Bulletin*, **53**, 187-209.

[2] Armor, D. J. (2001), On family size and intelligence. *American Psychologist* **55**, 521-522.

[3] Belmont, L., and Marolla, F. (1973). Birth order, family size, and intelligence. *Science*, **182**, 1096-1101.

[4] Blake, J. (1981), Family size and the quality of children. *Demography* **18**, 421-442.

[5] Downey, D. B. (1995), When bigger is not better: Family size, parental resources, and children's educational performance. *American Sociological Review* **60**, 747-761.

[6] Downey, D. B. (2001), Number of siblings and intellectual development. The resource dilution explanation. *American Psychologist* **56**, 497-504.

[7] Frick, K., and Lantz, P. (1996), Selection Bias in Prenatal Care Utilization: An Interdisciplinary Framework and Review of the Literature. *Medical Care Research and Review* 4(53), 371-396.

[8] Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, **95**, 29-51.

[9] Flynn, J. R. (1987). Massive Q gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, **101**, 171-191.

[10] Ghilagaber, G. (2004), Disentangling Selection and Causality in Assessing the Contribution of Health Inputs to Child Survival: Evidence from East Africa. *Research Report* **2004:7**, Department of Statistics, Stockholm University

[11] Heckman, J. J. (1979), Sample Selection Bias as a Specification Error. *Econometrica* 47: 153-161.

[12] Heckman, J. J., and Singer, B. (1984), A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica* 52(2): 271-320.

[13] Higgins, J. V., Reed, E. W. and Reed, S. C. (1962). Intelligence and family size: a paradox resolved. *Eugenics Quarterly*, **9**, 84-90.

[14] Jæger„ M. M. (2008), Do large sibships really lead to lower educational attainment? New evidence from quasi-experimental variation in couples' reproductive capacity. *Acta Sociologica* **51**, 217-235

[15] Jæger„ M. M. (2009), Sibship size and educational attainment. A joint test of the Confluence Model and the Resource Dilution Hypothesis. *Research in Social Stratification and Mobility* **27**, 1-12.

[16] Lillard, L. A., and Panis, C. W. A. (2003), *aML Multilevel Multiprocess Statistical Software, Version 2.0.* EconWare, Los Angeles, California.

[17] Nisbet, J. D., and Entwistle, N. J. (1967). Intelligence and family size, 1949-1965. *British Journal of Educational Psychology*, **37**, 188-193.

[18] Page, E. B., and Grandon, G. M. (1979). Family configuration and mental ability: two theories contrasted with U.S. data. *American Educational Research Journal*, **16**, 257-272.

[19] Rodgers, J. L (2001), What causes birth order – intelligence patterns? The admixture hypothesis, revived. *American Psychologist* **56**, 505-510.

[20] Rodgers, J. L., Cleveland, H. H., van den Oord, E., and Rowe, D. C. (2000). Resolving the debate over birth order, family size, and intelligence. *American Psychologist*, **55**, 599-612.

[21] Rodgers, J. L., and Wänström, L. (2007). Identification of a Flynn Effect in the NLSY: Moving from the center to the boundaries. *Intelligence*, **35**, 187-196.

[22] Upchurch, D. M., Lillard, L. A., and Panis, C. W. A.(2002), Nonmarital Childbearing: Influences of Education, Marriage, and Fertility. *Demography* **39**, 311-329.

[23] Vella, F. (1998), Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources* XXXII, 127-169.

[24] Velandia, W., Grandon, G. M., and Page, E. B. (1978). Family size, birth order, and intelligence in a large South American sample. *American Educational Research Journal*, **15**, 399-416.

[25] Wänström, L. (2007), Sibship Size and Cognitive Ability: Are Cognitive Abilities in Children Affected by the Birth of a Sibling? *Research Report* **2007:3**, Department of Statistics, Stockholm University

[26] Wichman, A. L., Rodgers, J. L., and MacCallum, R. C. (2006). A multilevel approach to the relationship between birth order and intelligence. *Personality and Social Psychology Bulletin,* **32**, 117-127.

[27] Winship, C., and Mare, R. D (1992), Models for Sample Selection Bias. *Annual Review of Sociology* **18**, 327-350

[28] Yamaguchi, K. and Ferguson, L. R. (1995), The stopping and spacing of childbirths and their birth history predictors: Rational Choice Theory and Event-History Analysis. *American Sociological Review* **60**, 272–298.

[29] Yoo, B. -K., and ,Frick, K. D. (2006), The Instrumental Variable Method to Study Self-Selection Mechanism: A Case of Influenza Vaccination. *Value in Health* 9, 114–122

[30] Zajonc, R. B. (2001), The family dynamics of intellectual development. *American Psychologist* **56**, 490-496.

[31] Zajonc, R. B., and Markus, G. B. (1975), Birth order and intellectual development. *Psychological Review* **82**, 74-88.

[32] Zajonc, R. B., Markus, G. B., Berbaum, M. L., Bargh, J. A., and Moreland, R. L. (1991). One justified criticism plus three flawed analyses equals two unwarranted conclusions: a reply to Retherford and Sewell. *American sociological Review*, **56**, 159-165.