Economics and Human Biology 33 (2019) 15–28

Contents lists available at ScienceDirect

Economics and Human Biology

journal homepage: www.elsevier.com/locate/ehb

Estimating the educational consequences of teenage childbearing: Identification, heterogeneous effects and the value of biological relationship information

Frank Heiland[a,b,c], Sanders Korenman[a,b,c,*], Rachel A. Smith[a]

[a] Marxe School of Public & International Affairs, Baruch College CUNY, United States
[b] CUNY Institute for Demographic Research, United States
[c] Economics Doctoral Program, Graduate Center, CUNY, United States

ARTICLE INFO

Article history:
Received 9 May 2018
Received in revised form 30 November 2018
Accepted 9 December 2018
Available online 28 December 2018

Keywords:
Teenage childbearing
Sibling-difference methods
Instrumental variables

ABSTRACT

Understanding the contribution of childbearing to social disadvantages of teenage mothers requires estimates that control for unobservables and generalize to teenage mothers. Sibling-differences and Instrumental Variables (IV) are common approaches to this end. Using the "Add Health" data, which oversampled siblings, and building on IV specifications from a widely-cited study, we compare various estimates of the consequences of teenage childbearing for schooling attainment. These IV-based estimates suggest moderate to large adverse impacts of teenage births (point estimates of −0.7 years of schooling or larger). However, the IV estimates are highly sensitive to choice of instrument and model specification. Estimates based on sibling and twin differences are consistently near zero—e.g., an estimated difference of −0.1 years between a teen mother and her biological full sister who did not have a teen birth—and are estimated with sufficient precision to exclude effects larger than −0.5 years. We review concerns about sibling methods and conclude that, despite their limitations, sibling estimates should be admitted along with other evidence on the consequences of teenage childbearing. Appreciation of the sensitivity of IV estimates and their other limitations would reinforce this conclusion.

1. Introduction

Despite rapid declines in teenage fertility in the United States since the early 1990s, U.S. rates remain high in international comparisons (Kearney and Levine, 2012). The consequences of teenage births have continued to concern advocates and policy-makers, despite mixed and inconclusive evidence on the effects of teenage births (Kearney, 2010; Geronimus and Korenman, 1992; Hoffman et al., 1993; Geronimus and Korenman, 1993; Ashcraft and Lang, 2006; Fletcher and Wolfe, 2009; Hotz et al., 2005; Levine and Painter, 2003; Lee, 2010).

Researchers have turned their attention to explaining the variation across estimates of the effect of a teen birth. They have taken two broad approaches: (1) analyzing the sensitivity of results to the choice of method used to address selection bias in estimates of the effects of teen childbearing (e.g., Kane et al., 2013; Herrera Almanza and Sahn, 2018), and (2) examining whether effects are heterogeneous—differ systematically by the propensity to have teen births—explaining variation across methods and samples by differences in effects across the populations that identify the effect of a teen birth (e.g., Yakusheva, 2011, and Diaz and Fiel, 2016).

Despite the intention of this literature to explain variation in estimates across analytical approaches, authors have excluded results from some estimation strategies in preference to others. Those who favor sibling methods and exclude instrumental variable (IV) methods cite weak instruments and imprecise estimates, while those who favor instrumental variable methods and exclude sibling difference (or within-family) methods, a strategy for reducing selectivity bias, cite concerns over small sample sizes (low statistical power or imprecise estimates), limited generalizability and within-family selectivity.

Duncan et al., (2018) focused on OLS and sibling and cousin fixed-effects estimates of the impact of maternal age on child development, using data from the Children of 1979 National Longitudinal Survey of Youth. They described, qualitatively, several IV estimates in their Appendix B. Duncan et al. (2018) found age at menarche (e.g., Ribar, 1994) to be a weak instrument for maternal age at first birth. Miscarriages (e.g., Hotz et al., 2005)

* Corresponding author.
E-mail address: sanders.korenman@baruch.cuny.edu (S. Korenman).

https://doi.org/10.1016/j.ehb.2018.12.003
1570-677X/

and state abortion laws (citing Bitler and Zavodny, 2001) were strong instruments but the IV estimates " . . . resulted in standard errors that were too large to detect significance given the effect sizes in our (OLS) regressions. So our estimates of the effects were neither significantly different from zero nor significantly different from our main OLS results." But what dictates a preference for more precise but possibly inconsistent estimates over less precise but potentially more consistent (less biased) estimates? In fact, it is common for meta-analyses of causal impacts to limit reviews to studies with experimental or strong quasi-experimental designs. More to the point, what would be the harm in presenting these IV estimates and confidence intervals in the online appendix?

Conversely, a prominent study by Kane, Morgan, Harris, and Guilkey (2013, hereafter "KMHG") used OLS regression and three "quasi-experimental" methods to reduce bias from non-random selection into teenage childbearing, but excluded results from sibling-based estimates. Specifically, KMHG obtained estimates from cross-section OLS, Propensity Score Matching (PSM), Parametric Maximum Likelihood with Instrumental Variables (PML-IV) and Semi-Parametric Maximum Likelihood with Instrumental Variables (SPML-IV) models. KMHG calculated sibling fixed effects (FE) models (i.e., family fixed-effects or sibling-difference estimates) but did not report the results in their paper or supplementary materials, writing "Sibling fixed-effects models were initially explored, but the small sample size called into question the robustness of the results" (p. 2138). In addition to small samples (and hence low power), they cited three other weaknesses of sibling methods (within-family selectivity/endogeneity, limited generalizability, and contamination). Again, it is questionable why sibling estimates and confidence intervals should not be included with online supplementary materials.

KMHG also considered the possibility that the effect of a teen birth on education is heterogeneous, that is, it may vary according to the propensity to have a birth as a teenager, but they did not attempt to estimate heterogeneous effects. Rather, they focused on and preferred estimates of the average effect of a teen birth on educational attainment in the population of *teenagers*.

Diaz and Fiel (2016) employed two methods (smoothing-differencing and inverse probability weighting) that allow heterogeneity in the effect of a teen birth. Using data from the Child and Young Adult Cohorts of the NLSY79, they found that effects are indeed heterogeneous, more adverse among women who are *less* likely to have teen births. However, they were unable to account for selectivity on unobservables: "Although we include a substantial number of background, personal, and contextual indicators in our propensity model, unobserved factors could introduce bias in our results; this is a weakness shared with other propensity-based studies" (p. 114).

Diaz and Fiel's results were consistent with those of Yakusheva (2011), who used High School and Beyond data to estimate effects on educational outcomes of a birth during high school that varied according to the propensity to have a birth. Controlling for educational and fertility expectations and test scores, effects of a birth in high school were small and not significant for high-risk teens; the few significant effects were found among teens at low risk of a high-school-aged birth (e.g., Yakusheva, 2011, Table 6).

We study sibling-based estimates in the context of method choice and heterogeneous treatment effects. Using recent data and large samples from the National Longitudinal Study of Adolescent to Adult Health (Add Health), we estimate sibling and twin fixed-effects (FEs) models of the educational consequences of teenage childbearing. We present our sibling estimates alongside estimates obtained from alternative approaches such as OLS and a Maximum Likelihood IV strategy employed by KMHG. We i) present new sibling FE estimates based on recent samples of siblings and twins and other relatedness information available in the Add Health; ii) present Instrumental Variables (IV) estimates building on KMHG's specifications; iii) evaluate the relative strengths and weaknesses of the sibling-based and IV approaches; and iv) discuss our findings in the context of the heterogeneous treatment effects literature and their implications for policy.

Our main purpose in writing this paper is to argue for inclusion of sibling-based estimates in studies of the effects of behaviors that differ greatly by family socioeconomic background, such as teenage childbearing (Geronimus, 1987). Including sibling-based estimates is especially important for studies that intend to replicate and reconcile differences in findings in a literature where sibling methods have been used to add evidence. Strengths and weakness of all methods should be noted. (Though it is not our focus, we would also argue that studies that have conducted IV estimations with strong, plausibly valid instruments, should report those results fully.) We illustrate these points in the context of a KMHG's study of disparate findings in the literature on educational consequences of teenage childbearing that excluded sibling estimates despite using a sample with a large sibling oversample. We conclude that including sibling estimates would have strengthened one of the study's chief conclusions: namely, that methods used to identify the effects teenage childbearing drive differences across studies, rather than differences in the data source or period of study. However, including sibling methods would also have altered the reading of the "weight of the evidence" on teenage births, resulting in a more cautious conclusion regarding the adversity of their causal impacts.

In the next section, we introduce the Add Health data and describe the samples and measures used in the analysis. Section III reports estimates from sibling and twin fixed effects methods. In Section IV, we analyze concerns commonly expressed about the sibling approach, emphasizing that they can only be assessed in comparison to an alternative approaches. Section V summarizes our findings and discusses the implications of method choice for policy in the presence of heterogeneous treatment effects.

2. Data, samples, measures, and replication

2.1. Data and samples

The data for this study come from the National Longitudinal Study of Adolescent Health ("Add Health") contractual dataset (Harris, 2009). Add Health is a longitudinal, nationally representative sample of over 20,000 U.S. adolescents in grades seven through twelve who attended 80 high schools and 52 middle schools in 1994-95. The first wave was collected through student in-school questionnaires (90,118), and a subsample was interviewed in-home (20,745) and included an in-home parent questionnaire (17,670). At Wave I the Add Health "provides a nationally representative sample of . . . adolescents in grades 7 to 12" (Chen and Chantala, 2014, p.4); this sample (12,105) is referred to as the "core in-home sample".

Supplemental samples were also drawn at Wave I, including oversamples of those of Cuban, Puerto Rican and Chinese ethnicity, black adolescents with highly-educated parents, adopted children, those with disabilities, as well as based on genetic relatedness (Chen and Chantala, 2014, p. 4). Specifically, Add Health identified adolescent pairs (or multiples) related to varying degrees, including twins (1981 adolescents), full siblings (1186 adolescents), half siblings (783 adolescents), unrelated adolescents living in the same household (415 adolescents), and siblings of twins (162 adolescents). Co-resident adolescent pair members were identified through reports on the in-school questionnaire or during the in-

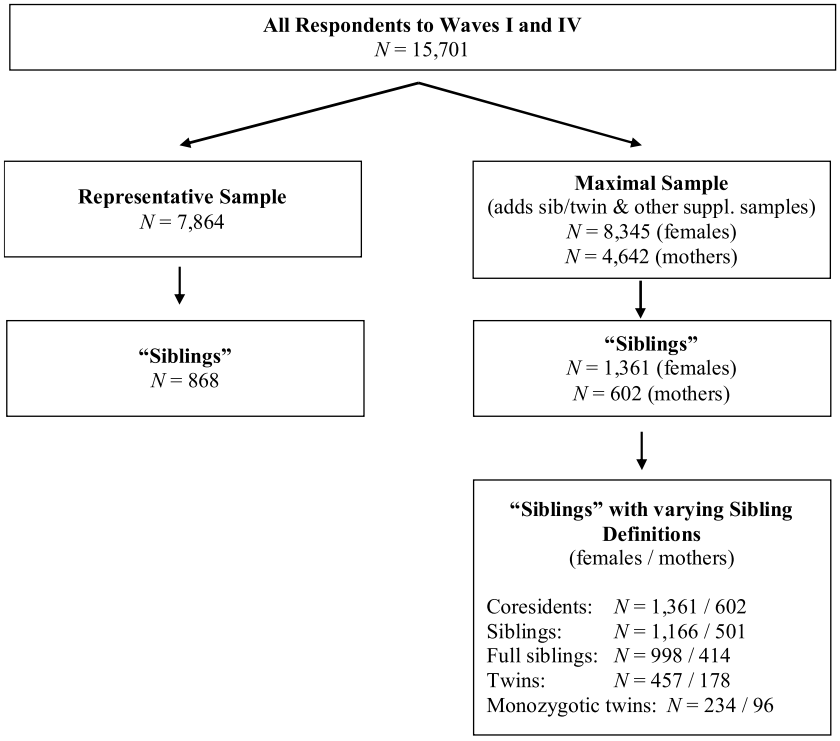

Fig. 1. Sample selection flow chart.

home interview. Members of pairs were both in grades seven through twelve at Wave I (see Harris et al., 2009, 2013).

Three further waves of data were collected from the original in-home adolescent respondents in 1996 (II), 2001-02 (III), and 2007-08 (IV). The data for our analyses come from Wave I and Wave IV. Wave IV included 15,701 respondents who were then aged 24–34.

The oversampling of siblings and twins in a longitudinal study with rich information on family background and school contexts are recognized strengths of the Add Health data. As Harris, Halpern, Haberstick and Smolen (2013) note, Add Health's large samples of twins and siblings, and inclusion of genetic information, support path-breaking research on genetic and social determinants of health and socioeconomic outcomes.

Our samples and measures were selected to facilitate comparisons of results across estimation strategies or for comparisons to estimates reported by KMHG. To this end, we briefly discuss our (largely successful) to attempt to replicate KHMG's samples, measures and results. Most of the paper concerns our main analyses which take advantage of sibling oversamples in the Add Health and to maximize sample sizes available for sibling and twin FE estimates.

We analyzed two Add Health samples that we will refer to as the "representative" sample and the "maximal" sample (see Fig. 1). The representative sample included 7864 females who responded to both Wave I and Wave IV questionnaires and have a valid Wave IV sampling weight ("gswgt4_2"). The maximal sample included all (8345) female respondents to both the Wave I and Wave IV questionnaires, including members of the core in-home sample as well as the supplemental samples, most notably the supplemental samples of siblings. We analyzed various "sibling" subsamples of the maximal sample, including female sample members who co-resided at baseline, siblings, biological full siblings, twins and monozygotic twins. We also conducted a more limited analysis of "siblings" in the representative sample.

2.2. Measures

For comparative purposes, we followed KMHG's variable selection and recoding (as needed) by matching the means and maxima and minima reported in their paper, which we have reproduced in the left-hand columns of Table 1, though missing data and imputation made exact replication difficult.[1] Missing data on family background variables suggest another advantage of sibling FE models: their ability to control for family background determinants of education even when missing.

When there was ambiguity about variable choice and coding, we conducted sensitivity analyses. For example, the reported income values indicated the presence of outliers, so for some analyses we top-coded and bottom-coded the extreme 2% of values (i.e., top and bottom coded 1%). Although this recoding increased the coefficient of income, the coefficient of teen birth was not affected.

Respondents' educational attainment at Wave IV is a categorical variable ("h4ed2") ranging from 1 to 13 (1 = 8th grade or less; 11 = completed a doctoral degree; 13 = completed post baccalaureate professional education). Our main results use a recoding suggested by Jason Fletcher (personal communication). For this "Fletcher" version, we coded "completed a doctoral degree" as 21 years of

[1] KMHG reported using a single-equation imputation model but provided no details. They also reported (p. 2137, note 3) that "4% or more" observations were missing data for parent's education and income. The sample sizes reported in the sixth column of numbers in Table 1 indicates that family income and parental education were missing for 25% and 18% of cases, respectively.

Table 1
Sample means, descriptive statistics, representative sample (includes "non-mothers").

Variable	KMHG, Table 1[a] With Imputation. N = 7,870			Representative Sample, Best Replication Without Imputation[b,c]				Siblings in Representative Sample[b,c,d]	
	Mean (SE)	Min	Max	Mean (SE)	Min	Max	N	Mean (SE)	N
Outcome									
Years of Completed Education (Wave IV) (Max = 26)	**14.39** (0.10)	8	26	**14.41** (0.09)	8	26	7,864	**14.19** (0.13)	868
Years of Completed Education (Wave IV) (Max = 21)	NA			**14.37** (0.09)	8	21	7,864	**14.15** (0.12)	868
Teen Birth (by age 19)	**0.12** (0.01)	0	1	**0.12** (0.01)	0	1	7,864	**0.13** (0.02)	868
Sociodemographic Variables									
Two-Parent Family (1-yes)	**0.53** (0.01)	0	1	**0.56** (0.01)	0	1	6,726	**0.60** (0.04)	766
Parent's Education (categorical)	**5.28** (0.09)	1	9	**5.36** (0.10)	1	9	6,706	**5.18** (0.19)	754
Income-to-Needs Ratio	**2.86** (0.10)	0	83.75	**3.12** (0.12)	0	84.90	5,870	**2.44** (0.18)	660
Income-to-Needs Ratio (1% Top & Bottom Code)	NA			**3.02** (0.10)	.04	18.37	5,870	**2.40** (0.18)	660
Race/ethnicity									
Non-Hispanic Black	**0.17** (0.02)	0	1	**0.16** (0.02)	0	1	7,864	**0.16** (0.04)	868
Hispanic	**0.12** (0.02)	0	1	**0.12** (0.02)	0	1	7,864	**0.11** (0.03)	868
Non-Hispanic Other	**0.04** (0.01)	0	1	**0.05** (0.01)	0	1	7,864	**0.04** (0.02)	868
Non-Hispanic White (ref.)	**0.66** (0.03)	0	1	**0.67** (0.03)	0	1	7,864	**0.69** (0.07)	868
Foreign-born	**0.04** (0.01)	0	1	**0.05** (0.01)	0	1	6,751	**0.04** (0.02)	756
PVT Score	**100.19** (0.67)	15	146	**100.59** (0.68)	15	146	7,529	**99.18** (1.16)	831
Age (Wave I)	**15.40** (0.12)	11	21	**15.36** (0.12)	11	21	7,864	**15.51** (0.21)	868
Age (Wave IV)	NA			**28.23** (0.12)	24	34	7,865	**28.36** (0.19)	868
Per Capita Income	**12.96** (0.42)	1.84	68.82	**12.96** (0.41)	1.84	68.82	7,803	**12.11** (0.40)	868

[a] KMHG report that Standard Errors (SEs) of means are adjusted for clustering, but do not report the clustering variable.
[b] To account for the complex design of the Add Health, our replication means for the full and representative sibling samples are weighted by the Wave IV sample weight (gswgt4_2) and clustered using Stata's svyset command: svyset psuid [pw= (gswgt4_2)], strata(region), as recommended in Chen and Chantala (2014).
[c] For the income/needs ratio we use Cutler and Katz's (1992) implicit equivalence scale for US poverty thresholds, Needs = (single adult threshold) * (Adults + 0.76*children)$^{0.61}$ and the 1994 US Census Bureau poverty threshold for one nonelderly adult living alone ($7710).
[d] The representative siblings sample is a subsample of the representative Add Health sample. As explained in the text it excludes sibling/twin oversamples and requires a Wave IV sampling weight.

schooling rather than 26. This version of the educational outcome variable also has a mean of 14.4 years, but a range of 8 to 21 years.[2]

Like KMHG, we defined a teen birth as a dichotomous variable indicating a young woman had a live birth before exact age 19 (henceforth: "Teen Birth (by age 19)"). We also examined the sensitivity of results to varying the cut-off age, using either age 18 or 20, and to restricting the sample to mothers (by Wave IV) only.

Additional covariates used are from the in-home and contextual data files in Wave I. The student in-home covariates include age, race/ethnicity, and Add Health Picture Vocabulary Test (PVT) standardized percentile score (henceforth: "PVT Score"). Parent-reported variables include parent's education, household income (1994), child nativity (henceforth: "Foreign-born"), and whether both biological parents are present (henceforth: "Two-parent Family"). We constructed an equivalized income (i.e., household income divided by the poverty line) using the parent's reported income and an estimated official poverty line for the household based on the number of adults and children resident in the household (henceforth: "Income-to-Needs Ratio").[3] Following KMHG, we also controlled for per capita income, using a census tract-level variable for 1989 ("tst90591") from the contextual file.

2.3. Brief replication of KMHG

Table 1 presents figures from KMHG's Table 1 and our best replication attempt using the representative sample. We required a non-missing educational attainment outcome at Wave IV and a valid Wave IV sample weight for weighted means. We did not attempt to impute values for missing data. We used Stata's *svyset* command to adjust standard errors (SEs) for clustering due to the sample design, and weighted using the Wave IV weight.

The first three columns of Table 1 reproduces descriptive statistics (means, SE, minima and maxima) reported in KMHG's Table 1. The fourth through sixth columns present our replication.

[2] KMHG recoded this variable and reported a mean number of years of education of about 14.4, with a range of 8 to 26. For analyses reported in Tables 1 and 2, but not in later analyses, we attempted to replicate their recoding. Stata code for our preferred version ("Fletcher") is as follows: rename h4ed2 educationw4; recode educationw4 96=. 98=. 13=19 12=18 11=21 10=18 9=18 8=17 7=16 6=14 5=14 4=13 3=12 2=11 1=8. The h4ed2 education variable is described here: www.cpc.unc.edu/projects/addhealth/documentation/ace/tool/variable?VariableId=6896.

[3] The poverty line is based on the equivalence scales implicit in the official US poverty thresholds, estimated by Cutler and Katz (1992), and adjusted for inflation.

Table 2
OLS Results and Sensitivity Analysis, Representative Sample (includes "non-mothers").

Variable	KMHG Table S1	Replication[a,b]	Income & Edu. Recoded, Wave IV Age[a,c]	Siblings in Representative Sample[a]			
				XSEC OLS[a] No Controls	XSEC OLS[a,c]	FE[c,d] No Controls	FE[c,d]
	1	2	3	4	5	6	7
Teen Birth (by age 19)	**−0.98**	**−0.89**	**−0.88**	**−1.39**	**−0.84**	**−0.05**	**−0.02**
	(0.10)	(0.06)	(0.06)	(0.23)	(0.19)	(0.19)	(0.19)
Two-Parent Family	**0.51**	**0.55**	**0.51**		**0.56**		
	(0.06)	(0.05)	(0.05)		(0.12)		
Parent's Education (categorical)	**0.33**	**0.22**	**0.20**		**0.16**		
	(0.03)	(0.02)	(0.02)		(0.04)		
Income-To-Needs Ratio	**0.09**	**0.04**	**0.12**		**0.16**		
	(0.03)	(0.01)	(0.02)		(0.04)		
Race/ethnicity							
NH Black	**0.81**	**0.62**	**0.64**		**0.81**		
	(0.12)	(0.10)	(0.10)		(0.22)		
Hispanic	**0.41**	**0.33**	**0.35**		**0.17**		
	(0.14)	(0.09)	(0.08)		(0.20)		
NH Other	**0.35**	**0.39**	**0.43**		**−0.40**		
	(0.17)	(0.14)	(0.14)		(0.37)		
Foreign-born	**0.85**	**0.44**	**0.48**		**0.53**		
	(0.19)	(0.22)	(0.21)		(0.33)		
PVT Score	**0.05**	**0.03**	**0.04**		**0.04**		**0.02**
	(0.004)	(0.002)	(0.002)		(0.005)		(0.009)
Age (Wave I or IV)	**0.06**	**0.04**	**0.03**		**0.07**		**0.02**
	(0.02)	(0.02)	(0.02)		(0.04)		(0.04)
Per Capita Income (census tract)	**0.04**	**0.04**	**0.03**		**0.03**		
	(0.01)	(0.01)	(0.01)		(0.01)		
Intercept	**5.68**	**13.16**	**13.19**	**14.51**	**11.45**	**0.00**	**0.00**
Top/bottom code income?	No	No	Yes	NA	Yes	NA	NA
Number of Obs. (N)	7,870	7,864	7,864	868	868	868	868

[a] Regressions are unweighted. SEs are clustered to account for the Add Health complex survey design (Chen and Chantala, 2014).
[b] As in KMHG, income is not top/bottom coded, max for education outcome is 26, and age is as of Wave I. See notes to Table 1.
[c] Income bottom-coded and top-coded at 1st and 99th percentile values. Maximum for education outcome is 21, age is age at Wave IV.
[d] In order to apply "svy" commands to account for the complex Add Health survey design, FE (fixed effects) estimates for the representative sample for this table only were calculated by taking differences from family means and dropping one observation per household.

We succeeded in matching closely means and proportions for the following variables: educational attainment; teen birth; Hispanic, non-Hispanic black, white and other racial identification; age, nativity, and county per capita income. We also came reasonably close to matching means for two-parent family, parental education, and adolescent Picture Vocabulary Test (PVT) score, but were further off for the family income/needs ratio (2.86 vs. 3.12), likely due to the high percentage of missing (imputed) values.

The eighth column of numbers in Table 1 shows means for a subset of 868 observations from the representative sample that have at least one "sibling" (co-resident) who is also in the representative sample. Characteristics for this subsample match well those for the overall sample, although their mean family income is somewhat lower (income/poverty ratio of 2.4 compared to 3.1).

Table 2 presents OLS regression results, including KHMG's results reported in their on-line supplement as Table S1. Our standard errors are adjusted for clustering and we did not weight data in the regressions. Unlike KHMG, we did not impute missing values but instead created dummy variables for non-missing response and interacted those dummies with the corresponding variable with missing values. While this has disadvantages relative to multiple imputation, likely understating standard errors (e.g., King et al., 2001), we wanted to facilitate replication of our results.

The first column shows KHMG's results from OLS regressions with controls. The adjusted effect of a teen birth on education is estimated to be -0.98 (0.10) years. The estimate from our best replication specification (second column) was -0.89 (0.06), which is reasonably close to theirs. With some exceptions, the covariate coefficients were similar as well.[4] The third column shows the OLS results from our preferred specification (income and education recoded, wave IV age). According to this specification, the OLS estimate of teen childbearing was -0.88 (0.06), which is nearly identical to the best replication estimate.

3. Sibling estimates

3.1. Sibling estimates from representative sample

Columns 4–7 of Table 2 shows results for the "sibling" subsample of the representative sample. Columns 4 and 5 present cross-section (pooled) OLS results ("XSEC OLS") and columns 5 and 6 results from models with sibling FEs (fixed effects). Analysis of this sample was intended to address concerns that results from the sibling oversample might not generalize to any real-world population. Results based on siblings from the representative sample should generalize to the corresponding subpopulation (i.e., persons with a co-resident sibling in the relevant age range). Furthermore, the descriptive statistics in Table 1 suggests that demographic and socioeconomic characteristics of the subsample

[4] The major differences (comparing columns 1 and 2) were our much smaller estimated effect of income/needs (0.09 vs. 0.04), foreign born (0.85 vs. 0.44) and age at Wave I (0.06 vs. 0.04). Column 3 shows the effects of top-coding and bottom-coding the highest and lowest 1% of reported income values, the education recode, and use of wave IV age instead of wave I. This markedly increased the coefficient of income-to-needs ratio (to 0.12), but had little effect on other coefficients, including the coefficient of teen birth.

Table 3
Sample Means, Descriptive Statistics by Type of Family, Maximal Sibling Sample.[a,b,c]

		Mixed Teen/Non-teen Families		
Variable	No teen births	Not teen moms	Teen moms	All sibs teen moms
Number of observations (outcome variable sample size)	1,018	146	143	54
Outcome				
Years of Completed Education, Wave IV (Max = 21 y.)	14.6	13.4	13.1	12.8
Covariates				
Two-Parent Family	51.6	28.6	26.8	35.4
Parent's Education (years, recoded from categorical)	13.2	12.3	12.3	12.4
Income-to-Needs Ratio (1% top & bottom-coded)	2.9	1.7	1.7	1.7
Race/Ethnicity				
NH Black	21.1	43.8	42.7	37.0
Hispanic	12.5	19.9	19.6	11.1
NH White	59.8	34.9	36.4	46.3
NH Other	6.6	1.4	1.4	5.6
Foreign-born	5.4	3.8	3.2	6.4
PVT score	99.4	92.9	91.9	91.5
Age, Wave IV	28.4	28.2	27.8	28.1
Per Capita Income (census tract) ($1000 s)	13.0	10.3	10.5	10.8
Additional descriptive information				
Age at first birth (conditional)	23.1	21.9	16.9	17.3
Age, Wave I	15.6	15.4	15.1	15.2
Number of teen moms among respondents in family	0.0	1.0	1.1	2.1
Number of moms among respondents in family	1.1	1.8	1.8	2.1

[a] Teen mother is defined having a live birth before exact age 19. Non-mothers are included in the non-teen mother category.
[b] Siblings are defined as co-resident female sample members at Wave I with educational outcomes at Wave IV.
[c] Total number of observations (based on outcome variable) is 1,361 (=sum of obs. across four types of families).

of siblings from the representative sample were similar to those of the full representative sample.

The adjusted educational differential by teen birth status for this subsample was nearly identical to the same differential in the overall representative sample (-0.88 vs. -0.84, comparing columns 3 and 5 of Table 2). The unadjusted OLS estimate of the education differential was larger at -1.39 (0.23) years (column 4). When we took "sibling" differences (i.e., included household FEs), the unadjusted differential fell to -0.05 (column 5). Controlling for (sibling differences in) PVT score and age reduced the coefficient further, to -0.02 (column 6). Although this estimate is somewhat imprecise (SE = 0.19), the 95% CI [-0.39, 0.35] includes neither -0.84 (SE = 0.19) nor KMHG's preferred point estimate of -0.7 (SE = 0.31) years of education.

3.2. Sibling estimates from maximal sample

We next report results for the various samples of siblings, using the maximal sample (representative sample plus sibling and twin oversamples) available from the Add Health as explained above.

Table 3 shows unweighted means for all young women who co-resided with at least one other young woman also in the sample. Of the 1361 young women in this sample, 1018 came from families in which no female sample member had a teen birth (before age 19), 289 came from "mixed" teen/non-teen families, i.e. families in which at least one sample member had and one did not have a teen birth, and 54 came from families in which all members had a teen birth.

Three things are apparent from this table. First, families in which no female sample member had a teen birth are more socially advantaged than those where at least one did. Their parents have more education (13.2 versus about 12.3 years), higher incomes (income/needs of 2.9 versus 1.7), they resided in wealthier areas (census tract per capita income of 13 thousand versus 10.5 thousand), they were less likely to be racially identified as non-Hispanic black (21% versus roughly 40%) and more likely to have two biological parents present at baseline (52% versus 25 to 35%). They also attained more education by Wave IV; adolescents from families where no sample members had a teen birth completed 1.5 to 1.8 more years of education than teen mothers. But they also had substantially higher PVT scores (99 vs. about 92) at Wave I. Note that adolescents from families with no teen births also completed about 1.2 years more education than *non-teen* mothers from "mixed" families (where a sibling had a teen birth), underlining the importance of controlling for family background in estimating effects of teen births on education.

Second, not surprisingly, siblings in mixed (teen/non-teen) families were very well-matched on background characteristics; they have virtually identical parental education, race/ethnicity, etc. (compare the second and third columns). The teen mothers scored slightly below their sisters who were not teen mothers (including non-mothers) on the PVT (91.9 versus 92.9), on average, and also completed about 0.3 fewer years of education. This unadjusted differential among matched (teen/non-teen) siblings of 0.3 years is far smaller than the corresponding unadjusted or regression-adjusted differential of 1.0 to 1.5 years reported in Table 2. Also notable is the five-year difference in average age at first birth between matched siblings who did and did not have a teen birth (conditional on having a birth). A delay of five years, from age 17 to age 22 on average, is not trivial, and takes place across ages with considerable (but by no means universal) educational enrollment. In other words, the within mixed family teen/non-teen difference in education is a modest 0.3 years despite a delay of first birth of at least five years over ages with high enrollment, and despite the fact that many of the non-teen mothers had not had a first birth as of the Wave IV follow-up.

Third, families in which all sample members had teen births do not appear particularly distinct or more disadvantaged than other families in which a female sample member had a teen birth (other than that they are somewhat more likely to be non-Hispanic white). This similarity in observed characteristics suggests that teens (and families) that identify sibling-difference estimates are not highly distinct from the families of other teen mothers.

Tables 4 and 5 present our main sibling regression results. Table 4 shows results for the maximal sample, comparing unadjusted cross-section OLS coefficients ("XSEC OLS") for the maximal sample (N = 8,345) and the "sibling" subsample (N = 1,361) to fixed-effects (sibling-differenced) estimates for the

Table 4
Estimated Effects of Teen Birth on Education, Maximal Sibling Samples.[a,b,c,d]

		Coefficients (SEs), Number of Observations (N)			
		Full Sample	"Sibling" Subsample		
		XSEC OLS	Sibs XSEC OLS	Sib Fixed Effects	
Teen Cut-off Age	Sample Includes Non-mothers?	No Controls	No Controls	No Controls	Age & PVT
19	Yes	**−1.46**	**−1.36**	**−0.24**	**−0.19**
		(0.06)	(0.15)	(0.17)	(0.17)
		N = 8,345	N = 1,361	N = 1,361	N = 1,361
19	No	**−0.89**	**−0.77**	**−0.22**	**−0.07**
		(0.06)	(0.17)	(0.20)	(0.20)
		N = 4,642	N = 602	N = 602	N = 602
20	Yes	**−1.50**	**−1.30**	**−0.20**	**−0.16**
		(0.05)	(0.13)	(0.15)	(0.15)
		N = 8,345	N = 1,361	N = 1,361	N = 1,361
20	No	**−0.95**	**−0.75**	**−0.18**	**−0.01**
		(0.05)	(0.16)	(0.18)	(0.18)
		N = 4,642	N = 602	N = 602	N = 602

[a] "Siblings": young women co-resident at Wave I interview.
[b] Robust SEs: for pooled sample, clustered on family ID, for fixed-effects: robust SEs, not clustered.
[c] Age is measured age at Wave IV; PVT score is the Wave I Add Health Picture Vocabulary Test.
[d] For results based on age cut-offs for teenage birth defined as age <18, see Table A1.

sibling subsample both unadjusted and adjusted for baseline PVT score and age at follow-up. As discussed above, we used age at follow-up, Wave IV, to control for the age at which the educational attainment outcome was measured, rather than age at Wave I, though results are not sensitive to this choice. We tested the sensitivity of results to the definition of teen age (<19 and <20) and to whether we restricted the sample to mothers only.[5]

The unadjusted effect of a teen birth in this sample was about −1.5 years of education, or −0.9 year if the sample is restricted to mothers. In the pooled cross-section of siblings, educational differentials were similar, ranging from −0.75 to −1.36 (rather than −0.89 to −1.50). When we included sibling FEs (and no covariates), the estimated impact fell to −0.18 to −0.24 with SEs ranging from 0.15 to 0.20. When we also controlled for age and baseline test scores, the coefficients generally shrank, ranging between approximately -0.2 and 0. Although FE estimates are somewhat imprecise, the lower bound of the 95% CI was most often above (less adverse than) −0.5 years of schooling. Excluding women who have not yet had births lessened the adverse impact by about 0.15 year, though it also reduced precision.[6]

Table 5 shows estimates for different samples of siblings: co-residing female sample members, sisters (including half, step, and full siblings), full siblings, twins and MZ twins. We produced

Table 5
Estimated Effects of Teen Birth (Age < 20) on Education.[a,b,c]

	Coefficient (SEs)		
	Sib XSEC OLS	Sib Fixed Effects	
All women "Sibling" Definition	No Controls	No Controls	Age & PVT
Co-residents	**−1.30**	**−0.20**	**−0.16**
N = 1,361	(0.13)	(0.15)	(0.15)
Siblings	**−1.33**	**−0.15**	**−0.13**
N = 1,166	(0.15)	(0.16)	(0.16)
Full Siblings	**−1.44**	**−0.11**	**−0.10**
N = 998	(0.16)	(0.18)	(0.18)
"Twins"	**−1.29**	**−0.02**	**−0.02**
N = 457	(0.23)	(0.22)	(0.23)
MZ Twins	**−1.37**	**+0.32**	**+0.31**
N = 234	(0.34)	(0.34)	(0.34)
	Sib XSEC OLS	Sib Fixed Effects	
Mothers Only "Sibling" Definition	No Controls	No Controls	Age & PVT
Co-residents	**−0.75**	**−0.18**	**−0.00**
N = 602	(0.16)	(0.18)	(0.18)
Siblings	**−0.78**	**−0.13**	**−0.01**
N = 501	(0.18)	(0.20)	(0.20)
Full Siblings	**−0.87**	**−0.10**	**−0.01**
N = 414	(0.20)	(0.23)	(0.22)
"Twins"	**−0.74**	**−0.03**	**−0.06**
N = 178	(0.31)	(0.31)	(0.31)
MZ Twins	**−1.06**	**+0.53**	**+0.35**
N = 96	(0.45)	(0.46)	(0.45)

[a] Robust SEs: for pooled sample, clustered on family ID; for fixed-effects, robust SEs, not clustered.
[b] Age is measured age at Wave IV; PVT score is the Wave I Add Health Picture Vocabulary Test.
[c] For results based on age cut-offs for teenage birth defined as age <19 and age <18, see Tables A2 & A3.

estimates for siblings with more distant and more proximate biological relationships. Estimates based on genetic relationship information has long been of interest to applied economists (e.g., Komlos and Kelly, 2016). More closely related siblings should be better matched on genetic and environmental factors, and therefore produce better counterfactual outcomes for those of teen mothers. For example, co-residents may be biologically unrelated and may have spent time in different households prior to

[5] Results with an age cut-off for "teen" births of 18 are reported in Appendix Table 1. They are similar to those where the cutoff is age 19, though less precisely estimated due to the smaller number of births under age 18.

[6] We also followed an anonymous referee's suggestion that, to increase precision, we run OLS regressions on the maximal sibling sample, controlling for whether or not there were any teenage births in the family. The coefficients and standard errors were similar to the fixed effects estimates reported in Table 4; the impact of a teen birth (age <19) was −0.24 (0.16). Note that the dummy variable for whether there were any teenage births indicates either that the family was "mixed teen/nonteen" or one in which all siblings had a teenage birth. Thus, including this dummy variable is not essentially equivalent to the family FE estimator, though it does control for a major source of family-background heterogeneity. Specifically, the teen mothers from families with all teenage mothers are not balanced in the sample by non-teen mothers from the same families, but from families where at least some sisters were not teen mothers. The estimated effect is a combination of the between-sibling estimator in the mixed-teen/nonteen families and the cross-section difference between teen mothers from "all teen" families and non-teen mothers from mixed-teen/nonteen families. If there are unobserved family background differences correlated with the propensity to have a teen birth between "all teen" and "mixed-teen/nonteen" families, this estimate will be biased.

the baseline survey; full-siblings are more genetically similar than half-siblings; and monozygotic twins share all rather than half their genetic material, as is the case with dizygotic twins and full siblings. This better matching comes at a cost of reduced sample sizes and larger standard errors and more limited generalizability since most people have siblings, but few have a twin.

Table 5 shows (unadjusted) pooled, cross-section estimates for the full sample of siblings, and then sibling fixed effects estimates, unadjusted and adjusted for PVT score and age. To increase precision, we show results where the age cut-off for a teenage birth is 20. Results for different age cut-offs (<18 or <19) are similar though less-precise (see Appendix Tables 2 and 3). The first panel shows results for all women (i.e., the non-teen category includes women whether or not they have had a birth) while the second panel is restricted to mothers.

Although Table 5 presents many coefficients, the results are easily summarized. First, sibling FE estimates of effects of teen births were far smaller than cross-section estimates. Second, narrowing the definition of "sibling" to more closely matched "sibling" types generally attenuated the estimated effects, though also reduced precision, and there were a few positive coefficients suggesting "protective" effects of a teen births. Finally, the results for the sample of siblings who have all had first births (i.e., mothers), provide less evidence of adverse effects of a teen birth on educational attainment (and more evidence of protective effects), though the estimates were also less precise than those that included women who have not had births in the "non-teen" category.

The overall impression left by the sibling fixed-effects estimates is one of small to no adverse effects of a teen birth relative to a sibling. Although the sibling estimates are much less precise (larger SEs) than the estimates from the representative sample, the confidence intervals are generally sufficiently narrow that they do not contain large or even moderate adverse effects (e.g., point estimates ranging from -0.7 years for SPML-IV to -1.87 years for PML-IV reported by KMHG, Table 2, p. 2141).

4. Relative merits of sibling estimates

It is well documented in the literature (e.g., Geronimus, 1987; Geronimus and Korenman, 1992; Fletcher and Wolfe, 2009; Hoffman et al., 1993) and readily apparent from our analysis of the Add Health data, that teen mothers are far more likely than women who do not have teen births to come from socioeconomically disadvantaged backgrounds (e.g., Table 3). The case for using sibling FE to control for selectivity into teen childbearing on the basis of unobservables is strengthened by the large differences in measured family characteristics. As Angrist and Pischke (2015, p.11) warn, " . . . when observed differences proliferate, so should our suspicions about unobserved differences."

Researchers have relied on sibling comparisons for their potential to control many difficult-to-measure attributes common to siblings—such as school and neighborhood quality, family resources, and genetic factors—that could bias estimates of effects of attributes and behaviors (including teenage childbearing) on educational attainment and other socioeconomic and health outcomes. For example Domingue et al., (2015, p.6) wrote in their study of effects of individual genotype variation on educational attainment in the Add Health data:

> A limitation of Model 2 [without sibling fixed effects] is that it cannot account for unmeasured features of families and neighborhoods that are correlated with children's genotypes. Therefore, we fit a third model that utilized the family structure of the data to generate a sibling fixed effect estimate that fully controls for parental genotype and attainments and also for any neighborhood or environmental characteristics that may vary across families.

Wehby (2014) used within-mother (between-sibling) differences in breastfeeding to estimate effects of breastfeeding on childhood disability. More recently, Langa and Nystedt (2018) compared within-monozygotic-twin pairs that are discordant in height to isolate the impact of variations in environmental conditions in childhood on adult earnings. And Nielsen et al. (2017) used twin-comparisons to study relationships between risk aversion and religious behavior. Despite recognizing limitations (p. 7), Magnuson et al., (2016) estimated and reported results of models with sibling fixed-effects in a study of the influence of skills and behaviors in middle childhood on educational attainment.

Although sibling (and twin) fixed effects methods control for any unobservable characteristics common to siblings/twins including similar neighborhood, family and school contexts, they have limitations. For example, KMHG point to four potential weaknesses: within-family endogeneity/selectivity on unmeasured characteristics, low power and precision (small sample size), limited generalizability, and contamination (effect of a teen birth on the sibling who did not have a teen birth). We analyze these four concerns in turn and, when possible, examine the relative merits of IV-based estimates including KMHG's preferred SPML-IV estimate.

4.1. Selectivity/endogeneity

While sibling differences control for confounders common to siblings, they cannot account for unobserved differences between siblings. However, this is a weakness of many non-experimental estimates. For example, KMHG's SPML-IV estimates capture person-type-specific unobserved heterogeneity, but they also

> " . . . use additional variables in the parametric and semi-parametric ML models **to identify teen childbearing** [effects]: statewide abortion laws regarding parental consent for and public funding of abortion; the abortion rate among women aged 15–44; the average Medicaid payment per recipient; number of Ob/Gyn physicians per 100,000 women aged 15–44; and the percentage of family planning clients younger than age 20. All variables except for the latter have been used as instruments in past research" (p. 2137, emphasis added).

The validity and superiority IV estimates rest importantly on the strength and validity of instrumental variables. KMHG argue that "[e]ach factor is a plausible instrument for teen birth because each is correlated with teen childbearing and arguably has no direct effect on educational attainment" (p. 2134). This description raises several methodological concerns.

First, the plausibility (theoretical validity) of the exclusion restrictions is questionable. For brevity, we focus on the two (of six) instruments that were significant predictors of a teen birth in the first stage from their preferred models: state Medicaid payments per recipient and the share of family planning clients in a county that is under the age of 20.[7] KMHG (p. 2135) wrote that "higher Medicaid payments may also proxy for more generous social safety net programs and have a negative association with teen births." If generosity of a state's Medicaid program is a proxy for the generosity of other safety net programs, it could very well be correlated with the generosity of state support for public education (primary, secondary or post-secondary) and affect educational outcomes through many channels other than teenage births. In

[7] The IV estimate with multiple instruments is a sort of weighted average of IV estimates based on each instrument independently, where the weights are the covariance of each instrument with teen birth in the first stage (Angrist and Pischke, 2009, pp. 174–5).

Table 6
Estimated Effects of Teen Birth (Age <19) on Education, Linear Instrumental Variables (2SLS) Regressions.

Model	Instruments	Covariates[a]	SE Clustering Adjustment[b]	Coefficient (SE) of Teen Birth	First Stage F-statistic for Excluded Instruments	Overidentification test p-value[c]
1	All six[d]	None	None	**−7.05**	11.6	0.00
			Add Health DEs	(1.0)	NA	
			State	(2.2)	3.5	
				(2.0)		
2	All six[d]	All but Per Capita Income	None	**−5.39**	7.0	0.07
			Add Health DEs	(1.1)	NA	
			State	(1.6)	8.2	
				(1.5)		
3	All six[d]	All	None	**−4.28**	5.7	0.20
			Add Health DEs	(1.1)	NA	
			State	(1.5)	7.1	
				(1.6)		
4	Two strongest[e]	All but Per Capita Income	None	**−6.64**	15.3	0.28
			Add Health DEs	(1.5)	NA	
			State	(1.8)	18.6	
				(1.9)		
5	Two strongest[e]	All	None	**−5.94**	12.3	0.32
			Add Health DEs	(1.5)	NA	
			State	(2.0)	14.8	
				(2.1)		
6	One strongest[f]	All but Per Capita Income	None	**−5.76**	22.4	
			Add Health DEs	(1.6)	NA	
			State	(2.0)	39.4	
				(2.1)		
7	One strongest[f]	All	None	**−5.14**	18.7	
			Add Health DEs	(1.6)	NA	
			State	(2.2)	28.4	
				(2.5)		

[a] For list of "All" covariates, see Table 2, column (3).
[b] "Add Health DEs" indicates that standard error estimates account for the complex design of the Add Health; SEs are clustered (but not weighted) using Stata's svyset command: svyset psuid, strata(region). See Chen and Chantala (2014). Alternatively, SEs are clustered by state ("State").
[c] Sargan and Basmann chi-square p-values do not differ.
[d] Six instruments are: statewide abortion laws regarding parental consent for and public funding of abortion; the abortion rate among women aged 15–44; the average Medicaid payment per recipient; number of Ob/Gyn physicians per 100,000 women aged 15–44; and the percentage of family planning clients younger than age 20.
[e] Two strongest instruments: State Medicaid expenditures per recipient; county percent of family planning clients under age 20.
[f] Strongest instrument: county percent of family planning clients under age 20.

fact, to our ears, this theoretical justification sounds like a warning not to use this variable as an instrument.

It is not clear why the percent of a county's family planning clients that is under age 20 should be related to the risk of a teenage birth, and since this instrument had not been used in a prior study, we can only speculate. In theory, this variable could measure the level of sexual activity among teenagers or the demand for or access to contraception and abortion among sexually-active teenagers. However, all else the same, this variable is certainly determined by the population age distribution. But age distributions (cohort sizes) affect per capita educational spending, and, in turn, educational attainment (e.g., Bound and Turner, 2007). This mechanism plausibly links the instrument to educational attainment through educational processes directly, rather than solely indirectly through teenage fertility rates.

4.1.1. Linear IV

Tests of overidentifying restrictions can sometimes be used to bolster the case for exclusion restrictions. KMHG (footnote 6) reported that "Results from a test of overidentifying restrictions indicate that these instruments are valid (chi-square = 4.57; p = .47). This test was performed using a linear IV regression model. It rests on the assumption that one instrument is valid and then goes on to test the validity of the remaining instruments."

KMHG argue for the excludability of their instruments based on overidentification (OID) tests in a linear IV regression model that differs from the SPML-IV model they intended to validate. They did not report the linear IV estimation results. Therefore, we estimated linear IV regressions to replicate their OID tests and for additional evidence on the validity of instruments. A minor difficulty is that their models included a contextual variable (per capita income in the census tract) in the second stage (predicting education) that was excluded from the first stage (predicting teen births). Although ML-IV methods allow this, linear IV methods (two stage least squares) do not, and the resulting estimates are inconsistent. Therefore, we estimated linear IV regressions two ways, including tract per capita income in both stages and excluding it from both. Table 6 shows the results.

Each row represents a different linear IV model specification. For each coefficient we report three SEs corresponding to i) not adjusting for clustering; ii) adjusting to account for the Add Health complex sample design; or iii) clustering by state.[8] Clustering is needed for correct standard errors and inference since instrumental variables are measured at the county or state level, and the Add Health sample is clustered within schools (Moulton, 1986, 1990). The model specifications (numbered in the first column) also differ according to whether census-track per capita income is included in both stages and according to the instruments employed. Models 1 through 3 include all six KMHG instruments. Models 4 through 7 include one or two instruments that were significant predictors of a teen birth in KMHG's preferred models.

The major feature of the table is the large, implausible estimates of the effect of a teen birth on years of education. The effect of a

[8] KMHG do not report whether standard errors for their statistical models are clustered. Ambiguity arises since the footnotes to the relevant tables report only that the analyses were unweighted. The notes to their Table 1 (descriptive statistics), however, report that the statistics for that table were weighted *and* adjusted for clustering.

Table 7
Estimated Effects of Teen Birth (Age <19) on Education, Dummy-Endogenous Variable Estimators.

Model Num.	Method of Estimation[a,b]	SE Clustering Adjustment[c]	Coefficient (SE) of Teen Birth	Rho (SE)
1	KMHG, PML-IV (Table S3)	Not Reported	**−1.87**	0.23
			(.23)	(.05)
2	PML, IV	None	**−1.67**	0.24
		Add Health DEs	(.30)	(.09)
		State	(.36)	(.10)
			(.37)	(.11)
3	PML, no IV	None	**+1.56**	−0.69
		Add Health DEs	(.12)	(.02)
		State	(.14)	(.03)
			(.15)	(.03)
4	Two-Step, IV	None	**−0.91**	0.01
			(0.48)	NA
5	Two-Step, no IV	None	**+0.65**	−0.45
			(0.59)	NA

[a] Education outcome models used the full set of covariates listed in Table 2, Column (3). Following KMHG, the model predicting teen birth includes their six instrumental variables and the full set of controls except that it does not include the control for census tract per capita income (even though the model predicting education does). Including per capita income in the teen birth equation changes results only very slightly.
[b] All estimates were computed used Stata14 endogenous treatment effects "etregress" commands.
[c] "Add Health DEs" indicates that standard error estimates account for the complex design of the Add Health; SEs are clustered (but not weighted) using Stata's svyset command: svyset psuid, strata(region). See Chen and Chantala (2014). Alternatively, SEs are clustered by state ("State").

teen birth is greater in absolute value than (minus) seven years of education in model (1) that includes no controls, and ranges between −7.7 and −4.4 in IV models with full controls. Unclustered SEs were 1.1–1.7 years of education while clustered SEs were 1.7–2.7 years.

Massive inflation of coefficient sizes to implausible values likely reflects violation of exclusion restrictions in combination with weak instruments. Although KMHG argue that use of multiple instruments is an advantage over prior studies ("Another shortcoming of these IV studies is that they use only one identifying variable" p. 2134), adding weak instruments increases bias. Only two of the six instruments were significant predictors of a teen birth in their preferred models (SPML-IV), so their IV estimates may be biased in the same direction as the OLS estimate (Bound et al., 1995).

As a test for weak instruments, we present a partial F-statistic for the joint significance of the instruments in the first stage; a value of 10 or more is considered desirable (Stock et al., 2002). With the full six-instrument set (Models 1, 2 and 3), this F-statistic was greater than 10 only when no controls were included *and* standard errors were *not* clustered by state; otherwise, it ranged from 3.5 to 7.5. Since validity of the IV estimates requires that instruments affect education only by increasing teen births, and since the instruments are only weakly related to teen births, assuming that the effects of the instruments on education work only "through" teen births when in reality there are other pathways greatly exaggerates effects of teen births.[9]

The rightmost column of the table shows OID test statistics. When tract per capita income was not controlled, the OID test provided moderately strong evidence against their exclusion restrictions ($p = .07$ for Model 2 with detailed controls, or $p < .01$ in Model 1 with no controls). However, when we controlled for tract per capita income in both stages, the OID tests do not reject (the p-value of 0.33 is similar to the p-value of 0.47 reported by KMHG). OID tests rely on the untestable assumption that at least one instrument is valid, and tend to have low power, especially when the IV estimate is imprecise. So failure to reject the exclusion restriction may provide little evidence as to its validity (e.g., Angrist and Pischke, 2009, p. 146). Given possibly low power, the (near) rejection in one of the two models with detailed controls (Model 1 and 3) is worrisome.

We experimented with dropping weaker instruments, retaining either the two empirically strongest (state average Medicaid payments or the percent of family planning clients under age 20) or the strongest (percent family planning clients under age 20). While these may be statistically strong instruments, we reiterate that the theoretical justification for excluding them from the first stage is questionable. The results are reported as Models 4 through 7. Throwing out the four weakest instruments dramatically increased the first-stage F-statistic (to 12 to 41) but continued to indicate that a teen birth resulted in five to seven year reductions in educational attainment, although these estimates were less precise (SEs of 2.2 to 2.7 when clustered by state).

4.1.2. ML-IV and two-step IV

Given the implausibility of the linear IV results and evidence of weak instruments, the role of the instruments in the ML-IV estimation is of particular interest. We did not attempt SPML-IV estimation since it requires specialized software. However, we did perform PML-IV estimation. Table 7 presents the results.

Model 1 reprints KMHG's PML-IV results and Model 2 is our replication. The three SEs reported beneath each of our coefficient estimates again correspond to no clustering; clustering to take account of the Add Health complex sample design, and clustering by state. Comparing Models 1 and 2, the coefficients of teen birth (with full controls) were similar, suggesting adverse effects on education of −1.9 (KMHG) or −1.7 (us). The estimate of the residual correlation between the teen birth equation and the educational outcome were nearly identical (0.22 with a SE of .05). The positive *rho* indicates that, conditional on observables, there is unexpected "positive" selectivity into teenage childbearing based on unmeasured educational promise.

When we dropped the instruments and relied on distributional and functional form assumptions alone for identification (Model 3), the effect of a teen birth became large, *positive*, and highly statistically significant. KMHG dismissed the distributional (joint normality) assumption in favor of the SPML-IV models (which assume a more flexible mixture distribution). Nonetheless, it is

[9] To see this, assume for simplicity one instrument Z affects whether or not a young woman has a teenage birth (indicated by the dummy variable T) and her years of education, E, potentially through multiple causal channels. Then $\Delta E/\Delta Z = (\Delta E/\Delta T) * (\Delta T/\Delta Z)$. Dividing through, the IV (and Wald) estimate $\Delta E/\Delta T$ equals the ratio $(\Delta E/\Delta Z) / (\Delta T/\Delta Z)$. If the instrument is a weak determinant of teen births (especially relative to its effect on education), the denominator will be small and the resulting ratio may produce a spurious, large IV estimate of the effect of teen births.

clear that the instruments have a major influence on the results. In models without instruments, the estimate of *rho* is strongly negative (-0.8), suggesting negative educational selectivity into teen childbearing. Also, clustering reduces SEs modestly in Model 2, with IVs, and increases them noticeably in Model 3, without IVs.

Finally, Models 4 and 5 repeat the analyses of Models 2 and 3 but use a two-step consistent estimator that relaxes the PML-IV joint normality assumption.[10] Compared to Models 2 and 3, these results suggest much weaker adverse effects of a teen birth; the coefficient was about one year using the KMHG instrument set or essentially zero without IV, although SEs were much larger (albeit not clustered). In sum, the results presented in Tables 6 and 7 raise concerns about the validity of the KMHG instrument set and show the sensitivity of results to the choice of instruments.

4.2. Power and precision

Statistical power and precision are empirical issues. KMHG's preferred point estimate is -0.7 (SE = 0.31) years of education. Are the sibling sample sizes in the Add Health adequate to detect a difference of that magnitude? There were 138 families with at least one teen birth and at least one sibling who did not have a teen birth (there are 143 teen mothers and 146 non-teen mothers in those families; Table 3). For those families, the standard deviation (SD) of the within-family difference in educational attainment of a woman who did not have a teen birth and her teen-mother-sister was almost exactly 2 years. For a sample size of 138 with a SD of 2, using a two-sided test and a 0.05 alpha (significance level), the power—the probability of rejecting the null (of no difference in education) when the true difference is 0.7 years or larger—is over 98%. If the strong presumption of adverse effects of a teen birth provides the appropriate null hypothesis (i.e., that teen births *raise* education), then a one-sided test is called for, and we have over 99% power. We also have 90% power to detect a difference as small as 0.5 years of education using a one-sided test.

4.3. Generalizability

Citing Allison (2009), KMHG (p. 2133) argued that sibling estimates "are not generalizable beyond the analytical sample." This claim is linked to the observation that sibling FE estimates are identified from within-family variation in education and teen childbearing, and require multiple siblings who co-reside at the time of the baseline interview. Thus, the identifying sibling sample may be selected relative to the population of all families, or relative to families with two or more co-resident sisters of the appropriate ages, or relative to families in which one young woman had a teen birth. On the last point, Table 3 shows that, first, nearly three-quarters (72.6%) of teen mothers in our maximal sibling sample come from families in which one young woman had and one did not have a teen birth. Second, judging by means of observable characteristics, families with variation in teen birth appear similar to those in which all young women had a teen birth. Third, unadjusted and regression-adjusted effects of a teen birth are similar in the sibling samples to those for the entire representative sample. So, even though the sibling sample may be unrepresentative in some ways, it does not mean that the results do not generalize to some population, say, families with multiple sisters in which one had a teenage birth and one did not.

A separate question is whether sibling differences in outcomes associated with a teen birth for this population are informative for other populations. Even if the estimates generalize only to a population that is in some ways selected, they may still be interesting or relevant to analyzing the social problem or efficacy of policy. After all, estimates from genuine field experiments may have good internal validity and unknown but not necessarily problematic external validity. In this case, one would not want to discard the evidence from sibling FE methods that control for important unobservables, but admit it for consideration along with evidence from other methods, and use it to test theories regarding generalizability, as one would experiments carried out in a limited set of environments (Banerjee et al., 2016).

Finally, we should also consider the generalizability advantages of the ML estimates. KMHG wrote (p. 2136, emphasis added) that: "... we draw conclusions from a nationally representative population-based sample that provides an estimate of the years of education lost as a result of having a teen birth *for all those at risk of teen childbearing*." However, it is not just the national sample but also untested functional form and distributional assumptions (joint normality for the PML-IV and a complex mixture distribution for the SPML-IV) that allow generalization to all those at risk of teen childbearing (Angrist and Pischke, pp. 197–205).

If effects are heterogeneous, ability to estimate an average treatment effect for the population of teenagers is not necessarily desirable. If we are motivated to understand why teen mothers have reduced educational attainment and high rates of poverty, for example, the effect of teen births on the average teen mother is more relevant than the effect on the average teenager. This distinction becomes more important the more selected the population of teen mothers and the more heterogeneous the effects of a teenage birth (Diaz and Fiel, 2016).

4.4. Contamination

Families may compensate for (help) the teen birth, thus reducing resources available for and education of the other sibling. The difference in outcomes between the teen mother and her sister would not, therefore, properly account for the full costs of the teen birth to the individual or family. This criticism is important and logically correct. It also has value in shifting the focus of analysis beyond the individual to the family, extended family or larger community (Wall-Wieler et al., 2016; Yakusheva and Fletcher, 2015; Geronimus and Korenman, 1993). However, the sign of any bias from parental or family processes and behavioral responses is uncertain and an empirical issue. Family and kin may respond to the teen birth by reinforcing differences, for example, by reducing educational investment in the teen mother and increasing educational investments in her sister who has not had a birth (Geronimus, 2003; Ben-Porath, 1980). Very disadvantaged families, in particular, may lack the resources necessary to equalize or reduce differences in educational outcomes. Reinforcing behavior by parents could lead sibling differences to exaggerate effects of a teen birth. Again, this is an empirical issue worthy of study, and the sign of any bias cannot be assumed.

4.5. Attenuation bias

Many scholars have pointed to limitations of twin/sibling methods; see, for example, Boardman and Fletcher (2015), who review estimates of the impact of education on health, and Bound and Solon (1999), who review estimates of the impact of education on earnings. Along with within-family endogeneity discussed above, measurement error in the explanatory variable is the main focus. With respect to endogeneity, Bound and Solon's (1999) arguments suggest that if, within family, siblings with the highest

[10] The two-step consistent procedure first estimates a birth equation probit, predicts the birth probability, and then estimates the treatment effect from an education regression augmented by a function (inverse Mills' ratio) of the birth probability (see Maddala, 1983, pp. 120–122).

(unobservable) educational potential are likely to be selected out of teen childbearing, then sibling FE estimates will be upward inconsistent (i.e., exaggerate the adverse impact of a teen birth).

On the other hand, attenuation bias from classical measurement error is exacerbated, possibly severely, in the sibling FE estimates relative to OLS estimates, leading sibling FE estimates to understate adverse effects of teen births (e.g., Nielsen et al., 2017). In the case of classical measurement error, the extent of attenuation bias in the sibling FE estimate depends on the reliability of the fertility measure and the intra-family correlation in fertility timing. The within-family correlation in teen birth that we observe in the Add Health data is modest (about 0.18). But the bias in the coefficient of teen birth from the sibling FE models also depends on the reliability and within-family correlation of other included variables with which teen births may be correlated. The PVT score, the most important covariate in sibling FE models, is highly correlated within families (around 0.55) and also measured imperfectly.[11] Since PVT and teen births are negatively correlated, all else the same, measurement error in the PVT score will bias downward (exaggerate) the adverse effect of teen births.

Harris (2013) reports that several fertility measures in the Add Health are consistently reported.[12] Nonetheless, we have no direct measure of reliability of teen fertility reports and no information on whether measurement error is classical. Efforts underway to link Add Health records to birth records could prove useful in establishing the reliability of Add Health fertility reports.[13]

5. Summary and conclusions

We have explored and presented results from sibling-based and IV approaches to estimating the consequences of teen births for educational attainment. We were motivated in part by studies that embraced this objective but did not fully report results from all of the approaches pursued: i.e., they employed both sibling and IV-based methods, but did not present estimates from one or the other. Another motivation was the emergent literature on heterogeneous effects of teenage births. In keeping with this literature, we paid particular attention to the ways in which method choice affected estimated treatment effects and their interpretation and generalizability.

Using Add Health data, which provides large sibling and twin oversamples (Harris, 2009), we began by estimating several sibling-difference models and then analyzed the relative strengths and weaknesses of the sibling FE estimates. We found the Add Health sibling samples adequately powered to detect an effect as large as −0.7 years. Our estimates from models with sibling FEs were far smaller than the corresponding OLS estimates. Point estimates were often far less than 0.25 years of education, and the lower bound of the corresponding 95% confidence interval generally above (closer to zero than) −0.5 years, a finding inconsistent with large negative consequences and even moderate adverse effects of teen births on education.

We examined the sensitivity of the results to definitions of "siblings," from all young women who co-resided at baseline (age 11 to 21), to sisters (including half- and step-sisters), to full sisters, to twins, to monozygotic twins. Narrowing the definition of siblings provided better match quality/covariate balance, and generally smaller estimated effects, but also reduced sample sizes and precision. The results were not sensitive to the definition of teen age (<18, <19, or <20), though using an older age limit increased precision since it increases the number of "teenage" births.

Selectivity is a challenge to all non-experimental methods, including sibling FE estimates that may be baised by within-family selectivity into teenage childbearing. While instrumental variables can, in theory, eliminate problems of selectivity/endogeneity, including within-family selectivity, much rests on the quality of the instruments. Kane et al. (2013) also used the Add Health data to examine differences in estimates across methods. They relied on instrumental variables in parametric and semi-parametric maximum likelihood (ML-IV) estimates to estimate causal impacts for the population of teenagers, preferring them to sibling FE methods which they conducted, but did not report. However, our analysis of PML-IV and linear IV models raised concerns about the validity and strength of their instruments. The IV point estimates indicated moderate to large adverse impacts of teenage childbearing on education but were highly sensitive to choice of instrument and model specification. The role of the instruments in the ML-IV estimates calls for additional analysis in the context of SPML-IV models.

The role of method choice for generalizability is a subtle but crucial issue. Diaz and Fiel (2016) used an inverse probability weighting strategy to estimate heterogeneous effects of teen births and attributed differences in estimates across methods employed in the literature to the different populations that the methods rely on to identify the effect of a teen birth. The treatment effects literature identifies these effects as the average treatment effect (ATE, the average effect of the "treatment", a birth, on the population of teenagers); the average effect of the treatment on the treated (TOT, the average effect of a teen birth on teenage mothers); and the local average treatment effect (LATE, the effect on "compliers": teenagers whose fertility behavior was determined solely by the quasi-experiment under study).

Maximum Likelihood IV estimates based on representative samples rely on distributional and functional form assumptions, and the validity and strength of the instruments, to identify ATEs for the population of teenagers. In the corresponding linear IV models that we analyzed, the resulting estimates are a kind of weighted average (across instruments) of LATE (on the treated) that generalize only to compliers (Angrist and Pischke, 2009). It is the distributional and functional form assumptions that allow extrapolation of ML-IV estimates beyond LATEs to estimate population average treatment effects.

Whether or not generalizability to the entire population of teenagers is desirable depends on both the plausibility of the assumptions required and whether the average causal effect for the entire population answers an important question about the consequences of teen childbearing. Diaz and Fiel (2016) reasoned that sibling FE estimates likely approximate the TOT estimate rather than the ATE estimate that "applies to typical young women in the population" (p. 89). Although they employed a rich set of covariates, Diaz and Fiel did not use methods to control for selection of teen childbearing on unobservables (nor did the heterogeneous effects study by Yakusheva, 2011). Geronimus (1987, 2003) and co-authors (e.g., Geronimus and Korenman, 1992, 1993), Hotz et al. (2005) and Ashcraft et al. (2013), among others

[11] The Add Health PVT is "a computerized, abridged version of the Peabody Picture Vocabulary Test—Revised (PPVT-R)" (www.cpc.unc.edu/projects/addhealth/design/wave1 accessed July 29, 2016). Reliability of the full Peabody Picture Vocabulary Test-R is around 0.8 (Campbell, 1998, p. 336).

[12] "As one indicator of data quality we compared respondents' summary reports (e.g., how many times have you ever been married?) with counts generated from completing relationship and pregnancy history tables. We found that 97% matched on total number of pregnancies, 95% matched on total number live births, and 95% matched on total number of living children. Further, 93% of respondents matched on all 3 reports. We also gave respondents opportunity to confirm information provided such as birth dates of children. Only 0.68% of baby birth dates were missing after respondents had the opportunity to correct information provided earlier in the interview."(Harris, 2013, p. 11).

[13] See www.cpc.unc.edu/projects/addhealth/news/add-health-study-funded-for-5th-wave-of-data-collection-to-study-developmental-origins-of-health-and-chronic-disease-in-the-u-s (accessed July 29, 2016).

have argued for the social and policy relevance of TOT estimates, though their methods allowed them to identify only "local" effects (e.g., LATEs).

Since teenage childbearing is highly selected, even if sibling FE estimates are not generalizable to the national population, they nonetheless produce informative estimates of the effect of teen births that control for unobservable family background factors for a group of young women who actually had births as teenagers. Sibling FE models offer an additional, compelling way of addressing endogenous fertility decisions. Thus, despite their weaknesses of reduced power and risk of contamination, estimates from sibling difference methods should be admitted as evidence on the consequences of teenage births. Similarly, we would also encourage full reporting of results based on IV methods that use strong and plausibly valid instruments. Our sibling FE estimates from recent Add Health data strengthen the conclusion that method choice matters. They are also broadly consistent with results of studies employing similar methods with earlier data (e.g., Geronimus and Korenman, 1992).

This conclusion regarding the utility of sibling and twin methods echoes those of Boardman and Fletcher (p. 198) who, while critical, wrote that sibling estimates have "many advantages in adjusting for potential genetic and environmental confounding," and Bound and Solon (p. 180) who wrote: "Nonetheless, we have argued that, even though the . . . between-sibling estimates are inconsistent, they still may be useful." Given evidence that selectivity into teen childbearing is strongly related to characteristics linked to family socioeconomic background (Geronimus, 1987; Kearney and Levine, 2012), sibling estimates are especially appealing for the study of the consequences of teen childbearing.

The view that sibling differences are quasi-experiments suggests a LATE interpretation of the sibling FE estimates. However, since in the Add Health sample observable characteristics of teen mothers from such families appear similar to characteristics of young women from families in which all sample members had teen births, these estimates may also approximate an average treatment effect on the treated (TOT). On the other hand, families in which any sibling had a teen birth appear quite distinct from those in which none did. And since most U.S. teenagers do not have babies, it would be inappropriate to extrapolate the sibling FE estimate to the typical teenager. The field awaits a study that could both estimate heterogeneous effects and convincingly control for selectivity on unobservables/endogeneity of teen births.

From this perspective, efforts such as those by Diaz and Fiel (2016) and Yakusheva (2011) to model formally heterogeneous effects are particularly welcome contributions to the literature on the consequences of teenage childbearing. Modeling heterogeneous effects is especially important because they produce TOT estimates relevant for social and policy analysis; beyond this, they can promote a deeper understanding and wider appreciation of the effect heterogeneity that may underlie disagreements between well-intentioned persons over the efficacy of preventing teenage births (Geronimus, 2003; Kearney, 2010).

Acknowledgements

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

We thank John Bound, Arline Geronimus, Dahlia Remler, Jonathan Bearak and two anonymous referees for their comments. Smith acknowledges support from the Eugene M. Lang Junior Faculty Research Fellowship program of Baruch College.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ehb.2018.12.003.

References

Allison, P.D., 2009. Fixed effects regression models. Sage Publications, Newbury Park, CA.
Angrist, J., Pischke, J.-S., 2009. Mostly Harmless Econometrics. Princeton University Press, Princeton, NJ.
Angrist, J., Pischke, J.-S., 2015. Mastering `metrics: the Path From Cause to Effect. Princeton University Press, Princeton NJ.
Ashcraft, A., Lang, K., 2006. The Consequences of Teenage Childbearing (NBER Working Paper No.12485). National Bureau of Economic Research., Cambridge, MA.
Ashcraft, A., Fernández-Val, I., Lang, K., 2013. The consequences of teenage childbearing: consistent estimates when abortion makes miscarriage non-random. Econ. J. 123, 875–905.
Banerjee, A., Chassang, S., Snowberg, E., 2016. Decision Theoretic Approaches to Experiment Design and External Validity (NBER Working Paper No. 22167). National Bureau of Economic Research., Cambridge, MA.
Ben-Porath, Y., 1980. The F-Connection: families, friends, and firms and the organization of exchange. Popul. Dev. Rev. 6 (1), 1–30.
Bitler, M., Zavodny, M., 2001. The effect of abortion restrictions on the timing of abortions. J. Health Econ. 20 (6), 1011–1032.
Boardman, J., Fletcher, J., 2015. To cause or not to cause? That is the question, but identical twins might not have all of the answers. Soc. Sci. Med. 127, 198–200 Comments.
Bound, J., Solon, G., 1999. Double trouble: on the value of twins-based estimation of the return to schooling. Econ. Educ. Rev. 18, 169–182.
Bound, J., Turner, S.E., 2007. Cohort crowding: how resources affect collegiate attainment. J. Public Econ. 91 (5-6), 877–899.
Bound, J., Jaeger, D., Baker, R., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous variables is weak. J. Am. Stat. Assoc. 90, 443–450.
Campbell, J., 1998. third edition Review of Dunn, L. M., and Dunn, L. M. (1997). Peabody Picture Vocabulary Test, 16. American Guidance Service, Circle Pines, MN, pp. 334–338 Journal of Psychoeducational Assessment.
Chen, P., Chantala, K., 2014. Guidelines For Analyzing Add Health Data. Carolina Population Center Updated March 2014.
Cutler, D.M., Katz, L.F., 1992. Rising inequality? Changes in the distribution of income and consumption in the 1980's. The American economic review, 82(2). Papers and Proceedings of the Hundred and Fourth Annual Meeting of the American Economic Association, , pp. 546–551.
Diaz, C.J., Fiel, J.E., 2016. The effect(s) of teen pregnancy: reconciling theory, methods, and findings. Demography 53, 85–116.
Domingue, B., Belsky, D., Conley, D., Harris, K., Boardman, J., 2015. Polygenic influence on educational attainment: new evidence from the National Longitudinal Survey of Adolescent to Adult Health. AERA Open 1 (3), 1–13.
Duncan, G.J., Lee, K.T.H., Rosales-Rueda, M., Kalil, A., 2018. Maternal age and child development. Demography Published online, November 1.
Fletcher, J.M., Wolfe, B.L., 2009. Education and labor market consequences of teenage childbearing. J. Hum. Resour. 44, 303–325.
Geronimus, A.T., 1987. On teenage childbearing and neonatal mortality in the United States. Popul. Dev. Rev. 13 (2), 245–279.
Geronimus, A.T., 2003. Damned if you do: culture, identity, privilege and teenage childbearing in the United States. Soc. Sci. Med. 57, 881–893.
Geronimus, A.T., Korenman, S., 1992. The socioeconomic consequences of teen childbearing reconsidered. Q. J. Econ. 107, 1187–1214.
Geronimus, A.T., Korenman, S., 1993. The socioeconomic costs of teenage childbearing: evidence and interpretation. Demography 30, 281–290.
Harris, K.M., 2009. The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002; Wave IV, 2007-2009 [machine-readable Data File and Documentation]. Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC doi:http://dx.doi.org/10.3886/ICPSR27021.v9.
Harris, K.M., 2013. The Add Health Study: Design and Accomplishments. Carolina Population Center, University of North Carolina at Chapel Hill.
Harris, K.M., Halpern, C.T., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., Udry, J.R., 2009. The National Longitudinal Study of Adolescent to Adult Health: Research Design [WWW document]. URL. . http://www.cpc.unc.edu/projects/addhealth/design.

Harris, K.M., Halpen, C., Habertsick, B., Smolen, A., 2013. The National Longitudinal Study of Adolescent Health (add health) sibling pairs data. Twin Res. Hum. Genet. 16 (1), 391–398.
Herrera Almanza, C., Sahn, D.E., 2018. Early childbearing, school attainment, and cognitive skills: evidence from Madagascar. Demography 55, 643–668.
Hoffman, S.D., Foster, E.M., Furstenberg Jr., F.F., 1993. Reevaluating the costs of teenage childbearing. Demography 30, 1–13.
Hotz, V.J., McElroy, S.W., Sanders, S.G., 2005. Teenage childbearing and its life cycle consequences exploiting a natural experiment. J. Hum. Resour. 40, 683–715.
Kane, J.B., Morgan, S.P., Harris, K.M., Guilkey, D.K., 2013. The educational consequences of teen childbearing. Demography 50, 2129–2150.
Kearney, M.S., 2010. Teen pregnancy prevention. In: Levine, P.B., Zimmerman, D.J. (Eds.), Targeting Investments in Children: Fighting Poverty When Resources Are Limited. University of Chicago Press, Chicago.
Kearney, M.S., Levine, P.B., 2012. Why is the teen birth rate in the United States so high and why does it matter? J. Econ. Perspect. 26 (2), 141–166.
King, G., Honaker, J., Joseph, A., Scheve, K., 2001. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. Am. Polit. Sci. Rev. 95 (1), 49–69.
Komlos, J., Kelly, I., 2016. The Oxford Handbook of Economics and Human Biology. Oxford University Press, New York.
Langa, E., Nystedt, P., 2018. Two by two, inch by inch: height as an indicator of environmental conditions during childhood and its influence on earnings over the life cycle among twins. Econ. Hum. Biol. 28, 53–66.
Lee, D., 2010. The early socioeconomic effects of teenage childbearing: a propensity score matching approach. Demogr. Res. 23, 697–736. doi:http://dx.doi.org/10.4054/DemRes.2010.23.25 (Article 25).
Levine, D.I., Painter, G., 2003. The schooling costs of teenage out-of-wedlock childbearing: analysis with a within-school propensity-score-matching estimator. Rev. Econ. Stat. 85, 884–900.
Maddala, G.S., 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.
Magnuson, K., Duncan, G.J., Lee, K.T.H., Metzger, M.W., 2016. Early school adjustment and educational attainment. Am. Educ. Res. J. 53 (4), 1198–1228.
Moulton, B., 1986. Random group effects and the precision of regression estimates. J. Econom. 32, 385–397.
Moulton, B., 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. Rev. Econ. Stat. 72 (2), 334–338.
Nielsen, J.S., Bech, M., Christensen, K., Kiil, A., Hvidt, N.C., 2017. Risk aversion and religious behaviour: analysis using a sample of Danish twins. Econ. Hum. Biol. 26, 21–29.
Ribar, D.C., 1994. Teenage fertility and high school completion. Rev. Econ. Stat. 76, 413–424.
Stock, J., Wright, J.H., Yogo, M., 2002. A survey of weak instruments and weak identification in generalized method of moments. J. Bus. Econ. Stat. 20, 518–529.
Wall-Wieler, E., Roos, L.L., Nickel, N.C., 2016. Teenage pregnancy: the impact of maternal adolescent childbearing and older sister's teenage pregnancy on a younger sister. BMC Pregnancy Childbirth 16 doi:http://dx.doi.org/10.1186/s12884-016-0911-2.
Wehby, G., 2014. Breastfeeding and child disability: a comparison of siblings from the United States. Econ. Hum. Biol. 15, 13–22.
Yakusheva, O., 2011. In high school and pregnant: the importance of educational and fertility expectations for subsequent outcomes. Econ. Inq. 49 (3), 810–837.
Yakusheva, O., Fletcher, J., 2015. Learning from teen childbearing experiences of close friends: evidence using miscarriages as a natural experiment. Rev. Econ. Stat. 97 (1), 29–43.