Confronting Data Challenges in Research on Early Conditions and Later Outcome Trajectories: Methods of Combining Samples of Panel Data*

Lingxin Hao
Department of Sociology
Johns Hopkins University
Email: hao@jhu.edu

Stephanie D'Souza
Department of Sociology
Johns Hopkins University
Email: sdsouza5@jhu.edu

March 2019

Abstract

Researchers often face data challenges of panel data that lacks early key information for analyzing the later panel data on outcomes. For example, estimating the impacts of early home environment on elementary school progress is impossible due to such a missing data problem. This paper extends the cross-survey multiple imputation (CSMI) method to ECSMI. The paper develops statistical rationale as well as detailed procedures in an empirical illustration of ECSMI. It borrows the joint distribution of variables in a full model from a donor sample (CNLSY) to impute two early home environment missing by design in a target sample (ECLS-K). The empirical application shows multiple utilities of ECSMI, which is to be further developed for broader application in a next step.

* Prepared for presentation at the 2019 Annual Meetings of Population Association of America, Austin, Texas, April.

Confronting Data Challenges in Research on Early Conditions and Later Outcome Trajectories: Methods of Combining Samples of Panel Data

Introduction

Researchers often face challenges of panel data on later outcomes that lack early conditions. For example, early home environment is typically unavailable in panel data on students' achievement at formal education stages. In essence, these panel data fail to provide the joint distribution of all variables considered given the theoretical rationale and empirical evidence of the importance of early childhood environment (Heckman 2003; Parcel and Dufur 2009; Raudenbush and Eschmann 2015). Data combination methods that "borrow" information on the complete joint distribution from a "donor" sample which could be small is an attractive method (Rendall et al. 2013). This paper extends this method and focuses on the additional challenges when we are interested in outcome trajectories influenced by early conditions. We ask: How do early life circumstances (e.g. birth weight, early family background, and home environment) shape child development and academic performance at various educational stages? What are the returns on public investments in early childhood development and education (e.g. Head Start Programs) at various stages of formal education conditional on parental investment in children?

The framework of the cross-survey multiple imputation (CSMI) method (Rendall et al. 2013) consists of two major principles: the available joint distribution of all variables in the substantive model in the donor sample and missing-by-design of one or few key variables in the target sample. This framework addresses limitations in merging data by individual identities and the accompanying inevitable substantial proportion not merged because the CSMI framework avoids the stringent requirement of ID matching.

The CSMI has certain limitations when applied to empirical data. First, to satisfy the missing-by-design principle, a usual practice is to list-wise delete missing cases in the donor and target samples before performing CSMI. This suggests that the missingness in either samples must be sufficiently small (5% or less). As such, it limits the substantive model to include few variables because small missingness is seldom possible for theory-guided model specifications that include more than just a few variables. Most survey data, and increasingly so, have greater missingness. This problem is acute in panel data.

Investigations into social processes and causal inference require the use of panel data. To take advantage of many existing panel surveys, this paper seeks to confront challenges of panel data combination by extending CSMI (called ECSMI). The primary objective is two-fold. It explicates ECSMI and it illustrates it in an empirical application that multiple imputes early home environment (cognitive stimulation and emotional support) that is missing by design in Early Childhood Longitudinal Study: Kindergarten Class 2010-11 (ECLS-K) from the donor survey Children of National Longitudinal Survey of Youth, 1979 (CNLSY).

Substantive Consideration

Our substantive questions concern how coupled parental and public investment in children impact child educational and developmental outcomes. Our conceptual model integrates sociology of education, economics, and public policy literatures, as summarized below.

The expansion of early childhood programs in recent years has created a complex and varied landscape of program offerings which vary in structure and quality for nearly two-thirds of children enrolled in preschool (Kena et al., 2016). Quality and duration of exposure to early childhood programs are factors influencing kindergarten readiness. High-quality public early education programs have demonstrated benefits in cognitive development particularly for lowincome children at the time of school entry (Camilli et al. 2010; Magnuson et al. 2007; Puma et al., 2012). The promise of early childhood programs to level the playing field for children from different family backgrounds has led to renewed interest in the public sphere. In the context of scaling-up public investment parents may update their investment in children through changing home environment (Parcel and Dufur 2009). The question of how public and parental investments in tandem reshape the educational trajectory through the elementary stage has received less attention.

This advantage of early childhood programs has been shown to fade out in the first few years of elementary school (Puma et al., 2012). Note though that the effects do not fade out uniformly for all programs (Bailey et al. 2017). Early childhood scholars argue that fade-out of academic gains result from attending lower quality elementary schools that are unable to sustain the gains achieved from attending preschool (Currie and Thomas, 1995, 2000). Over the long-term, early childhood programs are associated with impressive gains in high school graduation and young adult outcomes (Doyle et al. 2009; Duncan et al. 1994; Garces et al. 2000; Heckman, 2006, 2012; Nores et al. 2005; Ramey et al. 2000). As such, early childhood programs are increasingly viewed as a significant policy lever for reducing social inequality.

At the same time, much of parental investment is conceptualized in school choice and extracurricular activities (Holme 2002; Stein 2015) and less on the home environment as simultaneous inputs to child development. This highlights a gap in the literature in considering both public and parental investment in the school and home arenas from early childhood to school stages. The instructional regime theoretical framework, which argues that a child's entire learning landscape must be considered when attempting to explain achievement outcomes, provides guidance for such analyses (Raudenbush, 2008; Raudenbush & Eschmann, 2015). Since development and growth does not only occur in educational institutions, attempts to understand inequality in educational outcomes must take into account learning that occurs within school and non-school settings. By considering both sources of investment in children, this analytic approach offers improved insights for trends in student achievement.

A small but growing literature has considered the tandem effects of parental and school institution investments. Parcel, Dufur, and co-authors conducted a literature review of the joint effects on school and family investments in school-age children which suggested that interactions of the two sources of investment do, indeed, explain academic outcomes though findings on the relationship between the two sources varied (Parcel, Dufur, & Zito, 2010). We address the gap in the literature by directing attention to a more systemic characterization of early environment for child educational and developmental outcomes at formal educational stages. Our ECSMI enables empirical investigations under this conceptual model.

Methodological Rationale

Graphical Model Formulation

Figure 1 is a graphical formulation of the two methods. CSMI (the green demarcation) combines data in cross-sectional format. The Target Survey A observes y_{τ}^{A} (e.g., kindergarten readiness) and w (e.g., time-invariant parent and child characteristics) while missing $x_{\tau-1}^{A}$ (e.g., parental investment during ages 3-5). This information was fully observed in Donor Survey B. CSMI, which combines data across surveys A and B, was introduced in Rendall et al. (2013). The two related red circles define ECSMI that combines panel data. For this purpose, we need Donor Survey C with all variables observed over all time points under consideration. CNLSY serves such a purpose. The two sections below lay out CSMI as a foundation on which we develop ECSMI.

(Figure 1 about here)

CSMI

The Target Survey A of a random sample n_1 drawn from a population observed variables $\{y_i^A, x_i^A, w_i^A\}_{i=1}^{n_1}$ at a fixed age τ and the Donor Survey B of a random sample n_2 drawn from the same population observed variables $\{y_i^B, x_i^B, x_{i,\tau-1}^B, w_i^B\}_{i=1}^{n_2}$ at the same fixed age τ . To construct micro data $\{y_i, x_i, x_{i,\tau-1}, w_i\}_{i=1}^{n_1+n_2}$, we clarify three conditions under which CSMI is valid and offer ways to relax them. The first condition is *same* population, i.e., samples A and B are independent probability samples drawn from the same population. Second, for the *same sampling design*, surveys A and B should have used the same sampling design. If the first or second condition is not met, one could use a model-fitting approach (Burnham and Anderson 2002; Rendall et al. 2013; Weakliem 2004) to include an indicator for the surveys and the potential differential estimates by w in the substantive model estimation with complete data. Third, to meet the condition of *same measurement*, we could harmonize variables across surveys.

The two assumptions must hold if we use CSMI.

1. *The joint distribution principle* means that all needed variables must be available in the donor survey. Violating this assumption is equivalent to claiming the hard-to-defend conditional independence, i.e., conditional on the common variables, the missing variable is independent of all other variables in the model. Consequently parameter estimates are attenuated (Meng 1994; Schenker et al. 2010).
2. *The monotone missing pattern principle* states that Target A misses one or more variables purely by design so there is no self-selection at all, giving rise to monotone missing patterns if the common variables in Target and Donor and the unique variables in Donor have no missing values. This allows for sequential or chained imputation with well-defined

conditional probability and generalized linear modeling for categorical left-hand-side variables (Raghunathan et al. 2001; Rubin 1987; Shaffer 1997). This implies that all missing cases in A or B should be list-wise deleted, such that the monotone missing patterns ensure that the conditional distributions in the process of chained imputation are well defined.

ECSMI

Carrying the joint distribution principle and modifying the monotone principle, we extend CSMI to ECSMI. As Figure 1 shows that Target Survey A of a random sample n_1 drawn from a population observed variables $\{y_{it}^A, x_{it}^A, w_i^A\}_{i=1}^{n_1}$ at $t = \tau, ..., T$ over a short period of calendar years. Note that time *t* is defined by age rather than calendar years. Donor Survey C of n_3 children of a random sample of the mothers of the same population of children observed variables $\{y_{it}^C, x_{it}^C, x_{i,\tau-1}^C, w_i^C\}_{i=1}^{n_3}$ at $t = \tau - 1, \tau, ..., T$, over a much longer period of years. The objective is to prepare micro data $\{y_{it}^A, x_{it}^A, x_{i,\tau-1}^A, w_i^A\}_{i=1}^{n_1}$ at $t = \tau - 1, \tau, ..., T$. Here we only aim at "borrowing" the joint distribution in C to multiple impute the missing-by-design $x_{i,\tau-1}^A$ given that the donor data are drawn indirectly from the same population and the observation of the same ages spans a long period.

As for CSMI, ECSMI follows the similar ways to relax the three conditions (same population, same sampling design, and same measurement) underlying the method. Also as for CSMI, ECSMI will hold stronger the joint distribution principle, with observed joint distribution of all variables in the substantive model. We propose a modified principle to reduce the rigidity of the monotone principle, however. In particular, the monotone principle requires list-wise deletion of missing cases that involves a substantial share of cross-sectional samples and this problem is even graver in the case of panel data.

2'. *Missing-by-design and missing-at-random (MAR) principle.* The missing-by-design variables in Target will give rise to partial monotone missing patterns if we do not list-wise delete missing cases of common variables in Target and Donor and the unique variables in Donor. For these observed variables, the missingness is unrelated to the true values of the variable in question after controlling for observed variables in the analysis and the missing mechanism is ignorable (Rubin 1976, 1987). Because MAR principle is widely used in within-survey MI, we carry it onto ECSMI. Without list-wise deleting missing cases within Target and Donor, the ECSMI will identify the partial monotone missing patterns and place them before the non-monotone patterns so as to start the ECSMI from well-defined conditional probability.

A parametric method of within-survey MI is maximum likelihood estimates (Rubin 1987; Shaffer 1997). This method fits an arbitrary pattern of missingness in within-survey MI and requires an assumption of multivariate normal distribution of variables in MI.

For both CSMI and ECSMI, the decision of the number of complete datasets, M , follows Shaffer (1997). The ratio of the variance of the chained MI estimator to a corresponding ML estimator is $1+ f / M$, where $f = (N - n) / N$, the fraction of missing information (Schafer 1997:110). M should be sufficiently large such that the ratio approaches 1. That is, a larger fraction of missing information requires a larger M . If an increase of 2% in the variance is acceptable, then the upper bound for M is about 50. Applying Rubin's rule (1987), we obtain the full model estimates using the M complete datasets.

Empirical Illustration of ECSMI

The illustration of ECSMI answers the research question on how coupled parental and public investment in early childhood shape math and reading achievement trajectories from K to 4th grade, where ECLS-K is Target and CNLSY is Donor. Because CNLSY and ECLS adopt similar measurements of home environment and CNLSY survey instruments primarily draw from the 1988 National Education Longitudinal Study (NELS:88), we could harmonize measurements across ECLS-K and CNLSY without major issues. After harmonization we perform ECSMI to create complete datasets with $M = 50$ using chained imputation with monotone missing patterns ordered first. We estimate a random effect growth curve model for achievement trajectories. The final point estimates and standard errors are obtained with Rubin's rule based on the 50 complete datasets.

Table 1 summarizes the major commonalities and differences between ELCS-K and CNLSY. The two surveys are common in population, probability sampling design, repeated measures of achievement at the same age/grade and major survey instruments. They differ, though, in school-based vs. household-based survey, sample size, child vs. mother representativeness, as well as a lack of early home environment measures by design in ELCS-K.

(Table 1 about here)

Data Source

The ECLS-K: 2011 followed a nationally representative sample of 18,174 children from kindergarten entry in 2010 until fifth grade in 2016. The sampling followed a complex, multistage stratified design. Among a total of nine rounds of data collections, seven rounds tracked the entire sample, specifically: fall kindergarten, spring kindergarten, and in the spring from grades one through five. Data collection included parent interviews, teacher interviews, information on classroom and school environments, and child direct assessments. Data from three time points are selected in our analysis: spring of kindergarten, second grade, and fourth grade. The analytic sample included 9,993 children who were present across the three time points.

The National Longitudinal Survey of Youth Children and Young Adults 1979 (CNLSY) is a longitudinal study of 11,521 children born to women in the National Longitudinal Study of Youth 1979 study. In contrast to the ECLS-K which employed a school-based survey, CNLSY employed a household-based survey, leading to different contents of survey instruments. Biennial data collection began in 1986 and are available through 2014. Extensive data collection included home observations, child developmental outcomes, preschool participation, child demographic characteristics, and family background.

Data and Analytic Preparation for ECSMI

We outline three steps in data and analytic preparation: sample selection, variable harmonization, and pre-ECSMI descriptive statistics.

Sample selection. The sample selection meets the requirement of outcome measured at the similar ages of children. Taking into account the follow-up schedule of CNLSY and ECLS-K, we pick 3 time points: Spring K, Spring 2nd grade, and Spring 4th grade with 2 years apart corresponding to the biennial follow-up of CNLSY. Because CNLSY is a household-based survey, its follow-up schedule is by years rather than by grades. We determine the norm age range as 9.5-10.5 years old for Spring 4th grade. We define the study population as U.S. children in this age range and represented by the sample children in CNLSY. The repeated assessments are the two prior to the one at ages 9.5-10.5 years old, most falling in the age range of 5.5-6.5 years old for Spring K and 7.5-8.5 years old for Spring 2nd grade. This sample selection criterion by age range ensures no self-selection. It also means that the outcome variables and covariates in the study may have missing values. Out of the 11,521 CNLSY children in the total sample, 2,897 meet this sample selection criterion.

To select a comparable sample from ECLS-K 2011, we pick the three waves that correspond to Spring K, Spring 2nd grade, and Spring 4th grade but also restrict the age range to 9.5-10.5 years old at Spring 4th grade. This selection criterion leads to a sample of 9,993 students in ECLS-K 2011. By the same token, the outcome variables and covariates for the analytic sample include missing values.

Harmonization of measures between CNLSY and ECLS-K. The data harmonization does, however, require compromises to account for differences in study designs and data collection instruments. Outcomes are reading and math scores collected across three time points. The reading and math scores in ECLS-K 2011 come from various large-scale early childhood studies. The CNLSY administered the PIAT to measure reading and math skills. These outcomes were standardized based on the original sample. Covariates in common across the two datasets include low birthweight, child race/ethnicity, sex, and age in months at assessment as well as family background such as mother's education level, family poverty status, and parents in first marriage at the time of kindergarten attendance. Early childhood program participation includes three categories: Head Start, preschool which combines prekindergarten and private preschool programs since these programs could not be distinguished in public-use CNSLY data, and other types of care (i.e. parent care, daycare, and other types of informal care arrangements). Home environment in early childhood is only available in CNLSY data. The home environment consists of two dimensions—cognitive stimulation and emotional support—and were constructed using items from the short form of home observation (HOME-SF). Table 2 shows the common variables in Target and Donor and the unique variables in Donor. It also shows the timing of the observation.

(Table 2 about here)

Pre-ECSMI Descriptive Statistics. The weighted distributions of variables in the study from the two samples are shown in Table 3. The two samples are similar in the distribution of gender, low birth weight, mother's education, and child age at assessment of math and reading achievement. Compared to the ECLSK sample, the CNLSY sample includes a smaller percentage of Hispanics and larger percentage of White/Asian/other, a lower poverty rate, a lower Head Start participation rate, and higher mean math and reading scores. Some of these differences – percent Hispanic and poverty rate – may reflect broarder demographic and economic trends in the country. The increased Head Start rate may capture the scaled-up early childhood education intervention over the years.

(Table 3 about here)

The descriptive statistics suggest that in this case when the observation of the donor sample spans in a much wider period, ECSMI is better used for the full model analysis of the target sample only with the imputed one or few variables missing by design with the joint distribution of all variables in analysis from the donor sample. In other words, we do not combine two samples in the substantive analysis.

Post-ECSMI Comparison of Reduced and Full Models

Table 4 shows the random-effects growth curve estimates from the reduced and full models for math and reading, respectively. First, the significant, substantial promoting effects of home environment suggest the greater explanatory power of the full model than the reduced model. In particular, an increase in one standard deviation in home cognitive stimulation increase 0.18 standard deviation of math and reading scores. The magnitude of effects for home emotional support is about one third of the home cognitive stimulation effect. Second, the estimates for other covariates retain the same sign and significance level, consistent with what our conceptual model expected. While the magnitudes appear to be considerably different, we developed a procedure to evaluate and interpret the result below.

(Table 4 about here)

Figure 2 compares estimates from reduced models between CNLSY and ECLS-K 50 within-survey multiple imputations. The reduced model results are similar between the two datasets with exceptions. For math, mother's education and having two parents in the 1st marriage have significantly different effects. For reading, Head Start shows differential effects. These results capture the multivariate distribution difference between the two samples and should be taken into account in understanding and interpreting the full model results after ECSMI.

(Figure 2 about here)

Figure 3 compares estimates from reduced and full models using ECLSK data after ECSMI with CNLSY as the donor. Including early home environment in the full model improves the model fit. On reading, we could substantiate the weaker effect of mother's education in the full model than that in the reduced model for reading using the result from Figure 2 where mother's education has a similar effect on reading in the reduced model between CNLSY and ECLS-K, suggesting that there is no obvious influence by CNLSY sample. In contrast, we could not make the same statement for math because Figure 2 shows that the weaker mother education effect may be due to the influence of CNLSY sample.

(Figure 3 about here)

Concluding Remarks

This paper has developed ECSMI that extends CSMI to address the challenges facing researchers who use panel data to study early precursors of later outcome processes and trajectories. We have developed statistical rationale for ECSMI as well as implemented procedures through an empirical illustration. Motivated by the need to understand the role of early childhood home environment for children's achievement trajectories over formal education stages, we apply ECSMI to ECLS-K 2011 using CNLSY as the donor survey that supplies the joint distribution of a fuller specification of the substantive model, especially including early home environment. Results from analyzing the pre-ECSMI and post-ECSMI data show the stability of common covariate estimates between the reduced and full models, the greater explanatory power of the full model than the reduced model, and the correction for potential biased estimates in the reduced model. To broaden the application of ECSMI is our next step.

Table 1. Survey Description

	ECLS-K 2011	CNLSY
Survey organization	NCES	BLS
Type of survey	School-based	Household-based
Population	US Kindergarten students	US births to NLSY79 female respondents
Sampling design	Multistage, stratified	Multistage, stratified
Representativeness	Nationally representative	Children of a sample of nationally representative women 14-21 in 1978
Total sample size	18,174	11,521
Panel sample	9,993 (assessed in spring 4th grade and biennially before)	2,879 (assessed at appropriate age for spring 4^{th} grade and biennially before)
Observation	2010 - 2015	1986 – 2014
Survey instruments	Parent surveys; child assessments;	Parent surveys; child assessments; home observations

Table 2. Available and Missing Variables across CNLSY and ECLS-K 2011

Grade Age	 birth	 3-5	sp k 5.5-6.5	sp 1st grade 7.5-8.5	sp 2nd grade 9.5-10.5
CNLSY					
Birth weight	x				
Family background					
Home environment		x			
Preschool program		x			
Reading/Math			x	x	x
ECLS-K 2011					
Birth weight	x				
Family background					
Home environment		impute			
Preschool program		x			
Reading/Math			x	x	x

Table 3. Weighted Proportion/Mean of Variables in ECSMI

Variable	CNLSY	ECLS-K 2011
Male	0.50	0.51
Hispanic	0.07	0.26
Black	0.14	0.13
White/Asian/other	0.79	0.61
Birthweight < 5.5 oz	0.07	0.08
Mother years of schooling at K	13.14	13.90
Family poverty at K	0.16	0.24
Mother in 1st marriage at K	0.62	0.70
Head start	0.14	0.20
Preschool	0.56	0.59
Other care	0.30	0.21
Home cognitive stimulation at K	0.22	--
Home emotional support at K	0.16	--
Math score, spring K	0.09	0.03
Reading score, spring K	0.12	0.04
child assessment age in month, spring K	72.07	72.72
Math score, spring 2nd grade	0.19	0.03
Reading score, spring 2nd grade	0.15	0.05
child assessment age in month, spring 2nd grade	96.08	96.72
Math score, spring 4th grade	0.32	0.02
Reading score, spring 4th grade	0.17	0.05
child assessment age in month, spring 4th grade	119.81	120.18
n	2,879	9,993

Notes: The weighted distributions are for each sample before ECSMI.

Table 4. Estimates of Reduced and Full Models for ECLS-K 2011: ECSMI Imputed Early Home Environment in ECLS-K 2011 with CNLSY as the Donor

	Math		Reading	
Variable	M0	M1	M0	M1
Male	0.076***	0.048**	-0.184***	-0.206***
	(0.016)	(0.020)	(0.016)	(0.019)
Black (ref: Hispanic)	-0.249***	-0.177***	-0.034	0.023
	(0.030)	(0.036)	(0.030)	(0.035)
White/Asian/oth (ref: Hispanic)	0.287***	0.205***	0.176***	0.100***
	(0.021)	(0.025)	(0.020)	(0.025)
Birth weight<=5.5 oz	-0.232***	-0.228***	-0.178***	-0.175***
	(0.033)	(0.034)	(0.032)	(0.033)
Mother yrs schooling at K	0.082***	0.060***	0.092***	0.072***
	(0.004)	(0.005)	(0.004)	(0.005)
Family poverty at K	-0.203***	-0.114***	-0.252***	-0.174***
	(0.025)	(0.029)	(0.025)	(0.028)
Mother in 1st marriage at K	0.158***	0.122***	0.122***	0.092***
	(0.020)	(0.022)	(0.020)	(0.021)
Head start (ref: preschool)	-0.110***	-0.053*	-0.135***	-0.086***
	(0.025)	(0.028)	(0.024)	(0.026)
Parent, day, other care (ref: preschool)	-0.125***	-0.068***	-0.149***	-0.100***
	(0.025)	(0.027)	(0.023)	(0.025)
Early home cognitive stimulation	--	0.188***	--	0.178***
		(0.023)		(0.026)
Early home emotional support	--	0.076***	--	0.051**
		(0.024)		(0.023)
Child assessment age in months	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	-1.322***	-0.941***	-1.242***	-0.898***
	(0.061)	(0.080)	(0.060)	(0.079)
Observations	29,979	29,979	29,979	29,979
Number of children	9,993	9,993	9,993	9,993

Notes: Using ECSMI we imputed home cognitive stimulation and emotional support for ECLS-K 2011. Only the imputed ECLS-K data (50 completes) are used in the analysis. M0 is the reduced model and M1 the full model. Estimates from growth curve models with random effects are reported.
Standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

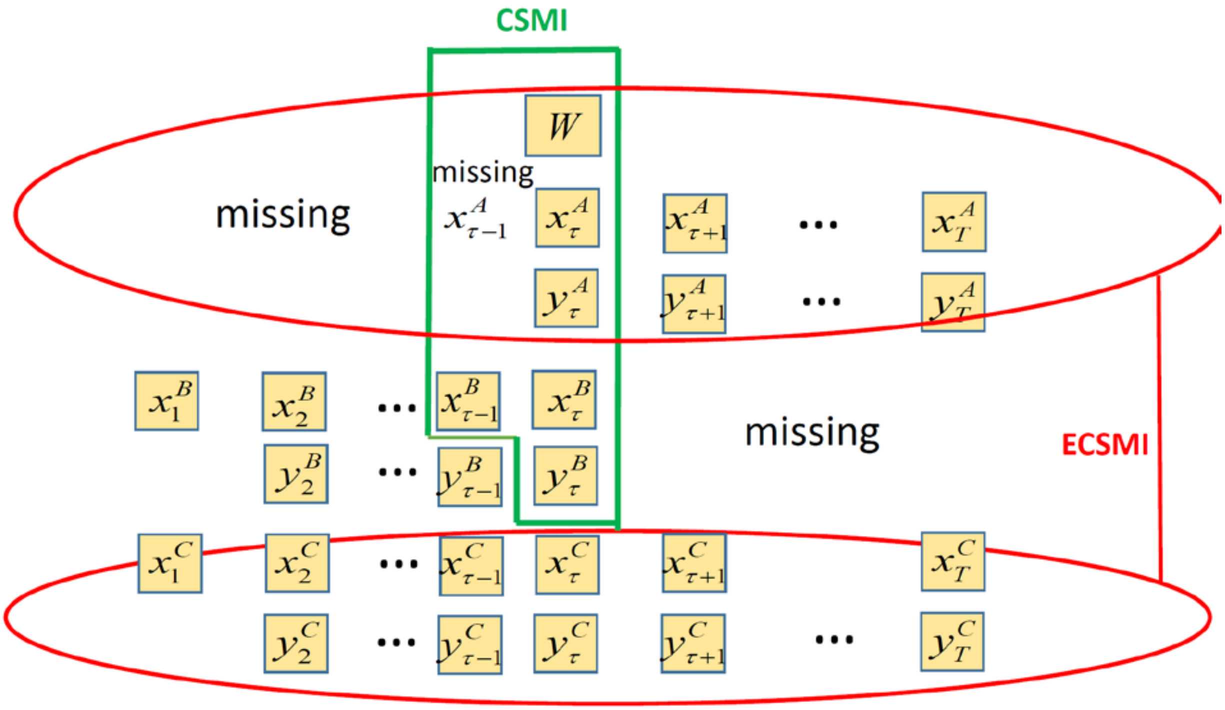
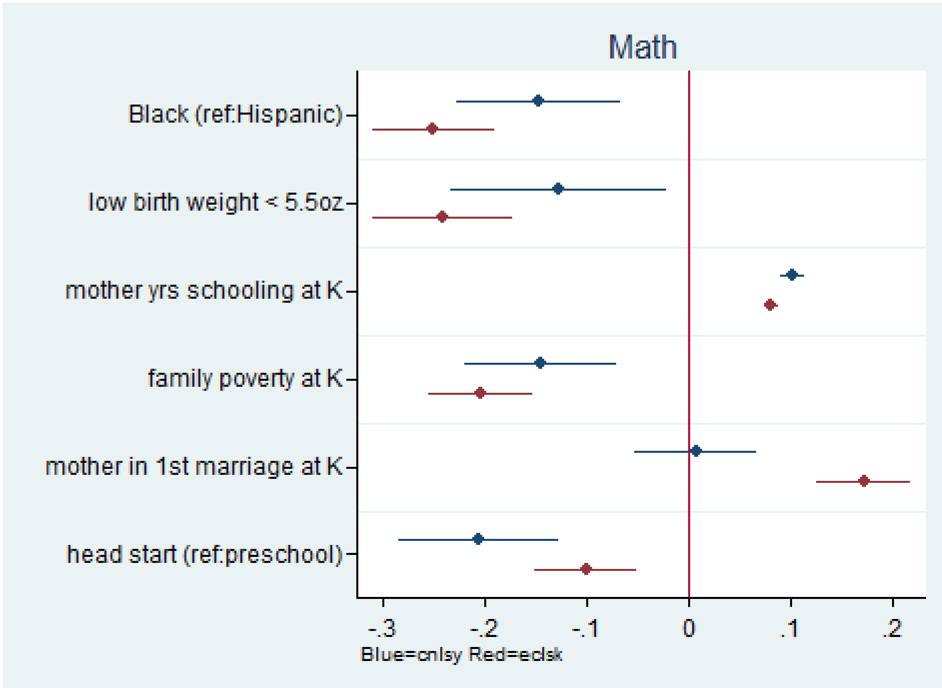
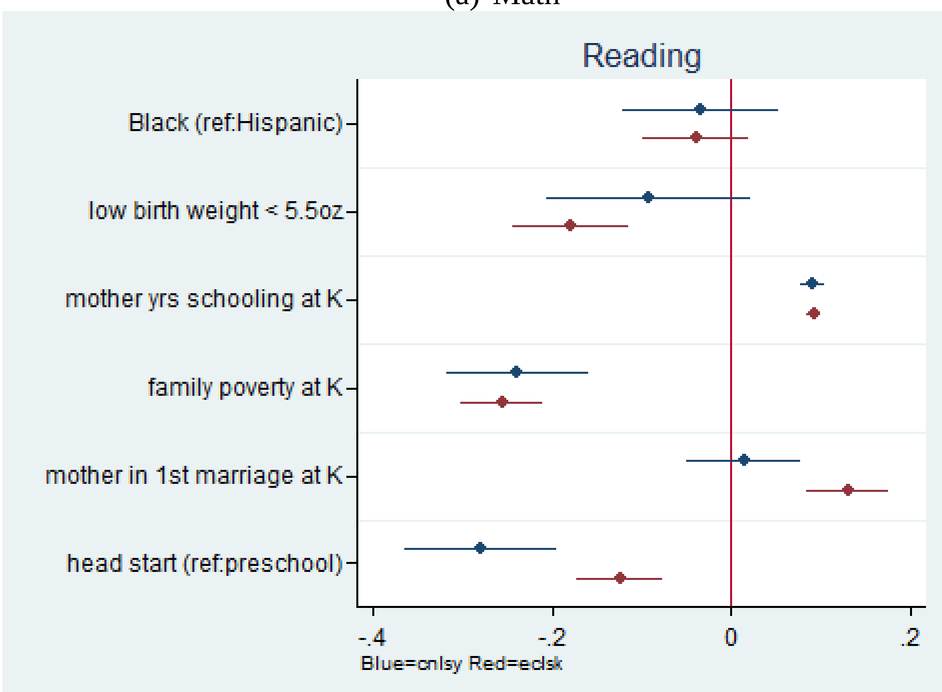


Figure 1. Graphical Formulation of Data Combination Methods
Target Survey A (ECLS-K)
Donor Survey B (ECLS-B)
Donor Survey C (CNLSY)
CSMI: cross-survey multiple imputation for data for cross-sectional data (Target Survey A and Donor Survey B)
ECSMI: extended cross-survey multiple imputation for panel data (Target Survey A and Donor Survey C)

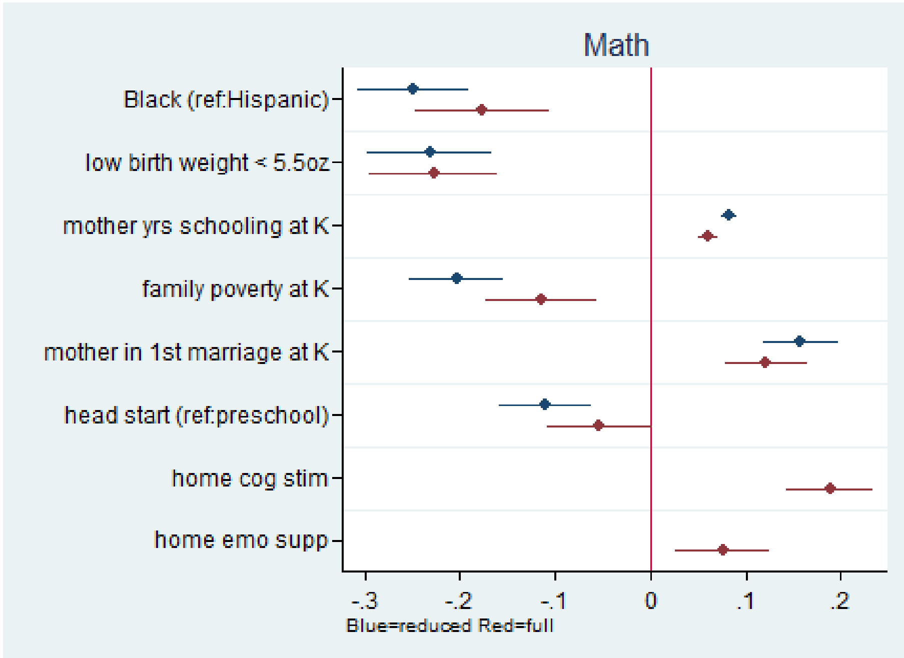


(a) Math

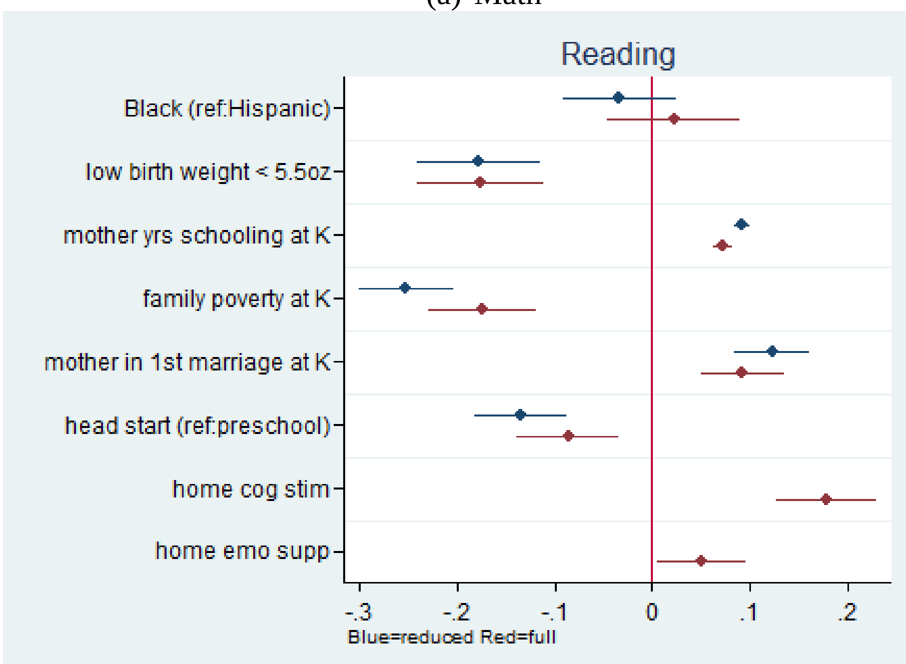


(b) Reading

Figure 2. Comparing Estimates from Reduced Models between CNLSY and ECLSK 50 within-survey multiple imputations. The reduced model results are similar between the two datasets with two exceptions. On math, mother's education and having two parents in the 1st marriage have significant different effects. On reading, Head Start shows differential effects. These results capture the multivariate distribution difference between the two samples and should be taken into account in understanding and interpreting the full model results after ECSMI.



(a) Math



(b) Reading

Figure 3. Comparing Estimates from Reduced and Full Models using ECLSK Data after ECSMI with CNLSY as the Donor. Including early home environment in the full model improves the model fit. On reading, we could substantiate the weaker effect of mother's education in the full model than that in the reduced model for reading using the result from Figure 2 where mother's education has a similar effect on reading in the reduced model between CNLSY and ECLS-K, suggesting that there is no obvious influence by CNLSY sample. In contrast, we could not make the same statement for math because Figure 2 shows that the weaker mother education effect may due to the influence of CNLSY sample.

References

Bailey, D, Duncan, G, Odgers, C, and Yu, W. 2017. Persistence and Fadeout in the Impacts of Child and Adolescent Interventions. *Journal of Research on Educational Effectiveness*, 10(1), 7–39.

Burnham, KP and Anderson, DR 2002. Model selection and multimodel interference. *Springer Verlag, New York*.

Camilli, G, Vargas, S, Ryan, S, and Barnett, W. 2010. Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record*, *112*(3), 579–620.

Currie, J, and Thomas, D. 1995. Does Head Start Make a Difference? *The American Economic Review*, 85(3), 341–364.

Currie, J, and Thomas, D. 2000. School Quality and the Longer-Term Effects of Head Start. *The Journal of Human Resources*, 35(4), 755.

Doyle, O, Harmon, C, Heckman, J, and Tremblay, R. 2009. Investing in early human development: Timing and economic efficiency. *Economics & Human Biology*, 7(1), 1–6.

Duncan, G, Brooks-Gunn, J, and Klebanov, P. 1994. Economic deprivation and early childhood development. *Child Development*, 65, 296–318.

Garces, E, Thomas, D, and Currie, J. 2002. Longer-Term Effects of Head Start. *American Economic Review*, 92(4), 999–1012.

Gelman, SA, Coley, JD, Rosengren, KS, Hartman, E, Pappas, A, and Keil, FC. 1998. Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child development*, i-157.

Heckman, J. 2006. Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*, 312(5782), 1900–1902.

Heckman, J. 2012. The case for investing in young children. *Defending Early Childhood*, 235–242.

Holme, J. 2002. Buying Homes, Buying Schools: School Choice and the Social Construction of School Quality. *Harvard Educational Review*, 72(2), 177-205.

Kena G, Hussar W, McFarland J, de Brey C, Musu-Gillette L, Wang X, Zhang J, Rathbun A, Wilkinson-Flicker S, Diliberti M, Barmer A. 2016. The Condition of Education 2016. NCES 2016-144. *National Center for Education Statistics.*

Magnuson, K, Ruhm, C, and Waldfogel, J. 2007. Does Prekindergarten Improve School Preparation and Performance? *Economics of Education Review*, *26*(1), 33–51.

Meng, XL. 1994. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 538-558.

Nores, M, Belfield, C, Barnett, W, and Schweinhart, L. 2005. Updating the Economic Impacts of the High/Scope Perry Preschool Program. *Educational Evaluation and Policy Analysis*, 27(3), 245–261.

Parcel, T, and Dufur, M. 2009. Family and school capital explaining regional variation in math and reading achievement. *Research in Social Stratification and Mobility*, *27*(3), 157–176.

Puma, M, Bell, S, Cook, R, Heid, C, Broene, P, Jenkins, F, and Downer, J. 2012. *Third Grade Follow-up to the Head Start Impact Study Final Report (OPRE Report 2012-45)*. Raghunathan, TE, Lepkowski, JM, Van Hoewyk, J, and Solenberger, P. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, *27*(1), 85-96.

Ramey, C, Campbell, F, Burchinal, M, Skinner, M, Gardner, D, and Ramey, S. 2000. Persistent

effects of early childhood education on high-risk children and their mothers. *Applied Developmental Science*, 4(1), 2–14.
Rendall, M, Ghosh-Dastidar, B, Weden, MM, Baker, EH, and Nazarov, Z. 2013. Multiple imputation for combined-survey estimation with incomplete regressors in one but not both surveys. *Sociological methods & research*, *42*(4), 483-530.
Rubin, D.B. 1987. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, *82*(398), 543-546.
Rubin, DB. 1997. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, *127*, 757-763.
Schenker, N, Raghunathan, TE, and Bondarenko, I. 2010. Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in medicine*, *29*(5), 533-545.
Schafer, JL. 1997. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.
Stein, ML. 2015. Public school choice and racial sorting: An examination of charter schools in Indianapolis. *American Journal of Education*, *121*(4), 597-627.
Tighe, E, Livert, D, Barnett, M, and Saxe, L. 2010. Cross-survey analysis to estimate low-incidence religious groups. *Sociological Methods & Research*, *39*(1), 56-82.
Van Hook, J, Bachmeier, JD, Coffman, DL, and Harel, O. 2015. Can we spin straw into gold? An evaluation of immigrant legal status imputation approaches. *Demography*, *52*(1), 329-354.
Weakliem, DL. 2004. Introduction to the special issue on model selection. *Sociological Methods & Research*, *33*(2), 167-187.