NIH Public Access
Author Manuscript
Sociol Methods Res. Author manuscript; available in PMC 2014 November 01.

Published in final edited form as:
Sociol Methods Res. 2013 November 1; 42(4): . doi:10.1177/0049124113502947.

Multiple Imputation For Combined-Survey Estimation With Incomplete Regressors In One But Not Both Surveys

Michael S. Rendall[1,2], **Bonnie Ghosh-Dastidar**[2], **Margaret M. Weden**[2], **Elizabeth H. Baker**[3,2], and **Zafar Nazarov**[4,2]

[1]University of Maryland, College Park

[2]RAND Corporation

[3]University of Alabama, Birmingham

[4]Cornell University

Abstract

Within-survey multiple imputation (MI) methods are adapted to pooled-survey regression estimation where one survey has more regressors, but typically fewer observations, than the other. This adaptation is achieved through: (1) larger numbers of imputations to compensate for the higher fraction of missing values; (2) model-fit statistics to check the assumption that the two surveys sample from a common universe; and (3) specificying the analysis model completely from variables present in the survey with the larger set of regressors, thereby excluding variables never jointly observed. In contrast to the typical within-survey MI context, cross-survey missingness is monotonic and easily satisfies the Missing At Random (MAR) assumption needed for unbiased MI. Large efficiency gains and substantial reduction in omitted variable bias are demonstrated in an application to sociodemographic differences in the risk of child obesity estimated from two nationally-representative cohort surveys.

I. INTRODUCTION

Frequently a social scientist has a choice of more than one survey that he or she could use to analyze a given social phenomenon occurring at a given time. The survey with the best set of predictor variables will typically be chosen, as to do otherwise would risk introducing omitted variable bias. This survey may suffer, however, from a sample size that is too small to detect true relationships between variables of interest to the researcher. For a recent review of studies facing this type of trade-off, see Rendall et al (2011). Standard methods for multivariate analysis rely on “rectangular” datasets (all predictor variables are present for all observations), thereby preventing analyses that pool observations across surveys without the same, complete set of predictor variables. The problem of missing predictor variables and consequent non-rectangular datasets, however, is not unique to analysis with pooled surveys. It also frequently confronts a researcher using a single survey, due to survey item non-response (Allison 2002; Little and Rubin 2002). Standard analysis methods for rectangular datasets require the discarding of entire observations if item non-response occurs for even one variable that belongs in the regression model, a practice sometimes referred to as “complete case analysis.” In response to this apparently wasteful treatment of survey information, “missing data” methods of analysis that combine incomplete observations with complete observations have been developed and are now used widely in the social and health sciences (Schafer and Graham 2002; Raghunathan 2004). The goal of the present

Please address correspondence to Michael S. Rendall at mrendall@umd.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

study is to show that missing data methods developed for handling non-response in single surveys can be profitably applied to pooled analysis of surveys in which predictor variables are “missing” from one or more surveys.

Among missing-data methods, multiple imputation (MI, Rubin 1987) offers a flexible and statistically rigorous option. Little and Rubin (1989) argued for social scientists to consider the efficiency advantages of MI over complete-case analysis, and to consider the implementation advantages of MI over “direct methods” that combine separate likelihoods for incomplete observations and complete observations within a single survey. These implementation advantages arise primarily from the separation of the imputation step from the target, post-imputation analysis. We refer to this standard use of MI as “within-survey MI.” Successful early adoptions of within-survey MI in sociology and demography include studies by Freedman and Wolf (1995), Goldscheider et al (1999), and Sassler and McNally (2003). Within-survey MI is now used frequently in the social sciences to handle item non-response, and MI software is available in the major statistical packages (Johnson and Young 2011).

A quite different context for the potential application of MI is to impute values from one survey to a second survey in which that variable is not present by design ---- that is, no question was asked and no other form of assessment was undertaken in the second survey. The value is then missing for every case in the second survey. We refer to MI undertaken in this circumstance as “cross-survey MI.” When additionally the observations from both surveys are pooled for the post-imputation analysis, we refer to this as “pooled cross-survey MI.” In the social sciences, we know of only one study that has implemented cross-survey MI ---- that of Gelman, King, and Liu’s (1998a) development of a Bayesian hierarchical model for MI across multiple public opinion surveys in a political science analysis. The two-survey context we address in the present study is crucially different from Gelman et al’s multiple-survey context, as only a multiple-survey context admits as a solution the parameterized hierarchical model they propose to account for survey design differences. We address the challenge of accounting for differences in survey design in the two-survey context with a model-fitting approach that compares pooled-survey models respectively with and without regressors that indicate in which survey the observation is found.

We also address explicitly the “variables never jointly observed problem” of cross-survey MI. The cross-survey MI method was first proposed by Rubin (1986), but in the context of the “statistical matching” of surveys, each survey with one or more variables not present in the other. The resulting problems of post-imputation analysis with variables never jointly observed have discredited the statistical matching approach in the social sciences (Rodgers 1984, Ridder and Moffitt 2007), and have left it largely on the margins of statistics (Judkins 1998; but see also Rässler 2002 and Moriarity and Scheuren 2003). By insisting that the post-MI analysis be specified from variables completely present in one of the two surveys (the “impute-from” survey), and imputing variables only to the survey that we designate to be incomplete (the “impute-to” survey), we propose a form of cross-survey MI that avoids the “variables never jointly observed” problem.

The remainder of the paper is structured as follows. In section II immediately below, we describe the relationship of cross-survey MI to both within-survey MI and to direct methods for combining data sources using Maximum Likelihood and Generalized Method of Moments estimators. We then describe our proposed adaptations of procedures and principles from both within-survey MI and direct data-combining methods to implement cross-survey MI most effectively. To assess the suitability of the two surveys for combined analysis, we propose a model-fitting approach using an analysis model specified from variables in common between the two surveys. In section III, we demonstrate cross-survey

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

MI in an application to sociodemographic differences in the risk of early childhood obesity estimated from two nationally-representative cohort surveys. Our model-fit statistics largely support the assumption that the two surveys sample from a common universe, though we also illustrate the use of a restriction on the target population to handle the circumstance in which that support is equivocal. Compared to estimation using only the smaller survey, we show empirically that large efficiency gains in estimates of coefficients are achieved for variables in common between the surveys and that a reduction in overall sampling bias is also achieved. Compared to estimation using only the larger survey, we show that notable reductions in omitted variable bias are achieved by our having imputed a key predictor variable available only in the smaller survey. Discussion follows in section IV.

II. RELATIONSHIPS OF CROSS-SURVEY MI TO WITHIN-SURVEY MI AND TO DIRECT METHODS FOR COMBINING DATA SOURCES

As noted by Hellerstein and Imbens (1999), multiple imputation (MI) may be viewed as an alternative method to the imposing of moment restrictions from a larger data source with fewer predictor variables to an equation estimated entirely from observations from a smaller data source but with more predictor variables. Although the analysis may be conducted from the smaller sample survey alone, more efficient estimation can be achieved by additionally exploiting information in the larger data source. The adverse effects of any sampling biases in the smaller sample survey may also be mitigated by anchoring estimates from the richer covariance structure of the smaller data source to the more population-representative larger data source.

Imbens and Lancaster (1994) reported large gains in efficiency by incorporating marginal moments from census data with sample-survey joint distributions using a Generalized Method of Moments estimator. Handcock, Huovilainen, and Rendall (2000) developed a Constrained Maximum Likelihood estimator (MLE) to impose restrictions from birth-registration and population-estimate data on a model estimated from sample-survey data and similarly reported large efficiency gains. Rendall et al (2008) showed that efficiency gains on a greater range of coefficients can be obtained by augmenting a small survey with both population-level data about bivariate relationships and additional surveys with data on a limited set of multivariate relationships. Hellerstein and Imbens (1999) considered the circumstance in which the data providing the moment restrictions are not from a population-level data source but instead from a large sample survey, and derived an expression for efficiency loss due to sampling error in the large survey.

Hellerstein and Imbens also considered in some detail the case in which the smaller and larger data sources do not sample from exactly the same population. They described the larger source (the cross-sectional Current Population Survey, CPS) as being a probability sample of the “target population” and the smaller source (the panel survey National Longitudinal Survey, Young Men’s Cohort, NLS) as being subject to attrition bias and describing it as being from the “sampled population” remaining after attrition. By weighting the NLS data to the CPS joint distribution of a dependent variable (log wages) and limited set of predictor variables, they showed that an unbiased larger data source can be used to partially correct for bias in the smaller data source. Handcock, Rendall, and Cheadle (2005) extended this approach to a constrained MLE estimator of differences in marital fertility between black and white women in which the smaller, panel survey (the Panel Survey of Income Dynamics, PSID) was similarly subject to non-response bias, whereas the data on the target population (registered U.S. births compiled by the National Center for Health Statistics, NCHS, matched to Census Bureau population estimates by marital status) were considered to be unbiased. Rendall, Handcock, and Jonsson (2009) considered the case in which the larger data source was also biased and, using a subjective Bayesian prior for the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

magnitude of that bias, showed that large reductions in mean square error can nevertheless be achieved by combining information from the larger data source in a regression using observations from the smaller data source.

Data-combining methods that impose moment restrictions or constraints, however, quickly become computationally unwieldy as successively more covariate information is included from the larger data source. Moreover, it is often the case that more covariate information may be added only when the larger data source is no longer very large. In this case there may be no clear designation of one of the two surveys as being the large, unbiased survey against which the other survey should be calibrated. More general methods that allow for more equal contributions to the model estimates from the observations of the two data sources are then needed. This argues for consideration of data-combining methods in which observations from the first survey, and not merely aggregate moments computed from it, are pooled with the observations from the second survey.

We argue for the extension of within-survey MI methods to the treatment of the observations from one survey as the "complete cases" and the observations from the other survey as the "incomplete cases." Compared to within-survey imputation, this cross-survey imputation requires larger numbers of imputations to compensate for the higher fraction of missing values for variables missing entirely in the larger survey. It also needs to address sample design and measurement differences between surveys. Cross-survey MI, however, has three desirable features not found in within-survey MI. Most obviously, major efficiency gains can be realized by pooling observations across surveys. This is analogous, but additional to, efficiency gains obtained by combining incomplete and complete observations from the same survey in within-survey MI. Second, in cross-survey MI the variables to be imputed are "missing-by-design" (Raghunathan and Grizzle 1995), meaning that the reason an individual is "missing" a value for a variable of interest is that the survey did not ask the question. This is very different from values that are selectively missing due to item non-response, or to dropout in a panel survey. It has the statistically important consequence that the "Missing at Random" (MAR) assumption will be more easily met in cross-survey MI than in within-survey MI. Third, the missing-by-design structure of the pooled observations used in cross-survey MI implies that missingness has a monotone rather than arbitrary pattern (Rubin 1987). An individual's being sampled in one survey and not the other means he or she will have missing values for all variables derived from questions not asked in that survey (but asked in the other). The resulting "monotone" missing pattern allows for sequential imputation with separate models for categorical and continuous variables (Raghunathan et al 2001). In within-survey MI the pattern of missingness is generally "arbitrary" (missing values on one variable does not invariably imply missing values on a second variable and vice versa). Consequently a convenient parametric multivariate model such as the multivariate normal (Schafer 1997) is assumed for joint-distribution imputation.

A cross-survey MI setup that allows for the application of methods and results both from within-survey MI and from direct methods of combining data

We describe data-combining in the typical case of a "smaller survey" (Survey 1) that has fewer observations but more regressors and a "larger survey" (Survey 2) that has more observations but fewer regressors. More generally, however, we can and do refer to Survey 1 as the "impute-from" survey and Survey 2 as the "impute-to" survey, and nothing in our results requires that Survey 1 actually be smaller than Survey 2. We assume that Survey 1 has outcome variable Y and predictor variables X_1 and X_2 for a sample of size N_1, and that Survey 2 has outcome variable Y and predictor variable X_2 for a sample of size N_2. We assume that no values of Y, X_1, or X_2 are missing in Survey 1 and that no values of Y or X_2 are missing, but all values of X_1 are missing, in Survey 2. The goal of the analysis is to estimate the parameters and standard errors of a multivariate regression model that includes

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

observations from both surveys and that is specified from the survey with the fullest set of available regressors. We consider the special case of a binary outcome variable Y and the logit model, $LOGIT[p] = \ln[p/(1-p)]$, for the regression:

$$LOGIT[\Pr\{Y|X_1, X_2\}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (1)$$

Although X_1 and X_2 are predictor variables that will be assumed first to be scalar (single regressors), they may easily be generalized to vectors of regressors.

We first make the assumption that the surveys randomly sample from a common universe using equivalent survey instruments and sampling designs. This is the context examined by Raghunathan and Grizzle (1995) in which sample components of a single survey are assigned survey questions according to a "missing-by-design" plan. Under these conditions, it is clear that the methods of within-survey MI with complete and incomplete cases apply equivalently to cross-survey MI. Survey 1 provides the complete cases $\{Y_i, X_{1i}, X_{2i}\}_{i=1}^{N_1}$ and Survey 2 provides the incomplete cases $\{Y_j, X_{2j}\}_{j=1}^{N_2}$. Standard within-survey imputation methods (Rubin 1987) may then be applied to derive estimates of the parameters and standard errors of Model (1), as follows. An imputation model for $E[X_1 \mid X_2, Y]$ is first estimated as a regression of X_1 on X_2 and Y using Survey 1 observations only. Second, using the parameters estimated in the imputation model from the data of Survey 1, together with the values of X_2 and Y in Survey 2, a value for X_1 is imputed by randomly drawing m times for each of the N_2 observations in Survey 2. Third, each of these versions of Survey 2 observations containing a different randomly imputed value for X_1 is concatenated with the N_1 ('complete') observations in Survey 1 to create m 'completed' datasets each of size $N_1 + N_2$. Fourth, the analysis model (1) is estimated on each of the m completed datasets. Indexing the m datasets by k, the analysis model produces m unique realizations $\hat{\beta}_k = \{\hat{\beta}_{1k}, \hat{\beta}_{2k}\}$ of parameter vector $\beta = \{\beta_1, \beta_2\}$. Standard multiple imputation algorithms (Schafer 1997, p. 109) are used to combine $\{\hat{\beta}_k\}_{k=1}^{m}$ to derive the final parameter estimates and their standard errors. The final parameter estimates are derived as simple averages over all m estimates $\hat{\beta} = \{(1/m)\sum_{k=1}^{m}\hat{\beta}_{1k}, (1/m)\sum_{k=1}^{m}\hat{\beta}_{2k}\}$. Let $\bar{U}$ represent the mean within-imputation variance $\overline{U} = (1/m)\sum_{k=1}^{m} Var(\hat{\beta}_k)$ and B represent the between-imputation variance $B = \frac{1}{m-1}\sum_{k=1}^{m}(\hat{\beta}_k - \hat{\beta})^2$. The final standard error estimates about the parameters are then derived as $S.E.(\hat{\beta}) = \sqrt{Var(\hat{\beta})} = \sqrt{\overline{U} + (1+1/m)B}$. The last term $(1+1/m)B$ represents the upward adjustment to the standard errors to account for the imputation of X_1 in the N_2 cases among the $N_1 + N_2$ cases pooled over Surveys 1 and 2 for the analysis.

Note that in the case that X_2 is a vector, sometimes in the practice of MI an imputation model to predict X_1 will be specified with regressors Y and a subset only of the variables in the X_2 vector. This is called an "uncongenial" method of imputation (Meng 1994) because variables in the analysis model are omitted from the imputation model. Uncongenial imputation has two theoretical drawbacks relevant to implementing and evaluating the performance of cross-survey MI. First, analytical expressions of variance reduction of Maximum Likelihood (ML) methods do not then apply directly to variance reduction in MI. Second, uncongenial imputation increases bias, as leaving out variables in the imputation model that are present in the analysis model may attenuate the analysis model regression coefficients (Schenker, Raghunathan, and Bondarenko 2010). Imputation may also be

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

described as uncongenial when variables are not used in the imputation model estimated from Survey 1 because they were not in Survey 1, but are in Survey 2 and are included in the analysis model on that basis. We discuss this below as the problem of having variables in the analysis model that are "never jointly observed" in either Survey 1 or in Survey 2. This is a circumstance in which we advise against the use of cross-survey MI.

Theoretical results for variance reduction when combining complete and incomplete observations over estimation with only the complete observations were derived in the linear regression case by Little (1992), subsequently extended by White and Carlin (2010). A key parameter in evaluating gains to cross-survey MI is the "fraction missing." Following the terminology developed for within-survey MI in White and Carlin, we define the fraction with missing values of X_1 by $\pi = N_2/(N_1 + N_2)$. This fraction missing can be considered either to be the fraction of incomplete cases in a single survey, or in our case to be the fraction of cases that come from the second of the two surveys.

The principal, or possibly only, variance reductions will then be in $Var(\beta_2)$, the parameter for which the regressor variable (X_2) is observed in both Survey 1 and Survey 2. Variance reduction about $Var(\beta_2)$ will depend not only on the fraction missing π but also on the correlation ρ_{12}^2 between X_2 and X_1, and on the partial correlation of Y and X_1 given X_2, $\rho_{1y.2}^2$. The expression for the proportion by which $Var(\beta_2)$ is reduced by adding observations from Survey 2 in the linear regression case is given by (White and Carlin 2010, p. 2922):

$$\pi(1-\rho_{1y.2}^2)[1-\rho_{12}^2(1-2\rho_{1y.2}^2)] \quad (2)$$

In the special case of no correlation between X_2 and X_1 and when X_1 has no association with of Y independent of variation in X_2, then $Var(\beta_2)$ reduces by the maximum amount, equal to the fraction of observations in Survey 2, π. In this case, however, we could estimate β_2 without the need for MI. Instead we would simply pool Survey 1 and Survey 2 observations and estimate $LOGIT[\Pr\{Y \mid X_1, X_2\}] = \beta_0 + \beta_2 X_2$. In the more relevant case of $0<\rho_{1y.2}^2<1$ and $0<\rho_{12}^2<1$, cross-survey MI will always result in a reduction in $Var(\beta_2)$ in the linear regression case because both ($1-\rho_{1y.2}^2$) and [$1-\rho_{12}^2(1-2\rho_{1y.2}^2)$] will always be less than 1.

In the general case, reductions in $Var(\beta_1)$ will be negligible unless X_1 and X_2 are very highly correlated, given that the observations from Survey 2, in which X_1 is always missing, contribute no direct information about the relationship of X_1 to Y. This result of negligible reductions in coefficients about non-common variables between the two data sources was also found empirically by Imbens and Lancaster (1994), and by Handcock et al (2005) who referred to these β_1-type coefficients as being "indirectly constrained" only. White and Carlin (2010, p. 2930) claim, moreover, that in the specific case of binary Y there is never any reduction in $Var(\beta_1)$ achieved through adding the observations from Survey 2. It is not clear, however, if this claim applies to the logistic model only or to all binary outcome models.

In addition to a "congenial" specification of the imputation equation, the number of imputations, m, needs to be sufficiently large for the MI variances $Var(\beta)$ to approach the variances of the Maximum Likelihood (ML) estimator. This approximation of the MI variances to those of ML as the number of imputations m becomes large follows from the expression for the ratio of the variance of the MI estimator to a corresponding ML estimator as $1+ f/m$, where f is "the fraction of missing information" for the parameter (Schafer 1997, p. 110). The larger is f, the higher is the number of imputations needed to make MI nearly as

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

efficient as ML, but given that f has an upper bound of 1 then f/m will always quickly converge to 0.

Cross-Survey MI with Differences in Survey Designs

We next relax the assumption that the surveys randomly sample from a common universe using equivalent survey instruments and sampling designs. Incorporating complex survey design features can be important for both within-survey MI for non-response (Reiter, Raghunathan, and Kinney 2006) and therefore potentially even more so for cross-survey MI in which clustering and strata designs will differ between the imputed-from and imputed-to surveys. No agreed-upon set of methods has been developed, however, for incorporating design effects into imputation modeling. von Hippel (2007, p. 88) notes that standard MI software does not allow for survey sample design effects to be taken into account in the imputation model, and that Rubin's (1986) recommended approach of including fixed effects for clusters can be problematic when clusters are small. He found in a practical example that failure to model clusters had little biasing effect. Reiter et al (2006) showed that including fixed effects for clusters reduced otherwise substantial imputation bias in an analysis of data generated for their simulation study but had a negligible effect in their real-data example. Concerning the incorporation of strata in the imputation equation, Reiter et al proposed a random effects model for sampling strata but cautioned that it is both more difficult to estimate computationally and is easier to mis-specify than is a fixed effect model for cluster effects. They described the fitting of hierarchical models for sequential MI in which there is a series of imputation equations as "…an area for future research" (p. 148). Schenker et al (2010) similarly deferred a comprehensive treatment of the differences in clusters and strata across surveys for future research, and noted that cluster and strata identifiers are frequently not available to researchers in public-use versions of survey data. Gelman et al (1998a) and Schenker et al (2010), also in a cross-survey imputation context, incorporated as many variables associated with cluster and strata identifiers as possible among their imputation predictor variables. This appears to be a reasonable compromise strategy, and is adopted in the present study.

Additional to differences in survey sample design are differences across surveys in variable definition, measurement instrument differences, and survey operations differences. When multiple surveys are combined, these differences in survey 'context' can be handled through a hierarchical modeling framework in which the effects of differences across surveys on the estimated model relationships can be parameterized. This is the circumstance of Gelman et al's (1998a) cross-survey MI development for estimation of a model jointly from 51 cross-sectional surveys in which some questions were not asked in some of the 51 surveys and were consequently multiply imputed by the researchers. They proposed a Bayesian hierarchical model with random effects in which survey is a level in the hierarchy. Tighe et al (2010) similarly used a hierarchical approach in their pooled-survey analysis with variables common across all surveys that they pooled. This is analogous to the pooling of data across countries in cross-national analyses in which each country has the same model variables but a different social and institutional context in which the relationships between the variables are played out (Western 1998). A hierarchical modeling approach, however, depends on pooling a sufficient number of surveys to allow for a parameterization of model-parameter variability across the surveys. In both the Gelman et al and Tighe et al cases, there were approximately 50 surveys whose observations were pooled, and for which hyperparameters for a distribution of variability across surveys could be estimated from the empirical variability across the 50 surveys. When only two surveys are combined, no parameterization of cross-survey variability is possible and therefore the Bayesian hierarchical model is not an option.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

To address survey differences in a cross-survey MI analysis with only two surveys, Schenker, Raghunathan, and Bondarenko (2010) subdivided their two surveys' samples into subgroups that were identified through propensity-score analysis to have similar covariate distributions across the two surveys. A disadvantage of this method is that smaller subsamples were then used to estimate the imputation models. Simpler imputation models had to be specified, which the authors suggested may have caused attenuation of the coefficients subsequently estimated in their analysis model. Only the observations of their "impute-to" survey were used in that model.

For our general social-science context of two surveys, we propose instead a pooled cross-survey MI method, preceded by a model-fitting approach (e.g., Burnham and Anderson 2002; Weakliem 2004) to evaluate the reasonableness of the assumption that two surveys are independent realizations of the same superpopulation. We also allow for this assumption to hold only up to a possibly non-zero scale factor allowing for difference in overall level of the outcome variable. The relevant statistical theory here is that two surveys whose samples were drawn in approximately the same period and geographical area are candidates for being treated as independent draws from either: (1) the same finite population, under a design-based paradigm; or (2) from the same superpopulation, under a model-based paradigm. The model-based paradigm is more flexible, as it allows for the comparability of the surveys to depend on the particular model being estimated. It therefore provides the more relevant criterion for determining whether observations from two surveys may reasonably be pooled.

A recommended set of procedures for cross-survey multiple imputation

In light of the preceding discussion, we propose three conditions and procedures to adapt within-survey MI methods successfully to cross-survey MI involving two surveys: (1) including in the analysis model only variables observed entirely within one of the two surveys; (2) use of sequential multiple imputation; and (3) testing for survey sampling and instrument differences using model-fit statistics calculated for an analysis model specified entirely from variables in common between the two surveys.

(1) Exclusion of Variables Never Jointly Observed—Special care must be taken in the specification of the variables to be included in the analysis and imputation models of a cross-survey MI study. We argue that a cross-survey imputation study should be designed such that the analysis model can be estimated with one of the surveys alone. This guarantees that "variables never jointly observed" (Rubin 1987) will be excluded. Violations of this exclusion condition discredited earlier attempts at cross-survey imputation and estimation conducted under the methodological heading of "statistical matching" (Rodgers 1984). A recent review and extension of that literature is found in D'Orazio, Di Zio, and Scanu (2006). The problem for the credibility of statistical matching techniques is in their handling of the variables not observed in common across the data sources. To give a simple illustration of the problem of variables never jointly observed, assume Survey 1 has outcome variable Y and predictor variable X_1 and Survey 2 has outcome variable Y and predictor variable X_2. The goal is to estimate $E[Y \mid X_1, X_2]$, for example in a multivariate regression model $Y = f(X_1, X_2)$. Without additional information on the joint distribution of Y, X_1, and X_2, estimation that combines observations across the two surveys leads to no additional knowledge about $Y \mid X_1$ that cannot be derived from estimation using observations from Survey 1 alone, and no additional knowledge about $Y \mid X_2$ that cannot be derived from estimation using observations from Survey 2 alone (Ridder and Moffitt 2007, pp. 5491–5494).

Additional information may come from an auxiliary data source in which Y, X_1, and X_2 are all observed, though this auxiliary data source will often be for a sample drawn from a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

different universe such as a segment only of the population (Singh et al 1993). This difference in universe may be handled by Bayesian methods that attach probability distributions to represent the degree of similarity of the joint distribution of Y, X_1, and X_2 in the auxiliary data to the true joint distribution in the target population. Such an approach was proposed by Rubin (1986) and was explored by Rässler (2002). No accepted methodology for implementing this, however, has taken hold in the social sciences.

The analysis of Gelman et al (1998a) is an extension of within-survey MI to a cross-survey context, but is not free of the "variables never jointly observed" problem. We consider their approach now in more detail. The setup of their problem is of 51 cross-sectional surveys (public opinion polls) conducted at various times preceding an election, and an analysis model predicting voter preference. They drew both on MI for within-survey non-response and on MI for split-survey "missing-by-design" data structures (Raghunathan and Grizzle 1995). Gelman et al developed and estimated a cross-survey multiple imputation model that combined the surveys and that introduced an additional diagnostic layer to understand any unique "survey" effects though a Bayesian hierarchical model. Although they noted that 5 of the 51 surveys included all the questions used to construct their model variables (Gelman et al 1998b, p. 870), a critical motivator of their study concerned variables not derived from the survey questions but instead from the period at which the poll was conducted. They handled this by including a variable for the date at which the survey was conducted (Gelman et al 1998a, p. 850), representing time until the election. They noted also, however, that particular events such as a party convention were likely to affect voter intentions separately from any overall time trend. A survey question for a key variable for their model, self-reported political ideology (liberal, moderate, or conservative), was not asked in the poll conducted around the time of the party convention. Therefore party convention, political ideology, and voter intention were never jointly observed. The authors argued that this was not problematic for their analysis because "…public opinion shifts are generally uniform across the population…" (Gelman et al 1998a, p. 855). This is a type of conditional independence assumption. The nature of this assumption, informed by previous studies and theory (see also Gelman et al 1998b), is no stronger than those commonly used to identify statistical models in the social sciences. Nevertheless, given the history of skepticism about cross-survey imputation approach used in the statistical matching literature due to its invoking of conditional independence assumptions about variables never jointly observed, it seems useful to propose a context in which to evaluate and illustrate the utility of the cross-survey multiple imputation approach that does not include variables never jointly observed.

(2) Use of sequential imputation—Both continuous multivariate normal joint imputation methods and chained sequential imputation methods for deriving the joint distribution have been used in multiple imputation (Lee and Carlin 2010). We recommend that cross-survey MI take advantage of the monotone missing pattern (Rubin 1987) that comes with the "missing-by-design" structure of the incomplete data in the cross-survey context to conduct sequential imputation. In our example application below, only one variable has missing values. This can be considered the simplest case of monotone missingness. If X_1 is instead a vector of regressors, they are assumed still to be missing only in Survey 2, and therefore missing values for any one of the elements of X_1 implies missing values for all other elements of X_1, a monotone missingness pattern. If the missingness pattern were instead "arbitrary," for example if in Survey 1 some cases had missing values for X_1 and other cases had missing values for X_2, then a model for the joint imputation of X_1 and X_2 would be needed. This requires that the relationships between Y, X_1, and X_2 be estimated jointly through a parameterized multivariate distribution, and in the practice of multiple imputation this has meant imposing a multivariate normal distribution (Schafer 1997). If either or both X_1 and X_2 are categorical or count, the standard procedure is to first transform these variables using a continuous normal approximation.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Considerable work has been conducted on evaluating the biases induced by imposing a continuous normal approximation on partly categorical or count data (Raghunathan et al 2001; Allison 2009; Lee and Carlin 2010; White and Carlin 2010). When the categorical variable values have close to equal probabilities, the approximation is very good and results in almost no bias. When the categorical variable values have very disparate probabilities (e.g., a probability of less than .05 for a binary variable), the approximation is much worse and substantial bias may be introduced. This has led to the development of sequential regression methods of multiple imputation (Raghunathan et al 2001) that allow for categorical regression equations for categorical variables and linear regression equations for continuous variables. A theoretical concern with sequential imputation is that the distribution resulting from a sequence of imputations may not converge to a true joint distribution. Simulation studies, however, have found this not to be a problem in practice (Raghunathan et al 2001; Lee and Carlin 2010). Moreover, the monotone missingness pattern of the cross-survey MI structure allows the joint distribution to be specified as a series of conditional distributions, whereas arbitrary missingness patterns typically used in simulation studies do not. For these reasons, we recommend that cross-survey MI take advantage of the monotone missingness pattern of the missing-by-design structure of the data and use the sequential regression method of multiple imputation.

(3) A model-fitting approach to testing for sampling from the same universe— Recommendations (1) and (2) above follow the method of Raghunathan and Grizzle's (1995) simpler case in which values are missing-by-design for sample components within a single survey. In practice, different surveys will almost never sample from (or be designed to generalize to) exactly the same finite population. Moreover, there will often be variations in variable definitions and in survey operations between the two surveys. To handle this, we recommend a model-fitting diagnostic approach with three sets of models: the first with no "survey" covariate; the second with a "survey" main effect only (the scale factor); and the third with a "survey" main effect plus full interactions between "survey" and the model covariates. For each of the three models, a penalized model-fit statistic is estimated. Weden, Brownell, and Rendall (forthcoming) demonstrated this model-fitting approach applied to two surveys in which the same variables were present in both surveys.

Returning to the simple example above, and using S_2 as an indicator variable for the observation's being from Survey 2 (S_2 =1) versus in Survey 1 (S_2 =0), the models whose fit should be compared exclude X_1 because this is missing for every observation in Survey 2. The three models to be compared are:

$$LOGIT[\Pr\{Y|X_2\}]=\beta_0+\beta_2 X_2 \quad (1a)$$

$$LOGIT[\Pr\{Y|X_2, S_2\}]=\beta_0+\beta_2 X_2+\beta_3 S_2 \quad (1b)$$

$$LOGIT[\Pr\{Y|X_2, S_2\}]=\beta_0+\beta_2 X_2+\beta_3 S_2+\beta_4 X_2 S_2 \quad (1c)$$

If Model (1b) has a smaller model-fit statistic than Model (1a), then the "survey" main-effect variable S_2 should be added to the analysis model (1). If Model (1c) has a smaller model-fit statistic than Model (1b), then we would conclude that the surveys differ also with respect to the relationship between the covariate and outcome variable. In that situation, although a hierarchical, objective Bayesian approach of the type used by Gelman et al (1998a) to handle more general survey differences is not feasible with only two surveys, a subjective Bayesian approach to the construction of priors about the survey differences,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

similar to that developed by Rendall et al (2009) to combine population and survey data, may be considered. Alternatively, defining a less broad target population, for which the sampling designs across the two surveys are more similar, may be considered. We illustrate the latter strategy in the example application described below.

Among model-fit statistics, the BIC, given by $BIC = -2\log L + p\log(N)$ where is the number of free parameters and N is the number of observations, is the standard statistic used by sociologists, whereas both the BIC and AIC, given by $AIC = -2\log L + 2p$, are frequently used by economists (Weakliem 2004: and see Burnham and Anderson 2002, chapter 6, comparing the BIC and AIC). A smaller model-fit statistic indicates a better-fitting model. The BIC and AIC differ only by their penalty term, with the BIC penalizing both an increase in the number of model parameters and an increase in the number of observations, whereas the AIC penalizes only an increase in the number of model parameters. Because the penalty term in the BIC is multiplicative with respect to and $\log(N)$, for a given pooled-survey sample size an increase in the number of parameters will increase the model-fit statistic more than will the AIC. Since the penalty for adding variables is less when using the AIC, use of this criterion makes it more likely that a pooled-survey model with a survey indicator and survey indicator and covariate interactions will have the better fit than a pooled-survey model without a survey indicator or survey indicator and covariate interactions. Weakliem (2004, p. 183) concludes that the BIC criterion is preferred when there is “a real chance” that the hypothesis that the simpler model is true. This suggests that the BIC criterion should be preferred for pooled cross-survey MI evaluation, since the hypothesis in question is that the two samples really do sample from the same target population. If we didn’t have an *a priori* belief that there was “a real chance” that this were at least approximately true, we would likely not initiate a combined-survey analysis.

III. APPLICATION TO SOCIODEMOGRAPHIC DETERMINANTS OF EARLY CHILDHOOD OBESITY

Overview

We illustrate the use of pooled cross-survey MI in an application to sociodemographic differences in the likelihood of obesity in kindergarten. We consider race/ethnicity, household income, and maternal education and marital status as markers of these sociodemographic differences. Our estimation combines two nationally representative, longitudinal surveys that we provisionally assume to sample from a common universe, defined in a model-based sense. We conduct diagnostics under a model-fitting framework to test this assumption. Sociodemographic and some biosocial variables are present in both surveys. Maternal height and weight, from which body mass index (BMI) are derived, are present in only one of the surveys. We use pooled cross-survey MI first to impute maternal BMI to the other survey and second to account for the additional variability of the parameter in the pooled-survey child obesity model due to their being estimated with multiply-imputed maternal BMI data.

The prevalence of child obesity in the U.S. is higher among Hispanic and black children than among white children, and these racial/ethnic disparities have widened with the development of the child-obesity epidemic (Freedman et al. 2006). Similarly, the prevalence of child obesity is higher among children living in families with lower household income, lower parental education, or with unmarried mothers, and disparities by these and other socioeconomic indicators have also been widening over the last several decades (Miech et al. 2006; Singh et al. 2010). Markers of children’s socioeconomic status like household income, maternal education and maternal marital status are important to assess not only because they describe differences in family circumstances, but also because they may be

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

associated with proximate environmental conditions amenable to policy intervention, such as underfinanced schools attended by children from low-income households (Anderson and Butcher 2006). High maternal body mass index (BMI) is one of the strongest risk factors for obesity in early childhood (Classen and Hoykayem 2005; Salsberry and Reagan 2005;), with both genetic and social factors figuring prominently in inter-generational correlations in obesity (Martin 2008). Because racial/ethnic minority children and children with poorer sociodemographic circumstances are disproportionately exposed to the risks of high maternal overweight and obesity (Kimbro et al. 2007; Weden et al. forthcoming), maternal weight status represents an important factor through which socio-demographic variables influence children's weight.

Data

Our two surveys are the Early Childhood Longitudinal Study 1998 Kindergarten cohort (ECLS-K) and the Early Childhood Longitudinal Study 2001 Birth cohort (ECLS-B). Both surveys were directed by the National Center for Educational Statistics (NCES) to assess children's early learning environments, health, and development (see, for example, Downey, von Hippel, and Broh 2004; and Mollborn and Morningstar 2009). Both surveys included observations of the child's height and weight in kindergarten, and these were measured by trained interviewers and not parent-reported. The ECLS-B and ECLS-K have many variables in common, measured using similar survey instruments. They are therefore good candidates for combined-survey analysis. Only one of the surveys, the ECLS-B, however, collected mother's height and weight. These are needed to calculate maternal BMI, which is strongly predictive of the child's obesity. This makes the use of pooled cross-survey MI potentially very valuable for analysis of other determinants of child obesity, controlling for maternal BMI. In the terminology of Section II above, the ECLS-B would serve as the "impute-from" survey and the ECLS-K as the "impute-to" survey.

The ECLS-K followed a nationally-representative cohort of children attending kindergarten in the U.S. in 1998 and assessed children periodically through eighth grade (U.S. Department of Education 2009a). The baseline kindergarten sample was selected using a three-stage probability-sampling design. Counties or groups of contiguous counties were first sampled as primary sampling units (PSUs), then schools within these PSUs, and finally students within schools. The ECLS-K used Census population estimates of five-year-olds in the PSUs by race/ethnicity to oversample Asians and Pacific Islanders. Additionally, during the selection of schools within PSUs, private schools were oversampled. An overall unweighted response rate of 61.9% was achieved for the baseline fall kindergarten child assessment. This overall response rate is a product of the cooperation rate by the 1,280 sampled schools nation-wide (68.8%), and the completion rate for children attending cooperating schools (89.9%)(U.S. Department of Education 2009b). Our analysis is of obesity measured at the fall kindergarten wave of the ECLS-K. Our analytical sample is restricted to U.S. born children whose biological mothers responded to the parent survey (89.2% of the fall kindergarten sample). This restriction excludes children who were born outside the U.S. (540 cases) to achieve comparability between the ECLS-K and the ECLS-B. An additional 10.9% of cases are excluded due to their missing information on one or more of the other study variables, for a final sample of 15,240 children. All ECLS-K counts are rounded to the nearest 10 to comply with NCES confidentially requirements.

The ECLS-B was designed as a nationally-representative sample of the cohort of children born in the U.S. in 2001 who survived to nine months, were not adopted from birth to nine months, and did not leave the country (Snow et al. 2009). Births were sampled within primary sampling units (96 counties or contiguous counties) using a sampling frame consisting of registered births obtained from the NCHS Vital Statistics system and from two hospitals. Additionally, a supplementary sample of 18 primary sampling units was selected

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

from a frame consisting of areas with a greater number of American Indian/Alaskan native births. Assessments were conducted at 9 months, 2 years, 4 years, and at kindergarten in either the 2006–2007 or 2007–2008 school year. Children who had not yet entered kindergarten when assessed during the 2006–2007 school year were re-contacted for assessment in the 2007–2008 school year. Using data recorded in the birth certificates, the ECLS-B oversampled children with low birth weights, twins, and children in the following racial/ethnic categories: American Indian/Alaskan Native, Chinese, and Other Asian/Pacific Islanders. Mothers younger than 15 years old when they gave birth to their child were excluded from the ECLS-B by sample design. Our analysis is of obesity measured at the kindergarten assessment. By the kindergarten wave, cumulative unit non-response was just over half the original sample. Specifically, an unweighted 47.0% of the children originally sampled from the birth certificates were assessed as kindergarteners. This overall unweighted response rate is a product of the baseline unweighted response rate at the 9 month wave (76.8%) and the unweighted retention rate for successful follow-up from the 9 month wave through the assessment upon entry into kindergarten in either the 2006–2007 or 2007–2008 school years (61.2%). We excluded ECLS-B cases where the child was homeschooled (2.0%), went straight to first grade (0.5%), or where the grade they were enrolled in was unknown or ungraded (1.7%). We excluded an additional 6.4% for whom the responding parent was not the biological mother. From this ‘eligible children’ sample, 11.7% of cases were excluded due to their missing information on one or more of the other study variables, for a final sample of 5,200 children. All ECLS-B counts are rounded to the nearest 50 to comply with NCES confidentially requirements.

Variables—The dependent variable in our study is child obesity in kindergarten, defined as a BMI at or above the 95th percentile using the U.S. Centers for Disease Control reference population and procedures (Kuczmarski et al. 2002) that account for developmental differences in growth by age and gender. The ECLS-K and ECLS-B used comparable measurement protocols for assessing child height and weight, using a Shorr board for height, a digital bathroom scale for weight, and requiring that children were wearing light clothing when weighed.

We include both sociodemographic and biosocial predictor variables in our analysis. Mother’s race/ethnicity is self-identified and coded for our analysis into the five categories of Hispanic and (non-Hispanic) white, black, Asian, and Other (which includes the self-reported categories of Native Hawaiian/Pacific Islander, Native American/Native Alaskan, and multi-racial). Maternal education, marital status, and household income are all measured in the child’s kindergarten year. Education and marital status are assessed using identical survey protocols in the ECLS-B and ECLS-K. Household income is a continuous measure in the ECLS-K but it is measured using a 13 category variable in the ECLS-B ($5,000 or less; $5,001 to 10,000; etc., up to $200,001 or more). In order to harmonize across the two datasets, we coded each of the categories of the ECLS-B to the middle of the range (with the open ended category coded to $408,500). We adjusted these values for inflation by measuring income in 1998 dollars (U.S. Department of Labor 2012), and transformed this inflation-adjusted income into the log of household income. Additional demographic variables are child’s age, child’s gender, and number of siblings. Child’s age is measured in months and corresponds to the age they were when the height and weight measurements were taken.

Biosocial variables include mother’s age at birth, whether the child was a singleton birth or part of a twin or higher order birth, child’s birth weight, and mother’s BMI. Mother’s age at birth is a continuous measure of the mother’s age in years when she gave birth to the study child. Birth weight is obtained from birth certificates in ECLS-B and parental reports in ECLS-K. We coded this into low birth weight (less than 2,500 grams, reference), average

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

birth weight (2,500 to 3,999 grams), and high birth weight (4,000 grams or heavier). In the ECLS-B, mothers reported their own weight at the time of the child's kindergarten assessment. We combined this with their height, self-reported in the 9 month wave, to calculate maternal BMI as weight (kg)/height (m)^2. No maternal weight or height measures are present in the ECLS-K and so maternal BMI is "missing" for all ECLS-K children.

Evaluation of comparability of the two surveys on outcome and predictor variables—Comparison of sample-weighted estimates of children's obesity and predictor variables in the two surveys allows for a first opportunity to assess their comparability, and therefore also their suitability for pooled analysis. The ECLS-K and ECLS-B are designed to be nationally representative with respect to the cohorts they sample from, respectively children who entered kindergarten in 1998 and children who entered kindergarten in 2006 or 2007. We know from previous analyses of National Health and Nutrition Examination Survey (NHANES) data, considered the 'gold standard' for U.S. prevalence estimates, that the prevalence of child obesity changed little between 1998 and 2007 (Ogden et al. 2010). We also conducted a direct comparison of the ECLS-B and ECLS-K to microdata from the NHANES (National Center for Health Statistics, no date), allowing us to match age and period from the ECLS-B and ECLS-K to the NHANES (see Table 1). We found that the ECLS-B kindergarten obesity prevalence of 16.4% (95% Confidence Interval (CI): 14.9, 17,9) was substantially and statistically higher than the 12.1% for children aged 4 to 6 in NHANES 2005–2008 (95% CI: 9.4, 14.7). In contrast, the ECLS-K kindergarten obesity prevalence of 11.6% (95% CI: 11.0, 12.2) was not statistically different from the 11.4% for children aged 4 to 6 in NHANES 1999–2000 (95% CI: 7.6, 15.2). Given the strength of both the ECLS-B's and ECLS-K's height and weight measurement protocols, we speculate that the higher cumulative non-response by kindergarten in the ECLS-B is in part responsible for the higher obesity prevalence than in the NHANES (also seen at the pre-school wave (Anderson and Whitaker 2009)), and therefore also higher than in the ECLS-K.

We next assessed whether there were survey differences between the ECLS-B and ECLS-K on observed predictor variables included in our regression model (see Table 2). Despite weighting for differences in sample design and non-response, and likely in part due to changing sociodemographic circumstances of children between 1998 and 2006/07, there were differences between the two surveys on sociodemographic predictor variables and on two of the biosocial variables. Although in both samples, a weighted 14% approximately of children were black and about 3% were Asian, a lower percentage of non-Hispanic white children (57.7%) and higher percentage of Hispanic children (23.2%) were estimated from the ECLS-B than from the ECLS-K (64.4% and 16.5%, respectively). Maternal education and household income were lower, and the prevalence of never-married mothers higher, when estimated from the ECLS-B than from the ECLS-K. There were no statistically-significant survey differences in mother's age at the child's birth or the child's gender, age, or number of siblings. Twin or multiple births, however, were estimated to be more prevalent, and high birth weight less prevalent, in the ECLS-B than in the ECLS-K.

The weighted percentages in Table 2 also provide the first opportunity to evaluate racial/ethnic differences in child obesity in both surveys and in maternal BMI in the ECLS-B. Racial/ethnic differences in the prevalence of child obesity are statistically significant at $p<0.001$ in both the ECLS-B and ECLS-K. The highest rates of child obesity are observed among Hispanic children (26.4% in ECLS-B and 16.2% in ECLS-K), followed by non-Hispanic black children (19.4% in ECLS-B and 12.3% in ECLS-K), while non-Hispanic Asian children have rates similar to non-Hispanic white children of about 10–12%. Differences in mean maternal BMI by race/ethnicity are also statistically significant in the ECLS-B at $p<0.001$. Maternal BMI was respectively just above and just below the BMI=30 threshold for adult obesity for black and Hispanic children. Whether the survey differences

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

we observe on the outcome variable and on some of the predictor variables leads to bias in the estimated relationships between our included regressor variables and the obesity outcome variable, however, needs to be evaluated specifically for our analysis model.

Model Specification, Model-Fit Diagnosis, Imputation, and Post-Imputation Analysis

The cross-survey imputation and analysis method proceeds in six steps. First, analysis and imputation models are specified to satisfy the "variables never jointly observed" requirement, and simultaneously "congeniality" between the imputation and analysis equations. This is achieved by specifying an analysis model that can be estimated on the ECLS-B alone, and by including in the imputation equation all the variables that are in the analysis equation. Since only maternal BMI is "missing" from the ECLS-K in our analysis model, a single imputation equation only is needed. Second, model-fit diagnostics for this analysis model are conducted by specifying a version of the model that uses only those variables observed in both surveys, and estimating this model on the pooled ECLS-B and ECLS-K data. Model-fit statistics are compared between alternative versions of this "variables-in-common" analysis model that respectively do and do not include an indicator variable for survey, and that respectively do and do not include variables for interactions between the covariates and this survey indicator variable. Third, the imputation model for maternal BMI is estimated from the ECLS-B. Fourth, using the estimated imputation model, an augmented ECLS-K dataset is constructed that contains multiple (20) versions of each ECLS-K observation, each with a random draw of maternal BMI from the estimated imputation equation parameters. Each of the 20 multiply-imputed ECLS-K datasets is concatenated to the ECLS-B dataset. Fifth, the full analysis model including values of the maternal BMI predictor variable for both ECLS-B and ECLS-K cases, and (in accordance with the model fit statistic results) including a variable indicating from which survey the case is drawn, is estimated on each of the 20 pooled ECLS-B and multiply-imputed ECLS-K datasets. Sixth, the estimated parameters and standard errors are combined using the standard multiple-imputation formulas. These adjust for the additional uncertainty in the estimates that is introduced by using imputed and not observed values of maternal weight status in the ECLS-K component of the pooled data.

Analysis Model and Model-Fit Diagnostics for Survey Differences—We specify a logit model for the probabilty that child i is obese, designated by the (0,1) variable Y_i, as a function of a vector $\boldsymbol{X_{ki}}$, of the sociodemographic and biosocial regressors described above. The analysis model also includes a regressor for maternal BMI:

$$LOGIT[\Pr\{Y_i\}]=\beta_0+\beta_1 MomBMI_i+\sum_k \beta_k X_{ki} \quad (3)$$

We conduct model-fitting diagnostics for the version of this analysis model that can be estimated from variables in common between the surveys. This is simply equation (3) but dropping the *MomBMI* variable. Following our recommended model-fit evaluation procedure of Section II above, we fit three sets of models, (3a), (3b), and (3c), corresponding to equations (1a), (1b), and (1c) of Section II. Model (3a) has no "survey" covariate; Model (3b) has a "survey" main effect only for presence in the ECLS-B data source; and Model (3c) has a "survey" main effect plus full interactions between the ECLS-B "survey" and the model covariates:

$$LOGIT[\Pr\{Y_i\}]=\beta_0+\sum_k \beta_k X_{ki} \quad (3a)$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$LOGIT[\Pr\{Y_i\}]=\beta_0+\sum_k \beta_k X_{ki}+\gamma_1 ECLSB_i \quad (3b)$$

$$LOGIT[\Pr\{Y_i\}]=\beta_0+\sum_k \beta_k X_{ki}+\gamma_1 ECLSB_i+\sum_k \gamma_k X_{ki} ECLSB_i \quad (3c)$$

Parameter estimates for each of these models are presented in Appendix Table A1. The model-fit statistics are summarized in Table 3. We use both the BIC and AIC model-fitting criteria. We discuss results first for the version estimated for the full target population that includes all race/ethnic groups (See Panel A of Table 3). According to the BIC, the best fitting pooled-survey regression specification is clearly model specification (3b) which includes all study variables observed in both surveys and a survey indicator variable for the intercept-shift for the ECLS-B relative to the ECLS-K (BIC=14,866.0). Using the AIC model fit statistics, however, the (3b) and (3c) specifications are essentially tied on model fit, with a very small worsening of fit over the survey main-effect variable only in specification (3b) (AIC=14,699.6) by adding the interactions between the survey and the vector of covariates in specification (3c) (AIC=14,700.5). Inspection of the survey-covariate interactions in specification (3c) (see Appendix Table A1) revealed statistically significant interactions between the survey indicator and each of the following predictor variables: Asian and Other race/ethnicity (respectively less and more likely to be obese in the ECLS-B than in the ECLS-K); several of the maternal education categories (i.e., high school and GED, some college, and bachelor's degree less likely to be obese in the ECLS-B than in the ECLS-K); and number of siblings (associated with a larger decrease in obesity in the ECLS-B than in the ECLS-K). In light of the potential role of survey sampling differences with respect to the Asian and Other groups noted earlier and suggested also by the AIC statistics and by the individual survey interaction variables, we therefore also conducted a second set of imputations and analyses restricting the ECLS-K and ECLS-B samples to black, white, and Hispanic children. In this restricted sample, both the BIC and the AIC are seen to be clearly minimized in the pooled model specification (3b) with the survey indicator but no survey-covariate interactions (See Panel B of Table 3).

Imputation Model

We impute maternal BMI to the ECLS-K observations from observed maternal BMI in the ECLS-B. Being sure to include the analysis model's outcome variable in the imputation equation, the imputation model for the maternal BMI variable is:

$$E[\{MomBMI_i\}]=\delta_0+\delta_1 ChildObese_i+\sum_k \delta_k X_{ki} \quad (4)$$

We include all the same variables in the X vector as in equation (3) to assure congenial imputation and analysis models. This equation is estimated unweighted from ECLS-B observations alone. The resulting parameter estimates are randomly perturbed in the imputation procedure to produce 20 realizations of maternal BMI for each ECLS-K observation. Although 5 imputations are standard in within-survey imputation (Rubin 1987), the large fraction of observations with missing values in our cross-survey imputation case (approximately three quarters, being every ECLS-K observation) leads us to create instead the 20 imputed datasets. We use SAS PROC MI with the MONOTONE option to implement the imputation model. Previous work [SELF-IDENTIFYING REFERENCE] has shown that the IVEware software for sequential multiple imputation (Raghunathan, Solenberger, and van Hoewyk 2000) and the PROC MI software we use here produce identical results.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

PROC MI outputs the imputation equation coefficient parameter estimates after standardizing continuous variables and using effects coding for categorical variables. In Table 4 we present imputation coefficient means and additionally standard deviations calculated over the 20 versions of the imputed coefficients. The standard deviations illustrate the additional uncertainty introduced by the imputation process. Their magnitudes approximate the analytical standard errors estimated for this imputation equation in a regular regression with standardized coefficients (results not shown). Coefficient means and standard deviations are presented for both the full sample and the sample restricted to white, black, and Hispanic children.

The coefficient estimates that are most readily interpretable are those for the dichotomous and continuous predictor variables. In both the full and restricted samples, child obesity is positively associated with mother's BMI, and this is one of the strongest predictors considered. Log household income is negatively associated with maternal BMI, while female child gender, number of siblings, and being a twin or higher-order multiple are all positively associated with maternal BMI. Effects coding, like dummy variable coding, involves specifying a reference category which is dropped in the estimation of the model; however, the interpretation of the parameter estimates differ. The advantage of effects coding is that these parameter estimates do not depend upon which group is nominated to be the reference category. The parameter estimates, however, must be interpreted with respect to an unweighted grand sample mean, calculated as the grand mean of the group (variable category) means on the outcome variable. Black race/ethnicity is thus seen to be positively associated with maternal BMI in both the full and restricted samples, and Asian is negatively associated in the full sample. Hispanic is weakly positively associated with maternal BMI in the full sample and weakly negatively associated in the restricted sample. The change in the sign is a direct implication of the increase in the unweighted grand mean (i.e., the increase in the intercept) when the Asian and Other racial/ethnic groups were excluded from the sample. There is a negative association between the highest levels of maternal education and maternal BMI in both the full and restricted samples. In addition, consistent with the positive associations between child obesity and maternal BMI, the child's low birth weight was negatively associated with maternal BMI and his or her high birth weight was positively associated with maternal BMI.

In the full sample, the coefficients for black and Asian are respectively the largest positive and negative associations observed. We can anticipate from the strongly positive multivariate association between the child's obesity and his or her mother's BMI, simultaneously with a strongly positive multivariate association between black race/ethnicity and maternal BMI, that maternal BMI will *mediate* (facilitate) black children's obesity. Analogously, we may anticipate from the strongly negative association between Asian and maternal BMI the opposite effect of including imputed maternal BMI for the ECLS-K observations: the lower BMI of mothers of Asian children will *suppress* the child's obesity. If the analysis equation would have been estimated on the ECLS-K cases without first multiply imputing maternal BMI differentially by child's race/ethnicity, much of the higher obesity of black than Asian children would go unexplained. By first multiply imputing maternal BMI, we allow for the difference between black and Asian children's likelihood of being obese in kindergarten to be estimated as a smaller residual after controlling for the opposite directions of influence of black and Asian children's mothers' BMI on their own obesity propensity.

Analysis Estimates

The analysis model (3), with the addition of an ECLS-B survey indicator, is estimated on each of the 20 datasets and the results combined using the standard MI algorithms given in section II above to account for the additional variance due to imputation. These pooled

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

cross-survey MI estimates of the regression parameters and standard errors are presented in Model 6 in Table 5. The validity of these estimates depends on the successful imputation of maternal BMI to every child in the ECLS-K. We assess this validity in two ways: by comparisons of the maternal BMI coefficient and standard error before and after imputation; and by comparisons of the coefficients and standard errors for variables in common between the ECLS-B and ECLS-K.

First, the magnitudes of the coefficient and standard error about the maternal BMI coefficient are expected to be materially unaffected by having imputed this variable to every case in the ECLS-K. We see that this holds by comparison with the parameter and standard error estimates for analysis model (3) estimated from the ECLS-B cases only (see Model 2 in Table 5). The maternal BMI coefficient is not substantially changed from 0.077 in the ECLS-B to 0.086 in the pooled ECLS-B and ECLS-K estimate (Model 6). Concerning the standard error, intuitively the accuracy of this coefficient estimate of the relationship of maternal BMI to the child's likelihood of being obese should not be improved by adding cases (from the ECLS-K) for which maternal BMI is not observed. This is confirmed by there being no material change in the maternal BMI standard errors between Model 2 and Model 6 (taken to 4 decimal places, the standard errors are respectively 0.0062 and 0.0066, for a ratio of 0.94). This is despite the quadrupling of total sample size when adding the multiply-imputed ECSL-K data.

Second, the nature of the change in the coefficients for variables in common between the ECLS-B and ECLS-K when adding the maternal BMI variable to the analysis model estimated for the ECLS-B observations should be reproduced in the ECLS-K observations with imputed maternal BMI. Intuitively, only the ECLS-B cases provide any information on that part of the multivariate relationship to child obesity of a given sociodemographic or biosocial variable that involves controlling for maternal BMI. This is best verified by comparing the change in coefficients between the model with and without maternal BMI in the ECLS-B (Model 1 to Model 2) with the change in coefficients between the model with and without maternal BMI in the ECLS-K (Model 3 to Model 4). Indeed this expectation of equivalent change holds across all the coefficients. Two of the largest changes in coefficient magnitudes when adding maternal BMI to the model, for example, are the reduction in the absolute values of the magnitudes for black race/ethnicity and for the log of household income. In the ECLS-B, adding maternal BMI to the model reduces the black coefficient from 0.368 to 0.209, and reduces the impact of log household income from −0.120 to −0.070. Adding multiply-imputed maternal BMI to the model estimated on only the ECLS-K observations reduces the black coefficient from 0.113 to −0.069, and reduces the impact of log household income from −0.091 to −0.036. As anticipated in the discussion of the imputation model coefficients above, maternal BMI is seen to *mediate* the relationship of black race/ethnicity to child obesity. Analogously, for the Asian coefficient, the anticipated *suppressing* effect of maternal BMI is also seen. In the ECLS-B, adding maternal BMI to the model reduces the negative (healthy) effect of being Asian on child obesity from a statistically-significant coefficient value of −0.477 to a non-significant coefficient value of −0.189. This direction and magnitude of coefficient change is reproduced in the ECLS-K, in which the coefficient of Asian on child obesity is positive but statistically non-significant (magnitude 0.170) without maternal BMI. Adding multiply-imputed values of maternal BMI reproduces the ECLS-B's positive change to the coefficient, here to a statistically-significant 0.425 after controlling for the lower than average maternal BMI of Asian mothers also in the ECLS-K.

We turn now to the improvements in substantive interpretation of the model estimates made possible by the pooled cross-survey MI method. The reduced effect of being black or Hispanic, or having higher household income on child obesity, and increased effect of being

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Asian on child obesity, once maternal BMI is added to the model, indicates that the omitted-variable bias in estimates of race/ethnicity and household income is corrected for by cross-survey imputation to the ECLS-K cases. This is analogous to the correction of omitted-variable bias achieved by adding observed maternal BMI to the equation estimated on the set of ECLS-B cases. Pooled ECLS-B and ECLS-K estimates of the effects of race/ethnicity and household income in the best-specified model (Model 6) may then be evaluated for their efficiency gains over estimates of these same coefficients from the ECLS-B only. The standard errors for log household income, black, and Hispanic coefficients are (see far right column of Table 5) around 1.8 to 1.9 times higher in the ECLS-B-only estimates (Model 2) than in the pooled ECLS-B and ECLS-K estimates (Model 6). The corresponding ratios of the standard errors for Asian and Other race/ethnicity children are in the 1.5 to 1.6 range, reflecting the oversampling of these groups in the ECLS-B and therefore the lesser proportionate increase in total sample size acheived by adding the ECLS-K cases.

The reduced standard errors in estimates of the coefficients for variables-in-common between the ECLS-B and ECLS-K without the omitted-variable bias that would have occurred had the model been estimated *only* with variables in common (that is, excluding maternal BMI) are, we argue, the main benefits of our pooled cross-survey MI estimation. For further evidence of these efficiency gains in the full model including maternal BMI, we note that the “married” coefficient is not statistically significant at the $p < .05$ level in estimates using either the ECLS-B or ECLS-K cases alone, but is statistically significant in the pooled ECLS-B and ECLS-K model. With respect to log household income, a more powerful test against the null hypothesis is made possible, as shown by the halving of the standard error between the ECLS-B-only model and the pooled model (from 0.052 to 0.027).

Given the mixed evidence we found with respect to ECLS-B and ECLS-K comparability for Asian and Other race/ethnicity children, we estimated the same models also for the restricted sample of white, black, and Hispanic children only (see Appendix Table A2). As in the full-sample estimates of Table 5, in the restricted samples in the pooled models that include maternal BMI among the predictors of the ECLS-B and ECLS-K, Hispanic but not black children were statistically-significantly more likely to be obese than white children. Findings on maternal education and household income were also similar in the samples restricted to exclude children with Asian and Other race/ethnicity. These socioeconomic variables were negatively associated with child obesity in both the ECLS-B and ECLS-K. By pooling the samples, moreover, we obtained statistically significant negative associations between increased maternal education and obesity across all educational categories (respective to the reference group with less than 9th grade). The lack of change in educational disparities after adjustment for maternal BMI, however, appears inconsistent with previous substantive findings of increased likelihood of exposure to high maternal BMI among children with lower educated mothers (e.g., McLaren 2007). In supplementary analyses in which we did not simultaneously adjust for household income, we determined that differences in child obesity by maternal education were reduced after adjusting for maternal BMI. Other associations between predictor variables and child obesity in both these restricted-sample models and the full-sample models are consistent with results found elsewhere in the literature (Classen and Hokayem 2005; Salsberry and Reagan 2005; Weden et al. Forthcoming). Factors negatively associated with obesity included low birth weight, female gender, and increased number of siblings. Factors positively associated with obesity included mother’s age at birth and high birth weight. In general, therefore, we found our pooled cross-survey MI estimates to be robust to whether estimates were generated for the full target population of U.S. children or for the target population restricted to black, white, and Hispanic children.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

IV. DISCUSSION

Within-survey MI has become a common and widely-accepted practice for improving estimation over that from complete-case analysis in sociology (e.g., Downey, von Hippel, and Broh 2004). We are not aware, meanwhile, of any successful implementations in sociology of *cross*-survey MI in the more than 25 years since Rubin (1986) proposed the method. Moreover, we know of no more than occasional, experimental implementations of cross-survey MI in the social and health sciences more generally (e.g., Gelman et al 1998a; Schenker et al 2010), even while the broader topic of combined-survey analysis is becoming an active area of new research (Roberts and Binder 2009). We argued that the benefits of cross-survey MI are potentially very large, and we proposed a set of implementation procedures to overcome previous objections in the social and statistical sciences to pooled-survey analysis with non-identical sets of regressors across the surveys. We illustrated these steps and the resulting improvement over single-survey analysis in an example application chosen to be representative of a common situation in sociology in which the researcher has at his or her disposal a first survey dataset of relatively small sample size and possibly some non-response sampling bias, and a second survey dataset of larger sample size and potentially smaller sampling bias but less covariate detail. These surveys were respectively the Early Childhood Longitudinal Survey 2001 Birth Cohort (ECLS-B) and the Early Childhood Longitudinal Survey 1998 Kindergarten Cohort (ECLS-K). We combined them using pooled cross-survey MI to analyze the sociodemographic associations with early childhood obesity. The analysis model was specified to include maternal BMI, which was derived from height and weight measures available only in the ECLS-B.

One way to view the benefits of combined-survey estimation through cross-survey MI is to compare the results to those that would have been possible from the larger survey (the ECLS-K) only. By estimating a model that included mother's BMI as an additional predictor, multiply imputed from the ECLS-B to the ECLS-K, we achieved substantial reductions in omitted variable bias. In particular, the magnitude and directions of the estimated racial/ethnic associations with child obesity were changed, and income associations were moderated. Both genetic and social factors figure prominently in explanations for strong inter-generational correlations of obesity (Martin 2008). Controlling for these inter-generational associations is therefore expected improve estimation of the roles of proximate environmental circumstances, such as local economic and political resources that may impact adversely on children's nutritional or physical exercise environments (Anderson and Butcher 2006). Crucially, we achieved the resulting reductions in omitted variable bias without losing the major sample-size gains of the ECLS-K relative to the ECLS-B. The typical trade-off between sample size and covariate richness was thus avoided.

A second way to view the gains to cross-survey MI is to contrast it with estimation from the smaller survey with the best possible model specification, in our case the ECLS-B. The principal advantage is likely to be efficiency gains. The standard errors about most of the regression coefficients for the variables in common between the ECLS-B and ECLS-K were almost twice as large in the estimates from the ECLS-B only as they were in our pooled cross-survey MI estimates. This large efficiency gain is an unsurprising consequence of the quadrupling of the sample size compared to using the ECLS-B alone. We submit, however, that efficiency gains in social science estimation continue to be undervalued, possibly in large part due to the predominance of a "p-value culture" that pays little attention to the size and precision of an estimated association after its "statistical significance" has been established (Taylor and Frideres 1972; McLoskey 1985; Fidler et al 2004). Even for researchers (and journals) focused on results that attain statistical significance, estimation with much larger sample sizes promises substantially enhanced potential to achieve that

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

goal. We gave an example of this in estimates of marital status associations with child obesity that were statistically significant in the pooled cross-survey MI estimates, but not in either of the single-survey estimates.

A second advantage of combining data from a larger survey (the ECSL-K) with the data from a smaller but covariate-rich survey (the ECLS-B) is to mitigate the effects of sampling biases in the latter. This advantage was illustrated by Hellerstein and Imbens (1999) in their labor economics application that combined the larger and more population-representative Current Population Survey (CPS) with the smaller and covariate-rich National Longitudinal Survey (NLS). The inclusion of the CPS corrected what were assumed to be substantial attrition biases in the NLS. The use in our study of the larger ECLS-K survey with, as at the kindergarten year, less attrition than in the smaller ECLS-B survey analogously corrected a substantial upward bias in overall level of child obesity in the smaller ECLS-B survey. The ECLS-B was already in its fourth wave by the kindergarten observation, and just under half of the original ECLS-B sample was still present. This level of cumulative attrition is typical in social surveys (see, for example, Fitzgerald, Gottschalk, and Moffitt 1998 and other articles in that special issue), but methods to deal with attrition are a challenge to develop and implement. Both Hellerstein and Imbens (1999) and Handcock et al (2005) proposed data-combining methods to mitigate bias from attrition through the imposing of constraints on the set of possible regressor values. The present study's method can be seen in part as an alternative approach to achieving this same goal. The computational demands of direct estimation methods such as those that impose constraints, however, are greater than for cross-survey MI, typically requiring code to be built that is specific to each application (Schafer and Graham 2002). Moreover, the assumption that bias is present only, or even primarily, in the smaller data source is limiting with respect to the application circumstances to which the method can be applied. In Schenker et al's (2010) cross-survey MI application, for example, measurement bias was assumed to be present only in the larger "impute-to" survey, and not in the smaller "impute-from" survey.

The principles of constraining approaches also differ from those of the present study's pooling approach. Pooled cross-survey MI gives equal weight in the analysis equation to each observation, irrespective of into which survey that observation was selected. This means that in the present example, the ECLS-K survey was effectively given more weight not because of an *a priori* designation as the less biased of the two with respect to covariate effects, but simply because its sample size was three times that of the ECLS-B by the Kindergarten wave. If the larger survey is indeed more population-representative than is the smaller survey, bias will be reduced by having the larger sample effectively dominate the estimation of the coefficients for which predictor variables are present in both the surveys. This weighted-averaging estimation obtained through pooling observations in cross-survey MI is more akin to the meta-analysis method in which findings across studies are averaged, though in that case typically using only a limited set of moments such as regression parameters without incorporating the full covariance structure of each data source. This involves implicit or explicit assumptions that differences in model specifications across studies can be ignored. Rao et al (2008) note the desirability of combining individual observations when conducting a meta-analysis that combines survey estimates. When discussing the problem of different regressors present across the different studies, however, they cite no meta-analysis studies and instead direct the reader to the "statistical matching" literature, which we noted has been discredited in the social sciences (e.g., Ridder and Moffitt 2007) due to its perceived failure to handle adequately the "variables never jointly observed" problem.

In thinking about what have been and continue to be barriers to the successful implementation of cross-survey MI in the social sciences, we consider the two largest of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

them to be: (1) overcoming the problem of estimation with variables never jointly observed; and (2) accounting for differences in survey sampling and measurement characteristics across the two (or more) surveys being pooled. The first of these problems is the easier one to solve in the pooled cross-survey MI approach, by simply specifying an analysis model that can be estimated with one of the surveys alone. This model is therefore no worse than the model that may be specified in a regular analysis in which the researcher first chooses the best available data source and proceeds to estimate a model with that data source.

The second problem, of accounting for survey design and operations differences, is ever-present. There is a high likelihood that any two surveys of a given population will have differences in survey instruments, sampling schemes, and survey operations that could affect the character of responses. Research into the development of diagnostic and estimation solutions applicable to the many different circumstances of survey differences is ongoing in the broader field of combined-data methods (e.g., Pfefferman and Sverchkov 2007). Selection of the best diagnostic and estimation methods for handling survey differences will vary according to the context. When many surveys are pooled, a hierarchical Bayesian method that parameterizes differences across surveys (Gelman et al 1998a) is a promising solution that has seen related applications in the small-area estimation literature (e.g., Assuncao et al 2005). Among the advantages of this approach is that it is empirically based, requiring no formal incorporation of outside information evaluating the relative quality of the surveys being pooled. That is, it is an “objective Bayesian” approach. With only two surveys being combined, however, the hierarchical Bayesian model is not feasible. A subjective Bayesian approach in which expert judgment informed by other data sources, as shown by Rendall, Handcock, and Jonsson (2009) in the constrained MLE case, is a plausible alternative. The challenges of developing defensible subjective Bayesian priors, however, increase with model complexity. It is noteworthy that Rendall et al’s prior adjusted only for biases in overall fertility rates.

We argued that for a two-survey context, a model-fitting diagnostics approach is an effective means of evaluating whether differences across surveys are large enough to demand additional statistical methods to correct for them. The modeling-fitting approach penalizes the adding of complexity to a model specification, requiring that additional variables be sufficiently informative about the social process being modeled to justify their inclusion in the statistical model. In our case, the additional variable indicates into which survey an individual was selected. Of course, this variable has no substantive meaning, and ideally we would like to be able to ignore it. We argued for a model-fitting approach to assess the degree to which it can reasonably be ignored, and that this model-fitting assessment be conducted by first adding only an intercept term for ‘survey,’ and second adding a full set of interactions of the covariates with ‘survey.’ If the model that includes the ‘survey’ intercept improves in a model-fit sense over the model estimated with the pooled data but without the intercept term, this implies a difference in the overall level of the outcome variable, but not in the relationships between the covariates and the outcome variable, between the surveys. A difference in the level of the outcome variable is unlikely to be a problem for most social science analysis, in which the goal is to understand relationships between covariates and the outcome variable. The analysis model using the pooled surveys may then be estimated with the addition only of an indicator variable for survey. Although the intercept is not generally considered to be a parameter of substantive interest, it nevertheless has two potentially important roles. First, it affects predicted values, and these may be of substantive interest. This was an early motivation for proposing combined-data methods in economics (Imbens and Lancaster 1994) and demography (Handcock et al 2000). Second, in nonlinear models the comparison of predicted values may be needed for valid statistical inference about covariate effects (Ai and Norton 2003).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

In our example application, we first established through comparison to outside data from the cross-sectional National Health and Nutrition Examination Survey (NHANES) that the overall level of obesity in the ECLS-K was unbiased whereas the overall level of obesity in the ECLS-B was upwardly biased. This conclusion was corroborated by our diagnostics-stage model-fit statistic being improved by including an “ECLS-B survey” indicator variable. Therefore we proposed that any predicted values should be generated omitting the coefficient for that survey indicator variable. This may be seen as being similar to constraining the results to the overall level of outcome variable to that of the larger survey. From a Bayesian viewpoint, ignoring the contributions of the ECLS-B observations to the overall level by excluding the intercept-shifting survey indicator coefficient value in calculating any predicted values is equivalent to imposing a “dogmatic prior” (Lancaster 2004) on the unbiasedness of the ECLS-K sample. This is analogous to the imposition of exact constraints on an overall outcome level, which again requires an a priori designation of one of the two data sources as being unbiased. For a Bayesian analysis that relaxes this dogmatic prior in a constrained estimation framework, see Rendall et al (2009). We suggest that this may be a relatively common situation in social science applications estimated across two or more surveys, that differences are found in the overall level of the variable being considered but not in the covariate relationships to the dependent variable. Weden et al (forthcoming) and [SELF-IDENTIFYING REFERENCE] give additional examples of an improvement in model-fit statistic with a survey indicator intercept-shift coefficient, but model-fit worsening after adding interactions between the survey indicator and the covariates present in both surveys.

The case in which the best-fitting model includes interactions with ‘survey’ is more complicated, requiring explicit additional statistical treatment to account for the differences. Without doing so, it may be unclear to which target population the parameter estimates apply (see Hellerstein and Imbens 1999 for further discussion of a similar case). The additional statistical treatment may involve restricting the samples and target population to those components for which comparability can be well established. This is the approach taken by Schenker et al (2010) in their propensity-score method for matching subsamples of two surveys based on equality of covariate distributions. Their sample-matching method, as opposed to matching on a target population, is consistent with their use of the smaller survey as an auxiliary data source whose observations are used for the imputation equation but not for the analysis equation. We suggest that one of the roles of a pooled analysis is to assure the integrity of the cross-survey imputation modeling process, in which the sample used in the imputation model can be shown to be drawn from the same target population as the sample(s) used in the analysis model. Schenker et al (p. 543) recognize the limitation of not pooling survey observations in the analysis equation under the heading of “issues of uncongeniality” between the imputation and analysis equations.

In our example application, we illustrated an alternative approach that explicitly restricted both the samples and target population for both the imputation equation and the subsequent analysis equation. After examining both the survey designs, and noting differences in the oversampling of Asian and selected ‘Other’ groups between the ECLS-B and in the ECLS-K, we experimented with restricting the samples and target population in the imputation and analysis steps to only the Hispanic and non-Hispanic white and black children. When the Asian and Other race/ethnicity groups were included, only for the BIC statistic was the simpler model without “survey” interactions with the covariates clearly superior to the model that included these interactions, whereas the AIC statistic was almost identical in the latter, “survey-covariate interactions” model compared with the “survey intercept-shift-only” model. Moreover, the sign of the Asian coefficient differed between the ECLS-B and ECLS-K samples, possibly as a result of different oversample designs between the two surveys. When Asian and Other race/ethnicity groups were excluded from both the ECLS-B

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and ECLS-K samples, both the AIC and BIC model-fit statistics indicated that the model that *excluded* the survey interactions with the covariates was clearly favored over the model that included these interactions. We argued that the BIC is probably the more appropriate fit statistic to use for pooled cross-survey MI. Nevertheless, in the case of different conclusions between the BIC and AIC, the researcher's judgment may be used to weigh the population-coverage and sample-size benefits of the larger target population (in our example case, including Asian and Other Race children) against the sample-comparability benefits of the smaller target population. Our substantive results were seen to be otherwise robust to the inclusion or exclusion of Asian and Other race/ethnicity children.

Our concluding guidance to social scientists is to view cross-survey MI primarily as a method for improving efficiency by adding *observations*, and secondarily for mitigating the adverse consequences of sampling bias by adding these observations. This differs from the original rationale in the statistical matching literature which views cross-survey MI as primarily a means to add *variables*. The "adding observations" viewpoint presumes the existence of a primary data source with all the needed variables but that suffers from sample-size limitations and possibly sampling bias. We suggest this situation is the norm rather than the exception in social science research. Given the choice between a data source containing most or all of the variables considered important to the substantive model but with a smaller sample size and potentially some sampling bias, and an alternate data source containing only a subset of variables considered important to a substantive model but with large numbers of observations and low sampling bias (in the limit, a census), the social scientist will usually opt for the more covariate-rich, smaller survey. Doing so will minimize omitted-variable bias. The researcher may even initiate a new survey data collection to obtain variables missing in existing data sources. Given the high costs of data collection, compromises leading to sample size limitations, and possibly also high non-response and non-response bias, are again likely to result. We additionally suggest, therefore, that the situation in which sample size limitations can be mitigated by the adding of "incomplete" observations from a larger survey, using cross-survey MI to "complete" the data in this larger survey, may also be viewed as the norm rather than the exception in social science analysis. The present study offers guidance of how this mitigation can be achieved using statistical package software allowing flexible specifications of both the analysis and imputation models, and provides an example of the substantial benefits that may be expected to result.

Acknowledgments

This work was supported by Research Project grants R21-OH009320 and R01-HD061967, Training Grant T32-HD007329, and Population Infrastructure grant R24-HD41041. The authors are grateful to Mark Handcock and Sanjay Chaudhuri for discussions about theoretical results on variance reduction and the model-fitting approach to testing for survey comparability, and to Chris Lau for valuable research assistance. The paper also benefitted from comments received at presentations to the 2011 Population Association of America Annual Scientific Conference and to the Department of Statistics at Brigham Young University and University of Maryland Joint Program in Survey Methodology seminar series, and from the authors' participation in the National Collaborative on Childhood Obesity Research (NCCOR) Envision Network.

References

Ai C, Norton E. Interaction terms in logit and probit models. Economics Letters. 2003; 80:123–129.

Allison, PD. Missing Data. Newbury Park: Sage publications; 2002.

Allison, PD. Paper 113–30, SUGI 30 Focus Session. 2009. Imputation of categorical variables with PROC MI.

Anderson PM, Butcher K. Reading, writing and refreshments: Are school finances contributing to children's obesity? Journal of Human Resources. 2006; 41:467–494.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Anderson SE, Whitaker RC. Prevalence of obesity among U.S. preschool children in different racial and ethnic groups. Archives of Pediatric and Adolescent Medicine. 2009; 163(4):344–48.

Assuncao RM, Schmertmann CP, Potter JE, Cavenaghi SM. Empirical Bayes estimation of demographic schedules for small areas. Demography. 2005; 42(3):537–58. [PubMed: 16235612]

Burnham, KP.; Anderson, DR. Model Selection and Multimodel Inference: A Practical Information-theoretic Approach. New York: Springer; 2002.

Classen T, Hokayem C. Childhood influences on youth obesity. Economics and Human Biology. 2005; 3:165–187. [PubMed: 15994141]

D'Orazio MB, Di Zio M, Scanu M. Statistical matching for categorical data: Displaying uncertainty and using logical constraints. Journal of Official Statistics. 2006; 22(1):137–157.

Downey DB, von Hippel PT, Broh BA. Are schools the great equalizer? Cognitive inequality during the summer months and the school year American Sociological Review. 2004; 69(5):613–635.

Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but they can't make them think: Statistical reform lessons from medicine. Psychological Science. 2004; 15(2):119–126. [PubMed: 14738519]

Fitzgerald J, Gottschalk P, Moffitt R. An analysis of sample attrition in the Michigan Panel Study of Income Dynamics. Journal of Human Resources. 1998; 33:251–99.

Freedman DS, Khan LK, Serdula MK, Ogden CL, Dietz WH. Racial and ethnic differences in secular trends for childhood BMI, weight, and height. Obesity. 2006; 14(2):301–308. [PubMed: 16571857]

Freedman V, Wolf DA. A case study on the use of multiple imputation. Demography. 1995; 32:459–470. [PubMed: 8829977]

Gelman A, King G, Liu C. Not asked and not answered: Multiple imputation for multiple surveys. Journal of the American Statistical Association. 1998a; 94(443):846–857.

Gelman A, King G, Liu C. Rejoinder. Journal of the American Statistical Association. 1998b; 94(443): 869–874.

Goldscheider F, Goldscheider C, StClair P, Hodges J. Changes in returning home in the United States, 1925–1985. Social Forces. 1999; 78(2):695–720.

Handcock MS, Huovilainen SM, Rendall MS. Combining registration-system and survey data to estimate birth probabilities. Demography. 2000; 37(2):187–192. [PubMed: 10836176]

Handcock MS, Rendall MS, Cheadle JE. Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. Sociological Methodology. 2005; 35:291–334.

Hellerstein J, Imbens GW. Imposing moment restrictions from auxiliary data by weighting. Review of Economics and Statistics. 1999; 81(1):1–14.

Imbens GW, Lancaster T. Combining micro and macro data in microeconometric models. Review of Economic Studies. 1994; 61:655–680.

Johnson DR, Young R. Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. Journal of Marriage and Family. 2011; 73:926–945.

Judkins DR. Not asked and not answered: Multiple imputation for multiple surveys. Comment Journal of the American Statistical Association. 1998; 94(443):861–864.

Kimbro RT, Brooks-Gunn J, McLanahan S. Racial and ethnic differentials in overweight and obesity among 3-year-old children American. Journal of Public Health. 2007; 97(2):298–305.

Kuczmarski RJ, Ogden CL, Guo SS, Grummer-Strawn LM, Flegal KM, Mei Z, Wei R, Curtin LR, Roche AF, Johnson CL. 2000 CDC growth charts for the United States: methods and development. Vital Health Statistics. 2002; 11(246)

Lancaster, T. An Introduction to Modern Bayesian Econometrics. Malden, MA: Blackwell; 2004.

Lee KJ, Carlin JB. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology. 2010; 171:624–632. [PubMed: 20106935]

Little RJA. Regression with missing X's: A review. Journal of the American Statistical Association. 1992; 87(420):1227–1237.

Little RJA, Rubin DB. The analysis of social science data with missing values. Sociological Methods and Research. 1989; 18:292–326.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2. Hoboken, NJ: Wiley; 2002.

Martin MA. The intergenerational correlation in weight: How genetic resemblance reveals the social role of families. American Journal of Sociology. 2008; 114(Suppl):S67–S105.

McCloskey DN. The loss function has been mislaid: The rhetoric of significance tests. American Economic Association Papers and Proceedings. 1985; 75(2):201–205.

McLaren L. Socioeconomic status and obesity. Epidemiological Reviews. 2007; 29(1):29–48.

Meng XL. Multiple-imputation inferences with uncongenial sources of input. Statistical Science. 1994; 9(4):538–573.

Miech RA, Kumanyika SK, Stettler N, Link BG, Phelan JC, Chang VW. Trends in the association of poverty with overweight among U.S. adolescents, 1971–2004. Journal of the American Medical Association. 2006; 295(20):2385–2393. [PubMed: 16720824]

Mollborn S, Morningstar E. Investigating the relationship between teenage childbearing and psychological distress using longitudinal evidence. Journal of Health and Social Behavior. 2009; 50(3):310–326. [PubMed: 19711808]

Moriarity C, Scheuren F. A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation. Journal of Business and Economic Statistics. 2003; 21(1):65–73.

National Center for Health Statistics. National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; no date. http://www.cdc.gov/nchs/nhanes.htm

Ogden CL, Carroll MD, Curtin LR, Lamb MM, Flegal KM. Prevalence of high body mass index in U.S. children and adolescents, 2007–2008. Journal of the American Medical Association. 2010; 303(3):242–49. [PubMed: 20071470]

Pfefferman D, Sverchkov M. Small-area estimation under informative probability sampling of areas and within the selected areas. Journal of the American Statistical Association. 2007; 102:1427–1439.

Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annual Review of Public Health. 2004; 25:99–117.

Raghunathan TE, Grizzle JE. A split questionnaire survey design. Journal of the American Statistical Association. 1995; 94(447):896–908.

Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology. 2001; 27(1):85–95.

Raghunathan, TE.; Solenberger, P.; van Hoewyk, J. IVEware: Imputation and Variance Estimation Software. 2000. http://www.isr.umich.edu/src/smp/ive/

Rao SR, Granbard BI, Schmid CH, Morton SC, Louis TA, Zaslavsky AM, Finkelstein DM. Meta-analysis of survey data: Application to health services research. Health Services Outcomes Research Methods. 2008; 8:98–114.

Rässler, S. Statistical matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. New York: Springer Verlag; 2002.

Reiter JP, Raghunathan TE, Kinney SK. The importance of modeling the sampling design in multiple imputation for missing data. Survey Methodology. 2006; 32:143–149.

Rendall MS, Admiraal R, DeRose A, DiGiulio P, Handcock MS, Racioppi F. Population constraints on pooled surveys in demographic hazard modeling. Statistical Methods and Applications. 2008; 17(4):519–539. [PubMed: 20668649]

Rendall MS, Handcock MS, Jonsson SH. Bayesian estimation of Hispanic fertility hazards from survey and population data. Demography. 2009; 46(1):65–84. [PubMed: 19348109]

Rendall MS, Weden MM, Favreault MM, Waldron H. The protective effect of marriage for survival: A review and update. Demography. 2011; 48(2):481–506. [PubMed: 21526396]

Ridder, G.; Moffitt, RA. The econometrics of data combination. In: Heckman, JJ.; Leamer, EE., editors. Handbook of Econometrics. Vol. 6b. Amersterdam: North Holland; 2007. p. 5469-5547.

Roberts, G.; Binder, D. Analyses based on combining similar information from multiple surveys. Proceedings of the Joint Statistical Meetings Section on Survey Methods Research; 2009. p. 2138-2147.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Rodgers WL. An evaluation of statistical matching. Journal of Business & Economic Statistics. 1984; 2(1):91–102.

Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business and Economic Statistics. 1986; 21(1):65–73.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.

Salsberry PJ, Reagan PB. Dynamics of early childhood overweight. Pediatrics. 2005; 116(6):1329–1338. [PubMed: 16322155]

SAS Institute. SAS 9.2 Product Documentation. no date. http://support.sas.com/documentation/92/index.html

Sassler S, McNally J. Cohabiting couples' economic circumstances and union transitions: A re-examination using multiple imputation techniques. Social Science Research. 2003; 32(4):553–578.

Schafer, JL. Analysis of Incomplete Multivariate Data. Boca Raton, FL: Chapman and Hall; 1997.

Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods. 2002; 7(2):147–177. [PubMed: 12090408]

Schenker N, Raghunathan TE, Bondarenko I. Improving on analysis of self-reported data in a large-scale health survey by using information from an examination-based survey. Statistics in Medicine. 2010; 29:553–545.

Singh AC, Mantel HJ, Kinack MD, Rowe G. Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. Survey Methodology. 1993; 19(1):59–79.

Singh GK, Siahpush M, Kogan MD. Rising social inequalities in U.S. childhood obesity, 2003–2007. Annals of Epidemiology. 2010; 20(1):40–52. [PubMed: 20006275]

Snow, K.; Thalji, L.; Derecho, A.; Wheeless, S.; Kinsey, S.; Rogers, J.; Raspa, M.; Park, J. Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Kindergarten 2006 and 2007 Data File User's Manual. Institute of Education Sciences, US Department of Education; 2009.

Taylor KW, Frideres J. Issues and controversies: Substantive and statistical significance. American Sociological Review. 1972; 37(4):464–472.

Tighe E, Livert D, Barnett M, Saxe L. Cross-survey analysis to estimate low-incidence religious groups. Sociological Methods and Research. 2010; 39(1):56–82.

U.S. Department of Education. Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Kindergarten through Eighth Grade Full Sample Public-Use Data and Documentation (DVD). Washington, D.C: National Center for Education Statistics; 2009a.

U.S. Department of Education. Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Base Year Public-Use Data Files and Electronic Codebook. Washington, D.C: National Center for Education Statistics; 2009b.

U.S. Department of Labor. Consumer Price Index. Washington, D.C: Bureau of Labor Statistics; 2012. Retrieved May 14, 2010 (ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt)

von Hippel PT. Regression with missing Y's: An improved strategy for analyzing multiply imputed data. Sociological Methodology. 2007; 37:83–117.

Weakliem WL. Introduction to special issue on model selection. Sociological Methods and Research. 2004; 33(2):167–187.

Weden MM, Brownell P, Rendall MS. Prenatal, perinatal, early-life, and sociodemographic factors underlying racial differences in the likelihood of high body mass index in early childhood. American Journal of Public Health. (Forthcoming).

Western B. Causal heterogeneity in comparative research: A Bayesian hierarchical model. American Journal of Political Science. 1998; 42(4):1233–1259.

White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. 2010; 29:2920–31. [PubMed: 20842622]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Percentage Obese in the Early Childhood Longitudinal Study, 1998 Kindergarten Cohort (ECLS-K) and 2001 Birth Cohort (ECLS-B), and National Health and Nutrition Survey (NHANES) 1999–2000 and 2005–2008 [1]

	Observations[2]	% Obese	95% Confidence Interval
ECLS-K 1999, Fall Kindergarten [1]	15,240	11.6	(11.0, 12.3)
NHANES 1999–2000, Age 4–6	497	11.4	(7.6, 15.2)
ECLS-B 2006–7, Kindergarten[2]	5,200	16.4	(15.0, 17.9)
NHANES 2005–8, Age 4–6	1,217	12.1	(9.4, 14.7)

Notes:

[1]All percentages are weighted using ECLS-K, ECLS-B, and NHANES sample weights; confidence interval estimates adjust for stratification and clustering in the sample designs.

[2]Observations are rounded to comply with NCES disclosure guidelines.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Weighted Percentages and Means for Study Variables from the Kindergarten Waves of the Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K) 1998 and Birth Cohort (ECLS-B) 2001, Percentages Unless Otherwise Noted[a]

	ECLS-B						ECLS-K						ECLS-B Total versus ECLS-K Total (p-value)
	Total	Non-Hispanic White	Non-Hispanic Black	Hispanic	Non-Hispanic Asian	Non-Hispanic Other	Total	Non-Hispanic White	Non-Hispanic Black	Hispanic	Non-Hispanic Asian	Non-Hispanic Other	
Child obese[b]	16.4	12.0	19.4	26.4	10.0	18.1	11.6	10.3	12.3	16.2	11.1	13.3	<.001
Mother's race/ethnicity													.017
Non-Hispanic White	57.7						64.4						
Hispanic	23.2						16.5						
Non-Hispanic Black	13.7						13.8						
Non-Hispanic Asian	3.2						2.7						
Non-Hispanic Other	2.3						2.6						
Mother's education													.002
<9th grade	4.3	0.7	1.3	15.3	3.6	4.0	4.1	0.8	1.4	18.8	6.1	2.5	
9th–12th grade (no diploma)	10.2	6.4	15.6	17.3	3.6	9.6	8.9	5.7	14.3	16.9	6.0	9.8	
High school or GED	27.9	23.4	39.8	32.3	20.6	32.7	30.8	29.7	37.5	30.6	19.4	37.2	
At least some college	30.6	32.2	31.9	26.6	19.2	38.2	32.9	34.5	35.4	25.4	23.5	36.7	
Bachelor's degree	16.1	22.8	6.7	4.2	28.9	7.3	15.7	19.3	8.4	5.9	30.7	9.9	
At least some grad. school	10.9	14.5	4.7	4.2	24.2	8.2	7.7	9.9	3.0	2.4	14.2	3.9	
Log adjusted household income	10.46	10.78	9.67	10.09	10.95	10.18	10.57	10.81	9.97	10.13	10.75	10.22	
Mother's marital status													<.001
Never married	19.5	8.9	55.6	26.1	5.1	23.7	13.4	6.4	43.7	15.5	5.8	22.6	
Currently married	69.4	80.9	32.5	61.6	88.5	53.3	72.3	80.3	38.0	69.5	88.0	58.6	
Formerly married	11.1	10.2	12.0	12.2	6.4	23.0	14.2	13.3	18.3	15.0	6.2	18.8	
Child's birthweight													<.001
Low (<2,500 grams)	7.3	6.3	12.1	7.0	8.1	8.6	7.4	6.0	13.9	7.2	8.8	6.7	
Normal (2,500–3,999 grams)	83.6	83.0	84.0	84.1	87.9	86.2	80.8	80.4	79.2	83.0	84.6	81.9	

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

	ECLS-B						ECLS-K						ECLS-B Total versus ECLS-K Total (p-value)
	Total	Non-Hispanic White	Non-Hispanic Black	Hispanic	Non-Hispanic Asian	Non-Hispanic Other	Total	Non-Hispanic White	Non-Hispanic Black	Hispanic	Non-Hispanic Asian	Non-Hispanic Other	
High (≥4,000 grams)	9.0	10.8	3.9	8.9	4.0	5.1	11.8	13.7	6.9	9.8	6.6	11.4	
Mother's age (in years) at child's birth	27.4	28.3	25.3	26.2	30.0	25.5	27.4	28.1	25.5	26.2	29.6	26.1	.763
Child multiple birth	3.2	3.7	2.9	2.2	2.4	2.6	2.4	2.6	2.6	1.5	2.4	1.9	.004
Child female	49.2	50.3	50.1	46.1	47.5	50.0	48.9	48.6	49.8	49.1	49.2	49.0	.720
Child's age (in months)	68.2	68.5	68.0	67.8	67.4	67.5	68.5	68.8	68.1	67.7	67.7	68.2	.055
Child's total number of siblings	1.52	1.48	1.62	1.60	1.25	1.46	1.46	1.39	1.60	1.57	1.54	1.74	.062
Maternal BMI [c]	28.3	27.5	30.6	29.1	24.2	28.7							
Observations [d]	5,200	2,350	800	950	700	350	15,240	9,570	1,970	2,350	830	510	

Notes:

[a] The ECLS-K sample is observed in the fall of kindergarten 1998; the ECLS-B is observed during kindergarten of the 2006–2007 or 2007–2008 school year. Statistical test of differences between ECLS-B and ECLS-K figures adjust for clustering and stratification in the sample designs.

[b] Differences in the prevalence of child obesity by race/ethnicity are statistically significant in the ECLS-B (p<.001) and ECLS-K (p<.001).

[c] Differences in mean maternal BMI by race/ethnicity are statistically significant (p<.001).

[d] Observations are rounded to comply with NCES disclosure guidelines and thus may not sum to the total.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

AIC and BIC Model Fit Statistics for Multivariate Logistic Regressions Predicting Child Obesity using (in Panel A) the Full Sample and (in Panel B) a Sample Restricted to White, Black, and Hispanic Race/Ethnicity, Pooling Data from the Early Childhood Longitudinal Study, 1998 Kindergarten Cohort (ECLS-K) and 2001 Birth Cohort (ECLS-B)

	BIC	AIC
Panel A: Full Samples of ECLS-B and ECLS-K		
Model "a" with no survey indicator variable for ECLS-B vs. ECLS-K	14,916.3	14,757.8
Model "b" with inclusion of survey indicator variable	14,866.0 *	14,699.6 *
Model "c" with interactions between the survey indicator and all study variables	15,017.4	14,700.5
Panel B: Restricted Samples of ECLS-B and ECLS-K [1]		
Model "a" with no survey indicator variable for ECLS-B vs. ECLS-K	13,152.9	13,012.5
Model "b" with inclusion of survey indicator variable	13,101.0 *	12,952.9 *
Model "c" with interactions between the survey indicator and all study variables	13,246.6	12,965.9

Notes:

[1] The restricted samples of ECLS-K and ECLS-B are retricted to children whose mother was non-Hispanic white, non-Hispanic black, or Hispanic; they thus exclude children whose mother was non-Hispanic Asian or non-Hispanic other (i.e., American Indian/Alaskan Native, Native Hawaiian Pacific Islander, and more than one race). All models include the following predictor variables: race/ethnicity, mother's education, log of household income, mother's marital status, mother's age at birth, child's age, gender, birth weight, number of siblings, and singleton status. All regressions are unweighted.

* Indicates the best-fitting model of the variations for the respective pooled-survey or individual-survey specifications; smaller AIC or BIC indicates better model fit.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 4

Multiple Imputation Equation Parameters for Estimating Mother's BMI Estimated using Effects Coding and Standardized Variables in the Full and Restricted ECLS-B Samples[1]

	Full ECLS-B Sample		Restricted ECLS-B Sample	
Parameter	**Observed Parameter Estimate**	**Standard Deviation of 20 Imputation Parameters**	**Observed Parameter Estimate**	**Standard Deviation of 20 Imputation Parameters**
Child obese	0.261	0.021	0.267	0.023
Mother's race/ethnicity				
Hispanic	0.065	0.024	−0.054	0.029
Non-Hispanic Black	0.291	0.019	0.171	0.031
Non-Hispanic Asian	−0.432	0.037		
Other	0.086	0.038		
Mother's education				
9th–12th grade	0.044	0.047	0.053	0.050
High school or GED	0.070	0.022	0.103	0.024
Some college	0.085	0.026	0.084	0.027
Bachelor's degree	−0.075	0.034	−0.084	0.037
Some graduate	−0.133	0.045	−0.158	0.040
Log Household income	−0.098	0.017	−0.112	0.018
Mother's marital status				
Currently married	0.014	0.029	0.035	0.019
Formerly married	−0.054	0.042	−0.063	0.038
Birthweight				
Low	−0.116	0.031	−0.111	0.038
High	0.226	0.046	0.224	0.052
Mother's age at child's birth	0.000	0.014	0.012	0.018
Multiple birth	0.079	0.019	0.085	0.021
Female	0.025	0.009	0.028	0.013
Child's age	0.001	0.012	−0.007	0.012
Number of siblings	0.039	0.015	0.023	0.014
Intercept	0.289	0.043	0.303	0.031
Observations[2]	5,200		4,100	

Notes:

[1]The restricted sample of the ECLS-B is retricted to children whose mother was non-Hispanic white, non-Hispanic black, or Hispanic; they thus exclude children whose mother was non-Hispanic Asian or non-Hispanic other (i.e., American Indian/Alaskan Native, Native Hawaiian Pacific Islander, and more than one race).

[2]All observations are rounded to comply with NCES disclosure guidelines.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 5

Log Odds of Obesity among Kindergarteners using Data from the Full Samples of the Early Childhood Longitudinal Study, 1998 Kindergarten Cohort (ECLS-K) and 2001 Birth Cohort (ECLS-B), Parameter Estimates (β) and Standard Errors (SE)[1]

Part 1

	ECLS-B				ECLS-K			
	Model 1. ECLS-B no maternal BMI		Model 2. ECLS-B with maternal BMI		Model 3. ECLS-K, no maternal BMI		Model 4. ECLS-K, with maternal BMI	
	β	SE	β	SE	β	SE	β	SE
Mother's race/ethnicity (Ref: non-Hispanic white)								
Hispanic	0.608 **	0.113	0.588 **	0.117	0.485 **	0.073	0.438 **	0.077
Black	0.368 **	0.135	0.209	0.140	0.113	0.090	−0.069	0.095
Asian	−0.477 **	0.155	−0.189	0.158	0.170	0.120	0.425 **	0.128
Other	0.561 **	0.156	0.520 **	0.162	0.134	0.144	0.077	0.153
Mother's education (Ref: < 9th grade)								
9th–12th grade	−0.609 **	0.207	−0.637 **	0.211	−0.181	0.144	−0.193	0.155
High school/GED	−0.669 **	0.185	−0.694 **	0.189	−0.104	0.127	−0.139	0.138
Some college	−0.790 **	0.187	−0.823 **	0.191	−0.175	0.129	−0.214	0.141
Bachelor's	−0.930 **	0.220	−0.831 **	0.223	−0.412 **	0.146	−0.346 *	0.159
At least some graduate school	−0.787 **	0.229	−0.670 **	0.234	−0.556 **	0.167	−0.454 *	0.183
Log household income	−0.120 *	0.049	−0.070	0.052	−0.091 **	0.028	−0.036	0.032
Mother's marital status (Ref: never married)								
Married	−0.142	0.123	−0.132	0.125	−0.194 *	0.088	−0.166 †	0.091
Formerly married	−0.097	0.153	−0.063	0.155	−0.120	0.099	−0.042	0.109
Birthweight (Ref: average)								
Low	−0.631 **	0.114	−0.649 **	0.118	−0.397 **	0.120	−0.395 **	0.124
High	0.508 **	0.144	0.327 *	0.148	0.654 **	0.070	0.447 **	0.088
Mother's age at birth	0.017 *	0.008	0.017 *	0.008	0.012 *	0.005	0.012 *	0.005
Multiple birth	−0.108	0.138	−0.198	0.144	−0.088	0.209	−0.186	0.215

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Part 1

	ECLS-B				ECLS-K			
	Model 1. ECLS-B no maternal BMI		Model 2. ECLS-B with maternal BMI		Model 3. ECLS-K, no maternal BMI		Model 4. ECLS-K, with maternal BMI	
	β	SE	β	SE	β	SE	β	SE
Female	−0.165 *	0.082	−0.194 *	0.083	−0.151 **	0.052	−0.188 **	0.056
Child's age	−0.001	0.009	−0.002	0.009	−0.002	0.006	−0.002	0.006
Siblings	−0.254 **	0.043	−0.273 **	0.044	−0.154 **	0.025	−0.174 **	0.027
Maternal BMI			0.077 **	0.006			0.090 **	0.009
Survey sample control for Intercept	−0.254 **	0.043	−2.394 **	0.871	−0.800	0.524	−3.964 **	0.653
Observations[3]	5,200		5,200		15,240		15,240	

Part 2

	ECLS-B and ECLS-K Pooled				Ratio of SE's	
	Model 5. Pooled, no maternal BMI		Model 6. Pooled, with maternal BMI		Ratio of Pooled versus ECLS-B, no maternal BMI	Ratio of Pooled versus ECLS-B, with maternal BMI
	β	SE	β	SE	(Model 5)/(Model 1)	(Model 6)/(Model 2)
Mother's race/ethnicity (Ref: non-Hispanic white)						
Hispanic	0.510 **	0.061	0.472 **	0.063	1.86	1.84
Black	0.184 *	0.074	0.008	0.078	1.83	1.79
Asian	−0.115	0.096	0.171 †	0.100	1.61	1.58
Other	0.320 **	0.102	0.265 *	0.108	1.53	1.50
Mother's education (Ref: < 9th grade)						
9th–12th grade	−0.321 **	0.116	−0.340 **	0.124	1.78	1.70
High school/GED	−0.287 **	0.103	−0.320 **	0.110	1.79	1.72
Some college	−0.373 **	0.105	−0.410 **	0.112	1.79	1.70
Bachelor's	−0.588 **	0.120	−0.516 **	0.128	1.83	1.74
At least some graduate school	−0.643 **	0.133	−0.536 **	0.142	1.72	1.65
Log household income	−0.099 **	0.024	−0.044 †	0.027	2.03	1.93
Mother's marital status (Ref: never married)						

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Part 2

	ECLS-B and ECLS-K Pooled				Ratio of SE's	
	Model 5. Pooled, no maternal BMI		Model 6. Pooled, with maternal BMI		Ratio of Pooled versus ECLS-B, no maternal BMI	Ratio of Pooled versus ECLS-B, with maternal BMI
	β	SE	β	SE	(Model 5)/(Model 1)	(Model 6)/(Model 2)
Married	−0.188 **	0.071	−0.165 *	0.073	1.75	1.72
Formerly married	−0.123	0.082	−0.057	0.088	1.87	1.75
Birthweight (Ref: average)						
Low	−0.502 **	0.083	−0.514 **	0.086	1.37	1.37
High	0.623 **	0.063	0.425 **	0.075	2.28	1.98
Mother's age at birth	0.013 **	0.004	0.013 **	0.004	1.86	1.82
Multiple birth	−0.139	0.112	−0.244 *	0.117	1.24	1.23
Female	−0.152 **	0.044	−0.187 **	0.046	1.87	1.81
Child's age	−0.001	0.005	−0.001	0.005	1.85	1.79
Siblings	−0.181 **	0.022	−0.201 **	0.023	1.95	1.89
Maternal BMI			0.086 **	0.007		0.94
Survey sample control for Intercept	0.404 **	0.052	0.402 **	0.054		
	−0.629 †	0.439	−3.635 **	0.518		
Observations[3]	20,440		20,440			

Notes:

[1] All regressions are unweighted.

[2] Mother's weight status in the ECLS-B is calculated from reported height and weight; mother's weight status in the ECLS-K is imputed.

[3] All observations are rounded to comply with NCES disclosure guidelines.

† $P<.10$;

* $P<.05$;

** $P<.01$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table A1

Multivariate Logistic Regressions Predicting Child Obesity using the Full Sample of Pooled Data from the Early Childhood Longitudinal Study, 1998 Kindergarten Cohort (ECLS-K) and 2001 Birth Cohort (ECLS-B)

	Model "a" with no survey indicator variable for ECLS-B vs. ECLS-K		Model "b" with inclusion of survey indicator variable		Model "c" with interactions between the survey indicator and all study variables	
	β	SE	β	SE	β	SE
Mother's race/ethnicity (Ref: non-Hispanic white)						
Hispanic	0.553 **	0.060	0.510 **	0.061	0.485 **	0.073
Non-Hispanic Black	0.207 **	0.074	0.184 *	0.074	0.113	0.090
Non-Hispanic Asian	0.005	0.093	−0.115	0.096	0.170	0.120
Non-Hispanic Other	0.417 **	0.102	0.320 **	0.102	0.134	0.144
Mother's education (Ref: < 9th grade)						
9th–12th grade	−0.304 **	0.116	−0.321 **	0.116	−0.181	0.144
High school/GED	−0.271 **	0.103	−0.287 **	0.103	−0.104	0.127
Some college	−0.347 **	0.105	−0.373 **	0.105	−0.175	0.129
Bachelor's	−0.548 **	0.120	−0.588 **	0.120	−0.412 **	0.146
At least some graduate school	−0.569 **	0.132	−0.643 **	0.133	−0.556 **	0.167
Log household income	−0.106 **	0.024	−0.099 **	0.024	−0.091 **	0.028
Mother's marital status (Ref: never married)						
Married	−0.220 **	0.070	−0.188 **	0.071	−0.194 *	0.088
Formerly married	−0.169	0.081	−0.123	0.082	−0.120	0.099
Birthweight (Ref: average)						
Low	−0.413 **	0.083	−0.502 **	0.083	−0.397 **	0.120
High	0.605 **	0.063	0.623 **	0.063	0.654 **	0.070
Mother's age at birth	0.013 **	0.004	0.013 **	0.004	0.012 *	0.005
Multiple birth	0.029	0.110	−0.139	0.112	−0.088	0.209
Female	−0.155 **	0.044	−0.152 **	0.044	−0.151 **	0.052
Child's age	−0.001	0.005	−0.001	0.005	−0.002	0.006
Number of Siblings	−0.179 **	0.022	−0.181 **	0.022	−0.154 **	0.025
Survey sample control for ECLS-B (Ref: ECLS-K)			0.404 **	0.052	1.055	0.983
Interactions with survey-indicator						
(Hispanic) * (ECLS-B)					0.124	0.134

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

	Model "a" with no survey indicator variable for ECLS-B vs. ECLS-K		Model "b" with inclusion of survey indicator variable		Model "c" with interactions between the survey indicator and all study variables	
	β	SE	β	SE	β	SE
(Non-Hispanic Black) * (ECLS-B)					0.255	0.162
(Non-Hispanic Asian) * (ECLS-B)					−0.647 **	0.196
(Non-Hispanic Other) * (ECLS-B)					0.427 *	0.212
(9th–12th grade) * (ECLS-B)					−0.429 †	0.252
(High school/GED) * (ECLS-B)					−0.566 *	0.224
(Some college) * (ECLS-B)					−0.615 **	0.227
(Bachelor's) * (ECLS-B)					−0.518 *	0.264
(Graduate school) * (ECLS-B)					−0.231	0.283
(Log household income) * (ECLS-B)					−0.028	0.057
(Married) * (ECLS-B)					0.053	0.151
(Formerly married) * (ECLS-B)					0.023	0.182
(Low) * (ECLS-B)					−0.234	0.166
(High) * (ECLS-B)					−0.146	0.160
(Mother's age at birth) * (ECLS-B)					0.005	0.009
(Multiple) * (ECLS-B)					−0.020	0.250
Interactions with survey-indicator						
(Female) * (ECLS-B)					−0.015	0.097
(Child's age) * (ECLS-B)					0.001	0.011
(Number of Siblings) * (ECLS-B)					−0.099 *	0.050
Intercept	−0.459	0.437	−0.629 †	0.439	−0.801	0.524
Observations [1]		20,440		20,440		20,440
BIC		14,916.3		14,866.0		15,017.4
AIC		14,757.8		14,699.6		14,700.5

Notes:

[1] All observations are rounded to comply with NCES disclosure guidelines.

All regressions are unweighted.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table A2

Log Odds of Obesity among Kindergarteners using Data from the Restricted Samples of the Early Childhood Longitudinal Study, 1998 Kindergarten Cohort (ECLS-K) and 2001 Birth Cohort (ECLS-B), Estimates (β) and Standard Errors (SE) [1]

Part 1

	ECLS-B				ECLS-K			
	Model 1. ECLS-B no maternal BMI		Model 2. ECLS-B with maternal BMI		Model 3. ECLS-K, no maternal BMI		Model 4. ECLS-K, with maternal BMI	
	β	SE	β	SE	β	SE	β	SE
Mother's race/ethnicity (Ref: non-Hispanic white)								
Hispanic	0.611 **	0.115	0.594 **	0.119	0.481 **	0.074	0.454 **	0.078
Non-Hispanic								
Black	0.386 **	0.139	0.228	0.143	0.128	0.091	−0.047	0.098
Mother's education (Ref: < 9th grade)								
9th–12th grade	−0.502 *	0.222	−0.533 *	0.225	−0.191	0.150	−0.189	0.166
High school/GED	−0.678 **	0.200	−0.722 **	0.204	−0.120	0.134	−0.158	0.142
Some college	−0.730 **	0.205	−0.759 **	0.207	−0.196	0.136	−0.220	0.145
Bachelor's	−0.841 **	0.244	−0.743 **	0.247	−0.488 **	0.156	−0.407 *	0.163
At least some graduate school	−0.868 **	0.262	−0.740 **	0.266	−0.604 **	0.177	−0.476 *	0.186
Log household income	−0.145 **	0.054	−0.089	0.057	−0.100 **	0.031	−0.035	0.036
Mother's marital status (Ref: never married)								
Married	−0.060	0.136	−0.075	0.137	−0.152 †	0.092	−0.158	0.098
Formerly married	−0.145	0.170	−0.118	0.170	−0.122	0.103	−0.066	0.114
Birthweight (Ref: average)								
Low	−0.658 **	0.121	−0.683 **	0.124	−0.411 **	0.126	−0.415 **	0.132
High	0.635 **	0.157	0.458 **	0.162	0.665 **	0.073	0.468 **	0.084
Mother's age at birth	0.017 *	0.008	0.017 *	0.009	0.014 **	0.005	0.012 *	0.006
Multiple birth	−0.098	0.145	−0.200	0.151	0.010	0.209	−0.091	0.217
Female	−0.034	0.091	−0.062	0.093	−0.139	0.054	−0.180 **	0.058
Number of Siblings	−0.307 **	0.048	−0.317 **	0.049	−0.191 **	0.028	−0.202 **	0.030
Child's age	−0.008	0.010	−0.008	0.010	−0.002	0.006	0.000	0.007
Maternal BMI			0.076 **	0.007			0.085 **	0.010
Survey sample control for ECLS-B (Ref: ECLS-K)								
Intercept	0.904	0.911	−1.778 †	0.953	−0.733	0.553	−3.950 **	0.738
Observations [3]	4,100		4,100		13,890		13,890	

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Part 2

	ECLS-B and ECLS-K Pooled				Ratio of SE's	
	Model 5. Pooled, no maternal BMI		Model 6. Pooled, with maternal BMI		Ratio of Pooled versus ECLS-B, no maternal BMI	Ratio of Pooled versus ECLS-B, with maternal BMI
	β	SE	β	SE	(Model 5)/(Model 1)	(Model 6)/(Model 2)
Mother's race/ethnicity (Ref: non-Hispanic white)						
Hispanic	0.507 **	0.061	0.482 **	0.064	1.88	1.85
Non-Hispanic						
Black	0.201 **	0.075	0.031	0.080	1.86	1.79
Mother's education (Ref: < 9th grade)						
9th–12th grade	−0.288 *	0.122	−0.300 *	0.132	1.81	1.70
High school/GED	−0.287 **	0.109	−0.328 **	0.115	1.84	1.77
Some college	−0.362 **	0.111	−0.391 **	0.117	1.84	1.77
Bachelor's	−0.615 **	0.129	−0.532 **	0.134	1.89	1.85
At least some graduate school	−0.694 **	0.144	−0.571 **	0.150	1.81	1.77
Log household income	−0.109 **	0.026	−0.046	0.030	2.07	1.92
Mother's marital status (Ref: never married)						
Married	−0.131 †	0.075	−0.139 †	0.079	1.80	1.73
Formerly married	−0.128	0.087	−0.078	0.094	1.95	1.82
Birthweight (Ref: average)						
Low	−0.532 **	0.088	−0.548 **	0.092	1.38	1.36
High	0.659 **	0.066	0.468 **	0.074	2.39	2.18
Mother's age at birth	0.015 **	0.004	0.014 **	0.005	1.92	1.77
Multiple birth	−0.130	0.116	−0.235 †	0.121	1.25	1.24
Female	−0.111 *	0.046	−0.149 **	0.049	1.95	1.89
Number of Siblings	−0.223 **	0.024	−0.234 **	0.026	2.02	1.92
Child's age	−0.003	0.005	−0.001	0.006	1.92	1.84
Maternal BMI			0.082 **	0.008		0.89
Survey sample control for ECLS-B (Ref: ECLS-K)	0.447 **	0.056	0.448 **	0.059		
Intercept	−0.475 **	0.468	−3.532 **	0.585		
Observations 3	17,990		17,990			

Notes:

[1]The restricted samples of ECLS-K and ECLS-B are retricted to children whose mother was non-Hispanic white, non-Hispanic black, or Hispanic; they thus exclude children whose mother was non-Hispanic Asian or non-Hispanic other (i.e., American Indian/Alaskan Native, Native Hawaiian Pacific Islander, and more than one race). All regressions are unweighted.

[2]Mother's weight status in the ECLS-B is calculated from reported height and weight; mother's weight status in the ECLS-K is imputed.

[3]All observations are rounded to comply with NCES disclosure guidelines.

†P<.10;

*P <.05;

**P <.01

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript