# A Case Study of Nonresponse Bias Analysis in Educational Assessment Surveys

**Yajuan Si** [iD]
**Roderick J. A. Little**
*University of Michigan*

**Ya Mo**
*Boise State University*

**Nell Sedransk**
*National Institute of Statistical Sciences*

*Nonresponse bias is a widely prevalent problem for data on education. We develop a ten-step exemplar to guide nonresponse bias analysis (NRBA) in cross-sectional studies and apply these steps to the Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011. A key step is the construction of indices of nonresponse bias based on proxy pattern-mixture models for survey variables of interest. A novel feature is to characterize the strength of evidence about nonresponse bias contained in these indices, based on the strength of the relationship between the characteristics in the nonresponse adjustment and the key survey variables. Our NRBA improves the existing methods by incorporating both missing at random and missing not at random mechanisms, and all analyses can be done straightforwardly with standard statistical software.*

## 1. Introduction

Surveys of educational assessment, such as the National Assessment of Educational Progress survey, the Program for International Student Assessment, and the Early Childhood Longitudinal Study (ECLS), have provided important data to inform policymakers and improve educational experiences. Such large-scale studies often implement complex probability sample designs to collect educational measurements on a sample representative of the target population. Survey variables are only collected for respondents to the study. However, statistical

analyses of the collected data are subject to nonresponse bias, especially given rapidly declining response rates. A variety of indicators of potential bias have been proposed, generally functions of the response rate and the difference between respondents and nonrespondents on variables measured for both respondents and nonrespondents, such as auxiliary variables available in the sampling frame or from external data sources (Andridge and Little, 2011, 2020; Brick and Tourangeau, 2017; Groves, 2006; Hedlin, 2020; Montaquila and Brick, 2009; Särndal and Lundquist, 2014; Schouten et al., 2009; Wagner, 2010). Results may vary depending on the choice of adjustment approaches for nonresponse.

Responding to a solicitation from the U.S. National Center for Education Statistics, we develop exemplars to help guide practices of the nonresponse bias analysis (NRBA) for educational assessment surveys. This article describes an exemplar developed on the ECLS, Kindergarten Class of 2010–2011 (ECLS-K:2011; National Center for Education Statistics, 2013). The ECLS-K:2011 study collects national data on children as they progress from kindergarten through the 2015–2016 school year, when most of them will be in fifth grade. The ECLS-K:2011 program is unprecedented in its scope and coverage of child development, early learning, and school progress, drawing together information from multiple sources, including school administrators, parents, teachers, early care and education providers, and children.

We focus here on a cross-sectional NRBA based on the first wave, 2010 fall data collection. Nonresponse in the ECLS-K occurs when schools in the sample are missing due to lack of cooperation and when children, parents, and teachers are nonrespondents within schools. We summarize the current NRBA implementation in the ECLS-K study, which has the limitation of the strong reliance on the missing at random (MAR) assumption (Rubin, 1976) and is described formally in Section 5. A major objective of our exemplar is the systematic formulation of NRBA steps to guide practice. These steps include a sensitivity analysis based on proxy pattern-mixture models (Andridge and Little, 2011, 2020), which allows for missing not at random (MNAR) missingness mechanisms. We present the NRBA measures and evaluate the quality of such measures based on the predictive performances of auxiliary variables in multivariate models.

This article is organized as follows. Section 2 discusses the background of missing data and NRBA. Section 3 describes the sensitivity analysis framework. The systematic NRBA with 10 detailed steps is summarized in Section 4. We demonstrate the NRBA steps with the ECLS-K:2011 study in Section 5. Section 6 concludes with challenges and future extensions.

## 2. Background

The *pattern* and *mechanism* of missing data play an important role informing the potential bias from unit or item nonresponse. The pattern refers to which values in the data set are observed and which are missing. Specifically, let
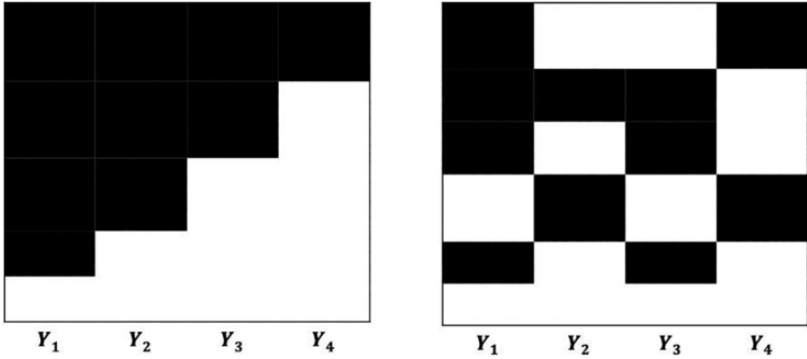
FIGURE 1. *Illustration of missing patterns and the implied response indicator matrices with four variables (black areas indicate response with* R = *1, and white areas indicate nonresponse with* R = *0). The left plot is a monotone "staircase" pattern, and the right shows a general "swiss-cheese" pattern of missingness.*

$Y = (y_{ij})$ denote an $(n \times p)$ rectangular data set without missing values, with the $i$th row $y_i = (y_{i1}, \ldots, y_{ip})$, where $y_{ij}$ is the value of the $j$th variable $Y_j$ for subject $i$, where $i = 1, \ldots, n$ total number of subjects and $j = 1, \ldots, p$ total number of variables. With missing values, the pattern of missing data is defined by the *response indicator matrix* $R = (r_{ij})$, such that $r_{ij} = 1$ if $y_{ij}$ is observed and $r_{ij} = 0$ if $y_{ij}$ is missing; equivalently, $1 - r_{ij}$ is the *missing-data indicator* for $y_{ij}$.

Unit nonresponse occurs when the survey variables $Y$ are missing for units subject to nonresponse and leads to a special case of the *monotone* missing data, where the variables can be ordered, so that $Y_{j+1}, \ldots, Y_p$ are missing for all subjects, where $Y_j$ is missing, for all $j = 1, \ldots, p - 1$, illustrated in the left plot of Figure 1. The monotone pattern often arises in longitudinal data subject to attrition, where once an individual drops out, no more data is observed for that person. In the cross-sectional ECLS-K data, the individual units of children will be missing if their school is missing, and a monotonic pattern arises for student-level and school-level data. Because student-level data are not observed for schools that do not respond to the survey, as an illustration, we can use $Y_1$ to denote the collected school characteristics and $Y_2$ to denote the children assessment variables. Item nonresponse occurs when the unit only responds to partial survey measures, and the right plot of Figure 1 shows that item nonresponse leads to a *general* "swiss-cheese" pattern of missingness.

The missingness mechanism addresses why values are missing and whether these reasons relate to values in the data set. For example, schools or pupils with schools that refuse to participate in the ECLS-K may differ in academic performances from schools or pupils that participate. Rubin (1976) treats $R$ as a random matrix and characterizes the missingness mechanism by the conditional distribution of $R$ given $Y$, say $f(R|Y, \psi)$, where $\psi$ denotes the unknown parameters.

Let $Y_{(1)}$ denote the observed components of $Y$, $Y_{(0)}$ denote the missing components of $Y$, and $X$ be the set of variables observed for respondents and nonrespondents. When missingness does not depend on the values of the data $X$ or $Y$, missing or observed, that is,

$$f(R|X, \ Y, \psi) = f(R|\psi) \text{ for all } Y, \ \psi.$$

The missingness is called missing completely at random (MCAR), and MCAR missing data lead to an increase in the variance of estimates but do not lead to bias. However, MCAR is a strong and unrealistic assumption in most survey settings, including the ECLS-K.

A less restrictive assumption is that missingness depends only on $X$ and the values $Y_{(1)}$ that are observed, and not on values $Y_{(0)}$ that are missing. That is, if $Y_{(0)}$ and $Y_{(0)}^*$ are any two sets of values of the missing data, then

$$f(R|X, \ Y_{(1)}, \ Y_{(0)}, \ \psi) = f(R|X, \ Y_{(1)}, \ Y_{(0)}^*, \ \psi) \text{ for all } X, \ Y_{(0)}, \ Y_{(0)}^*, \ \psi. \tag{1}$$

The missingness is called MAR at the observed values of $R$ and $Y_{(1)}$. If Equation 1 does not hold, the data are MNAR.

Most existing analyses, either by nonresponse weighting or imputation, make the MAR assumption, in part because MNAR analyses are often strongly reliant on untestable assumptions. The inclusion of variables predictive of survey variables strengthens the NRBA, providing more confidence that the MAR assumption is justified. In contrast, an NRBA based on variables that are weakly related to key survey variables provides weak evidence pro or con nonresponse bias; in other words, lack of evidence of bias from such an analysis does not imply lack of bias.

A simple expression of nonresponse bias for the mean of respondents $\bar{y}_R$ is

$$\text{Bias}(\bar{y}_R) = \frac{N - N_R}{N}(\bar{Y}_R - \bar{Y}_{NR}), \tag{2}$$

where $N$ is the population size, $N_R$ is the number of respondents, and $\bar{Y}_R$ and $\bar{Y}_{NR}$ are the respondent and nonrespondent means in the population, respectively. The bias is zero if missingness is MCAR, but that is generally a strong and untenable assumption.

If variables are measured for respondents and nonrespondents, either survey variables measured on other levels in the sample (e.g., collected school characteristics for nonresponding students) or auxiliary variables from the sampling frame and a census or large survey of the population, they can be used to attempt to reduce bias. The main approaches to bias adjustment are *nonresponse weighting*, where responding units are weighted by the inverse of an estimate of the probability of response, and *imputation*, where missing values are imputed by predictions based on observed variables. Weighting is commonly used to adjust for unit nonresponse, and imputation is usually applied to handle item

nonresponse because it is more effective than weighting for handling general patterns of missing data (Little and Rubin, 2019).

To construct response propensity weights for unit nonresponse, let $R_j$ be the indicator for response to $Y_j$, for $j = 1, 2$. With a monotone pattern of data with $Y_2$ less observed than $Y_1$, the MAR condition for missingness of $Y_2$ can be weakened to

$$Pr(R_2 = 1|X, \ Y_1, \ Y_2) = \ Pr(R_2 = 1|X, \ Y_1). \tag{3}$$

That is, the propensity to respond to the survey variable $Y_2$ can depend on the values of survey variables $Y_1$. Assuming MAR with the monotone pattern, $Pr(R_1 = 1|R_2 = 1) = 1$, the probability of response to $Y_2$ can be factored as

$$
\begin{aligned}
Pr(R_2 = 1|X, \ Y_1) = Pr(R_1 = R_2 = 1|X, \ Y_1) = Pr(R_1 = 1|X) \\
\times \ Pr(R_2 = 1|R_1 = 1, \ X, \ Y_1),
\end{aligned}
\tag{4}
$$

and the conditional probability of $R_2$ given $R_1$ can depend on the values of $Y_1$ as well as $X$. The response weight for variable $Y_j$ is then the product up to variable $Y_j$ of the inverse of these estimated conditional propensities (Little and David, 1983). In the ECLS-K study, the unit refers to a student. When the school refuses to participate in the study, all students in that school will be missing; and in participating schools, only a subset of students respond. Therefore, the school-level and child-level missing-data patterns are monotonic, where $R_1$ denotes the school-level response indicator and $R_2$ denotes the child-level response indicator. Applying this factorization in (4) to the ECLS-K study, we model the conditional response propensity of children given the observed variables $X$ and school characteristics $Y_1$ : $Pr(R_2 = 1|R_1 = 1, \ X, \ Y_1)$.

To handle item nonresponse, a drawback of single imputation is that it overestimates the precision of survey estimates. A recommended solution to this problem is *multiple imputation* (MI), where missing values are drawn from their predictive distributions, and multiple data sets are created with different draws of the missing data imputed. Although the theories are rooted in Bayesian statistics, MI can be applied with replication sampling methods, such as bootstrap and jackknife algorithms, to take into account imputation uncertainty. Estimates of the resulting data sets are then combined using Rubin's MI combining rules (see, e.g., Rubin, 1987 or Little and Rubin, 2019). A useful feature of MI for practitioners is that a wide variety of software for MI is now available, as summarized by Yucel (2011) and Si et al. (2022).

To reduce bias, auxiliary variables in the nonresponse adjustment must be predictive of both the survey variable of interest and nonresponse indicator (Little and Vartivarian, 2005). Table 1 presents the bias and variance of MI and inverse propensity weighting for estimates of means, compared to unadjusted analyses based on the complete cases. The variances are calculated based on large-sample approximations. Taken from Little et al. (2022), this table is a

TABLE 1.

*Bias and Variance of MI and IPW Relative to CC Analysis for Estimating a Mean by Strength of Association of the Auxiliary Variables With Response (*R*) and Outcome (*Y*; Little, Carpenter, and Lee, 2022)*

| Association of $X$ With Response $R$ | Association of $X$ With Outcome $Y$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Propensity: Low Other $X$s: Low | | Propensity: Low Other $X$s: High | | Propensity: High Other $X$s: Low | | Propensity: High Other $X$s: High | |
| | IPW | MI | IPW | MI | IPW | MI | IPW | MI |
| Low | Cell LLL | | Cell LLH | | Cell LHL | | Cell LHH | |
| Bias: | --- | --- | --- | --- | --- | --- | --- | --- |
| Var: | --- | --- | --- | ↓ | ↓ | ↓ | ↓ | ↓↓ |
| High | Cell HLL | | Cell HLH | | Cell HHL | | Cell HHH | |
| Bias: | --- | --- | --- | --- | ↓ | ↓ | ↓ | ↓ |
| Var: | ↑ | --- | ↑ | ↓ | ↓ | ↓ | ↓ | ↓↓ |

*Note*. For characterizing the association between $X$ and $Y$, $X$ is split into the propensity, which is the best predictor of $R$ in the regression of $R$ on $X$ (propensity), and components of $X$ orthogonal to the propensity, (other $X$s). The two columns represent two types of association between $X$ and $Y$ are distinguished, the strength of association between the propensity to respond and $Y$, and the strength of association between other $X$s and $Y$, respectively. With a single $X$, the propensity is a function of $X$ and other $X$ is a null set.

"CC" for complete case analysis, "IPW" for inverse propensity weighting, and "MI" for multiple imputation; "---" for bias (or Var) within a cell indicates that the estimate for the method has similar bias (or variance) to the estimate for CC; "↓" for Bias (or Var) within a cell indicates that the estimate for the method has less absolute bias (or variance) than the estimate for CC; "↓↓" for Bias (or Var) within a cell indicates that the estimate for the method has much less absolute bias (or variance) than the estimate for CC; "↑" for Bias (or Var) within a cell indicates that the estimate for the method has greater absolute bias (or variance) than the estimate for CC. In summary, "↓" indicates that a method is better than CC, "↑" indicates that a method is worse than CC, and "---" indicates that a method is similar to CC.

refinement of the simpler table in Little and Vartivarian (2005). Weighting is only effective when the auxiliary variables are related to the survey variables; otherwise, it increases the variance with no reduction in bias (Little et al., 2022; Little and Vartivarian, 2005).

## 3. Methods for NRBA and Sensitivity Analysis

A substantial difference between unadjusted estimates and estimates adjusted by weighting or imputation suggests that nonresponse adjustment is important and, hence, is often a component of NRBA. The key to a useful NRBA is to identify a rich set of auxiliary variables $X$ that are highly predictive for the survey variables $Y$. These might include variables in the sampling frame for bias

adjustments, from external data sources that are not included in the analysis of the data, and also available via data linkage.

We fit multivariate models for the survey variable $Y$ and response propensity $Pr$ $(R = 1)$ and obtain the predicted values for both respondents and nonrespondents. We fit a multivariate regression model with all the auxiliary variables because the variables are often adjusted simultaneously, and checking marginal relationships cannot account for the correlation among auxiliary variables. Variable selection procedures, such as the stepwise forward selection and LASSO (least absolute shrinkage and selection operator; Tibshirani, 1996), can be implemented to select predictive variables and handle multicollinearity among a large number of auxiliary variables. To assess the correlation between the survey variable and response propensity, we group respondents into strata based on the quintiles of the predicted response propensities, $\hat{P}_r(R_i = 1)$, and compare the distributions of survey variables across subgroups. To compare the mean differences of the survey variable between respondents and nonrespondents, we conduct sensitivity analyses under different missing data mechanisms. Conditioning on auxiliary variables and observed survey variables, we denote the predictions of $Y$ as a proxy variable $X$, where $X$ is available for both respondents and nonrespondents. Here, the proxy variable $X$ is a summary of available information denoted by $(X, Y_1)$ to predict $Y_2$ in Equations 3 and 4. We fit proxy pattern-mixture models for the distribution of $(X, Y, R)$ in the population (Andridge and Little 2011; Little 1994; Little et al., 2020): $f(X, Y, R) = f(X, Y|R)f(R)$, where the joint distribution of $(X, Y)$ varies between the respondents $(R = 1)$ and nonrespondents $(R = 0)$, specified as follows:

$$f(X, Y|R = r) = \text{Bivariate} - \text{Normal}\left[\begin{pmatrix} \mu_x^{(r)} \\ \mu_y^{(r)} \end{pmatrix}, \begin{pmatrix} \sigma_{xx}^{(r)} & \sigma_{xy}^{(r)} \\ \sigma_{xy}^{(r)} & \sigma_{yy}^{(r)} \end{pmatrix}\right], \ r = 0 \text{ or } 1,$$

$$Pr(R = 1|X, Y) = g(V), \text{ where } V = (1 - \phi)\sqrt{\frac{s_{yy}}{s_{xx}}}X + \phi Y, \tag{5}$$

which is a bivariate-normal distribution with mean $\left(\mu_x^{(r)} \text{ and } \mu_y^{(r)}\right)$ and variance-covariance parameters $\left(\sigma_{xx}^{(r)}, \ \sigma_{xy}^{(r)}, \ \text{and } \sigma_{yy}^{(r)}\right)$, and the constraint $g(V)$ is a link function (e.g., logit or probit) of $(X, Y)$, the prespecified constant $\phi$, and the estimated sample variance for respondents based on observed data, $s_{yy} = \sigma_{yy}^{(1)}$ and $s_{xx} = \sigma_{xx}^{(1)}$. This additive assumption of $(X, Y)$ in $g(V)$ requires that the effect of $X$ on the missingness is not moderated by $Y$. Since $X$ is the best prediction of $Y$ given the observed variables, we assume that the proxy variable $X$ and the survey variable $Y$ are positively correlated.

We can estimate the mean $\bar{Y}$ as $\hat{\mu}_y$ based on the proxy pattern mixture model. The NRBA index is the difference between $\hat{\mu}_y$ and the respondent mean $\bar{y}_R$ of $\bar{Y}$,

$$\text{NRBA}(\phi) = \hat{\mu}_y - \bar{y}_R = g(\hat{\rho},\ \phi)\sqrt{\frac{s_{yy}}{s_{xx}}}(\bar{x} - \bar{x}_R) = \frac{\phi + (1-\phi)\hat{\rho}}{\phi\hat{\rho} + (1-\phi)}\sqrt{\frac{s_{yy}}{s_{xx}}}(\bar{x} - \bar{x}_R), \quad (6)$$

where $\bar{x}_R$ is the respondent mean of $X$, and $\bar{x}$ is the sample mean of $X$. The function $g(\hat{\rho}, \phi) = \frac{\phi + (1-\phi)\hat{\rho}}{\phi\hat{\rho} + (1-\phi)}$ is a function of the respondent sample correlation $\hat{\rho}$ of $X$ and $Y$ and a sensitivity parameter $\phi$, $0 \le \phi \le 1$. Here, $g(\hat{\rho}, \phi)$ increases with the strength of the proxy, that is, $g(\hat{\rho}, \phi) \to 1$ as $\hat{\rho} \to 1$. With $\phi = 0$, $g(\hat{\rho}, 0) = \hat{\rho}$, the nonresponse is MAR. We will try different values of $\phi$ and compare the effects on the mean estimates. The NRBA index requires the calculation of: (1) $\hat{\rho}$, the correlation between the proxy and survey variables and (2) $d = \bar{x} - \bar{x}_R$, the mean differences of the proxy variable between the population and the respondents. Andridge and Little (2011) have discussed the effect of different correlation $\hat{\rho}$ values on the NRBA. Moderate values of correlation $\hat{\rho}$, such as .5, and even low correlation can provide useful evidence. The choices of the threshold values for both the correlation $\rho$ and the difference $d$ have to depend on the substantive application and whether they alter the key findings. As a subjective recommendation, we would suggest that a correlation of less than .4 is weak, a correlation from .4 to .7 is moderate, and a correlation of more than .7 is strong. We consider a difference of less than $0.1 \times s_{xx}$ as small, between $0.1 \times s_{xx}$ and $0.3 \times s_{xx}$ as medium range and larger than $0.3 \times s_{xx}$ as large. In a survey with multiple outcome measures of interest, we can use the ranking based on the list of $\hat{\rho}$ and $d$ values in the NRBA.

When the inferential interest is subgroup analysis, for example, educational assessments across different race/ethnicity groups, we modify the expression (6) for each subgroup $k$, for $k = 1, \dots K$, and obtain the subgroup mean estimates,

$$\text{NRBA}_k(\phi) = \hat{\mu}_{yk} - \bar{y}_{Rk} = \frac{\phi + (1-\phi)\hat{\rho}_k}{\phi\hat{\rho}_k + (1-\phi)}\sqrt{\frac{s_{yy.k}}{s_{xx.k}}}(\bar{x}_k - \bar{x}_{Rk}), \quad (7)$$

where $\hat{\rho}_k$ is the correlation between $(X, Y)$, and $s_{xx.k}$ and $s_{yy.k}$ are the respondent sample variances of $X$ and $Y$, in subgroup $k$. West et al. (2021) develop regression coefficient estimates with the proxy pattern-mixture models and generate implicit subgroup mean estimates. We extend their results and consider group-specific correlation and variance values. The underlying model we consider adds interactions between subgroup indicators and auxiliary variables in the model for $Y$, while the model in West et al. (2021) only includes main effects and results in the same values of partial correlation and variances across subgroups.

## 4. Main Steps of the NRBA

Figure 2 summarizes the steps of our proposed systematic NRBA.

1. **Analyze missing-data patterns.** Describe the missingness proportions of individual variables and the missingness pattern across variables. The size of bias is
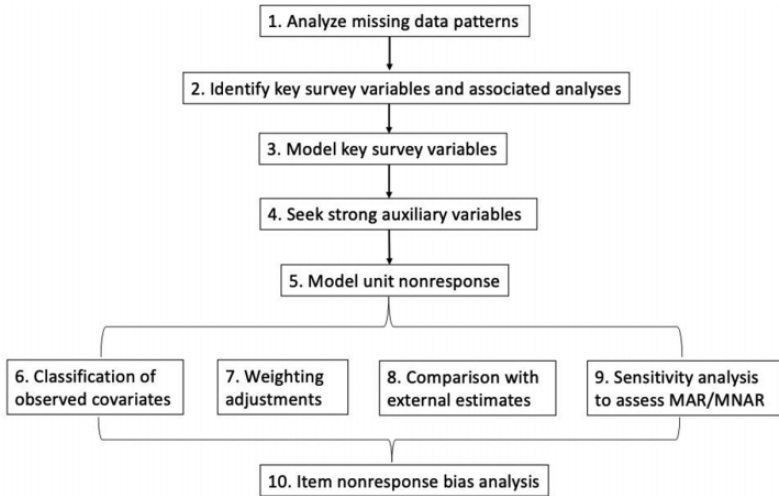
FIGURE 2. *Nonresponse bias analysis process.*

likely to be related to the extent of missing data. More generally, missingness patterns can indicate variables or sets of variables with high levels of nonresponse, which are likely to be vulnerable to nonresponse bias.

2. **Identify key survey variables and associated analyses.** Generally, surveys measure a large number of variables, and it is not feasible to include them all in an NRBA. Thus, identify a small set of "key" survey variables that represent the main subject–matter content of the survey. Also, identify several analyses of interest involving these variables. These could be descriptive or analytic in nature. For survey analysts, such as educational researchers, the NRBA will focus on the specific research questions of interest.

3. **Model key survey variables as a function of fully observed predictors.** The key to a successful NRBA is to find and include "strong" variables that are observed for respondents and nonrespondents and are predictive of key survey variables. Thus, key survey variables should be regressed on fully observed variables to identify predictors of one or more of the key survey variables.

4. **Seek strong observed predictors in auxiliary data.** Many existing NRBAs are limited by the absence of such variables in the data set (Kreuter et al. 2010). Particularly, if the analysis in Step 3 indicates that strong variables are absent in the data set, attempt to link the survey data to external information containing auxiliary variables observed for both respondents and nonrespondents and predictive of survey variables. Such variables are useful for the NRBA, whether measured on individuals in the survey or in aggregate forms, such as marginal proportions or means.

5. **Model unit nonresponse as a function of observed predictors.** Variables that are strongly related to nonresponse are important for nonresponse bias adjustment to the extent that they are also predictive of survey variables. Predictors that

are weakly related to nonresponse but strongly related to survey variables do not lead to bias adjustment but can improve the precision of survey estimates (e.g., Little and Vartivarian, 2005). If external variables are identified in Step 4 and measured at the unit level, separate models should be developed for: (a) variables restricted to those included in the database and (b) variables including external variables identified in Step 4.

6. **Assess observed predictors for the potential for bias adjustment.** Results from Steps 2 to 5 provide the basis for the classification of observed predictors according to the eight cells of Table 1. Key variables for nonresponse bias adjustment are predictive of both nonresponse and one or more key survey variables (Little and Vartivarian, 2005).

7. **Assess the effects of nonresponse weighting adjustments on key survey estimates.** Based on both statistical inferences and substantial findings, small changes between unweighted and weighted estimates suggest that bias may be small, particularly if the adjustment is based on variables that are strongly related to survey variables of interest, as identified in Step 4.

8. **Compare the survey with external data using summary estimates of key survey variables**. Ideally, the external data should be of high quality and are close to serving as the proxy of true values. The comparisons inform potential nonresponse bias, but differences in the estimates could be due to other sources of heterogeneity.

9. **Perform a sensitivity analysis to assess the impacts of deviations from MAR.** Based on the NRBA measure in Equation 6 or 7, we recommend trying different values of the sensitivity parameter and comparing the estimates under different missingness mechanisms. If the resulting confidence intervals largely overlap, the estimates are not sensitive about MAR assumptions. The correlation $\hat{\rho}$ between the proxy and survey variables indicates the quality of the NRBA measure, where larger $\hat{\rho}$ means stronger evidence.

10. **Conduct item nonresponse bias analyses for all analyzed variables.** As with unit nonresponse, the size of bias is related to the amount of missing information. Item nonresponse often results in general missing-data patterns, and an assessment of the degree of item nonresponse can be obtained from Step 1. Item NRBA using the above steps may be indicated for key survey items with high levels of item nonresponse. The values of items that are fully or close to fully observed can be included as additional predictors in item nonresponse models, and MI software for general patterns of missingness allows for fully exploiting information on the observed items. To yield valid inferences of the quantities of interest, imputation approaches that take into account the correlation structure with available data and propagate uncertainty due to missing data improve bias reduction and estimation precision (Si et al., 2022).

## 5. ECLS-K Analysis: Background and Application

We demonstrate the ten steps of our proposed NRBA in the ECLS-K:2011 study with 15,830 responding students out of 18,170 total eligible units. An NRBA needs to account for the design of the survey. We briefly introduce

the ECLS-K:2011 sampling design, weighting adjustment, and current NRBA procedures.

The ECLS-K:2011 adopts a three-stage sample design, with geographic areas as primary sampling units (PSUs), schools sampled within PSUs, and children sampled within schools. A stratified sample of PSUs is selected with probability proportional to size (PPS), where the measure of size is the estimated number of 5-year-old children in the PSU, with oversampling of Asians, Native Hawaiians, and other Pacific Islanders (APIs). All PSUs are grouped into 40 strata defined by metropolitan statistical area status, census geographic region, size class (defined using the measure of size), per capita income, and the race/ethnicity of 5-year-old children residing in the PSU. The sources for the school frames are the most recent Common Core of Data (2006–2007 CCD) and the Private School Survey (2007–2008 PSS). Schools are selected with PPS. The measure of size for schools is kindergarten enrollment adjusted to account for the desired oversampling of APIs. Schools are also sampled from the supplemental frame of newly opened schools and added kindergarten programs that are not in the original frames, and the selection probability for a new school in an existing PSU is conditional on the within-stratum probability of selecting that PSU. Public school substitution is conducted in nonparticipating districts assigned with the base weight of the original school, adjusted for school size differences. In the third stage of sampling, children enrolled in kindergarten of graded schools and 5-year-old children in ungraded schools are selected within each sampled school. Two independent sampling strata are formed within each school, one containing API children with a higher sampling rate than the second containing all other children. Within each stratum, children are selected using equal probability systematic sampling and the target number of 23 at any one school. Once the children are sampled from the school lists of enrolled kindergartners, parent contact information for each child is obtained from the school. The information is used to locate a parent or guardian, conduct the parent interview, and gain parental consent for the child to be assessed. Teachers who teach the sampled children and before- and after-care providers are also included in the study and asked to complete questionnaires.

For the base year of ECLS-K, weights are provided at the child and school levels as the inverse of the probability of the multistage selection. The ECLS-K applies raking to external control margins (Deming and Stephan, 1940). The base-year coverage-adjusted child base weight is raked to external control totals from the number of kindergartners enrolled in public schools in the 2009–2010 CCD and in private schools in the 2009–2010 PSS, the two most up-to-date school frames available at the time of weight computations that are also the closest to the time frame of the kindergarten year of the ECLS-K:2011. Raking cells are created using census region (Northeast, Midwest, South, and West), locale (city, suburb, town, and rural), school type (public, Catholic, non-Catholic

private, and nonreligious private), and kindergarten size (fewer than 85 and 85 or more). After raking, the extremely large weights are trimmed.

The response status is used to adjust the base weight for nonresponse to arrive at the final full sample weight. Nonresponse classes are formed separately for each school type (public/Catholic/non-Catholic private). Within school type, the analysis of child response propensity is conducted using child characteristics, such as date of birth and race/ethnicity to form nonresponse classes. The child-level nonresponse adjustment factor is computed as the sum of the weights for all the eligible (responding and nonresponding) children in a nonresponse class divided by the sum of the weights of the eligible responding children in that nonresponse class.

An NRBA of ECLS-K:2011 by Westat (Tourangeau et al., 2013) examines unit nonresponse with four approaches. The first approach reports school-level and student-level response rates for subgroups—an analysis related to Steps 1 and 5 above. The response rates show variation across school types, census regions, locale, kindergarten enrollment, percent minority, race/ethnicity, and years of birth, and large variation increases the potential for nonresponse bias. With a similar role, the R indicator (Schouten et al., 2009) measures the variability in the probability of responding to a survey as a function of auxiliary variables. Response rates and R indicators are agnostic with regard to specific survey variables of interest, failing to reflect the fact that selection bias depends on the strength of the relationship of selection with the survey variable. We recommend in Step 2 to select a few key variables of interest, which are child assessment outcomes in the ECLS-K:2011 study. The second and third approaches compare sample estimates to estimates computed from the sampling frame, the Census data, and other sources, similar to our recommendation in Step 8. The fourth approach compares ECLS-K:2011 estimates weighted with and without nonresponse adjustments, as recommended in Step 7. Larger differences could be indicative of substantial nonresponse bias; however, the strength of this evidence depends on whether the characteristics used in the nonresponse adjustment are strongly related to survey variables of interest.

Our proposed NRBA process distinguishes from the current practice (Tourangeau et al., 2013) with three aspects: (1) We explicitly conduct an outcome-specific NRBA and examine multiple key survey variables, (2) we calculate the NRBA measures and evaluate the quality of such measures based on the predictive performances of auxiliary variables in multivariate models, and (3) we conduct sensitivity analyses to assess the impact of deviations from MAR. The multistage sampling of PSUs, schools, and children results in nonresponse for schools and children, and there are school substitutes to replace the nonresponding schools. As an illustration, we focus on the child-level NRBA with interest in estimating the mean values of child assessment outcomes overall and across subgroups of interest.
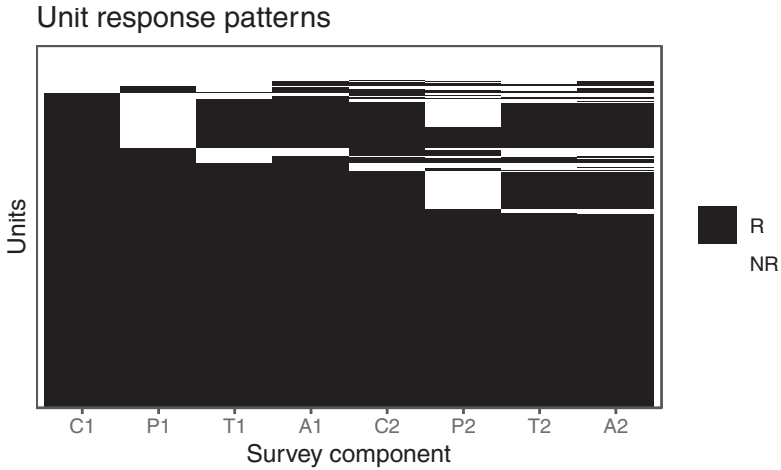
## Unit response patterns



FIGURE 3. *Unit response (R marked by black areas) and nonresponse (NR marked by white areas) patterns of child (C), parent (P), teacher (T), and teacher student-level assessment (A) survey instruments for fall 2010 (1) and spring 2011 (2) data collection. C1: child survey in fall 2010, P1: parent survey in fall 2010, T1: teacher survey in fall 2010, A1: teacher–student assessment survey in fall 2010, C2: child survey in spring 2011, P2: parent survey in spring 2011, T2: teacher survey in spring 2011, and A2: teacher–student assessment survey in spring 2011. Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011, 2010 Fall and 2011 Spring.*

### 5.1. Step 1: Analyze Missing-Data Patterns

The school-level and child-level missing-data patterns are monotonic, shown in the left plot of Figure 1, so school characteristics can be used in the child-level NRBA. In Figure 3, we present the unit nonresponse patterns of the child, parent, teacher, and teacher–student assessment surveys both for the fall and spring collection. The goal is to check response rates and whether nonrespondents in one variable could have information available from other variables that can be used in the NRBA. Conditional response propensities can be computed based on the factorization given in Equation 4.

Detailed response rates by school/child characteristics are reported in the User Manual of the ECLS-K:2011 Kindergarten Study (Tourangeau et al., 2013) and generally high during the base year, with an overall rate of 69% for schools, 87% for children, 74% for parents, and 82% for teachers. Looking into the missing data patterns of Figure 3, we do not have information on parents or teachers for most of the nonresponding children and for a small proportion of responding children. Characteristics of child respondents could be useful to inform the NRBA of parent and teacher interviews.

### 5.2. Step 2: Identify Key Survey Variables and Associated Analyses

Since the ECLS-K study focuses on child development, we have identified a few key survey variables on child assessment outcomes: reading scores estimated by the item response theory (IRT; Hambleton et al., 1991), mathematics IRT scores, child body mass index (BMI), scores on being impulsive/overactive and self-control based on parent interviews, and scores on externalizing and internalizing problems based on teacher interviews.

We are interested in estimates of means in the population and in mutually exclusive subgroups defined by race/ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, API, and Other) and school type (public and private).

### 5.3. Step 3: Model Key Survey Variables as a Function of Fully Observed Predictors

For the cross-sectional NRBA at the baseline, we have frame variables from the 2006–2007 CCD for public schools and the 2007–2008 PSS for private schools. We include the public and private schools and exclude schools selected from the supplemental frames. Because the sample size of private schools is smaller than that of public schools, we conduct the NRBA by combining the two sampling frames. That is, we identify overlapping variables between the CCD and PSS data and select one set of frame variables that are available for both respondents and nonrespondents in the sample.

The auxiliary variables include sex, year of birth, race/ethnicity, school type, census region, locale, the number of students, number of full-time-equivalent teachers, student to teacher ratio, lowest and highest grades offered, percentages of kindergarteners, American Indians, Asians, Hispanics, and Black in schools.

Given the auxiliary variables available for both respondents and nonrespondents, the survey variable is conditionally independent of the response indicator. Because the survey variables are continuous, we fit a linear regression model for $Y$ given the auxiliary variables. As alternatives to regression models, tree-based models, random forests, and gradient boosting algorithms can be used to model the survey variable, as well as the response propensity. Machine learning algorithms automatically detect interactions and nonlinear relationships and could yield good predictive performances. The nodes determined by a tree structure can be used as weighting classes for nonresponse adjustments. As an illustration, we use tree-based models for variable selection and regression models for prediction. We fit a tree model to select high-order interaction terms. Then, we include all main effects and the identified interactions into the linear regression for $Y$ and perform a stepwise variable selection based on the Akaike information criterion to determine the final models.

Using the reading IRT score as an example, the selected predictors in the final outcome model include race/ethnicity, year of birth, sex, school locale, region, lowest and highest grades offered, school type, the number of enrolled students,

the number of full-time-equivalent teachers, percentages of Hispanics, Asian, and Black in school, the two-way interactions between locale and school type, between locale and the percentage of Asian, between race/ethnicity and region, between race/ethnicity and the percentage of Hispanics, between race/ethnicity and the percentage of Black, between race/ethnicity and the lowest grades offered, and between locale and the number of full-time-equivalent teachers.

We predict the outcome values for both respondents and nonrespondents and obtain the proxy variable $X$. The correlation between the outcome $Y$ and the proxy variable $X$ for the respondents is $\hat{\rho} = .36$. Hence, $X$ is a moderately weak proxy with small $\hat{\rho}$.

### 5.4. Step 4: Seek Strong Observed Predictors in Auxiliary Data

The analysis in Step 3 indicates that strong variables are absent in the data set, and efforts should be made to add more predictors in the model for $Y$ and improve the prediction.

For the child assessment outcomes, the highly predictive variables include poverty level, socioeconomic status, the type of language use at home, parental education, parental marital status at the time of birth, and nonparental care arrangements during the year prior to kindergarten. The correlation $\hat{\rho}$ between the survey variable $Y$ and the proxy variable $X$ increases to .48 after adding them to the outcome model. However, they are only available for a small proportion of nonrespondents, 640 of 2,320 nonrespondents having such information.

### 5.5. Step 5: Model Unit Nonresponse as a Function of Observed Predictors

First, we model the school-level response propensities with logistic regression and find that the predictive variables include school type and percentages of kindergarteners and Asians in schools. The model for the school-level response propensities has a value of .60 for the area under the receiver-operating characteristic curve (AUC), an assessment of discriminatory ability. The small AUC value shows that the auxiliary variables are weakly predictive of the school-level response.

Next, we use a response indicator with $R_i = 1$ if the child $i$ responds to the study; otherwise $R_i = 0$. Including the school-level characteristics and auxiliary variables that are available for both responding and nonresponding children as predictors, we fit a logistic regression with the children response indicator $R_i$ as the outcome to estimate the conditional child-level response propensity. The final model has an AUC value of .67. The selected predictors include race/ethnicity, region, locale, the number of enrolled students, number of full-time-equivalent teachers, student to teacher ratio, highest grades offered, percentages of Hispanics and Black in schools, the two-way interaction between race/ethnicity and region, and the two-way interaction between race/ethnicity and the percentage of Black students in the school.
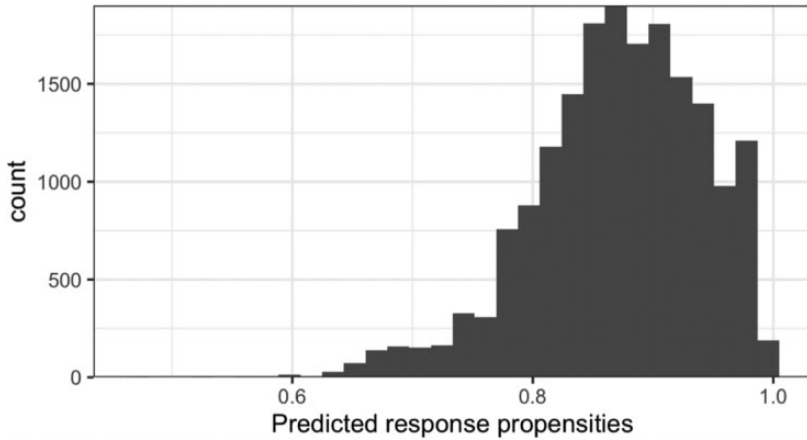
FIGURE 4. *Frequency distribution of the predicted response propensities by the logistic regression.*
Source. *U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011. 2010 Fall.*

Figure 4 depicts the frequency distribution of the predicted child-level response propensities $\hat{Pr}(R_i = 1)$ by the logistic regression. The predicted values are generally large, as the overall response rate is high (87%) and presents a modestly small amount of variation. We examine the relationships between predicted response propensities and key survey outcomes based on the respondents. Figure 5 presents the reading score distributions of respondents stratified by the quintiles of predicted response propensities. The boxplots show that the outcome distributions do not change across response propensity strata. That is, the outcome is weakly correlated with the response propensity, and the estimated correlation is $-.05$.

### 5.6. Step 6: Assess Observed Predictors for the Potential for Bias Adjustment

Results from Steps 2 to 5 provide the basis for the classification of observed predictors according to Table 1. Steps 2 and 3 show that the observed predictors are generally weakly related to the survey variables, and Step 5 finds that the response propensities are weakly related to the survey outcome. Referring to the first column in Table 1, weighting adjustment based on the observed predictors and the response propensities will not substantially affect nonresponse bias.

The resulting measure of deviation for the proxy variable is small with $d = .01$, $X$ is a weak proxy (with small $\hat{\rho}$), so the adjustment in the mean estimate $\hat{\mu}_y$ is small. This is some evidence against nonresponse bias, but this evidence is relatively weak since the correlation is weak.
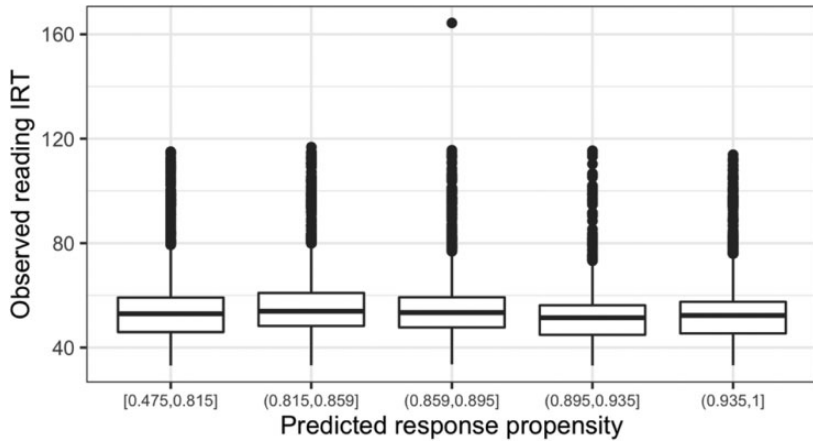
FIGURE 5. *Reading score distributions of respondents stratified by the quintiles of predicted response propensities.* Source. *U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011. 2010 Fall.*

## 5.7. Step 7: Assess Effects of Nonresponse Weighting Adjustments on Key Survey Estimates

We compare ECLS-K:2011 estimates of average reading IRT scores in Table 2 between unweighted and weighted estimates and between estimates using coverage-adjusted base weights and estimates using nonresponse adjusted weights.

Sampling variance estimation has to account for design features, such as clustering and survey weights. We use the Taylor series linearization approximation to obtain the standard errors with the PSU clustering effects and weights in the estimation (Binder, 1983). Table 2 also includes the design effects, calculated as the ratio of estimate variances under the complex design with PSUs and final weights, and the simple random sampling selection with the same sample sizes. For subgroup analyses, in addition to subgroups defined by race/ethnicity and school type, we add the estimates of subgroups defined by the quartiles of socioeconomic status. The estimates are significantly different across subgroups. API students tend to have better reading performances than those in different race/ethnicity groups. Students in private schools have higher scores than those in public schools. The reading scores are highly correlated with socioeconomic status.

The existing nonresponse adjustment of ECLS-K:2011 uses weighting classes defined by the cross-tabulation of date of birth, race/ethnicity, and school type (Tourangeau et al., 2013). This shows that the nonresponse weighting factors will

TABLE 2.

*Comparison of Unweighted, Nonresponse-Unadjusted Weighted Estimates and Weighted Estimates*

| Quantity | Unweighted | | Nonresponse-Unadjusted, Base Weighted | | Nonresponse-Adjusted, Weighted | | |
|---|---|---|---|---|---|---|---|
| | Mean | SE | Mean | SE | Mean | SE | Deff |
| Overall | 54.07 | 0.30 | 53.89 | 0.30 | 53.85 | 0.31 | 11.68 |
| Race/ethnicity | | | | | | | |
| White (not Hisp) | 55.45 | 0.34 | 55.54 | 0.36 | 55.54 | 0.35 | 7.62 |
| Black (not Hisp) | 52.25 | 0.58 | 52.26 | 0.48 | 52.17 | 0.48 | 5.13 |
| Hispanic | 50.32 | 0.49 | 50.19 | 0.49 | 50.13 | 0.49 | 10.05 |
| API (not Hisp) | 58.17 | 0.88 | 57.20 | 1.00 | 57.37 | 1.04 | 6.74 |
| Other | 56.25 | 0.65 | 55.45 | 0.59 | 55.36 | 0.60 | 1.70 |
| School type | | | | | | | |
| Public | 53.68 | 0.32 | 53.49 | 0.32 | 53.49 | 0.33 | 11.81 |
| Private | 56.64 | 0.48 | 56.89 | 0.49 | 56.76 | 0.49 | 3.90 |
| Socioeconomic status quartiles | | | | | | | |
| 25% below | 47.70 | 0.30 | 47.88 | 0.32 | 47.77 | 0.31 | 1.83 |
| 25%–50% | 50.28 | 0.27 | 50.26 | 0.29 | 50.22 | 0.29 | 5.01 |
| 50%–75% | 54.40 | 0.28 | 54.45 | 0.29 | 54.41 | 0.29 | 3.50 |
| 75% above | 60.15 | 0.37 | 59.71 | 0.36 | 59.78 | 0.36 | 3.33 |

*Note.* SE = standard error; Deff = design effect.
*Source.* U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011. 2010 Fall.

be equal for subjects that fall in each subgroup defined by race/ethnicity or school type. This is not the case for the subgroups defined by socioeconomic status. The unweighted, unadjusted, and adjusted estimates are not substantially different. The findings on nonresponse adjustment effects are consistent for overall and subgroup mean estimates. The complex sample design is considerably less efficient than simple random sampling. Because the weighted and unweighted estimates and standard errors are similar, the large design effects are mainly due to the PSU clustering effects, not due to weighting adjustments, which is confirmed by the large values when only accounting for clustering in the complex design.

### 5.8. Step 8: Compare the Survey With External Data Using Summary Estimates of Key Survey Variables

Current NRBA by Westat compare estimates of selected items from the base year ECLS-K:2011 parent interviews and the parent interviews in the 2007

TABLE 3.

*Nonresponse Bias Analysis for Mean Estimates (With Standard Errors in Parenthesis) Under Different Missingness Mechanisms for Different Survey Variables*

| Quantity | Correlation $\hat{\rho}$ | Deviation $d$ | $\bar{y}_R$ | $\hat{\mu}_y$: MAR ($\phi = 0$) | $\hat{\mu}_y$: MNAR ($\phi = 0.5$) | $\hat{\mu}_y$: MNAR ($\phi = 1$) |
|---|---|---|---|---|---|---|
| Child | | | | | | |
| Reading IRT score | .36 | .01 | 54.07 | 54.10 (.09) | 54.16 (.10) | 54.30 (.13) |
| Math IRT score | .44 | .01 | 35.56 | 35.61 (.09) | 35.68 (.09) | 35.83 (.12) |
| Body mass index | .18 | −.03 | 16.50 | 16.49 (.02) | 16.44 (.02) | 16.20 (.04) |
| Parent | | | | | | |
| Impulsive/ overactive | .18 | −.02 | 2.04 | 2.04 (.01) | 2.03 (.01) | 2.00 (.02) |
| Self-control | .13 | .01 | 2.89 | 2.89 (.01) | 2.87 (.01) | 2.75 (.02) |
| Teacher | | | | | | |
| Externalizing problems | .28 | −.01 | 1.60 | 1.60 (.01) | 1.59 (.01) | 1.54 (.01) |
| Internalizing problems | .14 | .01 | 1.46 | 1.46 (.004) | 1.46 (.005) | 1.45 (.01) |

*Note. MAR* = missing at random; MNAR = missing not at random; IRT = item response theory.
*Source.* U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011. 2010 Fall.

National Household Education Survey, subset to parents of kindergartners. The differences in the estimates between the two studies could be due to various sources of heterogeneity in data collection, such as time discrepancies, coverage, and sample design. These comparisons inform potential nonresponse bias but are not direct assessments. External data of high quality are crucial to validate the NRBA.

### 5.9. Step 9: Sensitivity Analysis

We set the sensitivity parameter $\phi$ as 0, 0.5, and 1, respectively, where $\phi = 0$ indicates MAR, $\phi = 1$ indicates an MNAR case where the missing-data mechanism only depends on the survey variable $Y$, and the midpoint $\phi = 0.5$ indicates MNAR, where both survey and proxy variables affect the response propensity. We compare the different estimates of $\hat{\mu}_y$ to assess the deviation from MAR. Table 3 displays maximum likelihood estimates of the mean estimates under different missing data mechanisms, and the standard error estimates are based on large-sample approximations based on Equation

TABLE 4.
*Nonresponse Bias Analysis of Reading IRT Scores for Subgroups*

| Quantity | Correlation $\hat{\rho}_k$ | Deviation $d_k$ | $\bar{y}_{Rk}$ | $\hat{\mu}_{yk}$: MAR ($\phi = 0$) | $\hat{\mu}_{yk}$: MNAR ($\phi = 0.5$) | $\hat{\mu}_{yk}$: MNAR ($\phi = 1$) |
|---|---|---|---|---|---|---|
| Race/ethnicity | | | | | | |
| White (not Hisp) | .26 | .01 | 55.45 | 55.49 (.13) | 55.61 (.13) | 56.09 (.23) |
| Black (not Hisp) | .27 | −.001 | 52.25 | 52.26 (.22) | 52.29 (.23) | 52.38 (.39) |
| Hispanic | .27 | −.05 | 50.32 | 50.19 (.16) | 49.83 (.16) | 48.54 (.27) |
| API (not Hisp) | .40 | −.01 | 58.17 | 58.07 (.41) | 57.93 (.43) | 57.59 (.60) |
| Other | .31 | .01 | 56.25 | 56.30 (.48) | 56.40 (.49) | 56.74 (.62) |
| School type | | | | | | |
| Public | .35 | .01 | 53.68 | 53.72 (.10) | 53.80 (.10) | 54.02 (.15) |
| Private | .36 | −.002 | 56.64 | 56.63 (.25) | 56.62 (.25) | 56.58 (.33) |

*Source.* U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011. 2010 Fall.

6. Alternatives approaches are Bayesian methods or multiple imputation (Andridge and Little, 2011).

For the reading IRT scores, the mean estimates in Table 3 slightly increase as $\phi$ increases, and the standard error of the case with $\phi = 1$ is larger than the remaining two values. The estimates are similar to each other and not sensitive to different missingness mechanisms, providing some evidence that the potential nonresponse bias is small. Again, the evidence is weak because the proxy variable is not strongly related to the outcome.

Table 3 also presents the NRBA results for other outcomes of interest. The correlations between the outcome and the proxy variables are generally low, ranging from .13 to .44, showing modest evidence for the NRBA. The sensitivity analyses find that the mean estimates for BMI, scores for self-control, and externalizing problems are significantly different under different missing data mechanisms, where the nonrespondents have substantially lower average scores than respondents. This indicates the potential for nonresponse bias for these outcomes, but the evidence is modest.

We estimate average reading IRT scores across subgroups defined by race/ethnicity and school type in Table 4. Similar to those overall estimates, subgroup estimates are not substantially different under different missing data mechanisms. Even with subtle differences, we observe different adjustment effects across subgroups as the missing data mechanisms change in the sensitivity analysis.

### *5.10. Step 10: Item NRBA*

Continuing the example with the reading IRT score, the outcome has 50 more null values than the number of unit nonresponses and 120 values of $-9$ indicating item nonresponse. Since the proportion of nonresponse mismatch is very small, we treat the 15,670 observed cases of the outcome variable as respondents. However, item nonresponse could also lead to substantial bias. Future extensions of this work will perform MI and assess the effects on inferences.

## 6. Discussion

We present a 10-step exemplar of the NRBA for cross-sectional studies. Our NRBA assesses the pattern of missing data, fits regressions of key survey outcomes and indicators of nonresponse on variables observed for both respondents and nonrespondents, compares estimates with and without nonresponse weighting adjustments, and implements sensitivity analyses based on proxy pattern-mixture models to assess the impact of deviations from MAR missingness. All analyses can be carried out straightforwardly with standard statistical software. We provide our example R codes in the Supplemental Appendix.

Overall, we do not find substantial evidence of nonresponse bias in the ECLS-K:2011 study, though modest differences present for several estimates, perhaps reflecting the high response rates at baseline. However, lack of evidence of bias in the NRBA does not necessarily mean lack of bias; the key to a strong NRBA is the existence of a rich set of auxiliary variables that are highly predictive of the survey variables. The strength of the evidence is generally weak in this application because the observed predictors are not strongly related to the survey outcomes. The auxiliary variables are collected from the frame, available for both respondents and nonrespondents, but a time lag exists between the frame (around 2007) and the actual data collection (around 2010), leading to weak correlations with the survey variables. The ECLS-K:2011 dataset has the children's zip codes and geographic information that can be linked to the census data for the neighborhood characteristics. Obtaining auxiliary information via geospatial linking will be future work.

As regard future work, data integration with multiple sources greatly enhances many ongoing survey research activities (National Academies of Sciences, Engineering, and Medicine, 2017). The NRBA requires information that is available for both respondents and nonrespondents. Combining multiple data sources and record linkage can provide auxiliary information for nonresponse adjustment and benchmark information for external validation. The proxy pattern-mixture models assume that the response mechanism depends on an additive effect of the survey variable and the proxy variable, and the assumption cannot be verified without external data. The mean estimates with large sample sizes are robust against the normality misspecification (Andridge and Thompson, 2015), but the

effect on other estimands with small samples is unclear and needs future work. Linking ECLS-K:2011 studies to other data could have great potential for the NRBA. Second, the NRBA of analytic inferences could have different findings from that of descriptive summaries.

We focus here on mean estimates for the population and population subgroups. The NRBA in regression models is important, and many of the steps outlined above can also be applied in the regression setting. Extensions of the proxy pattern-mixture approach to sensitivity analysis for regression are discussed in West et al. (2021). In the regression setting, the key to a strong analysis is the availability of strong auxiliary variables that are not predictors in the regression model of interest. Third, we mainly demonstrate the assessment and adjustment of estimates for unit nonresponse, which is the main concern in ECLS-K; assessment of item nonresponse may be important in surveys where the level of item nonresponse is greater. Si et al. (2022) have demonstrated the applicability of MI in handling various challenges on item nonresponse with massive data sets. MI has been used to simultaneously handle unit nonresponse and item nonresponse (Si et al., 2015, 2016). Combining weighting adjustment and imputation into one systematic process would be helpful for practical survey operation. Future work will also develop an NRBA exemplar for longitudinal studies.

## Declaration of Conflicting Interests

## Funding

## ORCID iD

Yajuan Si  https://orcid.org/0000-0001-8707-7374

## References

Andridge, R. R., & Little, R. J. A. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, *27*, 153–180.

Andridge, R. R., & Little, R. J. A. (2020). Proxy pattern-mixture analysis for a binary variable subject to nonresponse. *Journal of Official Statistics*, *36*(3), 703–728.

Andridge, R. R., & Thompson, K. J. (2015). Assessing nonresponse bias in a business survey: Proxy pattern-mixture analysis for skewed data. *The Annals of Applied Statistics*, *9*(4), 2237–2265.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*(3), 279–292.

Brick, J. M., & Tourangeau, R. (2017). Responsive survey design for reducing nonresponse bias. *Journal of Official Statistics*, *33*(3), 735–752.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*, 427–444.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*, 646–675.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Press.

Hedlin, D. (2020). Is there a "safe area" where the nonresponse rate has only a modest effect on bias despite non-ignorable nonresponse? *International Statistical Review*, *88*(3), 642–657.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., & Raghunathan, T. E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. *Journal of the Royal Statistical Society, Series A*, *173*, 389–407.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, *81*, 471–483.

Little, R. J. A., Carpenter, J., & Lee, K. (2022). A comparison of three popular methods for handling missing data: Complete-case analysis, weighting and multiple imputation. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221113873

Little, R. J. A., & David, M. (1983). *Weighting adjustments for nonresponse in panel survey*. U.S. Department of Commerce, Bureau of the Census. https://www2.census.gov/library/working-papers/1987/demo/SEHSD-WP1987-04.pdf

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.

Little, R. J. A., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, *31*, 161–168.

Little, R. J. A., West, B. T., Boonstra, P. S., & Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, *8*, 932–964.

Montaquilia, J. M., & Brick, J. M. (2009). The use of an array of methods to evaluate nonresponse bias in an ongoing survey program. *ISI Proceedings*. https://2009.isiproceedings.org/A5%20Docs/0215.pdf

National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in federal statistics: Combining data sources while protecting privacy*. The National Academies Press. https://doi.org/10.17226/24652

National Center for Education Statistics. (2013). Early Childhood Longitudinal Study, Kindergarten Class of 2010–11. https://nces.ed.gov/ecls/dataproducts.asp#K-5

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Särndal, C.-E., & Lundquist, P. (2014). Accuracy in estimation with nonresponse: A function of the degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, *2*, 361–387.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*, 101–113.

Si, Y., Heeringa, S., Johnson, D., Little, R. J. A., Liu, W., Pfeffer, F., & Raghunathan, T. (2022). Multiple imputation with massive data: An application to the panel study of income dynamics. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smab038

Si, Y., Reiter, J. P., & Hillygus, S. (2015). Semi-parametric selection models for potentially non-ignorable attrition in panel studies with refreshment samples. *Political Analysis*, *23*, 92–112.

Si, Y., Reiter, J. P., & Hillygus, S. (2016). Bayesian latent pattern mixture models in panel studies with refreshment samples. *The Annals of Applied Statistics*, *10*, 118–143.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Hagedorn, M. C., Daly, P., & Mulligan, G. M. (2013). *User's manual for the ECLS-K:2011 Kindergarten data file and electronic codebook*. National Center for Education Statistics. https://nces.ed.gov/pubs2015/2015074.pdf

Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly*, *74*, 223–243.

West, B. T., Little, R. J. A., Andridge, R. R., Boonstra, P. S., Ware, E. B., Pandit, A., & Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from non-probability samples, with applications to genetics and demographic surveys. *Annals of Applied Statistics*, *15*(3), 1556–1581.

Yucel, R. M. (2011). State of the multiple imputation software. *Journal of Statistical Software*, *45*(1), v45/i01.

## Authors

YAJUAN SI is a research associate professor in the Survey Research Center of the Institute for Social Research at the University of Michigan, ISR 4014, 426 Thompson St, Ann Arbor, MI 48104; Email: yajuan@umich.edu. Her research interests include Bayesian statistics, survey inference, missing data imputation, and data confidentiality protection.

RODERICK J. A. LITTLE is the Richard D. Remington Distinguished University Professor in the Department of Biostatistics, Research Professor in the Institute for Social Research, and Professor in the Department of Statistics at the University of Michigan, M4071 SPH II, 1415 Washington Heights, Ann Arbor, Michigan 48109; Email: rlittle@umich.edu. His research interests include incomplete data, sample surveys, Bayesian statistics, and applied statistics.

YA MO is an assistant professor of Curriculum, Instruction, and Foundational Studies at Boise State University and a research fellow at the National Institute of Statistical Sciences. 1750 K Street, NW, Suite 1100, Washington, DC 20006; Email: ymo@niss.org. Her research interests include quantitative methods, psychometric measures, survey statistics, and large-scale educational assessments.

NELL SEDRANSK is the director of the DC Office of the National Institute of Statistical Sciences and Professor of Statistics at North Carolina State University. 1750 K Street,

NW, Suite 1100, Washington, DC 20006; Email: NSedransk@niss.org. Her research interests include design of complex experiments, Bayesian inference, spatial statistics, and topological foundations for statistical theory.