Planned Missing Data Designs for Causal Inference in Large Surveys: Design and Imputation

By

Dan Su

A dissertation submitted in fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Educational Psychology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2019

Date of final oral examination: 08/22/2018

This dissertation is approved by the following members of the Final Oral Committee:

David Kaplan, Professor, Educational Psychology

Jee-Seon Kim, Professor, Educational Psychology

James Wollack, Professor, Educational Psychology

Felix Elwert, Associate Professor, Sociology

Manuscript 1: An Evaluation of Planned Missing Data Designs in Large Surveys

(Page 1-45)

Manuscript 2: Graphical Models for Planned Missing Data Designs

(Page 46-76)

This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation, under Grant #DRL-1749275. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies.

Abstract

Planned missing data designs in large surveys can efficiently reduce respondents' burden and lower the cost associated with data collection, without cutting down on the questionnaire items. If the missing data are not appropriately planned, it results bias in descriptive and potential causal parameter estimates. For a fixed sample size, the extend of bias depends on the three major characteristics of design and data: the missing percentage, the overlap percentage (i.e., the portion of the cases where two items are observed jointly), and the distributions of variables. My first two simulation studies investigate how the bias in marginal means, correlations and regression coefficients depends on the chosen planned missing data designs and the related characteristics.

Even if a planned missing data design allows researchers to recover parameters of interest without bias, an incorrect choice of covariates at the imputation stage might actually introduce bias. For example, if the missing data pattern of a specific form or booklet causes context effects on an auxiliary variable that is used for imputing missing values, bias can be introduced. Thus, including all measured variables in the imputation model is not necessary a good strategy and, given the huge number of items in large surveys, frequently is problematic. The question then is, how should researchers select the imputation variables to obtain valid parameter estimates? The simulation studies investigate which variables not necessary or should not be included in the imputation model. Graphical models provide the theoretical basis for my simulations and explanations.

An Evaluation of Planned Missing Data Designs in Large Surveys

Dan Su

University of Wisconsin-Madison

Abstract

Planned missing data designs in large surveys can efficiently reduce respondents' burden and lower the cost associated with data collection, without cutting down on the questionnaire items. If the missing data are not appropriately planned, the bias of parameter estimates results. This paper implemented two simulation studies to investigate the bias in the marginal means, correlations and regression coefficients using the planned missing data designs and multiple imputation of the missing data. The first study shows that for a fixed sample size, the extent of bias depends on the three major properties of design and data: overlap percentages, missing percentages, and distributions of variables. The second study applies the properties of design to illustrate that an optimal incomplete block design that ensures overlap can be a better choice than a multiple-form design. The issues and strategies of planning and imputing missing data are discussed.

Keywords

Planned missing data designs; large surveys; optimal incomplete block designs; multiple-form designs; three-form designs; missing data; multiple imputation.

Introduction

Large survey data are great resources for conducting social science research. However, a serious constraint with large surveys is that respondents can answer only a limited number of questionnaire items without being overwhelmed. Moreover, with long questionnaires the validity and reliability of measures very likely decreases. In order to overcome these limitations, researchers can use designs with carefully planned missing data where respondents answer only subsets of questionnaire items. This reduces respondents' burden and lowers the cost associated with data collection but nonetheless allows researchers to collect data on the full set of questionnaire items. Constructing the subsets is a big challenge because the resulting missingness structure in the data should not interfere with the researchers' aim to draw valid descriptive and, if possible, causal conclusions. Since the structure of the planned missingness strongly affects parameter estimates, carefully implemented simulation studies need to be conducted to compare different planned missing data designs with respect to valid and reliable parameter estimation. If the goal is to provide data to end-users, how should planned missing data be handled appropriately? In addition, what kind of missing data designs allow the parameter estimates of interest to be recovered without bias? This paper will address these questions with the focus on designs with two simulation studies.

The first simulation study investigates the optimal amount of missing data that occur in each variable (i.e., missing percentage) and the amount of observed data in each pair of variables jointly (i.e., overlap percentage) regarding the bias in the estimates of means, correlations and regression coefficients. To illustrate how the overlap and missing percentage apply to the specific designs, the second simulation study compares the two-form design, the three-form design, the optimal block design with 50% missingness, and the optimal block design with 33% missingness regarding the bias in the parameter estimates. The results show that given the same

amount of missing percentage, an optimal block design that ensures sufficient overlap and maximizes the efficiency at the same time can be a better choice than a multiple-form design that is frequently implemented in the psychological research.

This paper is organized as follows. In the background section, I give an overview of the designs used in large surveys in sociology, psychology and education. Then, I introduce design properties and commonly used planned missing data designs. Followed by the missing data methods section, I illustrate the missingness mechanism with design examples and discuss modern methods for dealing with planned missing data. The following two sections describe the two simulation studies, including the methods and results. Finally, in the conclusion section I discuss the findings of the two studies, the issues and strategies of planning and imputing missing data.

Background

Carefully planned missing data designs can dramatically reduce the cost associated with data collection, and even increase validity due to reducing participant burden (Rhemtulla & Hancock, 2016). Planned missing data designs are chosen according to the characteristics of data in specific research fields. In sociological research, researchers frequently use factorial surveys (also called vignette experiments) to measure social judgments. For example, researchers measure the perceived income by varying the factors such as gender, education and occupation (Steiner et al., 2016). The factor levels are combined and formed as a vignette or a hypothetical scenario for respondents to assess. Due to the large number of factors and factor levels, the number of full factorial combinations is frequently too large for respondents to assess. Thus, strategies for forming smaller subsets of vignettes such as randomly sampling vignettes (Rossi & Nock, 1982), confounded factorial designs (Kirk, 1995), or D-optimal designs (Atkinson at al., 2007) help to reduce respondents' burden. However, not all of these strategies can satisfactorily

deal with the missing vignette assessments created by design. Randomly selecting vignettes often results in biased parameter estimates because some effects of interest might be randomly confounded (Su & Steiner, 2018; Steiner et al., 2016).

In behavioral or psychological research, due to the financial burden associated with recruiting respondents, many survey questionnaires use existing survey instruments to obtain information that can be used for a variety of purposes. Long questionnaires inevitably increase respondents' burden, which likely increases nonresponse rates. Raghunathan and Grizzle (1995) implemented a split questionnaire survey design in an attempt to reduce the nonresponse rate in the Cancer Risk Behavior Survey. Other planned missing data designs like the three-form design outlined by John Graham and his colleagues (Graham et al., 2006) have gained popularity in psychological research. The split questionnaire survey designs and three-form designs have similar features. The complete questionnaires are split into multiple item sets and only a selection of item sets is assigned to respondents.

A common goal in large-scale educational assessments is to estimate the proficiency or achievement of students in different subject areas, for example in the National Assessment of Educational Progress (NAEP) and the Programme for International Students Assessment (PISA). Matrix-sampling designs (Shoemaker, 1973) and incomplete block designs (Frey et al., 2009; van der Linden et al., 2004) have been used to increase the number of test questions. Combined with a multidimensional Item Response (IRT) model, the proficiency scores of students can be estimated by drawing multiple plausible values from the distribution of the latent proficiency (Neal Thomas, 2004). Many studies have focused on item parameter estimates and proficiency estimates while incorporating the uncertainty due to missing data (Aßmann et al., 2015; Gonzalez & Rutkowski 2010; Hecht et al., 2015; Rutkowski, 2011; Weirich et al., 2014). Missing data designs for context questionnaires were also considered and investigated (Adams et

al., 2013; Kaplan & Su, 2016; authors, 2017; OECD 2014). It is worth noticing that across the different fields in the social sciences, only a few studies have addressed bias in the parameter estimates of substantive interest, while many studies focused on the efficiency or power issue of planed missing data designs (Graham et al., 2006; Pokropek 2011; Rhemtulla et al., 2016).

Planned Missing Data Designs

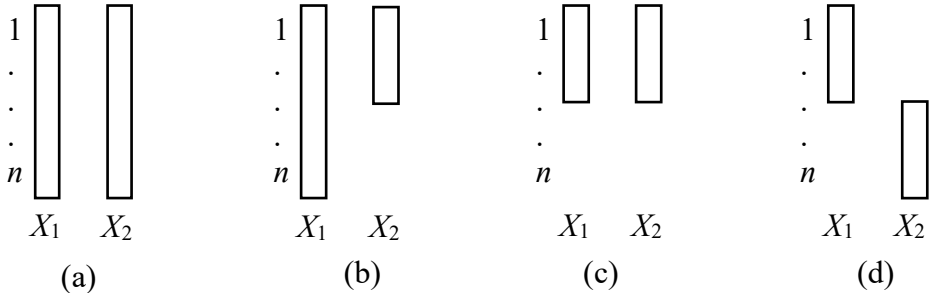
Design Properties

In planned missing data designs, an *item* is an individual task that is administered to a respondent. In this paper, I use the terms item and variable interchangeably. A *block* or *cluster* is a set of items that are blocked by design. I use the term block throughout this paper. A block of variables that contains no planned missing data is called a *common block*. In a large survey, demographic information of respondents represents important data for analysis. For example, gender and race information are collected from all respondents. Blocks with planned missing data are called *rotation blocks*. The variables that are assigned to rotation blocks are referred to as rotation variables. A *form* is the actual set of blocks that is administered to examinees. A form can contain either multiple blocks or only one block. Typically, a form contains a common block and at least one rotation block.

To systematically plan the missing data, the amount of missing data and where the missing data occur should be considered. The *missing percentage* of a single variable is the percentage of missing cases in this variable. If the missing percentage is 100%, the population means are not estimable. The *overlap percentage* of two variables is the percentage of simultaneously observed values in the two variables (relative to total number of cases). If the overlap percentage of two variables is 0%, correlations are not estimable, simply because of no data are available for estimation.

Figure 1 presents the examples of planned missing data in any two variables X_1 and X_2 of a survey questionnaire with n respondents. Imagine we have a dataset where rows are the respondents and columns are the variables. The vertical bars in Figure 1 represent the observed data in variable X_1 or X_2. Figure 1 (a) shows two full sized bars with no planned missing data in both X_1 and X_2. The missing percentage of both variables is 0%. The overlap percentage is 100%. This situation arises when both variables belong to the common block. In Figure 1 (b), the bar of X_2 is half size of X_1, meaning 50% of missing data are planned in X_2. In this case, the overlap percentage between X_1 and X_2 is 50%, because only half of the respondents have observed data in both X_1 and X_2. Such a situation occurs when one variable (X_1) comes from a common block and the other variable (X_2) from a rotation block. Figure 1 (c) shows that both X_1 and X_2 have 50% planned missing data. The overlap percentage is 50% as well. In this case the two variables come from the same rotation block. In Figure 1 (d), the missing percentage of X_1 and X_2 is 50%. However, the overlap percentage is 0%, since the observed data in X_1 and X_2 do not overlap. This example can represent that X_1 and X_2 come from the different rotation blocks.

Figure 1. The scenarios of overlap between two variables in a planned missing data design.



Multiple-form designs

Two-form designs. In two-form designs (Graham et al., 1996; Adams et al., 2013), all variables are divided into three blocks, a common block X and two rotation blocks A and B

(Table 1). Half of the subjects are assigned to form 1 which contains the common block X and rotation block A. The other half of respondents receives form 2, containing block X and B. None of the subjects respond to blocks A and B simultaneously. The missing percentage of the variables in either A or B is 50%. The overlap percentage between the variables from A and variables from B is 0%. Thus, with the two-form design, the correlations between variables from A and B are not estimable.

Table 1. The two-form design.

	Common Block	Rotation Blocks	
Form	X	A	B
1	1	1	0
2	1	0	1

Three-form designs. To avoid the limitation of the two-form design, researchers can use more than two rotation blocks. In three-form designs, variables are divided into four blocks, one common block X and three rotation blocks A, B, and C (Table 2). One third of the respondents get one of the three forms, XAB, XAC, or XBC. The missing percentage of variables in the rotation blocks is 33%. The overlap percentage of two variables across rotation blocks (e.g., one from A and one from B) is 33% as well. Thus, correlations of the variables across rotation blocks are estimable. Researchers can apply similar ideas and extend the number of rotation blocks. The form always contains the common block and any two of the rotation blocks. Raghunathan and Grizzle (1995) implemented a split questionnaire survey design with five rotation blocks. Five rotation blocks result in 10 forms, since there are 10 ways (5 choose 2) to combine any of the two rotation blocks. However, the larger the number of rotation blocks, the larger the missing percentage and the smaller the overlap percentage.

Table 2. The three-form design.

	Common Block	Rotation Blocks		
Form	X	A	B	C
1	1	1	1	0
2	1	1	0	1
3	1	0	1	1

Incomplete Block Designs

When variables are arranged into rotation blocks using efficiency criteria such as balancedness and optimality criteria, we call such designs *balanced*, *partially balanced*, or *optimal incomplete block designs.* Respondents are then assigned with a form consisting of one common block and one rotation block. In the following introduction to these designs, I focus on the rotation blocks only.

Balanced incomplete block designs. A balanced incomplete block design (BIB) divides the variables into multiple rotation blocks. Let t denote the number of variables, k the number of variables in each rotation block (also referred to as block size), b the number of rotation blocks, and r the replication times for each variable. The design is called *balanced* because each pair of variables is replicated the same number of times (λ), which is also referred to as the *associate class* (Montgomery, 2012). The BIB designs satisfy the following two equations:

$$bk = rt \tag{1}$$

$$r(k - 1) = \lambda\,(t - 1) \tag{2}$$

Consider a simple BIB design with four variables ($t = 4$) and a block size $k = 3$. Then, the above two equations hold if we choose $b = 4$, $r = 3$, and $\lambda = 2$, for instance. As Table 3 shows, each rotation block contains three variables. This design is balanced because each variable shows up with the same frequency ($r = 3$), and any pair of variables shows up equally often ($\lambda = 2$).

Each block is assigned to one quarter of respondents. In this design, the missing percentage of each variable is $(b - r) / b = 25\%$ and the overlap percentage is $\lambda / b = 50\%$.

Table 3. The BIB design with four variables and four rotation blocks.

Rotation Blocks	V1	V2	V3	V4
1	1	1	1	0
2	1	1	0	1
3	1	0	1	1
4	0	1	1	1

Partially balanced incomplete block designs. Although a BIB design can be constructed with any number of variables t and any block size k, the minimum number of blocks b is fixed by these two parameters (Cochran & Cox, 1957). In most cases, the number of required blocks b is too large to be implemented in practice. Thus, the balance criteria can be relaxed in order to obtain a smaller number of blocks. Instead of having one associate class λ, multiple associate classes $\lambda_1, \ldots, \lambda_s$ are used. Then the incomplete block design is called *partially balanced*. The fewer associate classes a PBIB design has, the closer it is to a BIB design. For a PBIB design, the following equations need to hold (with an integer a_i):

$$bk = rt \tag{3}$$

$$r(k - 1) = \sum_{i=1}^{s} a_i \lambda_i \tag{4}$$

Consider an example with six variables ($t = 6$). We can for instance construct a partially balanced incomplete block (PBIB) design with $b = 3$ blocks of size $k = 4$ and two association class $\lambda_1 = 1$ and $\lambda_2 = 2$ (Table 4). In this design, the missing percentage of each variable is $(b - r) / b = 33\%$ and the overlap percentages are $\lambda_1 / b = 33\%$ and $\lambda_2 / b = 67\%$. Notice if we group variables V1 and V2 into block A, V3 and V4 into block B, and V3 and V4 into block C, this

PBIB design is the same as the rotation part of the three-form design. Another example of a PBIB design with 19 variables can be found in Kaplan & Su (2016).

Table 4. The PBIB design with six variables and four rotation blocks.

Rotation Blocks	V1	V2	V3	V4	V5	V6
1	1	1	1	1	0	0
2	1	1	0	0	1	1
3	0	0	1	1	1	1

Optimal incomplete block design. With a large number of variables, finding an incomplete block design with maximum balance by hand is no longer an easy task. Instead a computer-generated-design that maximizing a specific optimality criterion can be used. Software packages like jmp from SAS (SAS Institute Inc., 2012) or the R function optBlock() from the AlgDesign package (R Core Team, 2014; Wheeler, 2014) search for an optimal incomplete block design. For a predefined number of variables, blocks and block size, the function optBlock() uses the design matrix **X** and searches for an incomplete block design such that the determinant of **X'X** is maximized. The design matrix **X** contains all the predictors in the rotation blocks including main and interaction effects of interest. The "D" in D-optimality reflects the determinant criterion (Atkinson et al., 2007; Wu & Hamada, 2009). Maximizing the determinant of **X'X** is equivalent to minimizing the volume of the joint confidence region of all effects, that is, all effects captured by the design matrix **X** can be jointly estimated with maximum efficiency. Another advantage of the optimal incomplete block design is the possibility of specifying higher order interactions to be estimable, which is not achievable via multiple-form designs.

The search algorithms are flexible enough to accommodate designs with any number of variables and block size. However, the algorithm does not automatically guarantee the minimum missing percentage and maximum overlap percentage. Thus, researchers should always check

the missing and overlap percentages for each generated design and adjust the number of blocks and block size to find the maximum overlap percentage.

Missing Data Mechanisms and Methods

Rubin (1976) originally defined missingness R as a random variable which has a probability distribution. The missingness mechanisms can be categorized as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The common belief is that with planned missing data designs MCAR or MAR is automatically met and thus the parameters can be recovered using multiple imputation. By reviewing the missingness mechanisms and methods, it is important to address that even if a planned misisng data design follows MCAR or MAR, the parameters are not garantteed to be recovered without bias.

Missing Data Mechanisms

Let D denote the hypothetical complete data matrix with n observations and k variables and D^* the realized data with the missing values. D contains two sets of variables U and V, $D = (U, V)$. The realized variables in U^*contain missing data, and the realized variables in V^* do not contain any missing data. Thus, $D^* = (U^*, V^*) = (U^*, V)$. R denotes the indicator matrix of planned missingness with respect to D^*, taking values of 1 if values in D^* are observed and 0 if values in D^* are missing. The planned missingness indicator R can have three relationships to the hypothetical complete data D: (1) independent of variables in U and V, (2) dependent on variables in V, but independent of variables in U, and (3) dependent on variables in U only, or dependent on variables in both U and V. Imagine D contains all the variables in a large survey. Variables in U belong to the rotation blocks that will be planned with missing data. Variables in V belong to the common block so they will not contain any missing data.

Missing completely at random. The missing data is missing completely at random when planned missingness R is independent of variables in U and V, thus independent of D (see equation (5)). In a planned missing data design, suppose the missing data in U are planned by randomly assigning only half of the items to each respondent. That is, respondents get different randomly sampled sets of items in U and each respondent always has missing values in half of the items. In this random sampling design, due to randomization, the probability of being observed in U is the same, which equals to 1/2 and does not depend on the distribution of U. Thus, the conditional probability of R given the data D is equal to the probability of missingness R (equation (6)).

$$R \perp D \tag{5}$$

$$P(R \mid D) = P(R) \tag{6}$$

For a multiple form design or an incomplete block design, MCAR is met when the blocks are randomized. Let B denotes a blocking variable that indicates which rotation block the variables in U belong to. In the two-form design, half of the variables in U belong to rotation block one (denote these variables as U_1) and the other half to rotation block two (denote these variables as U_2). One form that contains the V and U_1 and the other form that contains V and U_2 are randomly assigned to respondents. Respondents who get form one will have missing data in U_2^*, and respondents who get form two will have missing data in U_1^*. Thus, the missing data in $D^* = (U_1^*, U_2^*, V^*)$ is planned solely by randomizing B to respondents. The probability of being observed in U_1^*, U_2^* does not depend on the distribution of U_1 nor U_2. Each respondent has half chance of being assigned to form one or form two. Thus the conditional probability of being observed is the same as the marginal probability of being observed, $P(R = 1 \mid U_1, U_2, V) = P(R = 1) = ½$.

Missing at random. The missing data is missing at random when missingness R only depends on V. After conditioning on V, R is independent of U (equation (7)). The conditional probability of R given U and V is equal to the conditional probability of R given V alone (equation (8)).

$$R \perp U \mid V \tag{7}$$

$$P(R \mid U, V) = P(R \mid V) \tag{8}$$

A randomized block design follows the MAR mechanism when the blocking variable B is a variable in V. Consider the planned missing data designs blocked by school (the variable is denoted as V_{school}). Within each school, each respondent gets an independent random sample of the items in U. However, the number of items that are assigned to each respondent is different across schools. For instance, students in school A are assigned with 50% items while students in school B get 30% items, due to the fact that students in school A can assess more items without getting fatigue. Thus, the probability of an item being observed depends on schools, but does not vary within school, since within school all items in U have equal chance of being observed. In other words, the probability of being observed is the same conditioning on school V_{school}, $P(R = 1 \mid U, V_{\text{school}}) = P(R = 1 \mid V_{\text{school}})$. In this case, the missing data method or analysis procedure need to take into account the school variable to ensure the unbiased parameter estimates.

Missing not at random. The missing data is missing at random when missingness R depends on U, both U and V, or unobserved latent variable. Accordingly, no conditional independence holds, meaning that the missingness mechanism is nonignorable. Consider two variables income U_{income} and age V_{age}, if the missing data in U^*_{income} is due to U_{income} or the unmeasured variables, then the missing data is MNAR. For instance, the respondents with very high incomes tend not to report their incomes. Or respondents have missing values in U^*_{income} associated with higher anxiety which is an unobserved variable.

MNAR occurs frequently in practice even with planned missing data designs. For instance, due to convenience, instead of random assignment researchers use the administrative procedures that end up with more complex confounding between the variables planned with missingness and the outcome variable. If the confounding variables are not measured, the missing data will be MNAR. In addition, respondents assess a set of items in a form which can introduce context effects. When the context effects create spurious association between the missingness *R* and variables in *U*, the missing mechanism becomes MNAR. Thus, planned missing data designs do not guarantee the missingness mechanism to be MCAR or MAR. Even if they do, the parameter estimates are not guaranteed to be recovered without bias. The designs and methods that deal with missing data need to be considered carefully in order to obtain unbiased estimates.

Missing Data Methods for Planned Missing Data Designs

With the advances in the analysis of missing data, methods like multiple imputation (Rubin, 1987, 1996) allow researchers to analyze data from planned missing data designs without having to discard incomplete cases. Modern methods such as maximum likelihood or full information likelihood (FIML) produce accurate parameter estimates where traditional approaches (e.g., pairwise deletion and listwise deletion) fail when the missing mechanism is MAR. Researchers can choose FIML when they use statistical packages such as *sem* (Fox et al., 2014) or *lavaan* (Rosseel, 2012) in R (R core team, 2012) to deal with missing data. The procedure integrates missing data handling into the estimation process and no missing data are filled in. However, maximum likelihood is not flexible enough for researchers who want to use complete data sets for further data manipulation or analysis. For instance, when large survey data have item-level missing data but the analysis is conducted on the scale level, the process of

handling item-level missing data and the analysis cannot be treated separately using maximum likelihood method.

Research by Graham et al., (1996), Graham et al., (2006), and Rughunathan,(1995) has shown that multiple imputation performs well in imputing planned missing data of the cases studied. Multiple imputation is more flexible in dealing with planned missing data in large surveys. Since the imputation phase is separated from the analysis phase, researchers can use additional auxiliary variables in the imputation phase to impute missing data. Once the complete data sets are obtained, researchers can use a different set of variables for the analysis. Multiple imputation is easy to implement for large survey data with different types of data distributions. The approach that specifies the multivariate model by a series of conditional models, one for each incomplete variable, is called fully conditional specification approach (van Buuren, 2007). This approach is implemented in the *mice* package (van Buuren & Groothuis-Oudshoorn, 2010) in R. Researchers can choose the imputation method based on the types of variables, for example, Bayesian linear regression (*norm*) for normally distributed continuous variables, logistic regression (*logreg*) for categorical variables with two levels, and polytomous logistic regression (*polyreg*) for categorical variables with more than two levels. Other options such as data mining methods (e.g., random forest) or methods for multilevel data (Carpenter & Kenward, 2006) are available as well. Predictive mean matching (*pmm*) (Little, 1988) has shown good performance in imputing large-scale educational assessment data (Kaplan & Su, 2016). In addition to standard multiple imputation, a summary of adaptations of multiple imputation for large survey data can be found in Reiter & Raghunathan (2007). An adaptation that is applied in large-scale educational assessment is the nested multiple imputation (or two-stage multiple imputation, Rubin, 2003). Researchers use nested multiple imputations to combine the

imputation of plausible values and the missing data from the context questionnaire (Aßmann et al., 2015; Weirich et al., 2014).

The common belief is that multiple imputations can recover any parameter estimates from planned missing data designs when the missing mechanism is either MCAR or MAR. However, studies have shown (Kaplan & Su, 2016; authors, 2017) that the bias of the parameter estimates differ across planned missing data designs, especially for the estimates of correlation and regression coefficient. Consider the two-form designs which result in zero overlap between the rotation variables from the first form and the second form. Even though the software program (e.g., *mice* package in R) delivers the imputed data, the estimates of correlations between the rotation variables that have no overlap will be biased. Thus, the choice of planned missing data designs needs to be carefully considered depending on the parameter of interest.

Even if the designs ensure that the parameters are recoverable without bias, in practice there are likely other types of missing data which might induce bias. For instance, in addition to the missing data by design, other item-level nonresponses frequently appear. To use multiple imputation, the missingness mechanism of these item-level nonresponses should be ignorable. Furthermore, unit nonresponse (i.e., no single measure for a sampled respondent has been recorded.) might occur. In this case, methods like weighting adjustment can be applied (Kalton & Kasprzyk, 1986). Though multiple imputation has become an easy-to-use method for imputing planned missing data, many other issues such as choosing appropriate auxiliary variables in the imputation model should be considered carefully as well.

Study 1

Methods

This simulation study investigates how bias in parameter estimates of substantive interest depends on the properties of planned missing data designs. In the simulation, the planned missing data are generated assuming the absence of unit-nonrepsonse and additional item-level nonresponse. The simulation intends to answer four research questions: (1) What is the minimum overlap percentage required between any two rotation variables in order to recover the correlations? (2) How missing percentages affect the bias? (3) Do the results differ for continuous and categorical data? and (4) Do the sample sizes affect the results?

The simulation design consists of three simulation factors, the design settings including overlap percentages and missing percentages (nine variations listed in Table 5), type of independent variables distributions (multivariate normal distribution, skewed continuous distribution and categorical distribution), and sample sizes (100, 1000, and 10000 cases). In total, there are $9\times3\times3 = 81$ fully crossed simulation settings.

Design settings. When planning the missing data, the overlap and missingness of any two rotation variables X_1 and X_2 are considered. For achieving the balancedness, the missing percentages of the two variables are assumed to be the same. The missing percentage varies given an overlap percentage between X_1 and X_2.

Overlap percentages. The overlap percentages between the two rotation variables X_1 and X_2 are varied: 0%, 20%, 25%, 33% and 50% (corresponding to 0, 1/5, 1/4, 1/3 and 1/2 overlap cases of total cases). As an illustration (Figure 2 (a)), if the overlap percentage between X_1 and X_2 is 33%, it means that 1/3 of the sample has observations on both rotation variables. This also implies the two rotation variables lack overlap in 67% of the sample, meaning that 67% sample has observations in only one variable.

Missing percentages. For each overlap percentage, the missing percentage ranges between the minimum and maximum. For example, if the overlap percentage is 33%, the

minimum missing percentage of both X_1 and X_2 is 33% (Figure 2 (a)) and the maximum missing percentage is 67% (Figure 2 (c)). Figure 2 (b) shows that the missing percentage of X_1 and X_2 can take on values between 33% and 67%. The minimum and maximum missing percentages are chosen to see how extreme the factor affects the parameter estimates. The variations of overlap percentage and missing percentage are listed in Table 5.

Figure 2. Given 33% overlap between two rotation variables, the minimum missing percentage in (a) and maximum missing percentage in (c).

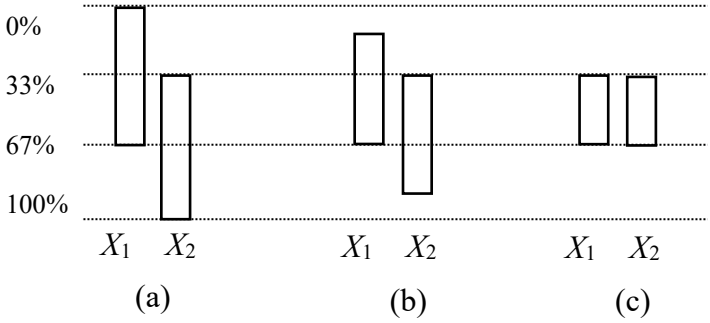


Table 5. The design settings of overlap percentages and missing percentages in study 1.

Overlap Percentage	0	20	25	33	50
Minimum Missing Percentage	50	40	37.5	33	25
Maximum Missing Percentage	-	80	75	67	50

Simulation data distributions. The simulated data consist of three variables, X_1, X_2, and Y. X_1 and X_2 represent any two variables from rotation blocks in a planned missing data design. The distributions of X_1 and X_2 are generated according to a multivariate normal distribution, skewed continuous distribution or a categorical distribution. For the skewed continuous distribution, X_1 and X_2 are log-transformed from the multivariate normal distribution. To

generate categorical variables, a cut-off point is chosen to transform X_1 and X_2 into two binary variables. Y is the dependent variable which is generated by regressing X_1 and X_2 on Y with a normally distributed error term. The true parameter values are set based on the PISA 2006 U.S. data (OECD, 2006). In the context of PISA data, X_1 and X_2 represent two scales and Y the achievement scores. The true values of the means of X_1 and X_2, the pairwise correlations among X_1, X_2, and Y, and the regression coefficients of regressing X_1 and X_2 on Y are listed in Table 6. The correlations among the variables are chosen to have a large range from 0.06 to 0.74.

Table 6. The true parameter values in study 1.

Parameters	Multivariate normal	Skewed	Categorical
Means			
μ_{X1}	0.30	1.53	0.09
μ_{X2}	0.30	-1.65	0.38
μ_Y	490	490	490
Variances			
σ^2_{X1}	0.90	0.05	0.08
σ^2_{X2}	1.00	0.04	0.24
σ^2_Y	714	714	722
Correlations			
ρ_{X1X2}	0.42	-0.42	-0.23
ρ_{X1Y}	0.12	0.12	0.06
ρ_{X2Y}	0.74	-0.73	-0.59
Regression coefficients			
β_1 (slope of X_1)	-6.74	-28.79	-7.91
β_2(slope of X_2)	22.59	-111.72	-33.57

Simulation procedures. The population data were generated according to the true parameter values of the distributions described above. From the population data, a random sample was drawn in each iteration. To plan the missing data in the sampled data, data in X_1 and X_2 were deleted according to the overlap percentage and missing percentage in each design

setting. For example, for an overlap percentage of 33% and the missing percentage for both variables of 33% (Figure 2 (a)), X_1 data were deleted for one third of randomly selected respondents, and X_2 data were deleted for another third of respondents. The remaining third of respondents thus has data in both X_1 and X_2. After creating the planned missing data, predictive mean matching was used to impute the missing data, which resulted in five imputed complete data sets. A regression analysis that regresses X_1 and X_2 on Y was conducted. Results were pooled over the five data sets. The marginal means, pairwise correlations and regression coefficients were extracted from the analysis results. This process is replicated for 5000 times.

Finally, the biases of means, correlations and regression coefficients were computed as the difference between the average estimates across simulations and the true parameter values. For the estimates of means and regression coefficients, 95% simulation confidence intervals were constructed as well. The standard errors used for the 95% simulation confidence interval were calculated as the standard deviation of the coefficients across simulations divided by the square root of the number of iterations.

Results

The biases of means, correlations and regression coefficients are plotted in Figure 3 to 7. In each figure, the plots from the first row to the third row present the results for the multivariate normal data, skewed continuous data and categorical data respectively. In each plot, the biases are presented in the order of the increased overlap percentage between X_1 and X_2 and increased sample size. The X-axis marks the combination of each overlap percentage with the minimum and maximum missing percentage. For the results of means and regression coefficients, the biases are standardized with the standard deviation of outcome Y and 95% simulation confidence intervals are plotted. To better see the effect of missing percentage, solid lines present the results for the maximum missing percentage and the dashed lines for the minimum missing percentage

of X_1 and X_2. The results for the correlation between X_1 and X_2 are in Figure 4, and the results of the correlation between X_1 and Y and X_2 and Y are in Figure 5. Notice the difference in Figure 5 that the X-axis in the plots of correlation bias for X_1 and Y (or X_2 and Y) marks the overlap between X_1 and Y and the missing percentage in X_1, but the order of biases shown is still the same as the order of the increased overlap percentage between X_1 and X_2.

Means. Figure 3 presents the biases of estimated means of X1 (the left column) and X2 (the right column). For the multivariate normal distribution, regardless of the missing percentages, overlap percentages, and sample sizes, the means are recovered without bias. For the skewed and categorical data, the bias in means is found for the maximum missing percentages (e.g., 80%, 75%, and 67%) and the small sample size (n = 100). However once the sample size increases to 1000, the bias in means is negligibly small. Overall, with large survey data (sample size over 1000), the mean estimates are robust against large missing percentage (i.e., 80%) and no overlap. Even though for the small sample size of 100 and large missing percentage the mean estimates are much less reliable, the bias is still within 0.001 standard deviation of the outcome variable Y. Consistent with the findings in Katherine and John (2010), even though predictive mean matching can deal with nonnormal data, for small sample sizes and large missing percentages, bias might appear. This is likely due to the first step of this imputation procedure which generates initial parameters using linear regression.

Correlations. Figure 4 presents the biases of estimated correlations between X_1 and X_2. Figure 5 presents the biases of estimated correlations between X_1 and Y (the left column) and between X_2 and Y (the right column). We first look at the correlations between the two rotation variables X_1 and X_2. For the multivariate normal and skewed continuous data, the trends of the bias in correlations are similar to each other. First, with no overlap, the bias in correlations is not shown in the plots because the bias is outside of the plotting range from -0.2 to 0.2. Second, as

the overlap increases, bias in correlations decreases. For sample sizes of 1000 and 10000, the bias becomes negligibly small (within the absolute value of 0.01) as overlap reaches 20%. Keeping missing percentages at the minimum largely help to reduce the bias for the small sample size of 100. However, for a sample size of 1000, the difference in bias between the minimum and maximum missing percentage becomes very small. Finally, for the categorical data, the bias decreases slower with an increasing overlap percentage as compared to the continuous data case. Bias remains even with 50% overlap. However, as long as there is overlap of 20%, the bias in correlations is still within the absolute value of 0.05. Moreover, increasing overlap to 50% only contributes to a small bias reduction, and keeping the missing percentage at the minimum does not help a lot to reduce the bias in this case. Overall, for large survey data with 20% or more overlap, bias is negligibly small for continuous data. This holds for various missing percentages.

For the correlation between a rotation variable (X_1 or X_2) and the fully observed variable Y, the overlap percentages range from 20% to 75% and the missing percentages range from 25% to 80%. For the multivariate normal and skewed continuous data, the minimum missing percentage in X_1 helps the bias reduction in the correlations between X_1 and Y. This is not surprising because the less missing data X_1 contains, the larger overlap between X_1 and Y since Y does not contain any missing data. When the sample size increases to 10000, the difference between the minimum and maximum missingness becomes negligible small. Even with 80% missingness in X_1 and a sample size of greater than 1000, the bias in the correlation between X_1 and Y does not exceed 0.03 in absolute value. In addition, as the overlap between X_1 and X_2 increases, it also helps to reduce the bias in the correlation between X_1 and Y. As shown for 50% overlap between X_1 and Y (indicated as 50% overlap and 50% missingness), the bias of their correlation is less when X_1 and X_2 have larger overlap. This suggests that overlap between two rotation variables helps to recover not only the correlation between these two variables but also

the correlation between the rotation variable and a fully observed variable such as a variable from the common block. For the categorical data, the trend in bias is similar as for the continuous data, because one variable is continuous. For the correlation between X_1 and Y, the biases from the minimum and maximum missing percentage are closer to each other than the correlation between X_2 and Y. This is so because the true correlation X_1 and Y is close to zero, while the true correlation between X_2 and Y is 0.59.

Regression coefficients. Figure 6 presents the biases of estimated regression coefficients, the slope of X_1 (the left column) and X_2 (the right column). When there is no overlap between X_1 and X_2, most of the confidence intervals are not shown in the plots because they are outside of the plotting range of 0.2 standard deviations of the outcome Y, indicating significant bias. For all types of data, the bias in regression coefficients has the tendency to decrease as the sample size or overlap percentage increases. For the multivariate normal data with sample sizes of 1000 or more, the bias is negligible small if overlap is 20% or higher. The bias difference between the maximum and minimum missing percentage becomes also very small. For skewed and categorical data, the coefficients are estimated with larger bias and less reliability than with multivariate normal data. With large survey data and an overlap of 20% or higher, the bias is within 0.09 standard deviations of the outcome Y. The bias of the skewed data further reduces below 0.02 standard deviation when the overlap reaches 33% or more. For categorical data, similar results as for the correlation between X_1 and X_2 are obtained. The bias in the slope of X_1 reduces slowly as the overlap increases. The bias reduces below 0.05 standard deviations when the overlap exceeds 33%.

Figure 3. The biases of estimated means of X_1 (the left column) and X_2 (the right column) under the multivariate normal (the first row), skewed continuous (the second row) and categorical distribution (the third row) in study 1.

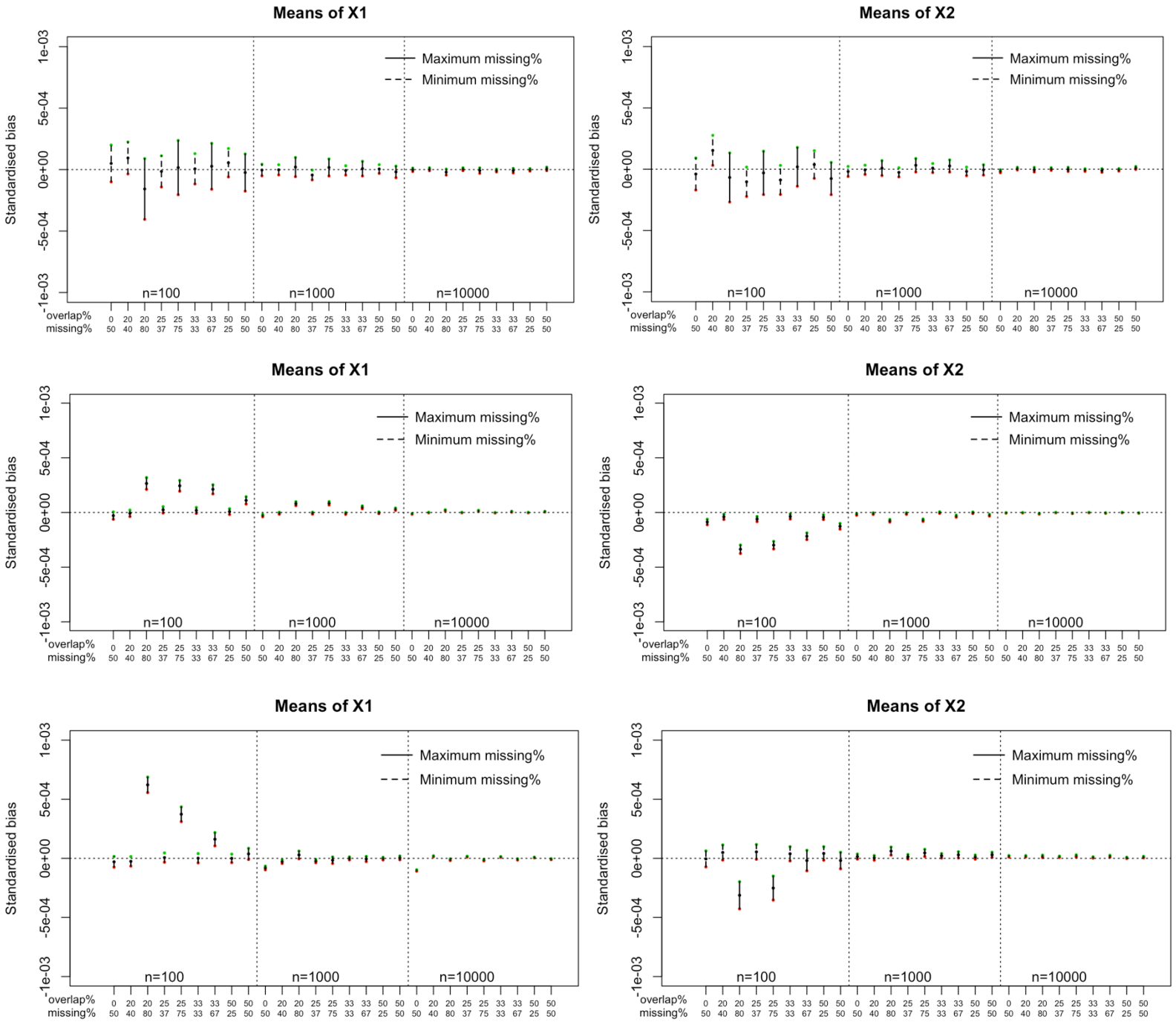


Figure 4. The biases of estimated correlations between X_1 and X_2 under the multivariate normal (the first plot), skewed continuous (the second plot) and categorical distribution (the third plot) in study 1.

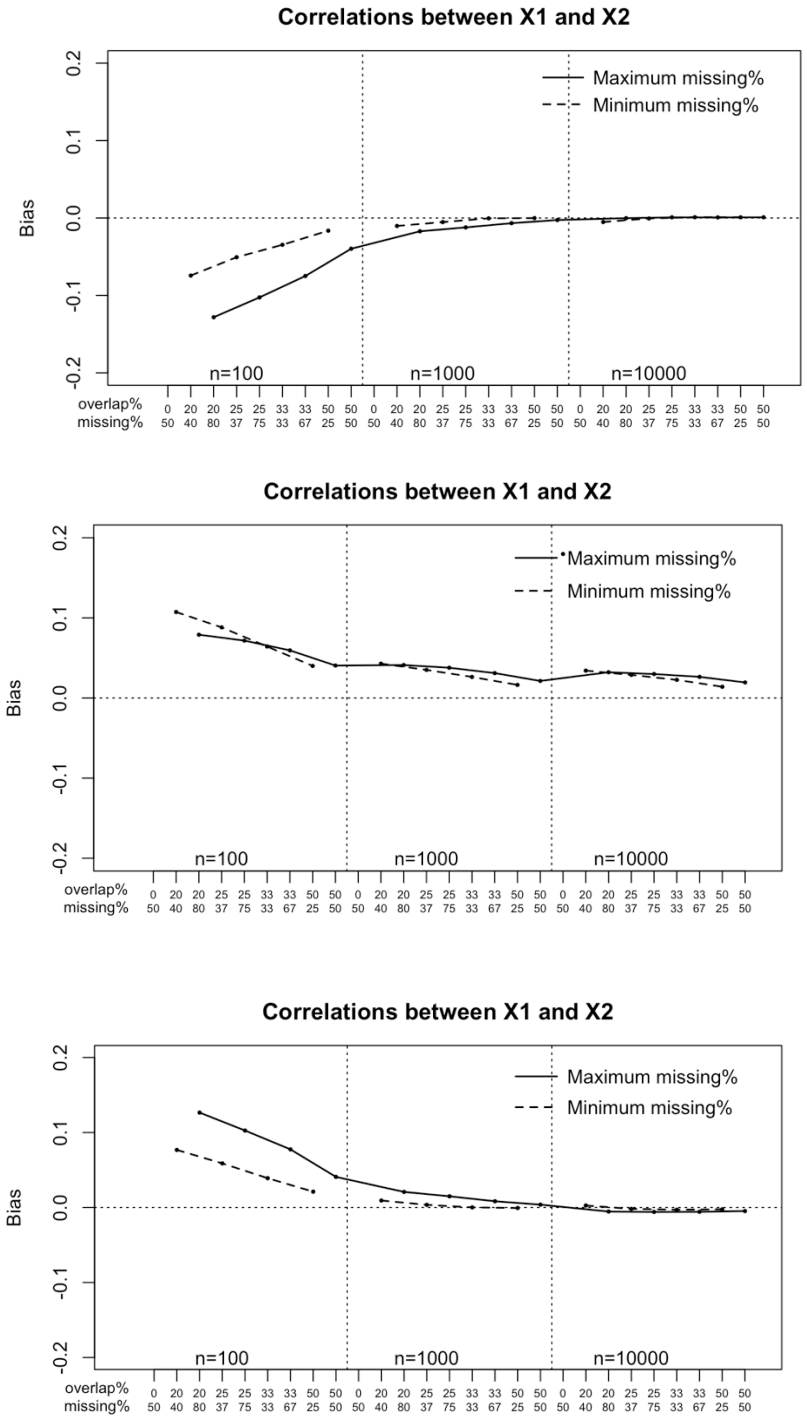


Figure 5. The biases of estimated correlations between X_1 and Y (the left column) and between X_2 and Y (the right column) under the multivariate normal (the first row), skewed continuous (the second row) and categorical distribution (the third row) in study 1.

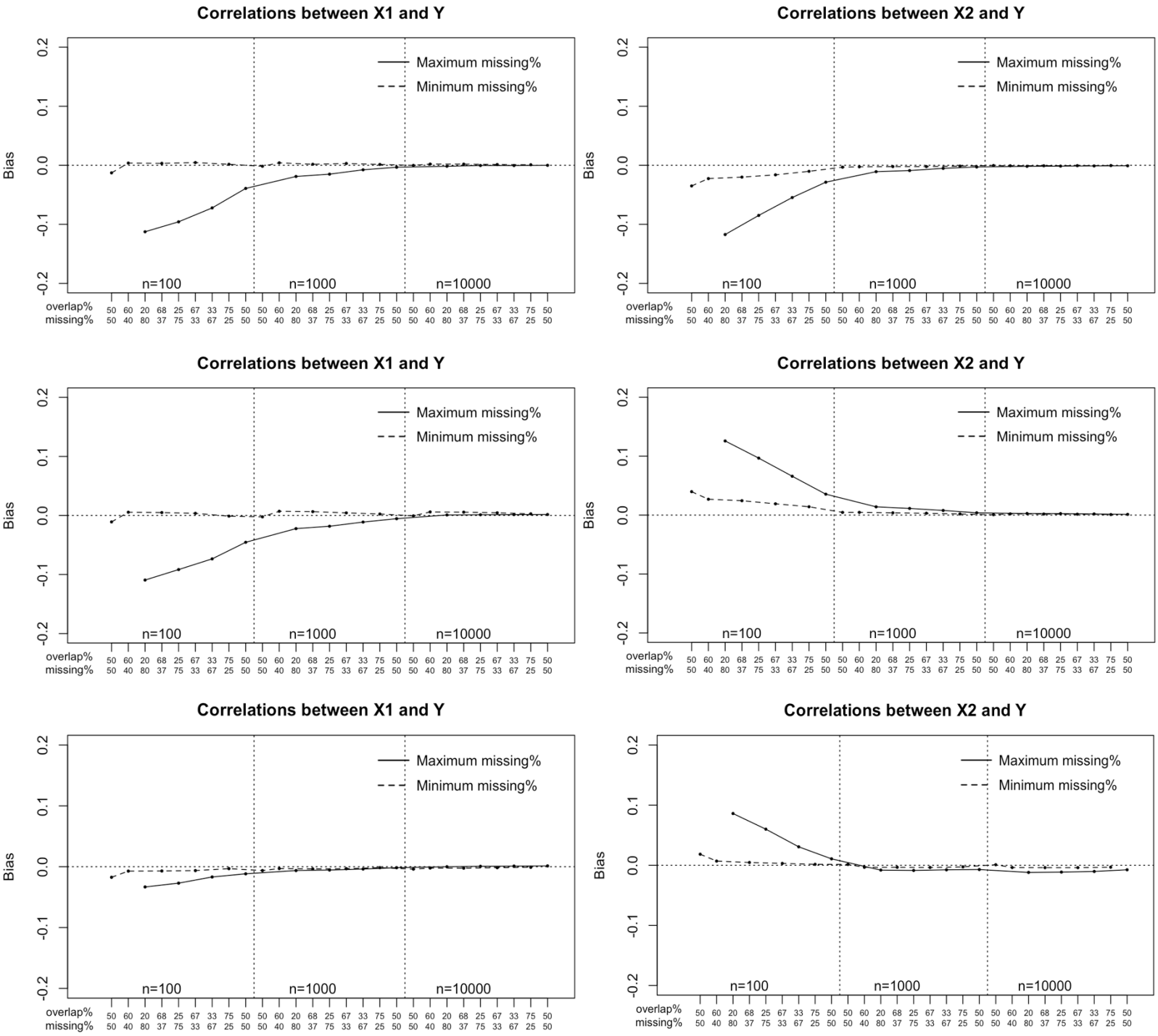
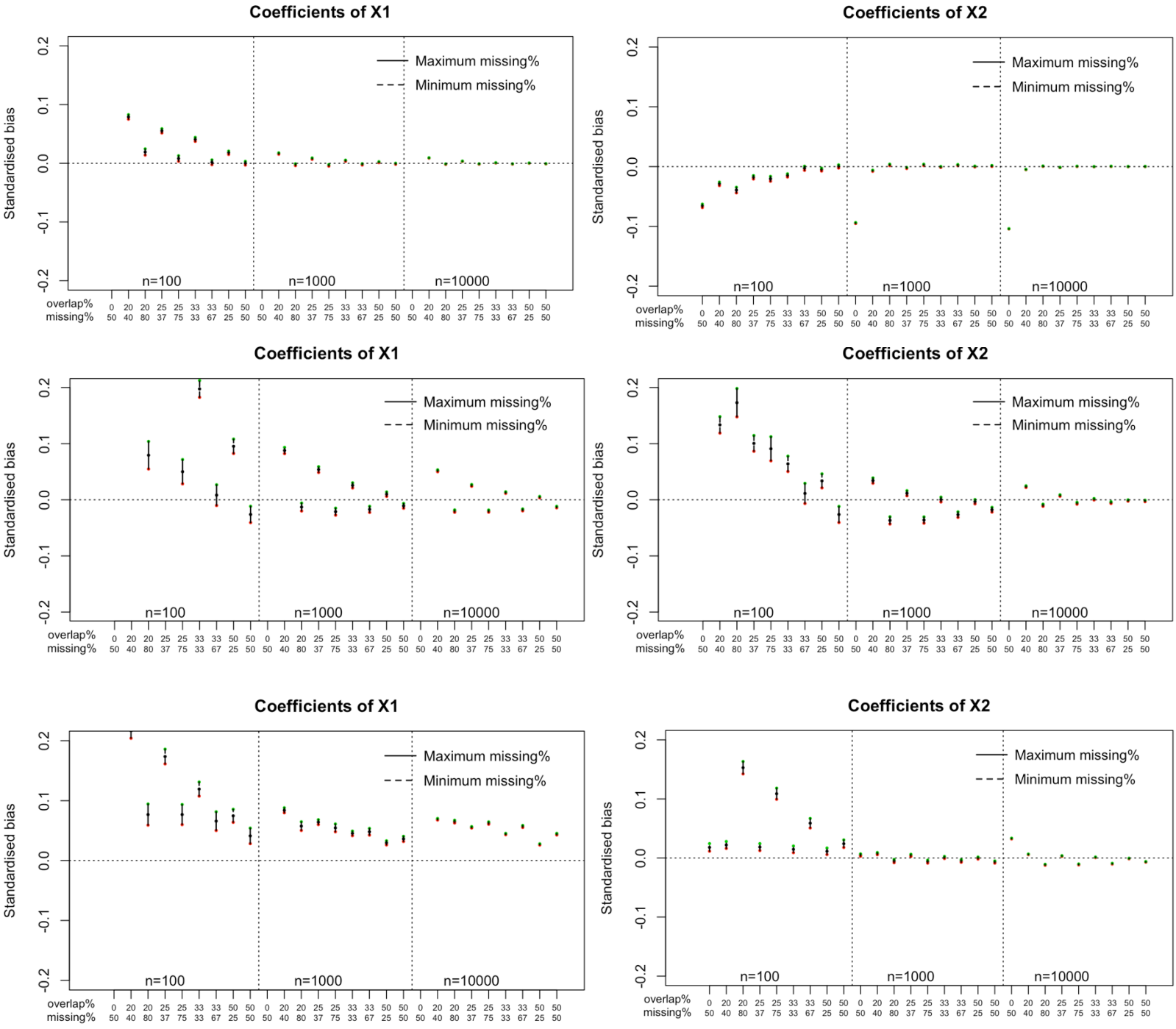


Figure 6. The biases of estimated regression coefficients, the slope of X_1 (the left column) and X_2 (the right column) under the multivariate normal (the first row), skewed continuous (the second row) and categorical distribution (the third row) in study 1.



Study 2

Methods

To illustrate how the overlap and missing percentage apply to the specific designs, the second study uses eight variables to construct a two-form design, three-form design, and two optimal incomplete block designs. The results show each design's performance in recovering unbiased marginal means, correlations and regression coefficients.

Data. Eight independent variables were generated according to a multivariate normal distribution with a sample size of 1000 cases. The independent variables and their parameter values were generated based on eight scales in PISA 2006 U.S. data (OECD, 2006). Table 7 lists the scales with their original names in PISA 2006 data set and their true parameter values (means, variances, correlations and regression coefficients). The pairwise correlations among the scales range between 0.16 and 0.80. The dependent variable (PVSCIE) is the plausible values of science performance and was generated according to the following regression model.

$$\mathrm{PVSCIE}_i = \beta_0 + \beta_1\mathrm{SCHANDS}_i + \beta_2\mathrm{INTSCIE}_i + \beta_3\mathrm{RESPDEV}_i + \beta_4\mathrm{SCIEEFF}_i + \beta_5\mathrm{PERSIE}_i + \beta_6\mathrm{GENSCIE}_i + \beta_7\mathrm{JOYSCIE}_i + \beta_8\mathrm{SCINTACT}_i + \beta_9\mathrm{PERSCIE}_i \times \mathrm{JOYSCIE}_i + \beta_{10}\mathrm{SCHANDS}_i \times \mathrm{SCIEEFF}_i + \beta_{11}\mathrm{GENSCIE}_i \times \mathrm{RESPDEV}_i + \varepsilon_i \quad (1)$$

Table 7. The independent variables in PISA 2006 US. data and true parameter values in study 2.

	Variable	Explanation	Mean	Variance	Regression
Common variables	SCHANDS	Science teaching: hands-on activities	0.68	0.80	-8.68
	INTSCIE	General interest in learning science	0.02	1.16	-12.40
Rotation variables	RESPDEV	Responsibility for sustainable [illegible]	-0.31	0.88	13.16
	SCIEEFF	Science self-efficacy	0.21	1.31	30.79
	PERSIE	Personal value of science	0.29	1.09	-9.02
	GENSCIE	General value of science	0.15	1.19	16.62

	JOYSCIE	Enjoyment of science	-0.04	1.03	21.90
	SCINTACT	Science teaching: interaction	-0.09	1.01	-9.25
Interactions	PERSIE×JOYSCIE				-4.22
	SCHANDS×SCIEEFF				6.75
	GENSCIE× RESPDEV				-5.25

Designs. The four planned missing data designs are a two-form design, a three-form design and two optimal incomplete block designs with missing percentages of 50% and 33% respectively. The design properties (overlap and missing percentage) for each design are summarized in Table 8. For each design, two of the eight independent variables (SCHANDS and INTSCIE) are assigned to the common block while the other six variables are assigned to the rotation blocks. The six rotation variables are planned with missing data.

Table 8. The overlap percentages between two rotation variables and missing percentages of the four designs in study 2.

Design	Overlap percentage	Missing percentage
Two-form design	0%	50%
Three-form design	33%	33%
Optimal block design-50%	20%	50%
Optimal block design-33%	33% or 50%	33%

Two-form design. In the two-form design, the six rotation variables are split into two rotation blocks with three variable each, (RESPDEV, SCIEEFF, and PERSCIE in the first block and GENSCIE, JOYSCIE, and SCIEACT in the second block). Subjects are randomly assigned to one of the two forms, each containing a common block and one of the rotation blocks. To create planned missing data, the data of the second rotation block are deleted for subjects who

get the first rotation block and vice versa. Thus, all the variables in rotation blocks have a missing percentage of 50% and a pairwise overlap percentage of 0% (but 50% with the variable of the common block).

Three-form design. In the three-form design, the six variables are first allocated into three sets with two variables each (RESPDEV and SCIEEFF in set one, PERSCIE and GENSCIE in set two, JOYSCIE and SCIEACT in set three). Then, three rotation blocks are formed according to the three possible combinations of two sets. Subjects are randomly assigned to one of the three forms, each with a common block and one rotation block. To create the planned missing data, data of the unassigned variable sets are deleted for each subject (e.g., for the first rotation block that contains variables of the first two sets, the data of the third set are deleted). Thus, all the variables in rotation blocks have a missing percentage of 33%. The pairwise overlap percentage of the rotation variable is 33% since any two rotation blocks have one set of variables that overlaps.

Optimal incomplete block designs. In the optimal incomplete block design with 50% missingness (each variable in the rotation blocks has 50% missing data), the six variables are assigned to 10 blocks according to the D-optimal criterion (Atkinson et al., 2007). Each block contains three variables as shown in Table 9. The overlap percentage of two variables across any two blocks is 20%. This optimal incomplete block design is a balanced incomplete block design. Subjects are randomly assigned to one of the ten forms, each with the common block and one rotation block. To create the planned missing data, data of the unassigned variables are deleted for each subject.

In the optimal incomplete block design with 33% missingness (each variable in the rotation blocks contains 33% missing data), the six variables are allocated into six blocks according to the D-optimal criterion. Each block contains four variables as shown in Table 10.

The overlap percentage of two variables across rotation blocks is either 33% or 50% (Table 11). This optimal block design is also a partially balanced incomplete block design.

Table 9. The variable assignment to ten blocks in the optimal incomplete block design with 50% missingness.

Block	1	2	3	4	5	6	7	8	9	10
SCHANDS	0	0	0	1	1	1	1	0	0	1
INTSCIE	0	1	0	0	0	1	0	1	1	1
RESPDEV	0	1	1	1	0	0	1	1	0	0
SCIEEFF	1	0	1	0	1	0	0	1	0	1
PERSIE	1	0	1	0	0	1	1	0	1	0
GENSCIE	1	1	0	1	1	0	0	0	1	0

Table 10. The variable assignment to six blocks in the optimal incomplete block design with 33% missingness.

	1	2	3	4	5	6
SCHANDS	1	1	0	1	0	1
INTSCIE	1	0	1	1	1	0
RESPDEV	0	1	1	1	0	1
SCIEEFF	1	1	0	1	1	0
PERSIE	0	1	1	0	1	1
GENSCIE	1	0	1	0	1	1

Table 11. The overlap percentage for the optimal incomplete block design with 33% missingness.

	SCHANDS	INTSCIE	RESPDEV	SCIEEFF	PERSIE	GENSCIE
SCHANDS	67%	33%	50%	50%	33%	33%
INTSCIE	33%	67%	33%	50%	33%	50%
RESPDEV	50%	33%	67%	33%	50%	33%
SCIEEFF	50%	50%	33%	67%	33%	33%
PERSIE	33%	33%	50%	33%	67%	50%
GENSCIE	33%	50%	33%	33%	50%	67%

Procedures. A random sample of 1000 cases was generated in each iteration. The data were planned with missingness according to the four designs. The rotation forms were randomly assigned to subjects. Predictive mean matching was used to impute the planned missing data, resulting in five complete data sets for each design. The imputation of interactions used *just-another-variable* approach (Seaman at al., 2012). Then, a pooled regression analysis using equation (1) was conducted. Finally, means, pairwise correlations, and regression coefficients were extracted from the analysis results. This process was replicated 5000 times.

The biases of means, correlations and regression coefficients were computed as the difference between the average estimates across simulations and the true parameter values. To assess how reliable the estimates are, the regular 95% confidence intervals for the bias of means and coefficients were constructed nonparametrically, using the 2.5% and 97.5% quantile of the estimates across all iterations.

Results

The biases of means, correlations and regression coefficients under each design are plotted in Figures 7 to 9. Figure 7 and 9 show the standardized bias in means and regression coefficients and the 95% confidence intervals. The biases in pairwise correlations among the eight variables of the four designs are shown in Figure 8.

Means. In Figure 7, the means of all the variables are recovered without bias even when there is no overlap for some pairs of variables in the two-form design. Moreover, the estimated means are overall more reliable in the optimal block design with 33% missingness, since the design has a lower missing percentage and stronger overlap (33% and 50%) than the optimal block design with 50% of missingness.

Correlations. In Figure 8, each dot represents the bias in the correlation between two variables. The correlation biases in the two-form design are larger than in other designs. In the

two-form design, five among the 28 correlation biases exceed 0.1 (in absolute values) with the maximum bias being 0.28. For the three-form design, all absolute correlation biases are within 0.04. For the optimal block designs with 50% and 33% of missingness, absolute biases never exceed 0.03 and 0.05, respectively.

Regression coefficients. In Figure 9, the biases in regression coefficients of the eight main effects and three interaction effects are shown. The two-form design again has larger bias in the estimated regression coefficients. With the same missing percentage (50%), the optimal block design recovers the coefficients with much less bias due to the 20% overlap. However, the estimates are less reliable compared to the optimal block design with 33% missingness due to its larger missing percentage. The three-form design and the optimal block design with 33% missingness produce similar results. With more overlap and a lower missing percentage, the interaction terms in these two designs are estimated with less bias compared to the two-form design and the optimal block design with 50% missingness.

Figure 7. The biases of estimated means in the four planned missing data designs.

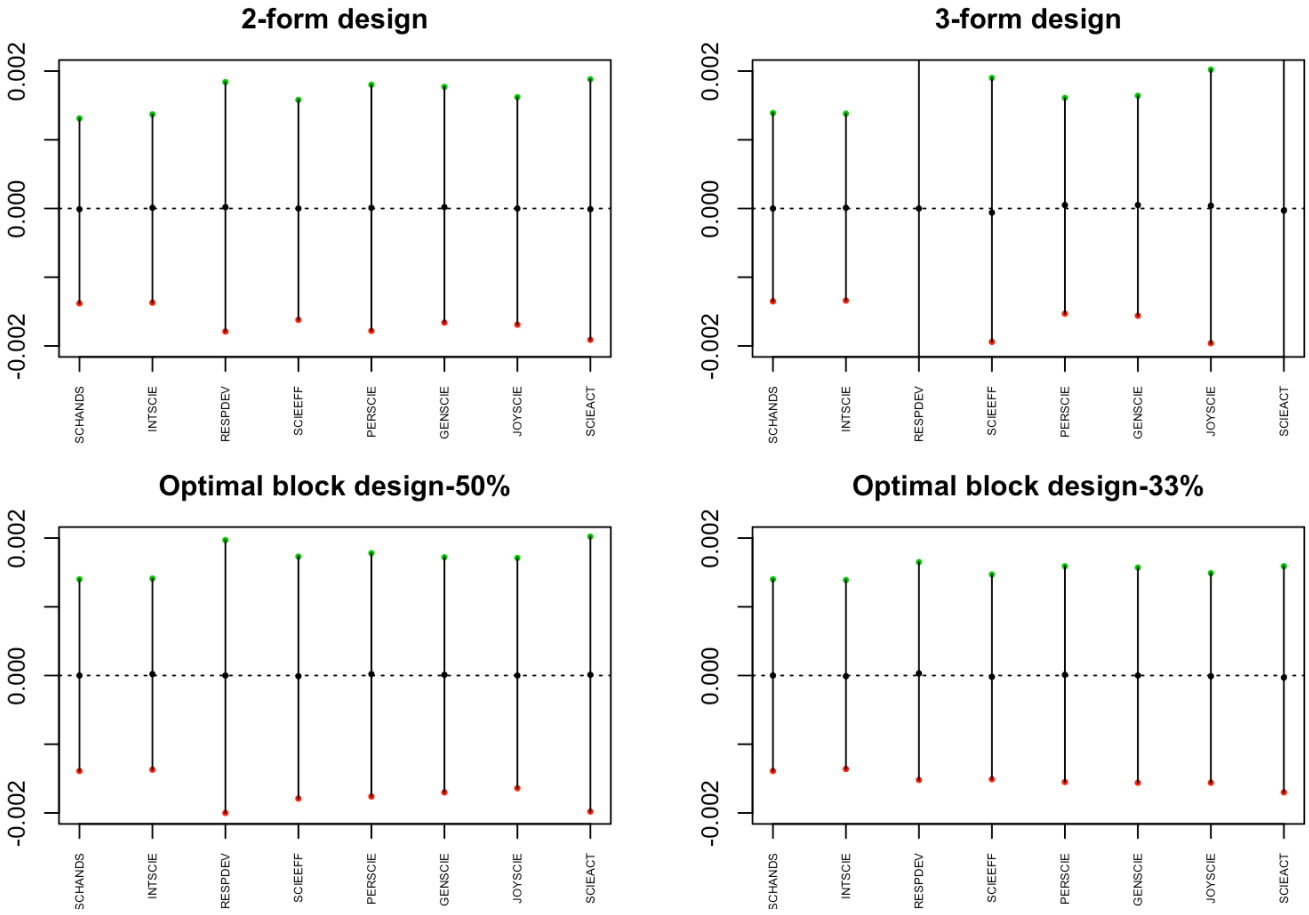


Figure 8. The biases of estimated correlations in the four planned missing data designs.

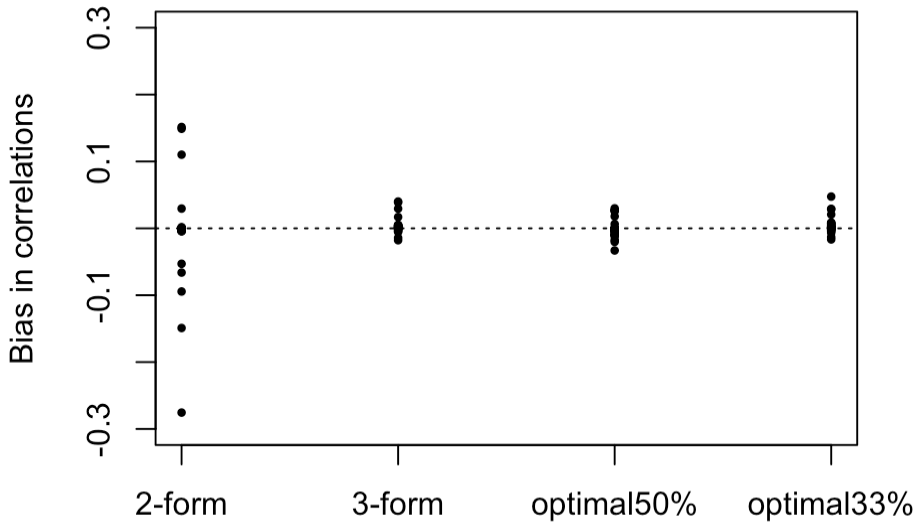
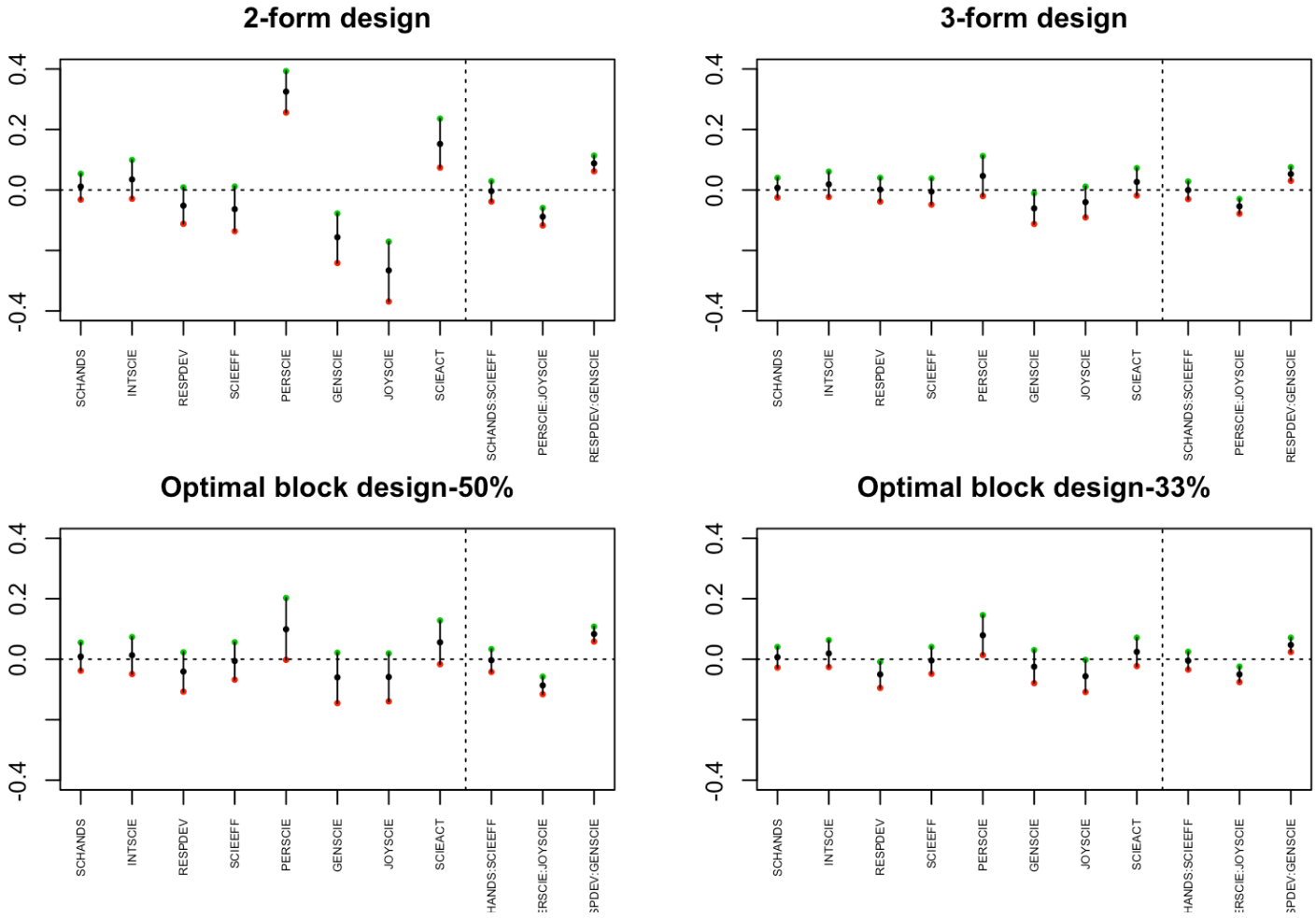


Figure 9. The biases of regression coefficients in the four planned missing data designs.



Conclusion

Using two simulation studies, this article investigates how the properties of missing data designs affect the bias of parameter estimates. The first simulation study investigates the bias in means, correlations and regression coefficients by systematically varying the overlap percentage, missing percentage, joint distribution of the data and sample size. The results show that the estimates of means are unbiased for large-scale survey data (i.e., sample sizes exceeding 1000 cases) even when overlap is zero and the missing percentage is high. However, the recovery of correlations and regression coefficients requires positive overlap. The bias in correlations is negligibly small when there is 20% or more overlap for continuous data. A low missing percentage is of minor importance for bias reduction as long as the sample size is large (at least 1000). Regarding the regression coefficients, the bias is negligibly small when overlap exceeds 20% for multivariate normal data. For skewed and categorical data, the coefficients are estimated with larger bias and less reliability than for multivariate normal data, though with 33% overlap or higher all biases are still within 0.05 standard deviations of the outcome variable.

The second simulation study compares a two-form design, a three-form design, and two optimal block designs with 50% and 33% missingness (Table 7). The results show that all designs recover the means of the eight variables without bias. The biases in correlations are negligibly small for all designs except for the two-form design which has no overlap across forms. For the regression coefficients, the two-form design again performs the worst due to no overlap. With the same amount of missingness, the optimal block design largely reduces the bias due to its 20% of overlap. Furthermore, the three-form design and the optimal block design produce negligibly small bias and more reliable estimates due to more overlap and less missingness.

To conclude, the choice of the design in a large survey strongly depends on the priority of the parameter estimates. If researchers are only interested in estimates of populations means, the two-form design is a good choice since it is simple and easy to implement. If preserving the correlations or regression coefficient is the main goal, enough overlap should be guaranteed. The choice of a specific design for creating sufficient overlap between variables depends on the percentage of items that are administered to respondents compared to the total number of items. It is advisable to first use a pretest to assess how many questions a respondent can answer without getting fatigue. Then, an optimal block design can be found to ensure the required overlap. If the number of items is too large to find an optimal design with sufficient overlap, it is advisable to reduce the number of items instead of increasing the number of questions administered to respondents. If hypothesis tests are the main interests, the amount of missingness should be kept as low as possible in addition to the sufficient overlap. In order to reduce a design's missingness percentage, it is better to again restrict the number of items instead of increasing respondents' burden.

Implementing a planned missing data design in practice requires many other considerations. For instance, should the missing data be planned on the item-level or on the scale-level? On the item-level, items are spread across the blocks or forms without considering the scales. On the scale-level, items are kept together within scales which are then assigned to blocks or forms. Studies showed that spreading the items across forms help to lower the standard error of regression coefficients, given that the procedure of handling the missing data (FIML or multiple imputation) can sufficiently account for the number of variables (Graham at al., 2006; Collins et al., 2001). However, in large surveys, imputation is challenging with hundreds of variables. From the design point of view, if the parameters of interest are on the item-level, the overlaps among items should be guaranteed. If the parameters of interest are on the scale-level,

an optimal block design can be found on the scale-level by treating the items within a scale as one unit.

From the imputation point of view, there are viable ways for solving the problem of simultaneously imputing too many variables. One solution is to reduce the number of variables for sequential imputations. That is, imputing groups of variables sequentially. This can be done in many ways. For example, in large-scale educational assessments, two-stage imputation is implemented by first imputing the achievement scale scores then using the achievement scores to impute other background data (authors, 2017; Weirich et al., 2014). One can also sequentially impute groups of variables that are formed naturally by design (Kaplan & Su, 2016). How to select variables and in which sequence they should be imputed without harming the recovery of parameter estimates are the challenges in imputing planned missing data in large surveys, which need to be addressed in future studies.

Reference

Authors (2017). On imputation for planned missing data in context questionnaires using plausible values: A comparison of three designs. Unpublished manuscript.

Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-Scale Assessments in Education*, 1, 5.

Atkinson, A., Donev, A., & Tobias, R. (2007). Optimum experimental designs, with SAS (Vol. 34). Oxford University Press.

Aßmann, C., Gaasch, C., Pohl, S., & Carstensen, C. H. (2015). Bayesian estimation in IRT models with missing values in background variables. *Psychological Test and Assessment Modeling, 57*(4), 595-618.

Carpenter, J. R., & Kenward, M. G. (2006). Multilevel Multiple Imputation. *Multiple Imputation and its Application*, 203-228.

Cochran, W. G., & Cox, G. M. (1957). *Experimental Designs*. Second Edition. New York: Wiley & Sons.

Collins, L., Schafer, J., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330–351.

Fox J., Nie Z., & Byrnes J. (2014). *sem: Structural Equation Models*. R package version 3.1-5. http://CRAN.R-project.org/package=sem

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, 28, 39–53.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph*, 3, 125 156.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31(2), 197-218.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). *Planned missing data designs in psychological research*. Psychological methods, 11(4), 323.

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of Design Properties on Parameter Estimation in Large-Scale Assessments. *Educational and Psychological Measurement*, 0013164415573311.

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey methodology, 12*(1), 1-16.

Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics*, 41(1), 57-80.

Katherine J. L., & John B. C. (2010). Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation, *American Journal of Epidemiology*, Volume 171, Issue 5, 1 March 2010, Pages 624–632, https://doi.org/10.1093/aje/kwp425

Kirk, R. E. (1995): *Experimental Design: Procedures for the Behavioral Sciences*. Third Edition. Pacific Grove: Brooks/Cole Publishing Company.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*(3), 287-296.

Montgomery, D. C. (2012). *Design and analysis of experiments* (8th ed.). Hoboken, NJ: Wiley.

Neal, T. (2004). Designs Producing Balanced Missing Data: Examples from the National Assessment of Educational Progress. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, 153-162.

Organization for Economic Cooperation and Development. (2014). *PISA 2012 technical report*. Paris, France: Author.

Pokropek, A. (2011). Missing by design: Planned missing-data designs in social science. *ASK. Research & Methods*, (20), 81-105.

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429), 54-63.

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association, 102*(480), 1462-1471.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software,* 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/.

Rhemtulla, M., & Hancock, G. R. (2016). Planned missing data designs in educational psychology research. *Educational Psychologist*, 51(3-4), 305-316.

Rhemtulla, M., Savalei, V., & Little, T. D. (2016). On the asymptotic relative efficiency of planned missingness designs. *Psychometrika, 81*(1), 60-89.

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

Rubin, D. B. (1987). *Multiple imputation in nonresponse surveys*. Hoboken, NJ: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473-489.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica, 57*(1), 3-18.

Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement, 48*(3), 293-312.

Rossi, P. H., & Nock, S. L. (Eds.). (1982). *Measuring Social Judgments: The Factorial Survey Approach*. Beverly Hills: Sage Publications.

SAS Institute Inc. (2012). Using JMP 10. Cary, NC: SAS Institute Inc.

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC medical research methodology*, 12(1), 46.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Oxford: Balinger.

Steiner, P. M., Atzmüller, C., & Su, D. (2016). Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap. *Journal of Methods and Measurement in the Social Sciences*, *7*(2), 52-94.

Su, D. & Steiner, P. M. (2018). An evaluation of experimental designs for constructing vignette sets in factorial surveys. *Sociological Methods & Research.* doi: 10.1177/0049124117746427

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3): 219–242

van Buuren, S., & Groothuis-Oudshoorn, K. (2010). *Multivariate imputation by chained equations, version 2.3*. http://www.multiple-imputation.com/.

Van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, 28(5), 317-331.

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education, 2*(1), 9.

Wheeler, B. (2014). *AlgDesign: Algorithmic Experimental Design*. R package version 1.1-7.2. http://CRAN.R-project.org/package=AlgDesign.

Wu, C. J., & Hamada, M. S. (2009). *Experiments: planning, analysis, and optimization* (Vol. 552). John Wiley & Sons.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Graphical Models for Planned Missing Data Designs

Dan Su

University of Wisconsin-Madison

Abstract

Planned missing data designs in large surveys can efficiently reduce respondents' burden and lower the cost associated with data collection, without cutting down on the questionnaire items. Imputing large amounts of planned missing data without harming the validity of causal parameter estimates is a big challenge. Contrary to the common belief that all auxiliary variables should be used to impute missing data when the missingness is ignorable, we use graphical models to illustrate that in some cases including the auxiliary variables is not necessary and in other cases it causes bias in parameter estimates. We implement simulation studies with different data distributions to show that whether an auxiliary variable should be included in the imputation model not only depends on the causal relationship between the auxiliary variable and the missingness of other variables but also on parameters of interest. Practical implications of imputing planned missing data are discussed.

Keywords

Graphical models, planned missing data designs, missing data, missing at random, multiple imputation, imputation model, auxiliary variable

Introduction

Planned missing data designs in large surveys can efficiently reduce respondents' burden and lower the cost associated with data collection, without cutting down on the questionnaire items. With the advances in the analysis of missing data, methods like multiple imputation (Rubin, 1987, 1996) allow researchers to analyze data from planned missing data designs without having to discard incomplete cases. The imputation methods should not interfere with the researchers' aim to draw valid descriptive and causal conclusions. Research has shown that methods such as predictive mean matching (Little, 1988) perform relative well in imputing planned missing data (Kaplan & Su, 2016, 2018; Su, 2018). One of the challenges in planned missing data designs of large surveys is to impute large quantities of items or variables. Is it necessary to use all or many covariates, so-called auxiliary variables, to impute? What's the impact of using these variables to impute regarding the validity of causal parameter estimates?

Prior research suggested to include as many auxiliary variables as possible (Schafer, 1997; Collin et al., 2001) when imputing missing data with ignorable missingness mechanism, namely the inclusive approach. More recently, studies have shown with several examples that including all variables in the imputation model can bias the causal parameter estimates (Thoemmes & Rose, 2014; Thoemmes & Mohan, 2015). Correctly specified imputation models not only guarantee unbiased parameter estimates but also reduce the number of unnecessary auxiliary variables and thus model complexity. How to identify such imputation models without harming the validity of parameter estimates, especially if the interest is in causal parameters? The theory of graphical models for missing data (m-graphs) has been laid out by Mohan et al. (2013), Mohan & Pearl (2014), and Pearl & Mohan (2013). This paper uses graphical models to discuss the auxiliary variables that are required or unnecessary for imputing planned missing data. This paper shows that contrary to the common belief that all auxiliary variables should be

used to impute missing data when the missingness is ignorable, in some situations including the auxiliary variables will cause bias. To be more specific, we lay out three typical cases to show when the auxiliary variable is not necessary or should not be used to recover unbiased parameter estimates when the missingness is ignorable. The three cases are when including the auxiliary variable is not necessary to obtain unbiased parameter estimates, when it is only necessary for some parameter estimates, and when it biases all parameter estimates. Furthermore, in order to find out how well the theory works in practice, simulations are implemented under the finite sample size and varied data distributions to examine the bias in means, correlations and regression coefficients.

Based on the theory of graphical models, for planned missing data designs we found that the inclusion of a fully observed variable in the imputation model strongly depends on the causal relationship between this variable and the missingness of other variables. The decision also differs for means, correlations and regression coefficients. The illustrated three typical scenarios in a planned missing data design guide practice to select imputation variables given parameters of interest. Based on the simulations, we found additional bias in parameter estimates can also be introduced by the limited number of imputations or an inadequate method for imputing categorical variables. The paper is organized as following. We first briefly discuss the standard missing at random (MAR) definition (Rubin, 1976) in relation to the graphical representation. Then we introduce the theory of identification of casual parameters using graphical models under the context of missing data. Followed by the introduction of missingess mechanisms using graphs, we illustrate with graphs the three typical cases in planned missing data designs. For each case we show and discuss simulation results. Finally, the practical implications of the suggested three cases are discussed with regard to planned missing data designs.

Standard Missing at Random in relation to Graphical Representation

The missingness in a planned missing data design needs to be ignorable when applying multiple imputation. However, the standard definition of MAR appears to be difficult to interpret in practice. There are a few versions of interpretations in the literature (Raghunathan, 2016; Enders, 2010; Schafer & Graham, 2002; Thomas & Mohan 2013). Some authors use Y_{obs} for the observed part of data in Y, and Y_{mis} for the missing part of data in Y. The question is that at which level does the "missing part" refer to. Does it refer to the missing values in Y or any of the variables that contain missing data in a data matrix Y? In the standard definition of MAR, it is hard to distinguish if the missingness refers to the occurrence at the event-level or the random-variable-level. Tian (2015) made a distinction between these two levels using graphs. Tian defined variable-level MAR (notated as MAR^{+}) and event-level MAR (notated as MAR^{*}). The original MAR assumption (Rubin, 1976) guarantees that likelihood-based inference can be performed while ignoring the missing mechanism. But from a modeling point of view, it is more natural to work with variable-level independencies. In addition, in many articles that work with variable-level independencies, another confusion arises when coming to determine which variables are referred to as MAR. Do we refer to the variable with missing data locally or do we refer to all the variables in a whole data set? Tian defined *Local MAR* in which one can identify MAR for each variable separately. Tian also defined *G-MAR* to identify the MAR condition for all variables in a graph.

For a planned missing data design, it is natural to consider the variable-level when assigning variables into blocks (e.g., in incomplete block designs or the three-form design). In the imputation stage, the approach for multiple imputation also automatically works at the variable-level. For example, the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2010) uses the fully conditional specification approach (van Buuren, 2007) which specifies the multivariate model by a series of conditional models for each incomplete variable. When

imputing planned missing data in large surveys, due to large amounts of variables, considering the missing mechanism for each variable (locally) helps to determine the imputation model case by case. Thus we refer to the planned missingness at the variable-level locally in the following discussion.

Graphical Representation of Missing Data

The graphs that we are going to use are also referred to as directed acyclic graphs (DAGs), or in the context of missing data, *m*-graphs (Mohan et al., 2013). The arrows in the graphs do not imply linearity but functional relationships with unknown form. The nodes in the graphs can represent fully observed, partially observed variables, unobserved variables or missingness indicators. Observed variables are often with error terms that represent other unobserved disturbance that have direct effects on this variable. The error terms notated as letter ε are omitted for simplicity. In *m*-graph, the nodes labeled *R* represent the causal mechanism that is responsible for missingness. For example, in planned missing data designs the missing indicator *R* is mainly caused by the blocking variable *B.* The *m*-graphs can be regarded as the data-generating mechanisms for any variables in the planned missing data designs, where the values of each variables are determined by the values of the variables that have direct arrows pointing into this variable.

Identification of Causal Parameters

In order to identify the causal effect, we rely on the *d-separation* criterion (Pearl, 2009) which determines whether two variables in a graph are statistically independent of each other conditional on a set of other variables. We only discuss the conditioning approach for

identification of the causal effect, although other approaches are available, for example, the instrumental variable approach (Angrist et al., 1996; Steiner et al. 2015). In order to illustrate the conditioning approach with graphs, we slightly modify the *m*-graphs that are used in Thoemmes and Mohan (2015). We use solid rectangles around the variable to indicate that a variable has been conditioned on.

For identifying the causal effect of *X* on *Y*, we would first need to find all the paths that connect *X* and *Y* through directed or bi-directed arrows. We can identify the causal effect between *X* and *Y*, when conditioning on a set of variables that block the *back-door* paths. The name *back-door* echoes conditioning, indicating that the paths with arrows pointing at *X* should be blocked; these paths can be viewed as entering *X* through the back door (Pearl, 2009). There are many possibilities how *X* and *Y* are connected. We summarize three typical scenarios. The first scenario is that *X* and *Y* are connected via a common cause variable *Z*, e.g., $X \leftarrow Z \rightarrow Y$. Conditioning on *Z* blocks the path from *X* to *Y.* The second scenario is that *X* and *Y* are connected via a mediator *M*, e.g., $X \rightarrow M \rightarrow Y$. Conditioning on *M* blocks this path from *X* to *Y.* The third scenario is that *X* and *Y* are blocked due to a collider *C*, e.g., $X \rightarrow C \leftarrow Y$. But conditioning on the collider will create a spurious association between *X* and *Y*, thus, opens the path.

Recoverability. Recoverability refers to the identification of causal effects in *m*-graphs. What recoverability means is that if the data are generated by any process compatible with a graph, a procedure exists that computes an estimator for the parameter of interest such that, in the limit of large samples, it converges to a bias-free estimate of the parameter. This procedure is called a "consistent estimator." Recoverability is the sole property of the graph and the causal relationships between the variables, not of the data available, or of any routine chosen to analyze or process the data (Mohan et al., 2013). One should be aware that recoverability is only related to identification not estimation. In other words, even if a causal parameter is recoverable with

regard to a specific graph, it does not automatically imply that the parameter is estimable without bias from finite data. In particular, conditioning on variables that induce bias, like collider variables, may result in biased parameter estimates.

Recovering Means. Thoemmes and Mohan (2015) illustrated the graphical criteria for recovering means and regression coefficients. To summarize, suppose variable Y contains missing data with its missing indicator denoted by R_Y, the mean of Y can be recovered if there exists a set of fully observed variables W (which can be treated as auxiliary variables) such that the following conditional independence holds:

$$Y \perp R_Y \mid W \tag{5}$$

Recovering Regression Coefficients. If we are interested in recovering the regression coefficient of Y on X. Both variables contain missing data and the missing indicators are R_Y and R_X. In order to recover the regression coefficient, Y has to be d-separated from R_Y and R_X, conditional on X and a set of fully observed variables W (equation 6):

$$Y \perp \{R_Y, R_X\} \mid X, W \tag{6}$$

For more complicated cases, namely when W is not fully observed, additional conditional independence between W and the missingness indicators is required, that is $W \perp \{R_W, R_X\} \mid X$. For theoretical proofs, readers can refer to Mohan et al. (2013) regarding *ordered factorization*. This paper only discusses cases with fully observed W.

Recovering Correlations. Thoemmes and Mohan (2015) did not discuss the criteria for identifying the correlation between X and Y. From the overlap assumption we discussed earlier we know that in order to recover the correlation between X and Y, we need to have observations on both X and Y. In addition, the correlation of X and Y is computed as the standardized

regression coefficient using the variance of X and the variance of Y. Thus in order to recover the correlation between X and Y, conditional independency in equation (6) needs to hold as well.

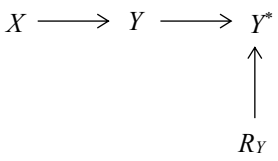
Missing Data Mechanism

Here we illustrate the identification of means and regression coefficients using simple illustrative examples. These examples also present different missingness mechanisms. Suppose we have two variables X and Y, and we are interested in recovering the mean of Y and the regression coefficient of Y on X. X is fully observed. We use a direct arrow pointing from X to Y to present the direct causal effect from X to Y. We use Y^* to represent the realized data of Y. We can think of Y as the variable with complete data in theory and Y^* as the variable with missing values in practice. We can also call Y^* a proxy variable of Y. Thus, the observed data in Y^* are directly obtained from Y, and we draw a directed arrow from Y to Y^*. The missing data in Y are determined by the missingness indicator R_Y which takes on values of 0 and 1. If R_Y is 0, Y^* is missing; if R_Y is 1, Y^* takes the value of Y. Thus Y^* is caused not only by Y but also R_Y. We draw a directed arrow from R_Y to Y^*. Graph 1 is the causal graph that presents the data generating mechanism of X, Y, Y^* and R_Y.

MCAR. The graph in Figure 1 shows the missing data in Y are missing completely at random. In this case, the missing values in Y^* are completely determined by R_Y (which itself is completely determined by a random error term which is not shown in the graph). The missingness R_Y is due to a random procedure which is independent of Y. From a graphical point of view, Y and R_Y are d-separated because the only path connecting Y and R_Y is naturally blocked by collider Y^*. Thus, without conditioning on any other variables the unconditional independency $R_Y \perp Y$ holds. Based on the graphical criterion, the mean of Y is recoverable. The regression coefficient of Y on X can be also recovered, because $R_Y \perp Y$.

The data generating mechanism that is encoded in this graph may represent a random attrition problem in a randomized experiment. X is the treatment or control condition, Y is the outcome and Y^* contain missing data due to random attrition of respondents. This graph implies that there are no confounders between the treatment and outcome, which is guaranteed by randomization. Respondents drop out from the experiment according to an independent random process. That means the attrition rates are the same for treatment and control group, and are not affected by respondents' characteristics or the experiment itself.

Figure 1. The graph of X and Y where Y is missing completely at random.



MAR. The graphs in figure 2 shows that the missing data in Y are missing at random. It has an additional causal path from X to R_Y. In this case, the missingness indicator R_Y is not only caused by some independent random process (which is not explicitly shown in the graph) but also by the fully observed variable X. Y and R_Y are no longer d-separated unless one conditions on X, since Y is connected with R_Y via X. We draw a solid rectangular box around X to indicate that the back-door path $Y \leftarrow X \rightarrow R_Y$ is blocked after conditioning on X. Based on the graphical criterion, the mean of Y and regression coefficient of Y on X are recoverable conditional on X, because $R_Y \perp Y \mid X$.

This graph can represent the data generating mechanism of a randomized experiment with attrition that is affected by the treatment. For example, if the treatment is a medication for curing a disease, the patients who are assigned to the treatment show strong side effect and they drop out from the study. The attrition rate in the treatment group is thus much higher than in the

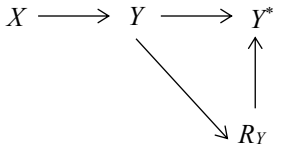
control group. But within the treatment group or control group, the patients drop out randomly. The graph also implies that the reason for patients' attrition is due to some independent random process and the treatment, not other variables such as the characteristics of patients.

Figure 2. The graph of X and Y where Y is missing at random.

MNAR. The graph in Figure 3 shows that the missing data in Y are missing not at random. The missingness indicator R_Y is directly caused by Y. Y and R_Y are no longer d-separated. Conditioning cannot help us to block the path between Y and R_Y because there is no observed variable between the two variables on which we could condition. Thus, based on the graphical criterion, the mean of Y and regression coefficient of Y on X are not recoverable.

An example for this graph would be a situation where respondents refuse to provide their outcomes because the outcome itself. For example, let X be a randomly assigned math training program and Y a math achievement score collected in a survey. If at the end, students who obtained lower math scores more likely refuse to reveal their scores, the missing data are directly caused by the outcome. This graph also implies that the missingness does not depend on the treatment status, that is, whether students choose to reveal their scores does not depend on the program assigned.

Figure 3. The graph of X and Y where Y is missing not at random.



Graphical Representation of Planned Missing Data

In planned missing data designs, an *item* (also referred to as *variable*) is an individual task that is administered to a respondent. A *block* is a set of items that are blocked by design. A block of variables such as demographic information that contains no planned missing data is called a *common block*. Blocks with planned missing data are called *rotation blocks*. The variables that are assigned to rotation blocks are referred to as *rotation variables*. A *form* is the actual set of blocks that is administered to examinees. A form can contain either multiple blocks or only one block. Typically, a form contains a common block and at least one rotation block. The *missing percentage* of a single variable is the percentage of missing cases in this variable. The *overlap percentage* of two variables is the percentage of simultaneously observed values in the two variables (relative to total number of cases). If the overlap percentage of two variables is 0%, correlations cannot be recovered. We now focus on the causal relationships among the two rotation variables X_1 and X_2, and one variable X_3 from the common block. The two rotation variables X_1 and X_2 can come from any two different rotation blocks. This implies that cases with missing values in X_1 are different from cases with missing values in X_2. But X_1 and X_2 have common observed cases (positive overlap). X_3 contains no missing data.

Cases 1-3 (Figures 4-7) represent the causal relationships among X_1, X_2, X_3, and Y in a planned missing data design. Y is the outcome variable that does not contain any missing data. X_1

and X_2 are planned with missing data through the block indictor B which generates the missingness indicators R_{x1} and R_{x2}. $X_1^*$and X_2^* are the realized data of X_1 and X_2. They take on the values of X_1 and X_2 when R_{x1} and R_{x2} indicate that the data are observed, and have missing data (i.e., NAs) when R_{x1} and R_{x2} indicate the data are missing. In all graphs, both X_1 and X_2 cause Y. We use PISA data as the example. Y is the math proficiency score, X_3 is the motivation of learning math that is measured for all students, thus it is in the common block. X_1 and X_2 are variables from two different questionnaire forms, for example, X_1 is the math self-efficacy measure from rotation block one and X_2 is the number of hours of studying from rotation form two. We are interested in recovering the means of X_1 and X_2, the partial regression coefficients of Y on X_1 and X_2, and the correlations between each pair of variables. We will show that the conditions for recovering the parameters differ as the causal relationships among X_3, X_1, and X_2 changes.

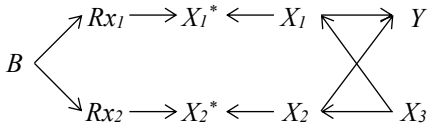
Case 1

Figure 4 shows the causal relationships among X_1, X_2, X_3, and Y in a planned missing data design where X_3 is a common cause of X_1 and X_2. With the PISA data example, motivation of learning (X_3) not only causes math self-efficacy (X_1) but also study hours (X_2).

The missingness mechanism of X_1 and X_2 is missing completely at random (Figure 4), because X_1 is d-separated from R_{x1}, and X_2 is d-separated from R_{x2} as well without conditioning on other variables, meaning $R_{x1} \perp X_1$ and $R_{x2} \perp X_2$. Given this unconditional independence, the means of X_1 and X_2 are recoverable. Moreover, the partial regression coefficients of X_1 and X_2 are also recoverable because $R_{x1} \perp Y \mid X_1$, $R_{x2} \perp Y \mid X_2$. Then the pair-wise correlations between

variables are recoverable as well, as long as we have overlap between each pair of variables. Since Y and X_3 are fully observed we have overlap between Y and X_1, Y and X_2, X_1 and X_3, and X_2 and X_3. In order to recover the correlation between X_1 and X_2, we should make sure that there is sufficient overlap between X_1 and X_2 when planning the missing data.

Figure 4. The graphs of case 1: Planned missing data design for X_1, X_2, X_3 and Y when X_3 is a common cause between X_1 and X_2.

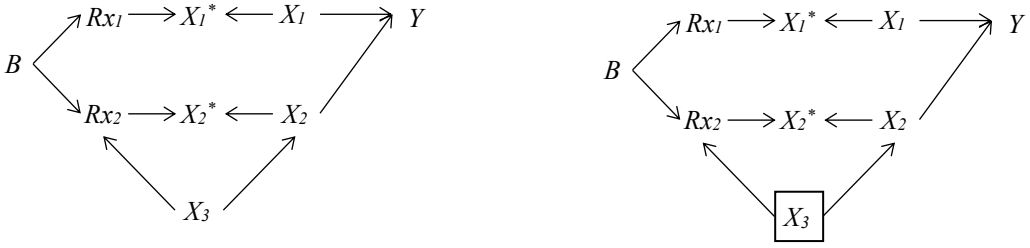


Case 2

Figure 5 shows the causal relationships among X_1, X_2, X_3, and Y in a planned missing data design where X_3 is a common cause of X_2 and its missingness R_{x2}. With the PISA data example, motivation of learning math (X_3) directly causes study hours (X_2). In addition to the planned missing data in study hours (X_2), motivation of learning math (X_3) causes other missing data in study hours (X_2). This can be the case that when students have low motivation of learning math, they tend to not report their study hours.

The missingness mechanism of X_2 is missing at random. In order to d-separate X_2 and R_{x2}, we need to condition on X_3 so that the back-door path $R_{x2} \leftarrow X_3 \rightarrow X_2$ is blocked. We draw a solid rectangle to indicate the conditioning approach. The mean of X_2 is recoverable once we condition on X_3, because $R_{x2} \perp X_2 \mid X_3$. The partial regression coefficient of X_2 is recoverable without conditioning on X_3, because $R_{x2} \perp Y \mid X_2$. Then the pair-wise correlations between variables are recoverable as well, as long as we have overlap between each pair of variables.

Figure 5. The graphs of case 2: Planned missing data design for X_1, X_2, X_3 and Y when X_3 is a common cause between R_{x2} and X_2.



Case 3

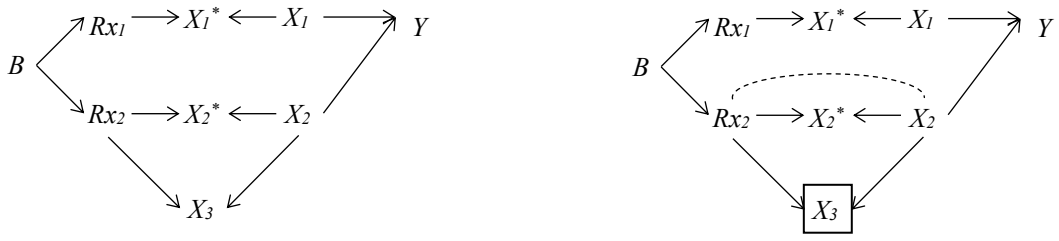
Figure 6 shows the causal relationships among X_1, X_2, X_3, and Y in a planned missing data design where X_3 is a common descendent of X_2 and its missingness R_{x2}. X_2 causes X_3, and the missingness R_{x2} also causes X_3. Thus X_3 is a collider between X_2 and R_{x2}. The direct causal path between R_{x2} and X_3 implies that how the missing data planned in X_2 or how the missing data are planned in the rotation block where X_2 locates directly cause the response values in X_3. For example, the rotation block where the variable study hours (X_2) locates might create a context that causes some students to mis-report their motivations of learning math (X_3). The missingness in X_2 is planned as the same way as the other items in this rotation block. When students answer the question on study hours, they also need to answer the other questions in this block. If all the items in this rotation block mainly measure how much time or energy students devote to learn math, the context that is created by this specific group of items can affect the answers of following questions. In other words, being exposed to the question on study hours might results in a different response when answering the question on motivation of learning math.

The missingness mechanism of X_1 and X_2 is missing completely at random. Because X_3 is a collider that naturally blocks the path $R_{x2} \rightarrow X_3 \leftarrow X_2$. Thus the unconditional independence

assumption, $R_{x2} \perp X_2$ holds. Based on the inclusive approach for imputing missing data (Collins et al., 2001; Schafer, 1997), X_3 is suggested to be included in the imputation model to help impute the missing data in X_2. However, we argue that including X_3 for imputing the missing data in X_2 will induce bias not only in mean of X_2, but also the partial regression coefficients of X_2.

We draw a solid rectangular to indicate the conditioning approach (Figure 6). Conditioning on collider X_3, it opens the collider path $R_{x2} \rightarrow X_3 \leftarrow X_2$. Thus X_2 and R_{x2} are no longer *d*-separated. This is because conditioning on the collider introduces the spurious association between X_2 and R_{x2} (as indicated with the dashed line between R_{x2} and X_2). This spurious association is responsible for collider bias in parameter estimates. Without conditioning on X_3, the mean of X_2 is recoverable, since the unconditional independence holds, $R_{x2} \perp X_2$. The partial regression coefficient of X_2 is also recoverable, since conditional independence holds $R_{x2} \perp Y \mid X_2$. However, once we condition on X_3, the mean of X_2 will be biased, because a spurious association is introduced which *d*-connects R_{x2} and X_2. The partial regression coefficient of X_2 will also be biased, because R_{x2} and Y are *d*-connected via the spurious association.

Figure 6. The graphs of case 3: Planned missing data design for X_1, X_2, X_3 and Y when X_3 is a collider between X_1 and X_2.



Simulation Studies

We implement simulation studies to investigate the three cases illustrated above. They are when including the auxiliary variable is not necessary to obtain unbiased parameter estimates (case 1), when it is only necessary for some parameter estimates (case 2), and when it biases all parameter estimates (case 3).

Data

We create X_1, X_2, X_3, and Y according to the graphs (Figure 4-7). The structural equations of all variables are in simple linear parametric forms. The population means, variances, and weights that are used to create the linear functional forms of X_1, X_2, X_3, and Y are chosen with reference to the PISA 2006 U.S. data (OECD, 2006). The true values of means and variances of X_1, X_2 and X_3, pairwise correlations among X_1, X_2, X_3 and Y, and regression coefficients of Y on X_1 and X_2 are listed in Table 1-3 for each case.

The distributions of X_1, X_2 and X_3 are varied as multivariate normal distributions, skewed continuous distributions, and categorical distributions. To generate skewed continuous distributions, we log-transform the normally distributed variables X_1, X_2 and X_3. To generate the categorical variables, we choose cut-off points to transform X_1, X_2 and X_3 into binary variables.

In order to plan missing data in X_1 and X_2, the block indicator B is created as a three-level categorical variable. B takes on the values of 1, 2 and 3. When B equals 1, the value of X_1 is missing (accordingly R_{x1} equals 0 and R_{x2} equals 1). When B equals 2, the value of X_2 is missing (accordingly R_{x2} equals 0 and R_{x1} equals 1). When B equals 3, both X_1 and X_2 are observed (R_{x1} and R_{x2} are 0). 20% of cases are randomly assigned to block 1, and another 20% are randomly assigned to block 2, and the rest to block 3. Thus, both $X_1^*$and X_2^* contain 20% of missing data. The overlap percentage between $X_1^*$and X_2^* (the ratio between the number of jointly observed cases and the number of total cases) is 60%.

The data in Case 2 and 3 (Figure 5 and 6) differ in the following aspects. For Case 2 (Figure 5), in addition to the 20% planned missing data in X_2^*, more missingness in X_2^* is caused by X_3. We create the additional missingness by choosing a cut-off value of X_3. If X_3 is greater than this value, X_2^* is missing. To be more specific, for the normally distributed data, if X_3 is greater than 2.5, X_2^* is missing. For the skewed continuous data, if X_3 is greater than 1.8, X_2^* is missing. This procedure produces 3% to 7% more missing data in X_2^*. The overlap percentages between $X_1^*$and X_2^* are from 53% to 57%. For Case 3 (Figure 6), the values of X_3 are altered according to the missingness of X_2. To be more specific, when the data are normal or skewed continuous distributions, if X_2^* is missing, X_3 is divided by 10. When the data are categorical distributions, if X_2^* is missing, X_3 equals 1.

Table 1. Population parameter values of variables in case 1.

Parameters	Multivariate normal	Skewed	Categorical
Means			
μ_{X1}	0.939	1.687	0.383
μ_{X2}	0.876	1.447	0.381
μ_{X3}	0.798	1.744	0.482
μ_Y	479.934	484.697	487.698
Variances			
σ_{X1}^2	1.568	2.287	0.236
σ_{X2}^2	1.324	2.239	0.236
σ_{X3}^2	0.893	0.028	0.250
σ_Y^2	772.510	905.920	777.222
Correlations			
ρ_{X1X2}	0.292	0.014	-0.064
ρ_{X1X3}	0.604	0.086	0.306
ρ_{X2X3}	0.488	0.069	0.224
ρ_{X1Y}	0.387	0.495	-0173
ρ_{X2Y}	-0.391	-0.247	-0.597
ρ_{X3Y}	-0.082	-0.024	-0.179
Regression coefficients			
β_0 (intercept)	475.090	475.346	503.551
β_1 (slope of X_1)	9.964	9.927	-7.778
β_2(slope of X_2)	-5.149	-5.115	-33.761

Table 2. Population parameter values of variables in case 2.

Parameters	Multivariate normal	Skewed	Categorical
Means			
μ_{X1}	0.503	1.690	0.367
μ_{X2}	0.879	1.860	0.390
μ_{X3}	0.799	1.744	0.290
μ_Y	475.611	482.622	476.439
Variances			
σ^2_{X1}	0.808	0.029	0.232
σ^2_{X2}	1.329	0.006	0.238
σ^2_{X3}	0.901	0.029	0.206
σ^2_Y	214.033	103.216	543.484
Correlations			
ρ_{X1X2}	0.001	0.002	-0.002
ρ_{X1X3}	-0.003	0.001	-0.005
ρ_{X2X3}	0.491	0.208	0.142
ρ_{X1Y}	0.612	0.172	0.416
ρ_{X2Y}	-0.391	-0.040	-0.308
ρ_{X3Y}	-0.197	-0.011	-0.042
Regression coefficients			
β_0 (intercept)	479.967	474.883	474.784
β_1 (slope of X_1)	9.967	10.299	20.123
β_2(slope of X_2)	-4.974	-5.197	-14.700

Table 3. Population parameter values of variables in case 3.

Parameters	Multivariate normal	Skewed	Categorical
Means			
μ_{X1}	0.300	1.647	0.252
μ_{X2}	0.600	1.708	0.335
μ_{X3}	-0.941	1.621	0.511
μ_Y	475.060	482.928	475.824
Variances			
σ^2_{X1}	1.098	0.044	0.188
σ^2_{X2}	0.894	0.031	0.223
σ^2_{X3}	0.019	0.042	0.250
σ^2_Y	1016.982	401.432	41.657
Correlations			

ρ_{X1X2}	0.303	0.597	0.392
ρ_{X1X3}	0.212	0.007	0.022
ρ_{X2X3}	0.686	0.024	0.042
ρ_{X1Y}	0.289	0.077	0.533
ρ_{X2Y}	-0.050	0.021	-0.104
ρ_{X3Y}	-0.035	0.005	-0.011
Regression coefficients			
β_0 (intercept)	475.056	474.696	474.975
β_1 (slope of X_1)	10.177	9.553	10.072
β_2(slope of X_2)	-5.094	-4.393	-5.059

Procedures

We generated the population data of 50000 cases. Then we drew a random sample of 1000 cases in each iteration. To plan the missing data in the sampled data, missing values in $X_1^*$and X_2^* were set according to the block indicator B. After creating the planned missing data, we used predictive mean matching to impute the missing data five times, resulting in five complete data sets. Two imputation models were chosen, one which includes X_3 as an auxiliary variable to impute $X_1^*$and X_2^* and the other one without X_3. We conducted regression analysis by regressing Y on X_1 and X_2. Results were pooled over the five data sets. The marginal means, pairwise correlations and regression coefficients were extracted from the analysis results. This process was replicated for 5000 times. Finally, the biases of means, correlations and regression coefficients were computed as the difference between the average estimates across simulations and the population parameter values. For the estimates of means and regression coefficients, 95% simulation confidence intervals were constructed as well. The standard errors used for the 95% simulation confidence interval were calculated as the standard deviation of the coefficients across simulations divided by the square root of the number of iterations.

Results

Case 1

Means. Table 4 presents the biases and 95% simulation confidence intervals of estimated means of X_1 and X_2, with and without including X_3 in the imputation model. For all types of data distributions, the means of X_1 and X_2 are estimated without bias regardless of including X_3 in the imputation model or not. Given the unconditional independence, results show that, the means of X_1 and X_2 are recoverable and estimated without bias. This implies that if researchers are

interested in the unbiased means of X_1 and X_2, when X_3 is a confounder between X_1 and X_2, X_3 is not required in the imputation.

Table 4. The biases and 95% simulation confidence intervals of estimated means of X_1 and X_2 in case 1.

Model	Parameter	Normal			Skew			Categorical		
	Mean	Bias	CIL	CIH	Bias	CIL	CIH	Bias	CIL	CIH
With	X1	0.000	-0.001	0.001	-0.001	-0.002	0.001	0.000	-0.001	0.000
X3	X2	0.000	-0.001	0.001	0.001	-0.001	0.002	0.000	-0.001	0.000
Without	X1	0.000	-0.001	0.001	0.000	-0.002	0.001	0.000	0.000	0.001
X3	X2	0.000	-0.002	0.001	0.000	-0.001	0.002	0.000	0.000	0.001

Regression coefficients. Table 5 presents the biases and 95% simulation confidence intervals of estimated regression coefficients of X_1 and X_2. For all types of data distributions, all coefficients are estimated without bias regardless of including X_3 in the imputation model or not. Given the unconditional independence, the coefficients of X_1 and X_2 are recoverable and estimated without bias. Again, X_3 is not necessary to be included in the imputation model for obtaining unbiased regression coefficients.

Table 5. The biases and 95% simulation confidence intervals of estimated partial regression coefficients of X_1 and X_2 in case 1.

Model	Parameter	Normal			Skew			Categorical		
	Coef	Bias	CIL	CIH	Bias	CIL	CIH	Bias	CIL	CIH
With	X1	-0.012	-0.032	0.009	-0.010	-0.026	0.006	0.006	-0.031	0.063
X3	X2	0.014	-0.010	0.038	0.016	-0.001	0.034	0.031	-0.012	0.073
Witho	X1	-0.003	-0.024	0.018	-0.007	-0.023	0.009	0.043	-0.005	0.091
ut X3	X2	0.015	-0.009	0.038	0.004	-0.014	0.021	0.005	-0.038	0.048

Correlations. Table 6 shows the biases of pairwise correlations among X_1, X_2, X_3 and Y. Without including X_3 in the imputation model, the correlations between X_1 and X_3, and X_2 and X_3

are biased. This is obvious because without X_3 there is no overlap between these two pairs of variables. In this case if researchers are interested in recovering correlations among all pairs of variables, X_3 has to be included in the imputation model. If the priority is not the correlations that involve X_3, it can be excluded to simplify the imputation model. It will not affect the means and coefficients of X_1 and X_2 regarding bias.

Table 6. The biases of pairwise correlations among X_1, X_2, X_3 and Y in case 1.

Model	Parameter Cor	Normal Bias	Skew Bias	Categorical Bias
With X3	X1X2	-0.001	0.002	0.000
	X1X3	-0.001	-0.001	0.000
	X2X3	-0.001	-0.001	0.001
	X1Y	-0.001	-0.001	0.001
	X2Y	0.000	0.002	0.001
Without X3	X1X2	0.000	-0.001	0.000
	X1X3	-0.075	-0.013	-0.057
	X2X3	-0.060	-0.013	-0.024
	X1Y	-0.001	-0.001	0.001
	X2Y	0.001	0.000	0.001

Case 2

Means. Table 7 presents for case 2 the biases and 95% simulation confidence intervals of estimated means of X_1 and X_2, with and without including X_3 in the imputation model. For all types of data distributions, the means of X_2 are estimated with bias when X_3 is not used in the imputation. The size of bias for normally distributed data is larger than the other two distributions, mainly because the size of correlation between X_2 and X_3 is much larger. Conditioning on X_3, the means of X_2 are recoverable and estimated without bias. When X_3 is the common cause of the missingness in X_2 and X_2 itself, X_3 should be included in the imputation model in order to avoid the bias in mean of X_2.

Table 7. The biases and 95% simulation confidence intervals of estimated means of X_1 and X_2 in case 2.

Model	Par	Normal			Skew			Categorical		
	Mean	Bias	CIL	CIH	Bias	CIL	CIH	Bias	CIL	CIH
With	X1	0.000	-0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
X3	X2	0.001	-0.001	0.002	0.000	0.000	0.000	0.000	0.000	0.001
Witho	X1	0.002	0.001	0.003	0.000	0.000	0.000	-0.001	-0.001	0.000
ut X3	X2	-0.037	-0.038	-0.036	-0.001	-0.001	-0.001	-0.007	-0.008	-0.007

Regression coefficients. Table 8 presents the biases and 95% simulation confidence intervals of estimated partial regression coefficients of X_1 and X_2. For all types of data distributions, all coefficients are estimated without bias except for the case with categorical distribution. Based on the unconditional independence, the coefficient of X_2 are recoverable without conditioning on X_3. The bias that occurred in the case of categorical distribution is largely due to the inadequacy of imputation method. Nonetheless, the results imply that if researchers are not interested in means but only in unbiased regression coefficients, including X_3 in the imputation is not necessary.

Table 8. The biases and 95% simulation confidence intervals of estimated partial regression coefficients of X_1 and X_2 in case 2.

Model	Par	Normal			Skew			Categorical		
	Coef	Bias	CIL	CIH	Bias	CIL	CIH	Bias	CIL	CIH
With	X1	0.004	-0.007	0.015	-0.018	-0.077	0.041	-0.025	-0.066	0.016
X3	X2	-0.003	-0.005	0.012	0.085	-0.030	0.200	0.075	0.032	0.118
Witho	X1	0.007	-0.004	0.018	-0.030	-0.089	0.030	0.028	-0.014	0.070
ut X3	X2	0.001	-0.007	0.010	0.078	-0.039	0.194	0.009	-0.033	0.052

Correlations. Table 9 shows the biases of pairwise correlations among X_1, X_2, X_3 and Y. Without including X_3 in the imputation, the correlations between X_2 and X_3 are biased. The estimated correlations between X_1 and X_3 do not show strong bias because the true correlations between X_1 and X_3 are very small and almost negligible. In this case, if the means and correlations are the priorities, X_3 should be included in the imputation.

Table 9. Population parameter values of variables in case 2.

Model	Parameter Cor	Normal Bias	Skew Bias	Categorical Bias
With X3	X1X2	-0.001	-0.001	0.000
	X1X3	0.000	0.000	0.000
	X2X3	0.000	0.000	0.001
	X1Y	-0.001	-0.001	-0.001
	X2Y	0.000	0.000	0.002
Without X3	X1X2	0.001	-0.001	-0.001
	X1X3	0.000	0.001	0.002
	X2X3	-0.038	-0.018	-0.045
	X1Y	-0.001	-0.001	0.000
	X2Y	0.002	0.000	0.001

Case 3

Means. Table 10 presents for case 3 the biases and 95% simulation confidence intervals of estimated means of X_1 and X_2, with and without including X_3 in the imputation model. For all types of data distributions, the means of X_2 are estimated with bias when including X_3 in the imputation. For the multivariate normal distribution, the mean of X_1 is also slightly biased. This might be largely due to the uncertainty of imputation. Given that the size of bias in mean of X_2 is much larger and imputing the missing data in X_1 needs to borrow the information from X_2, the estimated mean of X_1 can be slightly biased with limited sample size and limited number of imputation. Without X_3 in the imputation, the mean of X_2 is estimated without bias. But after

including X_3, X_2 is estimated with bias due to the collider X_3. This implies that to ensure the unbiased estimates of means of X_2, X_3 should not be included in the imputation model when X_3 is the collider between X_2 and its the missingness R_{x2}.

Table 10. The biases and 95% simulation confidence intervals of estimated means of X_1 and X_2 in case 3.

Model	Parameter	Normal			Skew			Categorical		
	Mean	Bias	CIL	CIH	Bias	CIL	CIH	Bias	CIL	CIH
With	X1	-0.004	-0.005	-0.003	0.000	0.000	0.000	0.000	-0.001	0.000
X3	X2	0.360	0.358	0.362	-0.005	-0.005	-0.005	0.003	0.002	0.003
Without	X1	0.000	-0.001	0.001	0.000	0.000	0.000	0.000	-0.001	0.000
X3	X2	0.001	0.000	0.002	0.000	0.000	0.000	0.000	-0.001	0.000

Regression coefficients. Table 11 presents the biases and 95% simulation confidence intervals of estimated partial regression coefficients of X_1 and X_2. For all types of data distributions, the coefficients of X_2 are biased after including X_3 in the imputation. The coefficients of X_2 are recoverable and estimated without bias when X_3 is excluded from the imputation. One exception with the categorical distribution is that the partial regression coefficient of X_1 is biased without X_3. This is again somehow expected due to the imputation method. The interesting finding is that by including X_3, the coefficients of X_1 are also strongly biased. Though based on the theory with infinite large sample size, the coefficients of X_1 should be recovered without bias, given the conditional independence holds, $R_{x1} \perp Y \mid X_1$ (Figure 6). However, including the collider biases the coefficients of X_1. This is largely due to the fact that X_1 and X_2 are correlated. If the coefficients of X_2 are biased, it is not surprising to see the bias in the coefficients of X_1. The results imply that X_3 should not be included in the imputation if the goal is to obtain unbiased regression coefficients.

Table 11. The biases and 95% simulation confidence intervals of estimated partial regression coefficients of X_1 and X_2 in case 3.

Model	Parameter	Normal			Skew			Categorical		
	Coef	Bias	CIL	CIH	Bias	CIL	CIH	Bias	CIL	CIH
With X3	X1	-0.931	-0.960	-0.902	-0.204	-0.327	-0.082	-0.051	-0.064	-0.038
	X2	2.552	2.526	2.578	0.287	0.141	0.432	0.021	0.009	0.038
Without X3	X1	-0.025	-0.055	0.006	-0.011	-0.138	0.117	-0.030	-0.043	-0.017
	X2	0.033	-0.002	0.068	0.068	-0.084	0.219	-0.011	-0.023	0.001

Correlations. Table 12 shows the biases of pairwise correlations among X_1, X_2, X_3 and Y. For the continuous variables, without including X_3 the correlations between X_1 and X_3 , and X_2 and X_3 are biased due to the lack of overlap. The size of bias depends on the size of true correlation values. After including X_3 in the imputation, the bias is not improved. Since X_3 introduces collider bias, we found additional bias in correlations between X_1 and X_2, X_1 and Y, and X_2 and Y, compared to the bias when excluding X_3. For the categorical distribution, the biases in correlations do not have strong systematic changes. In this final case, regardless including X_3 or not, the correlations that involve X_3 are biased. But in order to preserve other correlations, X_3 should not be included in the imputation.

Table 12. Population parameter values of variables in case 3.

Model	Parameter	Normal	Skew	Categorical
	Cor	Bias	Bias	Bias
With X3	X1X2	-0.114	-0.014	0.024
	X1X3	-0.151	-0.006	-0.004

	X2X3	0.073	-0.035	-0.003
	X1Y	-0.002	-0.001	-0.013
	X2Y	0.017	-0.001	0.016
Without X3	X1X2	-0.002	-0.003	0.026
	X1X3	-0.147	-0.005	-0.005
	X2X3	-0.476	-0.018	-0.007
	X1Y	-0.000	0.000	-0.013
	X2Y	0.000	0.000	0.016

Conclusion

This paper uses graphical models to investigate the specification of imputation models regarding obtaining unbiased parameter estimates of planned missing data. Contrary to the common belief that all auxiliary variables should be used to impute missing data when the missingness is ignorable, we show that in some cases including the auxiliary variables is not necessary and in other cases it causes bias in all parameter estimates. Even if the missingness mechanism is ignorable as frequently the case in planned missing data designs, whether an auxiliary variable should be included in the imputation model not only depends on the causal relationship between the auxiliary variable and the missingness of other variables but also on the parameters of interest. The illustrated scenarios guide researchers to identify these cases in the setting of planned missing data designs and decide if the auxiliary variable should be used in the imputation.

To summarize, the first case shows that when the auxiliary variable is neither a cause nor a descendant of the missingness of another variable, it is not necessary to be used in the imputation model to recover unbiased mean and partial regression coefficient of this variable, unless the interest is on the correlations that involve this auxiliary variable. In the second case,

when the auxiliary variable is a common cause of a variable X and the missingness of this variable R_x, it should be used to recover the mean of X. But it is not necessary for recovering the partial regression coefficient of X. In the third case, when the auxiliary variable is a common descendant of X and R_x, it should not be used to recover any parameters, because conditioning on the auxiliary variable introduces collider bias. The cases examined by no means exhaust all possibilities. For example, future studies can look at cases with planned missing data in auxiliary variables, or when the outcome variable has missing data, or when the auxiliary variable is a cause or a descendent of the outcome variable.

There are many challenges in obtaining unbiased causal parameter estimates in planned missing data designs. First the main challenge is to lay out the graphical model for all variables based on theory. However, if the causal model is laid out and the priority of parameters of interest is clear, we can identify if the parameters are recoverable. Second, even if the parameter estimates are recoverable by theory, it does not ensure that they are estimated without bias. Correctly specified imputation models have to be guaranteed to obtain unbiased estimates. Third, given a correctly specified imputation model, imputation methods need to be adequate on dealing with different types of data distributions and limited sample sizes, especially with categorical data. Bias can be reduced by increasing the sample size, but will not vanish given a finite sample since multiple imputation only delivers consistent parameter estimates. Further studies can look into the imputation procedures and methods that are capable of adapting to each variable case by case with regard to the search of imputation models.

Reference

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444-455.

Collins, L., Schafer, J., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330–351.

Enders, C. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics*, 41(1), 57-80.

Kaplan, D. & Su, D. (2018). On imputation for planned missing data in context questionnaires using plausible values: a comparison of three designs. *Large-scale Assessments in Education*, 6:6, doi:10.1186/s40536-018-0059-9

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*(3), 287-296.

Mohan, K., & Pearl, J. (2014). On the Testability of Models with Missing Data. In *AISTATS* (pp. 643-650).

Mohan, K., Pearl, J., & Tian, J. (2013). Graphical models for inference with missing data. In *Advances in neural information processing systems* (pp. 1277-1285).

Organization for Economic Cooperation and Development. (2006). *PISA 2006 technical report*. Paris, France: Author.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J., & Mohan, K. (2013). Recoverability and testability of missing data: Introduction and summary of results. Available at SSRN 2343873.

Raghunathan, T (2016). *Missing data analysis in practice*. CRC press.

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

Rubin, D. B. (1987). *Multiple imputation in nonresponse surveys*. Hoboken, NJ: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473-489.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2), 147.

Steiner, P. M., Kim, Y., Hall, C. E. & Su, D. (2015). *Graphical models for quasi-experimental designs. Sociological Methods & Research*. doi: 0049124115582272

Su, D. (2018). An evaluation of planned missing data designs in large surveys. Unpublished manuscript.

Thoemmes, F., & Rose, N. (2014). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate behavioral research*, 49(5), 443-459.

Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4), 631-642.

Tian, J. (2015). Missing at random in graphical models. *In Artificial Intelligence and* Statistics, (pp. 977-985).

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3): 219–242

van Buuren, S., & Groothuis-Oudshoorn, K. (2010, January). *Multivariate imputation by chained equations, version 2.3*. http://www.multiple-imputation.com/.