

## ON THE ASYMPTOTIC RELATIVE EFFICIENCY OF PLANNED MISSINGNESS DESIGNS

MIJKE RHEMTULLA

UNIVERSITY OF AMSTERDAM

VICTORIA SAVALEI

UNIVERSITY OF BRITISH COLUMBIA

TODD D. LITTLE

TEXAS TECH UNIVERSITY

In planned missingness (PM) designs, certain data are set *a priori* to be missing. PM designs can increase validity and reduce cost; however, little is known about the loss of efficiency that accompanies these designs. The present paper compares PM designs to reduced sample (RN) designs that have the same total number of data points concentrated in fewer participants. In 4 studies, we consider models for both observed and latent variables, designs that do or do not include an “X set” of variables with complete data, and a full range of between- and within-set correlation values. All results are obtained using asymptotic relative efficiency formulas, and thus no data are generated; this novel approach allows us to examine whether PM designs have theoretical advantages over RN designs removing the impact of sampling error. Our primary findings are that (a) in manifest variable regression models, estimates of regression coefficients have much lower relative efficiency in PM designs as compared to RN designs, (b) relative efficiency of factor correlation or latent regression coefficient estimates is maximized when the indicators of each latent variable come from different sets, and (c) the addition of an X set improves efficiency in manifest variable regression models only for the parameters that directly involve the X-set variables, but it substantially improves efficiency of most parameters in latent variable models. We conclude that PM designs can be beneficial when the model of interest is a latent variable model; recommendations are made for how to optimize such a design.

Key words: planned missing, missingness by design, incomplete data, missing data, efficiency, power.

### 1. Introduction

In planned missingness (PM) designs, certain data are randomly determined *a priori* to be missing. Reasons to use these designs include reducing participant burden in long surveys (Graham, Hofer, & MacKinnon, 2006; Saris, Satorra, & Coenders, 2004), reducing unplanned missing data rates (Harel, Stratton, & Aseltine, 2011), reducing cost over longitudinal assessments (Graham, Taylor, & Cumsille, 2001), and reducing retest effects and otherwise increasing data quality (Harel et al., 2011). In general, PM designs can increase validity and/or reduce the cost of data collection; however, little is known about the power loss associated with these designs.

Correspondence should be sent to Mijke Rhemtulla, Programme Group Psychological Methods, Department of Psychology, University of Amsterdam, Weesperplein 4, Room 208, 1018XA Amsterdam, Netherlands. Email: m.rhemtulla@uva.nl

A wide range of PM designs have been proposed, including designs for multi-trait multi-method research (Bunting, Adamson, & Mulhall, 2002; Saris et al., 2004; Revilla & Saris, 2013), longitudinal growth curves (Graham, Taylor, & Cumsille, 2001; McArdle & Woodcock, 1997; Mistler & Enders, 2012), cross-sectional survey research (Graham et al., 2006; Raghunathan & Grizzle, 1995; Thomas, Raghunathan, Schenker, Katzoff, & Johnson, 2006), case-control research (Wacholder, Carroll, Pee, & Gail, 1994), and educational assessment (e.g., “matrix sampling,” Shoemaker, 1973; Sirotnik & Wellington, 1977). The multi-form PM design (variously referred to as the “partial questionnaire,” “split questionnaire,” or “split ballot” design) is one of the simplest and most versatile of these, where each participant is randomly assigned to receive only a subset of all variables (Arminger & Sobel, 1990; Graham et al., 2006; Raghunathan & Grizzle, 1995; Wacholder et al., 1994). The multi-form design is most useful when one wants to collect data using a large number of variables but is faced with time constraints or concerns about respondent burden and fatigue. The most popular of these designs is the 3-form design (Graham et al., 2006; Graham, Hofer, & MacKinnon 1996; Graham et al. 1994), in which the variables are divided into 4 subsets including a common set (X) and three partial sets (A, B, and C). The variables assigned to X include those that are central to the study’s main hypotheses and are part of the model as well as those that are indicative of potential MAR missing processes and thus serve as auxiliary variables. The partial sets of variables in A, B, and C plus the common set X form three combinations: XAB, XBC, and XAC. Each combination is administered to one third of the participants. Random assignment to forms results in data that conform to the missing completely at random (MCAR; Little & Rubin, 2002) assumption.

Missing data, even if they are MCAR, will have a detrimental effect on the efficiency of the resulting estimates by reducing the amount of information available in a dataset (Orchard & Woodbury, 1972). This loss of efficiency in PM designs is relative to the design where all data are observed on all participants. A different comparison case would be a Reduced N (RN) design, where complete data are gathered on fewer total cases, resulting in the same total number of data points. It is not known how the efficiency of PM designs compares to such RN designs. It is possible that a larger dataset with some amount of planned missingness will result in more efficient parameter estimates than a smaller, complete dataset (Graham et al., 2001; Mistler & Enders, 2012).

We conducted four studies to examine the asymptotic relative efficiency of parameter estimates under PM designs relative to complete data (CD) designs across a range of parameter types and models. The studied designs included full PM designs (all cases have missing data), reduced N designs (RN; all cases are complete, and the sample size is reduced such that the total number of data points is the same as in a PM design), and a full range of intermediate (INT) designs having some proportion of cases with complete data and the remainder with missing data, with the sample size adjusted such that the total number of data points is the same as in the PM and RN designs.

For PM designs, relative efficiency (RE) was defined as the ratio of the squared standard errors for a parameter estimated in a PM design versus a complete data (CD) design with the same N. RE ranges from 0 to 1, where lower values reflect more efficiency loss as a result of missing values. RE thus provides the researcher with valuable information about the extent to which significance tests and confidence intervals are affected by missing data. For example, RE can be used to compute a width inflation factor,  $WIF_{\theta} = 1/\sqrt{RE_{\theta}}$ , which reflects the extent to which a confidence interval around the estimate of parameter  $\theta$  is expected to be inflated relative to complete data, and it can be used to compute an effective sample size,  $N_{\theta}^* = N \cdot RE_{\theta}$ , the complete data sample size that would result in the same efficiency for parameter  $j$  (Savalei & Rhemtulla, 2012). Given the results of a power analysis based on complete data, RE can be translated directly into power (more details are given in the next section).

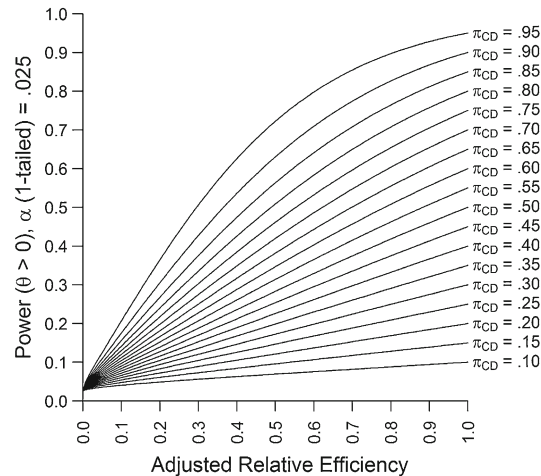


FIGURE 1.

Power to test the hypothesis that  $\theta_i > 0$  based on a Wald test.  $\pi_{CD}$  is power based on the complete data (CD) design.

For RN and INT designs, *adjusted* relative efficiency (ARE) was studied, defined as the ratio of the squared standard errors for a parameter estimated in an RN or INT design versus a CD design while *equating the total number of data points*. ARE allows direct comparisons of PM, RN, and INT designs. For example, ARE provides a direct comparison of a PM design that measures 75 % of variables on a sample of size  $N$ , an RN design that measures 100 % of variables on a sample of size  $.75N$ , and an INT design that measures 100 % of variables on 50 % of a sample size of  $.86N$  and 75 % of variables on the other 50 % (see Figure 1). For PM designs, ARE simply equals RE. Evaluating efficiency per data point is critical for determining the real cost savings of PM designs, under the assumption that research cost is a linear function of the number of data points collected. Of course, in many situations, cost is not linearly related to the number of data points (e.g., cost per participant might double when an assessment gets too long to complete in one sitting). If the cost of collecting complete data on a participant relative to the cost of collecting incomplete data (on a PM design) is known, then RE estimates can be transformed to accurately reflect cost efficiency (more details are given under “cost efficiency” in Sect. 7).

There is some indication in the literature that PM designs can result in more efficient estimates than RN designs. First, in the context of longitudinal PM designs, Graham et al. (2001) examined the power to detect the effect of group membership (a complete binary predictor) on the latent slope in a linear growth-curve model with five waves of data and planned wave-level missingness. They studied five PM designs, where participants missed up to two waves out of five (e.g., missingness was distributed equally across waves or concentrated in the middle waves, with the total amount of missing data varying from 17 to 55 %). Every PM design resulted in an estimate of the effect of group membership (a binary observed variable) on slope that was more efficient than in the corresponding RN design. Graham et al. did not report the results for any other parameters.

Second, Mistler and Enders (2012) extended Graham et al.’s finding by examining power to detect the mean of linear and quadratic latent slopes. They investigated two PM designs in which participants missed two out of six waves: missingness was either distributed equally across waves or confined to the middle four waves. Both linear and quadratic slope means were more efficiently estimated when missingness was confined to the middle waves than in the RN design, and the RN design produced more efficient estimates than when missingness was distributed equally across all six waves.

Third, Raghunathan and Grizzle (1995) examined the efficiency of a single logistic regression coefficient estimate in a multi-form design where 1,000 (simulated) participants completed 3 out of 5 item sets, leading to 40 % missing data on every variable. The model had a binary dependent variable, a binary predictor, and two continuous predictors. The parameter of interest was the regression path of the binary predictor on the binary dependent variable; these were assigned to different item sets. Missing data were dealt with using multiple imputations, and efficiency was measured via the width of the confidence interval surrounding the target regression coefficient relative to the width of the CD design confidence interval. The lowest observed ratio was 1.006 (reflecting a confidence interval that was a scant .6 % wider than that obtained from complete data); this ratio was observed when the strength of correlations among pairs of items between item sets was high (.8), while the strength of correlations among pairs of items within a set was low (.1). The highest observed ratio was 2.179 when the strength of all correlations (both between and within item sets) was .1. This study did not include the comparable RN condition (i.e., a complete data design with 20 % fewer cases). Had the RN condition been included, the expected value of its confidence interval ratio would be  $1/\sqrt{.80} = 1.12$ . That is, the RN design would be less efficient than PM when correlations are high ( $1.12 > 1.006$ ) and more efficient than PM when correlations are low ( $1.12 < 2.179$ ).

The finding that correlation strength affects efficiency is supported by other research as well. Graham et al. (1996) examined the average standard error of variance and covariance estimates resulting from either assigning whole scales to the same set or to different sets in a 3-form PM design. Items on a scale typically have higher correlations with each other than with items on other scales; thus, assigning entire scales to sets results in high within-set correlations, whereas splitting scales across sets results in higher between-set correlations. Graham et al. generated 3-item scales, with intra-scale item correlations ranging from .3 to .9, and inter-scale item correlations ranging from .1 to .5. The results reflected the same pattern found in Raghunathan and Grizzle (1995): Average efficiency was higher when scale items were highly correlated with each other, and they were assigned to different sets (resulting in high between-set correlations). In their most dramatic finding, when intra-scale correlations were .9 and inter-scale correlations were .1, average standard errors (averaged across parameter type) were five times larger when scale items were assigned to the same set than when they were split across sets. Graham et al. did not report efficiency relative to RN or complete data designs, however, so it is not known which, if any, of these observed standard errors represented an efficiency advantage relative to an RN design. They also did not report efficiency gains for the two parameter types separately, so the differential effect of correlation strength on efficiency of variances and covariances cannot be assessed.

This review suggests three factors that may affect the efficiency of parameter estimates in PM designs. The first is the type of parameter (e.g., covariances vs. regression coefficients). Even a reparameterization of the same mean and covariance structure will change the efficiency of the corresponding parameters. The second is the absolute and relative strength of between-set and within-set correlations. Third is the proportion of fully observed cases included in the design (e.g., a full 3-form PM design where all rows have missing data vs. an INT design where 20 % of rows are fully observed but the total sample size is smaller). While several studies have examined the impact of using a PM design on efficiency of certain types of parameters in certain types of models, no study has systematically examined how introducing planned missing data impacts the relative efficiency of parameter estimates in different models.

In the present study, we thoroughly investigate the asymptotic relative efficiency ratios of parameter estimates when data are collected via a PM design and compare them to the corresponding relative efficiency ratios under RN and INT designs, controlling for the total number of available observations. We employ a novel methodology of examining the exact asymptotic relative efficiency values obtained mathematically, rather than generating simulation data from finite sample sizes. This method allows us to remove the impact of sampling error from the comparison

of the designs. We examine all types of parameters within a variety of models, including saturated models, manifest variable regression models, and latent variable models. We vary the between-set and within-set correlations across their full permissible range. We also vary the proportion of complete data cases from 0 (PM) to 1 (RN), studying the full range of intermediate (INT) designs. Thus, we examine the impact of all three factors on efficiency simultaneously.

## 2. Method

Four studies were conducted, each evaluating several different models. Studies 1 and 2 examined the standard saturated model (i.e., a model with freely estimated means and covariances, Raykov, Marcoulides, & Patelis, 2013) and two 2-predictor regression models. Study 3 examined latent variable versions of two of the models in Studies 1 and 2, where each latent variable had three or four indicators (one from each item set). Study 4 compared two methods for assigning factor indicators to item sets in a 2-predictor latent regression model. Studies 1 and 4 used a 3-form PM design with no X set, Study 2 used a 3-form PM design with an X set, and Study 3 used a 3-form PM design with and without an X set.

For each model, two main variables were varied: (1) correlation strength (from .01 to .99 by .01 increments)—correlations among variables within and between sets were varied independently in Studies 1 and 2, whereas in Studies 3 and 4, the size of correlations among latent factors were varied while factor loadings were held constant; and (2) the proportion of complete cases included in the design, from 0 (PM) to 1 (RN), by .01 increments. Fully crossing these factors resulted in 980,100 conditions per model in Studies 1 and 2, and 9,900 conditions per model in Studies 3 and 4. The full results thus cannot be displayed in this paper or reasonably included in supplementary materials. We present a representative sample of data and summaries across selected conditions; results from other conditions are available from the authors on request.

All computations were done with the theoretical information matrices associated with the normal theory full information maximum likelihood estimator, evaluated at the true parameter values and assuming MCAR data. The advantage of our asymptotic approach is that it does not require generating or analyzing raw data; it also produces stable results that remove the influence of sample size on relative efficiency estimates. Technical details are given next. Several conditions were verified using simulated data with  $N = 500$ , and the results were highly similar.

### 2.1. Asymptotic Adjusted Relative Efficiency Computations

To compute asymptotic relative efficiency (RE) and adjusted relative efficiency (ARE) ratios comparing PM, RN, and INT designs to CD designs, asymptotic standard errors for both complete and incomplete data are needed. We assume that the data are multivariate normal with mean  $\mu$  and covariance matrix  $\Sigma$ , and that full information maximum likelihood was used to obtain parameter estimates (Arbuckle, 1996). Asymptotic SEs can be obtained from the diagonal of the inverse of the corresponding complete and incomplete data information matrices. These information matrices are greatly simplified under the MCAR assumption, which is met in these designs. All computations were done in R (R Statistical Core Team, 2011).

*2.1.1. Complete Data Computations* Let  $p$  be the number of variables and let  $p^* = .5p(p+1)$ . For the standard saturated model that freely estimates all variances, covariances, and means (Models 1 and 4 in the present study), the  $(p^* + p) \times (p^* + p)$  complete data information matrix is given by

$$H = \begin{pmatrix} .5D_p' (\Sigma^{-1} \otimes \Sigma^{-1}) D_p & 0 \\ 0 & \Sigma^{-1} \end{pmatrix},$$

where  $D_p$  is the duplication matrix of Magnus and Neudecker (1999), and the partitioning corresponds to covariance structure parameters and mean structure parameters, which are asymptotically independent (e.g., Bentler, 2007; Savalei, 2010). The asymptotic covariance matrix of parameter estimates is then given by  $\Omega = H^{-1}$ , and the asymptotic standard errors for complete data are obtained from the diagonals of  $\Omega$ . It will become important for the interpretation of some of the results to consider the  $p^* \times p^*$  submatrix of  $\Omega$  corresponding to the covariances:  $\Omega_{11} = 2D_p^+(\Sigma \otimes \Sigma)D_p^{+'}$ . The typical diagonal element of this matrix, corresponding to the asymptotic variance of an estimate of  $\sigma_{tk}$  ( $t = 1, \dots, p; k = 1, \dots, p$ ), is given by  $\{\Omega_{11}\}_{tk,tk} = \sigma_{tt}\sigma_{kk} + \sigma_{tk}^2$ . When  $\Sigma$  is set to have unit variances, as in this study, this expression simplifies to  $\{\Omega_{11}\}_{tk,tk} = 1 + r_{tk}^2$ .

For a just-identified model that is parameterized differently (Models 2, 3, and 5), let  $\theta$  be the  $q \times 1$  vector of model parameters and let  $\beta(\theta)$  be the model. Define the  $(p^* + p) \times q$  matrix of model derivatives  $\Delta = \frac{\partial \beta(\theta)}{\partial \theta'}$ . The information matrix for the model is then given by  $H_M = \Delta' H \Delta$ , and the asymptotic covariance matrix of parameter estimates is obtained by  $\Omega_M = H_M^{-1} = (\Delta' H \Delta)^{-1} = \Delta^{-1} H^{-1} \Delta'^{-1}$ , where the last step is possible because  $\Delta$  is invertible as the model is saturated (i.e.,  $q = p^* + p$ ). Algebraic expressions for  $\Delta$  for each of the models were used. We checked our derivations of these expressions against the results of a large-sample simulation. Variants of these expressions are also available in many other sources (e.g., Bentler & Lee, 1978; Nel, 1980; Mooijaart & Bentler, 1991).

For structured models, where  $q < p^*$  (Models 6, 6X, 7, 7X, 8a, and 8b), the matrix  $\Delta$  is defined as above but is no longer square, and thus no further simplification is possible for the expression  $\Omega_M = H_M^{-1} = (\Delta' H \Delta)^{-1}$ . All studied models of this type have a saturated mean structure, and thus  $\Omega_M$  is block-diagonal. Computations were done using the reduced-size information and derivatives matrices, involving only covariance structure parameters. We obtained an algebraic expression for the matrix of model derivatives for Model 6/6X, for two types of identification (factor variances fixed to 1 and first factor loadings fixed to the population value of .7). The derivative matrices for Models 7/7X were obtained by combining the derivatives for Model 6/6X (using identification via fixing loadings to .7) and Model 3 via the chain rule, and those for Models 8a/8b were done analogously.

*2.1.2. Incomplete Data Computations* Let  $j = 1, \dots, J$  indicate the missing data patterns,  $p_j$  be the number of variables observed in pattern  $j$ , and  $q_j$  be the probability of pattern  $j$ . In 3-form PM designs,  $J = 3$  and  $q_j = 1/3$  for all  $j$ . In INT designs,  $J = 4$  and  $q_j = (1 - q_1)/3$  for  $j = 2, 3, 4$ , where  $q_1$  is the probability of the complete data pattern.

For the standard saturated model (Models 1 and 4), the  $(p^* + p) \times (p^* + p)$  incomplete data information matrix assuming MCAR data is given by

$$H_{inc} = \begin{pmatrix} .5D_p' \sum_{j=1}^J q_j (\tau_j' \Sigma_j^{-1} \tau_j \otimes \tau_j' \Sigma_j^{-1} \tau_j) D_p & 0 \\ 0 & \sum_{j=1}^J q_j \tau_j' \Sigma_j^{-1} \tau_j \end{pmatrix},$$

where  $\Sigma_j$  is a  $p_j \times p_j$  submatrix of  $\Sigma$  with rows/columns corresponding to those variables missing for the  $j$ th pattern removed, and  $\tau_j$  is a 0–1 matrix of dimension  $p_j \times p$ , which can be obtained by removing, from the  $p \times p$  identity matrix, those rows corresponding to missing variables for pattern  $j$  (Yuan & Bentler, 2000; Savalei, 2010). The asymptotic covariance matrix of parameter estimates is then given by  $\Omega_{inc} = H_{inc}^{-1}$ , as before, but this inverse no longer has a simple form. The expression for the asymptotic covariance matrix of any other model is given by  $\Omega_{M,inc} = (\Delta' H_{inc}^{\Delta})^{-1}$ , where  $\Delta$  is the derivatives matrix for the model, defined exactly as

before. As MCAR data ensures the asymptotic independence of mean parameters and covariance structure parameters, computations for Models 6/6X, 7/7X, and 8a/8b were again done using the covariance structure components of the relevant matrices only.

*2.1.3. Adjusted Relative Efficiency Computations* For the  $i$ th parameter  $\theta_i$ , the asymptotic relative efficiency of an incomplete data design relative to the complete data (CD) design with the same  $N$  is defined as  $RE_i = \frac{\{\Omega_M\}_{ii}}{\{\Omega_{M,inc}\}_{ii}}$ , the ratio of the complete data asymptotic variance to the incomplete data asymptotic variance for that parameter.

If comparing across types of designs were not of interest, then the asymptotic relative efficiency for INT designs would be defined in the same way as above. However, of primary interest in this study was the comparison across types of designs (e.g., INT vs. PM), while controlling for the total number of observations in any finite sample  $N$ . For this reason, adjusted (for total number of observations) asymptotic relative efficiency of INT designs was computed. The adjustment constant  $c$  is derived below. For a PM design with sample size  $N$ , this adjustment gives the new sample size  $N^* = cN$  that the corresponding INT design should use to keep constant the total number of data points in the data matrix.

In a 3-form PM design with a finite  $N$  that is divisible by 3, the total number of observations is  $T = Np_J$ , where  $p_J$  is the length of any missing data pattern (patterns have the same length by design). In Study 1,  $p = 6$ ,  $p_J = 4$  and thus  $T = 4N$ . In Study 2,  $p = 8$ ,  $p_J = 6$ , and  $T = 6N$ . In Study 3, Models 6 and 7,  $p = 18$ ,  $p_J = 12$ , and  $T = 12N$ ; for Models 6X and 7X,  $p = 24$ ,  $p_J = 18$ , and  $T = 18N$ . In study 4, Models 8a and 8b,  $p = 9$ ,  $p_J = 6$ , and thus  $T = 6N$ . Now consider an INT design with the sample size  $N^*$ , which includes a complete data pattern that has probability  $q_1$ . The total number of observations is  $T^* = N^*q_1p + N^*(1 - q_1)p_J$ . In Study 1,  $T^* = (2q_1 + 4)N^*$ . In Study 2,  $T^* = (2q_1 + 6)N^*$ . In Study 3, Models 6 and 7,  $T^* = (6q_1 + 12)N^*$ , and for Models 7X and 8X,  $T^* = (6q_1 + 18)N^*$ . In Study 4, Models 8 and 8b,  $T^* = (3q_1 + 6)N^*$ . Thus, to ensure  $T = T^*$ , the adjustment constant is set to  $c = \frac{2}{q_1+2}$  for all models in Study 1 and for Models 6 and 7 in Study 3, and all models in Study 4. The adjustment constant is set to  $c = \frac{3}{q_1+3}$  for all models in Study 2 and for Models 6X and 7X in Study 3. In any finite sample size, the efficiency of parameter estimates under the PM and an INT design, while controlling for the total number of observations can be compared by setting  $N^* = cN$ . In asymptotic computations employed in this article, the adjusted relative efficiency for INT designs is computed as

$$ARE_i = c \frac{\{\Omega_M\}_{ii}}{\{\Omega_{M,inc}\}_{ii}}, \quad (1)$$

which adjusts the relative efficiency downward. Setting  $q_1 = 1$  results in an RN design, which, for example, produces  $ARE_i = 2/3$  for all parameters in Study 1. That is, when the sample size is two-thirds of the original sample size, estimates are two-thirds as efficient (see Figure 2). In the PM design,  $q_1 = 0$  and  $c = 1$ , so the constant drops out and  $ARE = RE$ . The ARE value for the RN design will be used as a benchmark for comparing PM and INT designs. We are interested in knowing whether the introduction of planned missing data, via PM or INT designs, ever results in increased efficiency compared to a complete data design with reduced number of observations, i.e., an RN design.

*2.1.4. Power* Unlike ARE, power to detect that a parameter is different from zero is affected by sample size. However, ARE can be used to determine power to test the hypothesis:  $H_1 : \theta_i > 0$  based on the Wald test for any parameter in a PM, INT, or RN design given the corresponding power for the CD design  $\pi_{CD}$  and 1-tailed alpha level,  $\alpha : \pi_{INC} =$

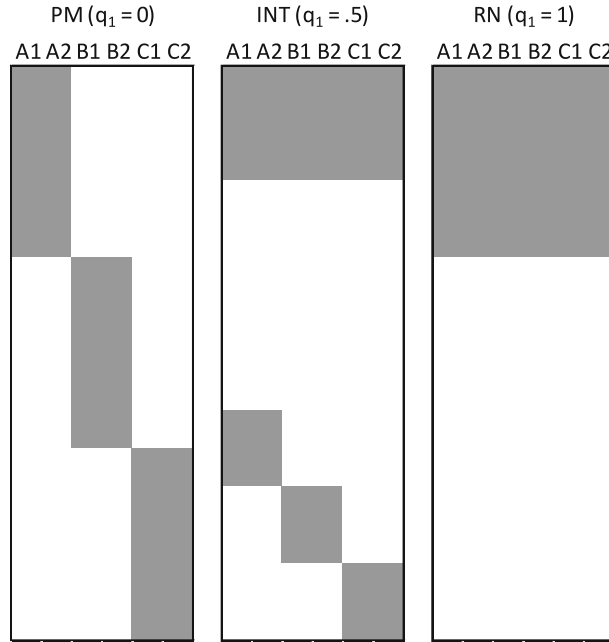


FIGURE 2.

Designs. *CD* complete data, *PM* planned missing, *INT* intermediate, *RN* reduced N. Variable sets are designated by letter (e.g., A1 and A2 are two variables in missing data set A, etc.). *Dark gray* represents missing data. *PM* designs have 0 % complete cases (all participants have 1/3 missing data). *INT* designs have between 1 and 99 % complete cases (i.e.,  $q_1 = .01$  to  $.99$ ; the design shown has  $q_1 = .5$ ) with the sample size reduced to keep the total number of observed data points constant. *RN* designs have 100 % complete cases, and a sample size reduced by 1/3.

$1 - F\left(F^{-1}(1 - \alpha) - \sqrt{\text{ARE}}\left(\left(F^{-1}(1 - \alpha)\right) + \left(F^{-1}(1 - \pi_{\text{CD}})\right)\right)\right)$ , where  $F(x)$  is the cumulative distribution function of the normal distribution, and  $F^{-1}(x)$  is its inverse. The derivation is given in Appendix A.

That is, power for any *PM*, *INT*, or *RN* design is a function of the *CD* design power, alpha level, and *ARE*. Figure 1 plots this function for values of  $\pi_{\text{CD}}$  ranging from .05 to .95, given 1-tailed  $\alpha = .025$ . Figure 1 can thus be used to translate any value of *ARE* into power based on an estimate of power in the *CD* design, which will, of course, be based on a discrete sample size.

### 3. Study 1

Study 1 examined the performance of a 3-form *PM* design with 6 variables, with two variables assigned to each set (A, B, and C; see Figure 3). While methodologists typically recommend including an X set in *PM* designs, the X set was omitted from this study to establish a baseline performance of all parameter estimates and also to explicitly evaluate the impact of the X set (added in Study 2) on efficiency.

All studied populations had zero means and unit variances on all variables. Thus, the off-diagonal elements of  $\Sigma$  could be considered covariances or correlations. We refer to them as correlations for the remainder of the paper when they are discussed as population parameters. The estimated models, however, were all covariance structure models, and we refer to the estimated parameters as estimated *covariances*. The population correlations were divided into two types: those between two variables in the same set (within-set correlations) and those between two vari-



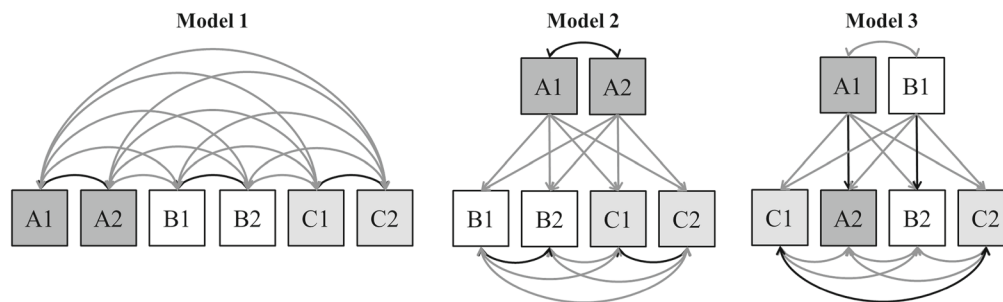


FIGURE 3.

Study 1, Models 1–3. Variable sets in the planned missing design are designated by background and letter (e.g., A1 and A2 are the two variables in missing data set A, etc.). *Black paths* are “within-set” correlations and regressions; *gray paths* are “between-set.” Variances of all variables are set to 1, and means are set to 0. All models are fit to data where the within-set correlations vary independently of the between-set correlations (both sets of correlations vary from 0 to 1).

ables in different sets (between-set correlations). Both types of correlations were independently varied from .01 up to .99 or to the highest value that ensured  $\Sigma$  remained nonnegative definite.<sup>1</sup>

Figure 3 displays the models studied. The results are summarized separately by type of parameter. Within each parameter type, we distinguish between within-set parameters (e.g., regression paths connecting two variables in the same set; black lines in Figure 3) and between-set parameters (e.g., regression paths connecting two variables in different sets; gray lines in Figure 3). In Model 3, we further distinguish between-set regression coefficients where the DV shares a set with one of the predictors (e.g., B2 on A1) from those where it does not (e.g., C1 on A1).

### 3.1. Results

*Model 1.* Model 1 includes four types of parameters: means, variances, within-set covariances, and between-set covariances. Selected results are presented in Figure 4. The top panel of Figure 4 displays the ARE (see Eq. 1) of parameter estimates when all variable correlations are equal, at  $\rho = .2$  (left),  $\rho = .5$  (middle), and  $\rho = .8$  (right). On the x-axis is  $q_1$ , the probability of the complete data pattern, ranging from  $q_1 = 0$  (PM design) to  $q_1 = 1$  (RN design), with intermediate values corresponding to INT designs. In any finite sample,  $q_1$  is the proportion of complete data cases in a sample of size  $N^* = \frac{2}{q_1+2}N$ , because adjusted efficiency ratios are being plotted. The sample size at which efficiency is compared thus varies, correspondingly, from  $N^* = N$  (PM design) to  $N^* = \frac{2}{3}N$  (RN design), with intermediate sample sizes corresponding to INT designs. The gray horizontal line at  $2/3$  represents the ARE of all parameter estimates in the RN design. Curves that begin (on the left) higher than  $2/3$  indicate that the corresponding parameter estimate is more efficiently estimated by the PM design than the RN design.

Figure 4 illustrates that means, variances, and within-set covariances are more efficiently estimated in PM and INT designs than RN designs, controlling for the total number of observations. This advantage is minimal at  $\rho = .2$  but increases as correlation size increases. Thus, having planned missing data is beneficial for these types of parameters. On the other hand, between-set covariances are estimated poorly with PM designs when the correlations among variables are low to moderate. At the lowest depicted correlation value ( $\rho = .2$ ), between-set covariances have  $ARE \approx 1/3$  under a PM design. In other words, when correlations are low, ARE values simply reflect the amount of data available for each estimate. As the correlations among variables increase,

<sup>1</sup> For between-set correlation values of .5 or lower, all within-set correlation values (between .01 and .99) were possible. Above .5, each .01 increment in between-set correlation value corresponded to a .02 increment in the minimum within-set correlation value (i.e., when the between-set correlation was .51, and the minimum within-set correlation was .02; when the between-set correlation was .99, the minimum within-set correlation was .98).

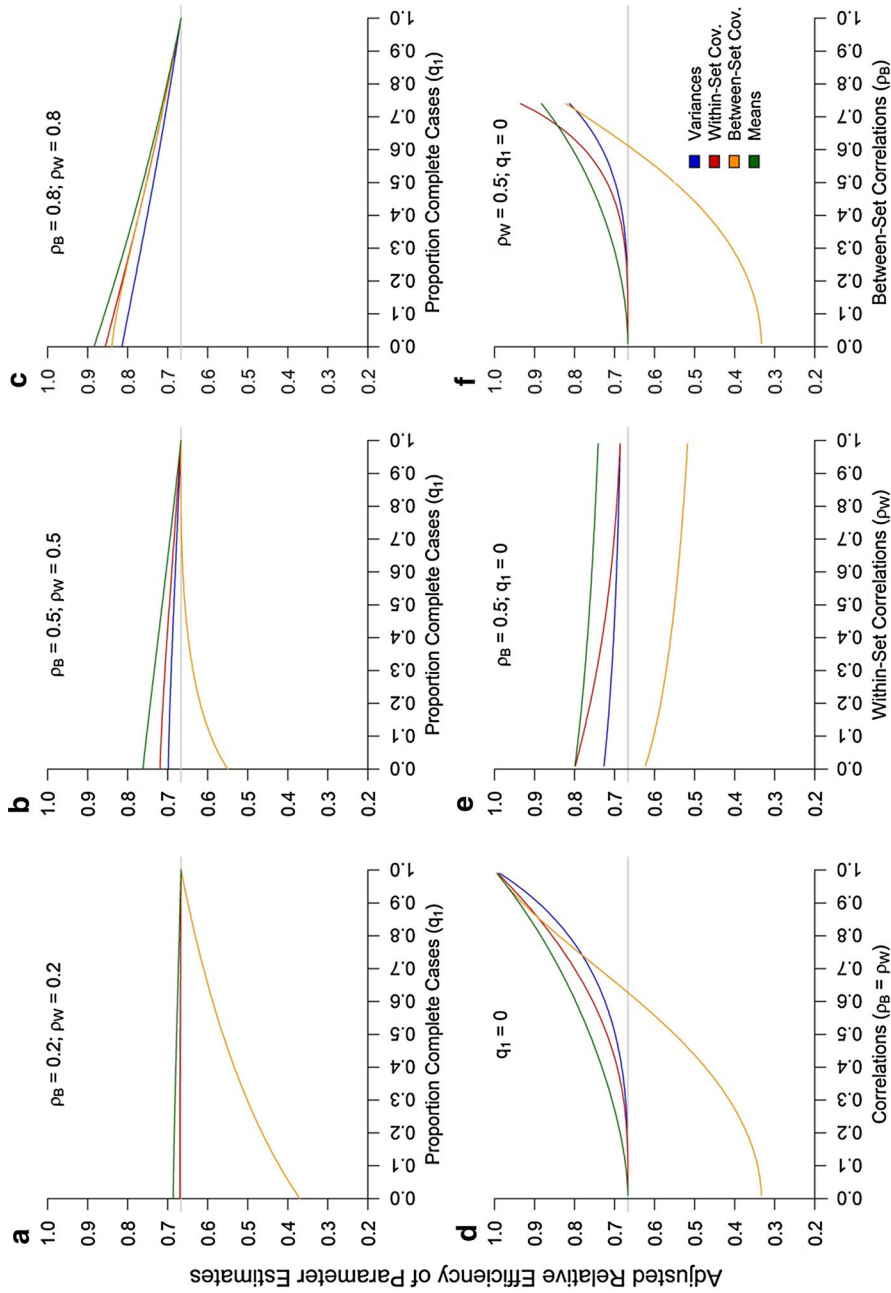


FIGURE 4.

Adjusted relative efficiency of parameter estimates in Model 1. The Y axis displays adjusted relative efficiency (ARE), which is the ratio of asymptotic variances of parameter estimates relative to the complete data design adjusted for the total number of data points (see Eq. 1).  $\rho_B$  is the value of all correlations among items between item sets, and  $\rho_W$  is the value of all correlations among items within the same item set. **a-c** the proportion of the sample which provides complete data (i.e.,  $q_1$ ) is varied along the X axis. **d**  $\rho_B$  and  $\rho_W$  are varied concurrently along the X axis; **e**  $\rho_W$  is varied along the X axis, while  $\rho_B$  is held constant at .5; **f**  $\rho_B$  is varied along the X-axis while  $\rho_W$  is constant at .5 (values of  $\rho_B$  above .74 result in a non-positive-definite matrix). Each line shows the RE of a single parameter and is the same for all parameters of the same type (e.g., all variances have the same population value and the same RE).

the ARE of between-set covariances under the PM design increases, presumably because more information from other variables becomes available. When correlations are very high ( $\rho = .8$ ), even the between-set covariances are much more efficiently estimated under the PM design, with ARE values in the .8 range, compared to  $2/3$  for the RN design.

Additionally, when correlations are moderate ( $\rho = .5$ ), adding a small proportion of complete cases (20–30 %) substantially improves the ARE of between-set covariance estimates so that it gets closer to  $2/3$  (Figure 4b). When variable correlations are moderate (specifically, between .49 and .77; not shown), REs of between-set covariances are actually at their highest when some proportion of cases have complete data (i.e., the yellow curves are no longer monotone). Including some proportion of complete cases in the PM design may thus maximize efficiency for the between-set covariance parameters for moderate size correlations; in the case of low correlations, the most efficient design for the between-set covariance parameters appears to be the RN design.

Figure 4d plots some of the same data differently: it shows ARE of parameters in the PM design ( $q_1 = 0$ ) as a function of correlation strength: along the X-axis, between- and within-set correlations are varied jointly from .01 to .99. This figure illustrates that the ARE of between-set covariances increases dramatically with correlation size, particularly for higher correlations. They become more efficiently estimated under the PM design than the RN design (the curve crosses the  $2/3$  line) when their true value is around .6. Other parameters are also more efficiently estimated with higher correlations but are always more efficiently estimated under a PM design than an RN design (their curves are always above  $2/3$ ). To further understand the effect of correlation size on ARE, Figure 4e, f shows the ARE plots when either between- or within-set correlation value is held at .5, while the other value is varied within its allowed range.

These two types of correlations affect ARE in opposite directions, although not to the same extent. As within-set correlations increase independently of between-set correlations, the ARE of all parameter estimates in PM designs decreases slowly (Figure 4e). In contrast, as between-set correlations increase independently of within-set correlations, the ARE of all parameter estimates in PM designs increases, and quite rapidly. The general pattern of relationships observed in Figure 4e, f holds when the correlation type held constant is fixed to values other than .5.

The finding that high between-set correlations helps efficiency makes intuitive sense: the higher these correlations, the more redundant each variable set is. At one extreme, if between-set correlations were 1, each set would provide no new information, and no information would be lost if one set was missing. At the other extreme, if between-set correlations were 0, each set's information is perfectly unique, so missing a set means missing all of the information contained within it—it is, therefore, not surprising that when correlations among variables are 0, ARE for many parameters simply reflects the amount of missing data on that parameter (e.g., between-set covariances have  $ARE = 1/3$ , and within-set covariances have  $ARE = 2/3$ ).

The finding that ARE of covariances is lower when within-set correlations are high may be understood by similar logic. As within-set correlations increase independently of between-set correlations, the amount of information shared across sets decreases. For example, holding between-set correlations constant at .5, when within-set correlations are 0, each variable in set A uniquely predicts 25 % of the variance in each variable in set B, such that 50 % of variable B1's variance can be predicted by the two variables in set A. But when within-set correlations are 1, each variable in a set predicts the same 25 % of variance in each variables in another set, such that just 25 % of item B1's variance can be predicted by the two variables in set A.

To gain insight into these results, it helps examine the two components of the relative efficiency computation separately (see Eq. 1, ignoring the  $c$  constant, which is 1 for PM designs). For the saturated model, the complete data asymptotic variances of mean and variance estimates are invariant to the strength of between- and within-set correlations, and the complete data asymptotic variances of covariance estimates increase as the parameter value increases. In contrast, the incomplete data asymptotic variances are functions of other parameters in the model. Examining

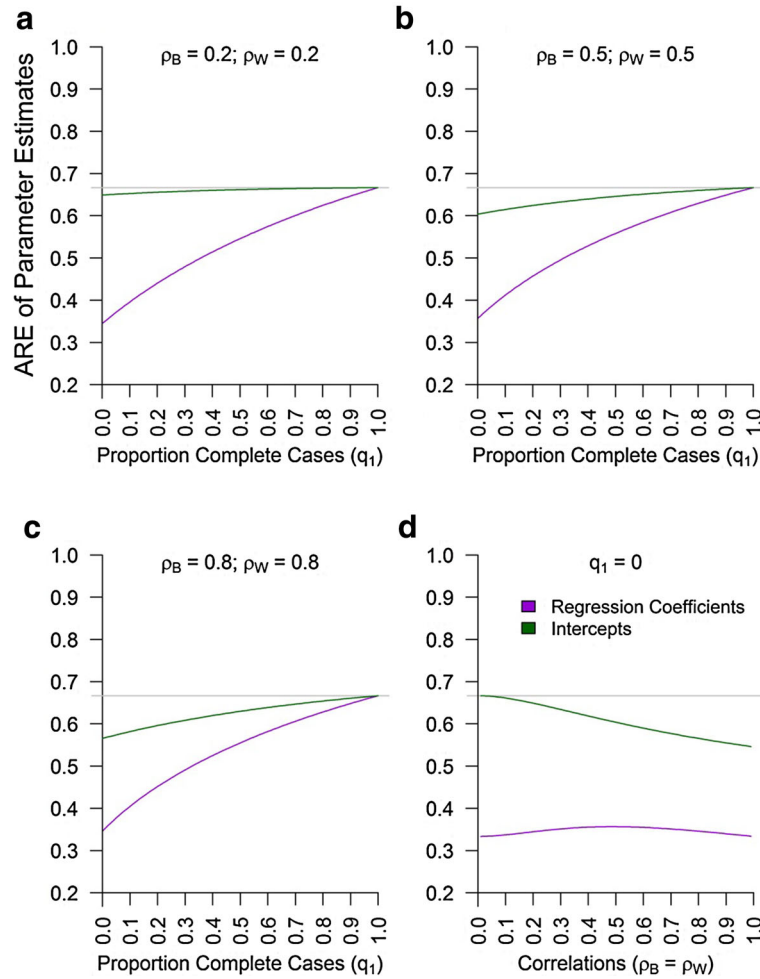


FIGURE 5.

Adjusted relative efficiency (ARE) of parameter estimates in Model 2. The Y axis displays ARE, which is the ratio of asymptotic variances of parameter estimates relative to the complete data design adjusted for the total number of data points (see Eq. 1).  $\rho_B$  is the value of all correlations among items between item sets, and  $\rho_W$  is the value of all correlations among items within the same item set. **a–c** The proportion of the sample which provides complete data (i.e.,  $q_1$ ) is varied along the X axis. **d**  $\rho_B$  and  $\rho_W$  are varied concurrently along the X axis. Each line shows the RE of a single parameter and is the same for all parameters of the same type.

the behavior of the incomplete data variances reveals that as the within-set correlations increase, incomplete data variances of all parameters increase, lowering efficiency. In contrast, as between-set correlations increase, incomplete data variances of all parameter estimates decrease, improving efficiency.

*Model 2* Figure 5 displays the adjusted relative efficiencies for regression coefficients, intercepts, residual variances, and within- and between-set residual correlations for Model 2. As shown in Figure 3, Model 2 is a regression model with two predictors (belonging to the same set) and four dependent variables (from two additional sets), and thus all regression coefficients are between sets. Means and variances of the predictor variables as well as the covariance between them have the same AREs as in Model 1; these results are not displayed.

Figure 5 illustrates that both regression coefficients and intercepts are less efficiently estimated under PM and INT designs than under an RN design. Regression coefficients (arguably the most important parameters in a regression model) have around .35 ARE in the PM design across all values of the correlations (the starting point of the purple curves in Figure 5a–c, the entire purple curve in Figure 5d). This means that to estimate between-set regression coefficients with comparable power to the complete data design with full N, a PM design could require almost three times as many participants. This dismal ARE can be raised to almost .50 by adding 30 % complete cases (Figure 5a–c), but adding this many or more complete cases would almost certainly defeat the practical purposes of using a PM design. ARE of the intercept estimates is much higher, hovering around .55 to .65 for the PM design, which is fairly close to the RN design (the starting point of the green curves in Figure 5a–c, the entire green curve in Figure 5d). The curves for the residual variances and covariances are omitted because they are generally not of interest in regression models, but we note briefly that between-set residual covariance estimates have the lowest REs, falling at around .30 or lower in the PM design, while residual variances and within-set residual covariances range from about .40 to .60 in the PM design. These results suggest that PM and INT designs do not fare well when the model of interest is a regression model, and regression coefficients are between-set, unless other practical factors such as participant fatigue strongly outweigh efficiency loss. If so, larger sample sizes than originally planned would be recommended.

Figure 5d illustrates that for PM designs, the ARE of intercepts (and also residual variances and covariances, not shown) decreases as correlations increase, while that of regression coefficients stays roughly stable. This trend is opposite of that observed for parameters in Model 1 (Figure 4d). The regression coefficients in multiple regressions are complex functions of the various parameters of the saturated model (Model 1), namely of the variances and covariances of exogenous variables as well as of the covariances between the exogenous variables and the dependent variable. Examining the numerator and denominator of the efficiency ratio separately (not shown) reveals that in the case of regression coefficients and intercepts, both the complete and the incomplete asymptotic variances decrease with the size of correlations. In the case of intercepts, incomplete data variances decrease more rapidly than complete data variances as the correlations increase, because they start at a higher value. In the case of regression coefficients, incomplete data variances decrease quickly initially and then slow down, whereas complete data variances decrease at a more steady pace, creating an ARE function that increases slightly then decreases again.

Figure 5 does not display the unique effects of varying between- and within-set correlations separately, because we find that these have a much smaller effect on the AREs of regression parameters than the corresponding effects on the parameters of the saturated model (Figure 4e, f).

*Model 3* Figure 6 presents the results for Model 3, which is a 2-predictor regression model but the two predictors belong to different sets (see Figure 3). This model allows us to examine the ARE of within-set regression coefficients (e.g., the regression of variable A2 on variable A1). Additionally, between-set regression coefficients can be grouped into two types: those where the dependent variable shares a set with the other predictor (e.g., B2 on A1), and those where the DV does not share a set with either predictor (e.g., C2 on A1). The behavior of intercepts is very similar to their behavior in Model 2 (as shown in Figure 5); these results are, therefore, not shown.

As Figure 6 illustrates, PM and INT designs again do not fare well relative to the RN design when it comes to the efficiency of all three types of regression coefficients. However, within-set regression coefficients have much higher ARE than between-set regression coefficients, around .50 or higher in most cases, although it drops sharply with higher correlations. For instance, for PM designs, this drop is from .61 to .43 as correlations increase from .2 to .8 (Figure 6d). Examining the behavior of the components of the efficiency ratio separately reveals that as correlation strength rises, complete data estimates of within-set regression coefficients become more efficient, and

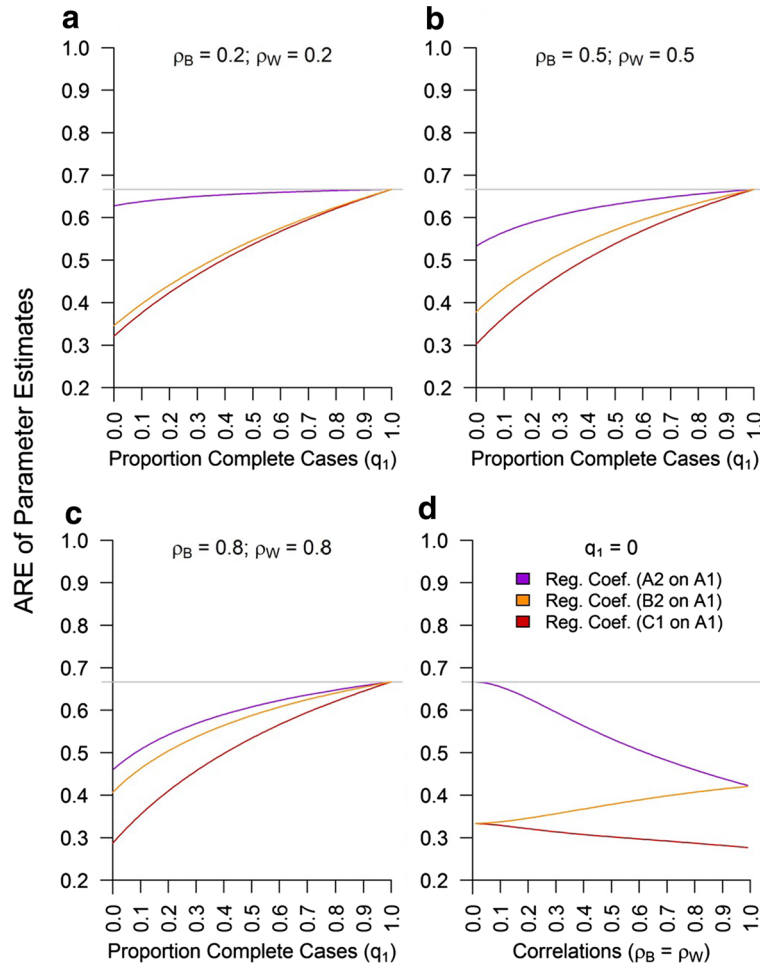


FIGURE 6.

Adjusted relative efficiency (ARE) of parameter estimates in Model 3. The Y axis displays ARE, which is the ratio of asymptotic variances of parameter estimates relative to the complete data design adjusted for the total number of data points (see Eq. 1).  $\rho_B$  is the value of all correlations among items between item sets, and  $\rho_W$  is the value of all correlations among items within the same item set. **a–c** the proportion of the sample which provides complete data (i.e.,  $q_1$ ) is varied along the X axis. **d**  $\rho_B$  and  $\rho_W$  are varied concurrently along the X axis. Each line shows the RE of a single parameter and is the same for all parameters of the same type.

simultaneously incomplete data estimates become less efficient. That is, the additional uncertainty in estimation that is due to missing data increases with increasing correlations.

The two types of between-set regression coefficients have very similar patterns of ARE to those observed for Model 2 (Figure 5), and are ordered relative to each other as one would expect: when the DV shares a set with one of the predictors (yellow curves), ARE is higher than when the DV shares a set with neither predictors (red curves). That is, the more the data overlap there is among the variables involved the regression (the more they come from the same set), the higher the efficiency. Examining the behavior of the numerator and denominator of the efficiency ratio separately reveals that both types of between-set regression coefficients decrease in absolute efficiency as correlations increase; when the DV shares a set with the other predictor (e.g., B2 on A1), this decrease is more rapid, which translates to increasing ARE. As with Model 2, the

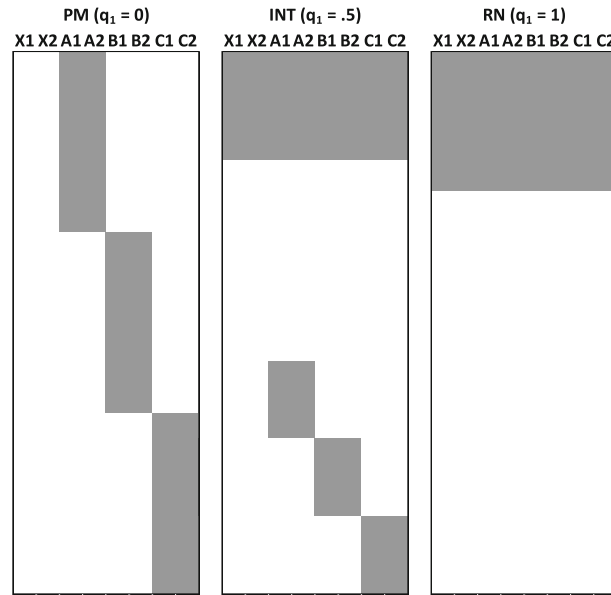


FIGURE 7.

Study 2 designs. *PM* planned missing, *INT* intermediate, *RN* reduced N. Variable sets are designated by letter (e.g., A1 and A2 are two variables in missing data set A, etc.). *Dark gray* represents missing data. PM designs have 0 % complete cases (all participants have 1/4 missing data). INT designs have between 1 and 99 % complete cases (i.e.,  $q_1 = .01$  to  $.99$ ; the design shown has  $q_1 = .5$ ) with the sample size reduced to keep the total number of observed data points constant. RN designs have 100 % complete cases, and a sample size reduced by 1/4.

independent effects of within- and between-set correlations on ARE are not shown as they are quite small.

The results from Models 2 and 3 suggest that the efficiency of regression coefficients, particularly between-set regression coefficients, suffers badly under the PM design, and is not rescued much by the INT designs unless the number of complete data cases is so high as to not be practical. However, when other reasons to use PM designs are compelling, the most efficient PM design is to place predictors and dependent variables in the same set.

#### 4. Study 2

In Study 2, we examined the performance of a 3-form PM design and corresponding INT and RN designs when an X set is present in addition to the three missing data sets (see Figure 7). Graham et al. (2006) argued that “the X set provides a kind of hedge against the possibility that a hypothesis critical to the research is actually tested with lower power than expected. Although situations may arise for which the X set may be safely excluded, ...including the X set is almost always a good idea.” (p. 326). Study 2 examines the extent to which ARE is improved by the inclusion of an X set. Model 4 was created by adapting Model 1 from Study 1; it is a saturated model that contains an X set with two fully observed variables (see Figure 8). Models 5a and 5b were created by adapting the regression Model 2 from Study 1: an X set with two fully observed variables was either the new set of independent variables, with the original six variables becoming dependent variables (Model 5a), or it was added to the set of dependent variables (Model 5b). These models allow us to investigate ARE when complete data are collected on key predictors or key dependent variables, as is frequently recommended (e.g., Graham et al., 2006). Population

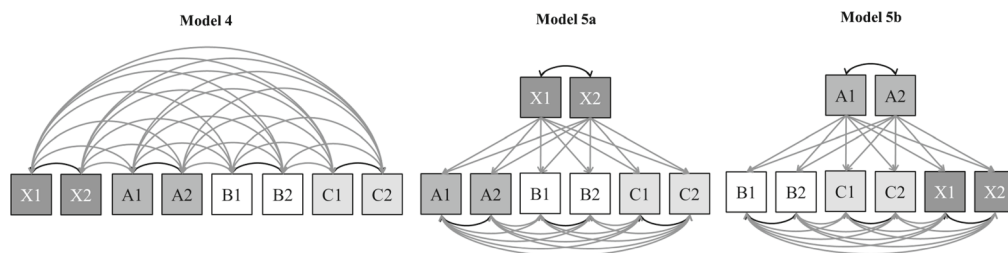


FIGURE 8.

Study 2, Models 4–5. Variable sets in the planned missing design are designated by background and letter. *Black paths* are “within-set” correlations and regressions; *gray paths* are “between-set”.

means and covariance matrices were identical to those used in Study 1, but extended in dimensions to include an additional set. As in Study 1, we considered all possible pairs of positive within- and between-set correlations that resulted in a positive definite covariance matrix. Note that with the X set added to the model, the ARE of the RN design increases from  $2/3$  to  $3/4$  (see Sect. 2.1).

#### 4.1. Results

*Model 4* Figure 9 displays the ARE ratios for all types of parameter estimates in the saturated model. Means, variances, and covariances corresponding to the fully observed X-set variables are fully efficient (because adjusted ratios are plotted, the efficiency ratios for these parameter estimates begin at 1 and end at  $3/4$  as  $q_1$  increases from 0 to 1). As this figure shows, adding an X set to Model 1 does not greatly affect the efficiency of PM model parameter estimates involving only variables in sets A–C. Variances, means, and covariances among variables in these sets (including both within-set and between-set covariances) have about 1 % greater ARE than they did in Study 1 across all conditions (see Figure 4). This miniscule increase in ARE actually translates into *poorer* performance relative to the RN design, which is now 75 % efficient relative to the complete data (CD) design with full  $N$ . In Study 1, within-set covariances in the PM design had ARE ratios that were at least as high as the ARE of the corresponding RN design (red curves in Figure 4); in Study 2, within-set covariances do not become more efficient under the PM design until the correlations among the variables reach .55 (red curves in Figure 9).

Covariances between an X variable and an ABC variable are estimated more efficiently than covariances among variables within an ABC set (dotted yellow lines in Figure 9). This efficiency gain is greatest when variable correlations are moderate to high ( $\rho = .4$  to  $.8$ ), when these X-to-ABC covariances are up to 10 % more efficient than within-ABC-set covariance estimates. The X-to-ABC covariances are estimated more efficiently with the PM design than with the RN design once the correlations among all variables reach .35.

This finding can be understood by considering the data that are present on the X-set variable but not the ABC-set variable. In a within-ABC-set covariance (e.g.,  $cov(A_1, A_2)$ ), both variables are missing the same  $1/3$  of cases, and the missing information cannot be “borrowed” from the other variable in the pair because it is also missing there. In contrast, in an X-to-ABC-set covariance (e.g.,  $cov(X_1, A_1)$ ), one variable in the pair is complete. When the correlation between variables is low, the extra information on  $X_1$  does not overlap with the missing information on  $A_1$ , so this information is not relevant to the estimation of the covariance. As the correlation strength increases, however, the information that is missing on  $A_1$  becomes increasingly redundant with the corresponding information on  $X_1$ , so the fraction of missing information decreases and the covariance is estimated with higher precision.



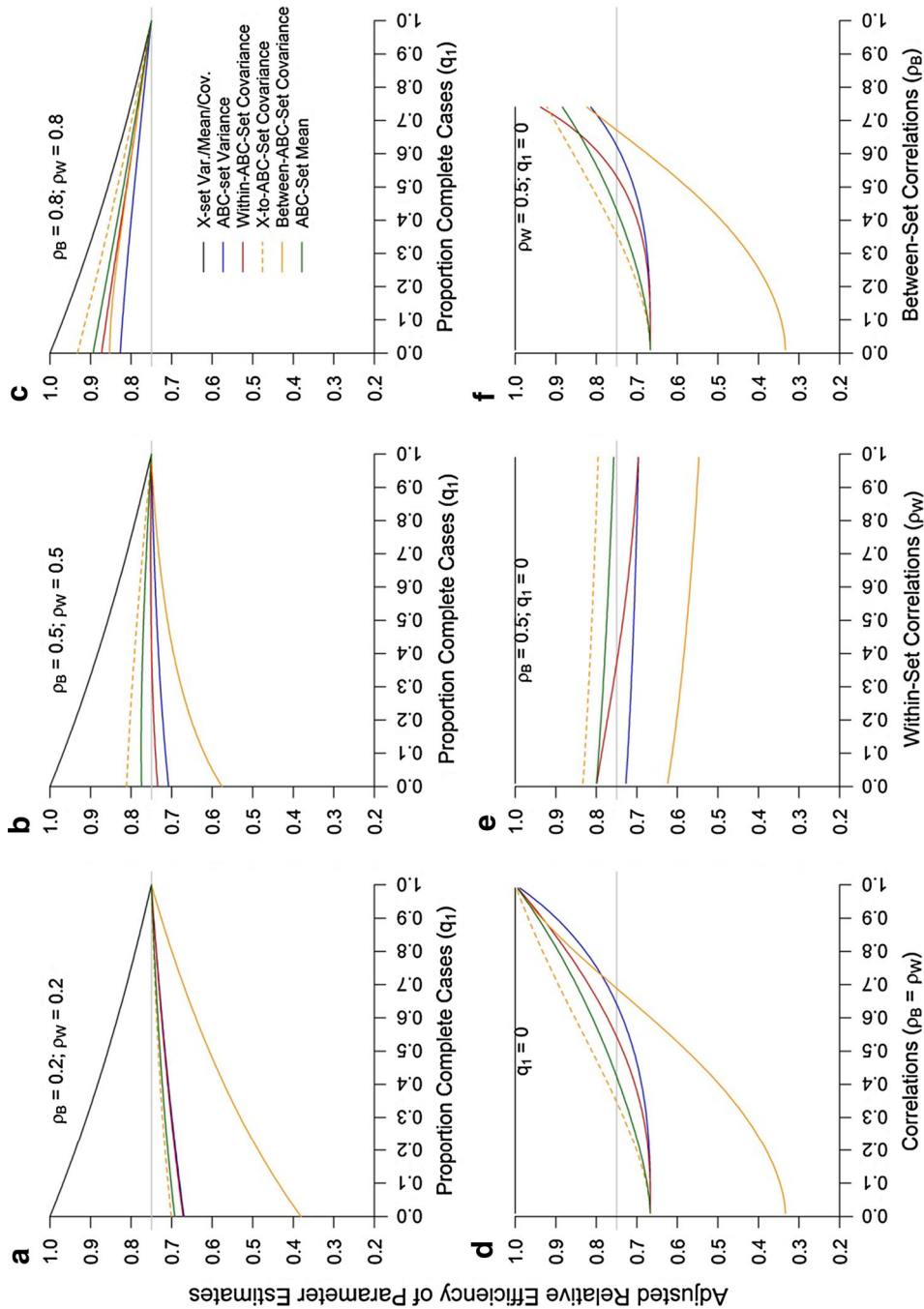


FIGURE 9.

Adjusted relative efficiency (ARE) of parameter estimates in Study 2, Model 4. The Y axis displays ARE, which is the ratio of asymptotic variances of parameter estimates relative to the complete data design adjusted for the total number of data points (see Eq. 1).  $\rho_B$  is the value of all correlations among items between item sets, and  $\rho_W$  is the value of all correlations among items within the same item set. **a-c** the proportion of the sample which provides complete data (i.e.,  $q_1$ ) is varied along the X axis; **d**  $\rho_B$ , and  $\rho_W$  are varied concurrently along the X axis; **e**  $\rho_W$  is varied along the X axis, while  $\rho_B$  is constant at .5; **f**  $\rho_B$  is constant at .5 (values of  $\rho_B$  above .74 result in a non-positive-definite matrix).

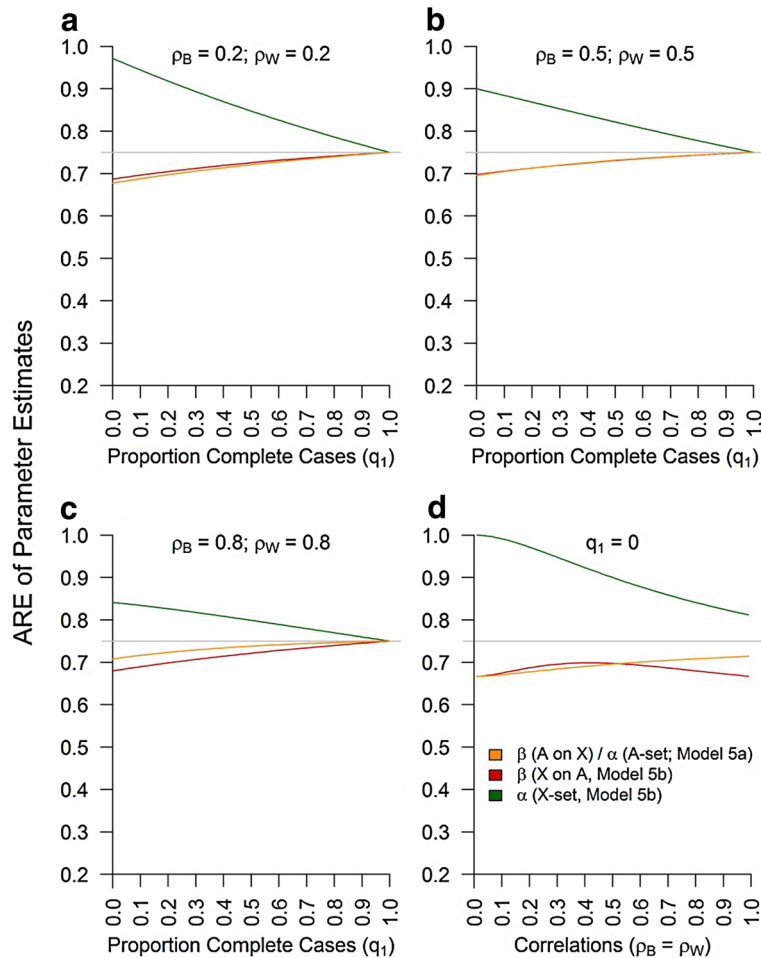


FIGURE 10.

Adjusted relative efficiency (ARE) of parameter estimates in Study 2, Models 5a and 5b. The Y axis displays ARE.  $\rho_B$  is the value of all correlations among items between item sets, and  $\rho_W$  is the value of all correlations among items within the same item set. **a–c** the proportion of the sample which provides complete data (i.e.,  $q_1$ ) is varied along the X axis. **d**  $q_1 = 0$  and  $\rho_B$  and  $\rho_W$  are varied concurrently along the X axis.

*Models 5a and 5b* Model 5a modified Model 2 by replacing the predictors with X variables; Model 5b added a pair of X variables to the set of dependent variables in Model 2. Figure 10 shows ARE of the four new parameters in Models 5a and 5b: regression coefficients of A–C on X (Model 5a), regression coefficients of X on A (Model 5b), intercepts of A–C variables (Model 5a), and intercepts of the X variables (Model 5b). Regression coefficients and intercepts in Model 5a had identical ARE in all conditions, so their results are plotted in one line (i.e., the orange line in Figure 10)<sup>2</sup>. The ARE of regression coefficients and intercepts involving only the variables in the A–C sets in Model 5b did not benefit from the inclusion of the X set and were very similar to the values obtained for Model 2; these results are not plotted.

<sup>2</sup> The finding that ARE of regression coefficients and intercepts falls in line with previous evidence that when exogenous variables have complete data, information loss may be characterized more simply. Savalei and Rhemtulla (2011) reported that when a single complete variable predicts a single dependent variable with 20 % missing data, both regression coefficients and intercepts have 20 % missing information, regardless of the correlation between the two variables. In Model 5A, we do see that ARE depends on correlation strength, but the effect is the same for both parameters.

Comparing Figures 10 and 5 reveals the sharp efficiency gain that results when either the dependent variables or the predictor variables are in the X set (i.e., have complete data). Regression coefficients display increases in ARE of greater than 30 %, and they are only slightly less efficiently estimated in the PM design than the RN design, hovering around 70 % ARE. Intercepts of the X-set variables (Model 5b) are now always more efficiently estimated in the PM design than the RN design. Intercepts of ABC variables when the predictors are in the X set (Model 5a) are less efficient in the PM design than the RN design, as they were in Model 2, but these intercepts increase in ARE as correlations increase, instead of decreasing as they did in Model 2.

Taken together, the results of Study 2 suggest that including an X set in a PM design can greatly improve the ARE of parameter estimates that involve the X variables, but does not have much effect on those estimates that do not directly involve the X variables. In addition, regression coefficient estimates that include X variables as either dependent variables or predictor variables, though still slightly less efficient in PM than RN designs, have high enough ARE that using a PM design should not seriously compromise the power to test hypotheses about these coefficients.

## 5. Study 3

A strong impetus for using PM designs comes from surveys that are so long that participants are overburdened: for example, when many constructs are assessed and each is measured by a long scale. In this case, the model of interest is often a confirmatory factor analysis (CFA) model, where each scale is assumed to measure a single underlying construct, and the constructs are allowed to be correlated, or a latent regression model, where some latent constructs are posited to cause other latent constructs. In Study 3, we examined the asymptotic ARE of parameter estimates within CFA models and latent regression models. Efficiency was studied for both measurement model parameters (e.g., factor loadings) and structural model parameters (e.g., factor covariances and latent regression coefficients).

Study 3 considered four models, as shown in Figure 11. Models 6 and 6X were CFA models, whereas Models 7 and 7X were latent regression models. Models 6 and 7 had 18 observed variables and 6 latent constructs; one indicator of each construct was included in each of sets A–C. Models 6X and 7X additionally included an X set, for a total of 24 observed variables. The X-set variables were distributed so that each factor had an additional indicator. All factor loadings were .7, and residual variances were .51. Variances of the latent variables were 1. All 15 correlation values among the 6 latent variables were equal to each other and were varied from .01 to .99 in increments of .01.

For Models 6 and 6X, two different identification methods were used, and their differences were studied. Latent factors were identified by either fixing the first loading on each factor to .7 (its true value; for models with an X set, the marker variable was the one in the X set) or by fixing factor variances to 1. For Models 7 and 7X, only the first identification method was used, because of the presence of latent dependent variables, whose variance is a function of other model parameters and cannot be fixed directly.

### 5.1. Results

*Model 6* The top panel of Figure 12 shows the ARE ratios for the parameter estimates for both types of model identification. Residual variances are not shown in the figure but their AREs are very similar to those of factor loadings. Under either method of identification, loadings estimates are less efficient under PM and INT designs as compared to the RN design. This is not surprising given that each factor's loadings are estimated primarily from covariances among its indicators, especially when factor correlations are low to moderate, but indicators of each factor come from

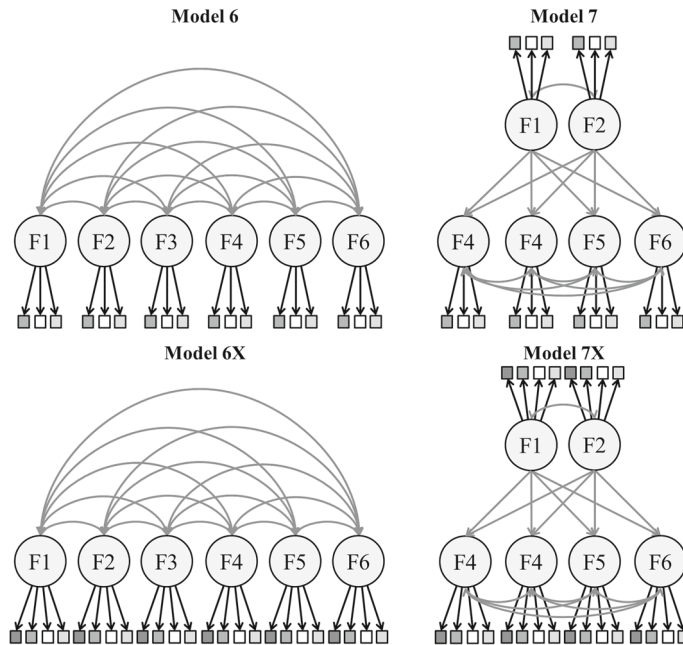


FIGURE 11.

Study 3 Models. Each latent factor has an indicator from each missing data item set. Models 6X and 7X include the X set.

different sets. Such between-set covariances are estimated with very low ARE (see Figure 4). Figure 12d shows that when factor correlations are zero, the ARE of factor loadings drops below .30 under a PM design. As factor correlations increase, factor loading ARE increases substantially, although the RN design is still more efficient. This rapid increase can be explained by the fact that indicators of different factors become available to estimate the loadings associated with each factor, and those indicators share sets with each factor's own indicators. Finally, loadings under the two methods of identification display similar patterns (blue and purple curves in Figure 12d), but identifying the factor by fixing the variance to 1 produces loadings estimates with higher ARE; this difference is around 10 % for PM designs.

Even though factor variances are rarely of interest, we note that under the marker variable method of identification (when these parameters are estimable), the ARE of factor variances is identical to that of loadings under the fixed-variance method of identification, and thus the discussion above applies to factor variances as well. Unlike factor loadings, factor covariances are always more efficient in the PM and INT designs than the RN design (Figure 12a–c, yellow curves). ARE of factor covariances follow a curvilinear pattern (Figure 12d), such that for a PM design factor, covariance estimates are most efficient when factor correlations are very low or very high, while the other two types of parameters have the highest efficiency when factor correlations are high. Factor correlations behave quite similarly to factor covariances when the factor correlations are low to moderate (Figure 12a–b, red curves), but are less efficiently estimated when the factor correlations are high (Figure 12c). Examining this pattern more continuously for PM designs (Figure 12d) reveals that factor correlation estimates are much more dependent on population values of factor correlations than are factor covariance estimates: when factors are nearly uncorrelated, factor correlations are estimated with almost .80 ARE, but as factor correlations increase above about .5, ARE decreases rapidly.

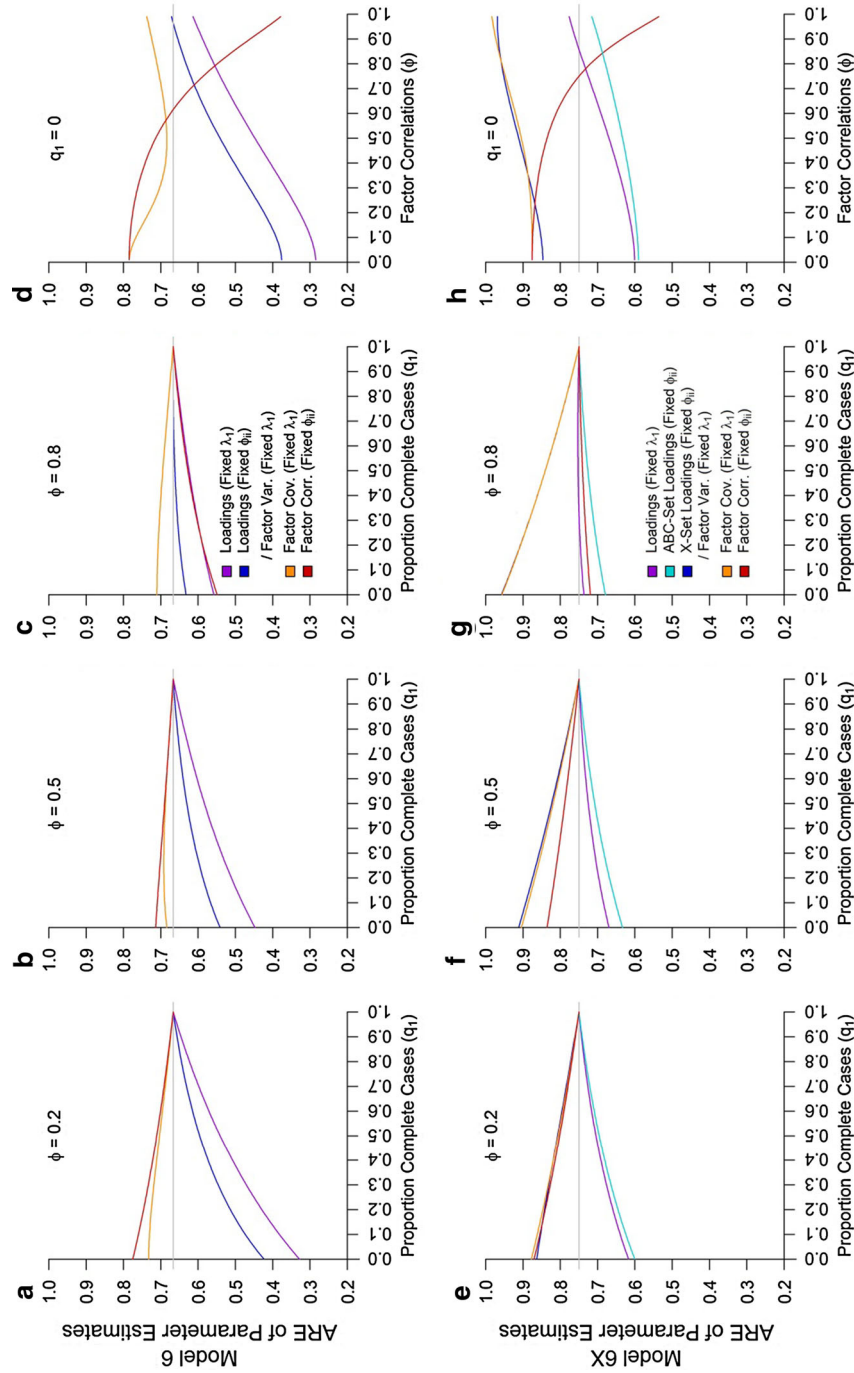


FIGURE 12.

Adjusted relative efficiency (ARE) of parameter estimates in Study 3, Model 6 (*upper panel*) and 6X (*lower panel*). The Y-axis displays ARE.  $\phi$  is the value of all factor correlations. **a–c, e–g** the proportion of the sample which provides complete data (i.e.,  $q_1$ ) is varied along the X axis. **d, h**  $q_1 = 0$  and factor correlation value is varied along the X axis. Factor loading values are all .7, and residual variances are all .51. *Fixed*  $\lambda_1$  the model is identified by fixing the first loading of each factor to .7 (its population value); *Fixed*  $\phi_{11}$  the model is identified by fixing the factor variances to 1.

*Model 6X*. The bottom panel of Figure 12 shows ARE ratios for Model 6X. It is clear that including an additional X-set indicator per factor substantially improves efficiency of PM design estimates<sup>3</sup>. In particular, factor variances and covariances have ARE around .90 under the PM design, much higher than the RN design (which is now 3/4). The ARE of X-set factor loadings under the fixed-factor-variance method of identification is identical to that of factor variances under the marker variable method of identification (the blue lines in the bottom panel of Figure 12), and thus also have very high efficiency. ARE of factor correlations under the fixed-factor-variance method of identification is consistently about .10 higher than in Model 6. Factor loadings continue to be less efficient under a PM design than the RN design, but their ARE is between .15 and .30 higher than in Model 6, and never drops below .60. Identification method appears to matter very little for the ABC-set loadings (i.e., the turquoise and purple lines in the lower panel of Figure 12 are very similar). Together, these results suggest that it is possible to achieve highly efficient latent variable model estimates under a PM design when an X set is included, and these indicators are distributed across all factors.

*Model 7* Model 7 reparameterizes the structural part of Model 6 from the saturated model to a latent regression model. The new parameters are latent regression coefficients, variances of latent disturbances, and covariances among latent disturbances. The top panel of Figure 13 shows ARE ratios for these parameters. Latent regression coefficients, which are arguably the most important structural parameters, are estimated more efficiently in the PM design than in the RN design when factor correlations are lower than .5, but ARE drops to around .55 under the PM design when factor correlations are very high (.8). Covariances among disturbances follow the same pattern as latent regression coefficients. In contrast, disturbance variances do not do well under PM designs, and their ARE ranges from just under .40 to around .55.

*Model 7X* The bottom panel of Figure 13 shows ARE ratios for Model 7X. As in Model 6X, adding an X-set indicator to each latent factor substantially improves efficiency. ARE of all three parameter types is now higher under the PM design than the RN design as long as factor correlations are below .75. When factor correlations are lower than .5, ARE of all three parameter types is around .85. These findings reinforce the suggestion that PM designs including an X set can be cost-effective, per observation, if the model of interest is a latent variable model, and the parameters of interest are structural parameters.

## 6. Study 4

In Study 3's PM design, indicators of each latent factor were split across sets. An alternative PM design strategy for latent variable models would be to assign all indicators of a factor to the same set. Recommendations in the literature on this topic are mixed: Graham et al. (1996) recommended splitting items across sets on the basis of efficiency differences in variance and covariance estimates resulting from the two strategies. Reversing this recommendation, however, Graham et al. (2006) noted that splitting items across sets is impractical because researchers cannot use available case analysis (i.e., pairwise deletion) on the resulting scale data (i.e., scale

<sup>3</sup> The addition of an X-set indicator to each factor also changes the model simply by introducing a fourth indicator to each factor. To investigate whether the observed benefits of the X set are due to having four indicators rather than 3, we also ran a version of Model 6 that used a 4-form design (sets A–D) where 25 % of the sample was missing each item set. Compared to this PM model, the PM model 6X had factor loadings with .05 higher RE, factor variances with .20 higher RE, and factor covariances that had between .02 and .15 higher RE (increasing with increasing factor correlations). Thus, though the 4-form PM design performed better than the 3-form PM design with no X set, the 3-form design with X set outperformed both of these.

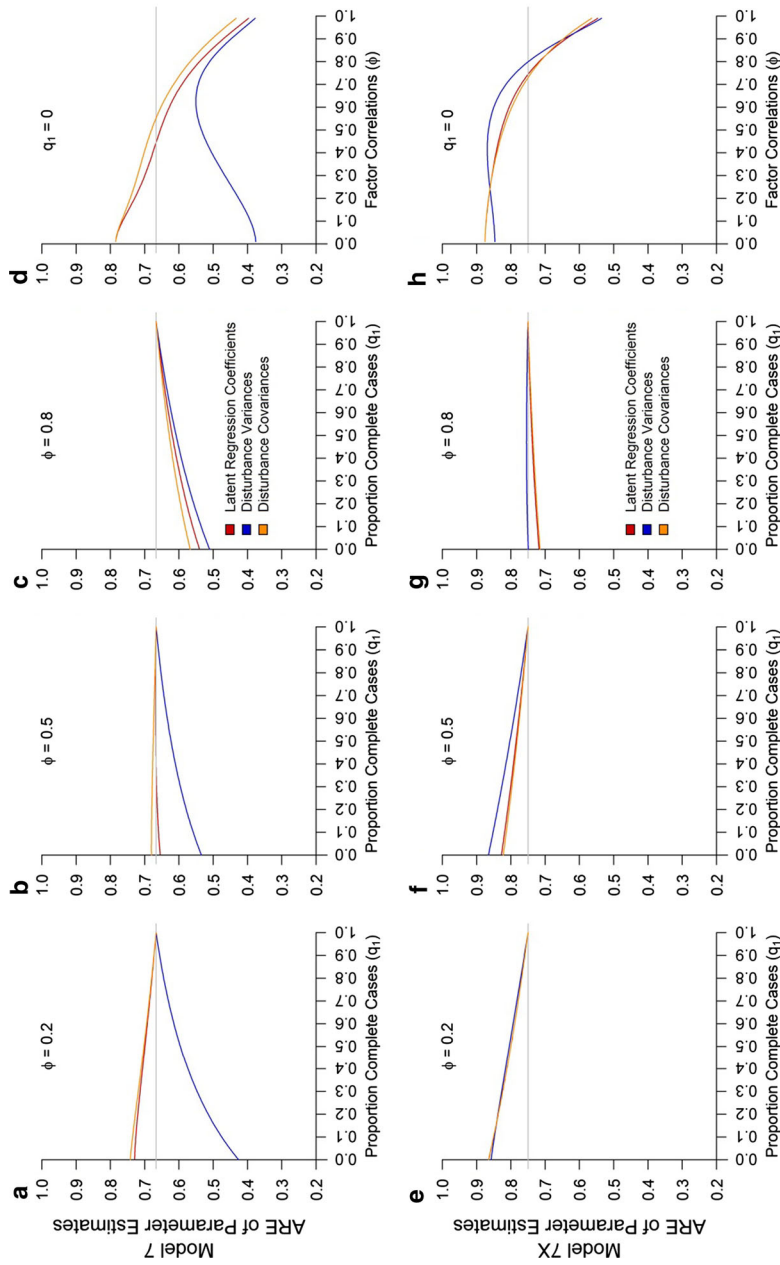


FIGURE 13.

Adjusted relative efficiency (ARE) of parameter estimates in Study 3, Model 7 (*upper panel*) and 7X (*lower panel*). The Y axis displays adjusted relative efficiency.  $\phi$  is the value of all factor correlations. In graphs **a–c** and **e–g**, the proportion of the sample which provides complete data (i.e.,  $q_1$ ) is varied along the X-axis. In graphs **d**, **h**,  $q_1 = 0$  and factor correlation value is varied along the X axis. Factor loadings values are all .7, and residual variances are all .51. The model is identified by fixing the first loading of each factor to .7 (its population value). The relative efficiency of factor loadings is shown in Figure 12.

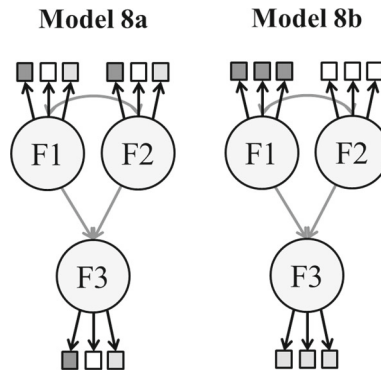


FIGURE 14.

Study 4 Models. In Model 8a, each latent factor has one indicator from each missing data item set. In Model 8b, each latent factor has three indicators from the same missing data item set.

scores cannot be formed by summing across items when different items have different missingness patterns). Thus, they recommended assigning all scale items to the same set, “even if it means somewhat less efficient estimation” (p. 327). In latent variable estimation, the logistical hassle is not an issue because FIML can be used to deal with missingness at the item-level (where scales are long and fewer indicators are desired, items can be parceled to create a single indicator per set).

In Study 4, we explicitly contrast the two approaches to explore the extent to which efficiency is affected by these strategies. If the primary goal of the research is to evaluate the factorial structure of an instrument (i.e., to evaluate the measurement model), then the “indicators within sets” approach may be able to provide more information to evaluate the strength of the relationship between each indicator and its factor. Study 4 compared the ARE of measurement model parameters (factor loadings and residual variances) and structural model parameters (latent variances and covariances, latent regression coefficients, and latent disturbances) under these two PM designs.

Models 8a and 8b (Figure 14) have two latent predictors and one latent dependent variable, and each latent variable has three indicators. In Model 8a (“indicators across sets”), these indicators come from each of the three sets. In Model 8b (“indicators within sets”), all three indicators of a factor come from a single set. The X set was not used. For simplicity, INT designs were also not studied.

### 6.1. Results

*Model 8* Figure 15 shows ARE under the PM design for 5 parameter types. Residual variances are not pictured, because their REs overlapped almost entirely with those of factor loadings. Figure 15 makes clear that the two PM design strategies lead to dramatically different results. Splitting indicators across sets leads to high ARE of factor covariances (across all levels of the factor correlation) and latent regression coefficients (particularly when factor correlations are less than .5). In contrast, keeping indicators together within sets leads to high ARE of factor loadings and factor variances. Thus, the choice of indicator assignment strategy should depend strongly on the research goals: if the goal is to study the measurement model, then keeping indicators together within a set will allow for the most precise estimation of the measurement model parameters. If the goal is to study relations between constructs, then splitting indicators across item sets will lead to the most precise estimation of structural model parameters. An exception to this pattern is the behavior of the variances of the latent disturbance terms, which, while poorly estimated by both PM designs, appear to fare better under the “indicators within sets” approach, at least for small factor correlation values.



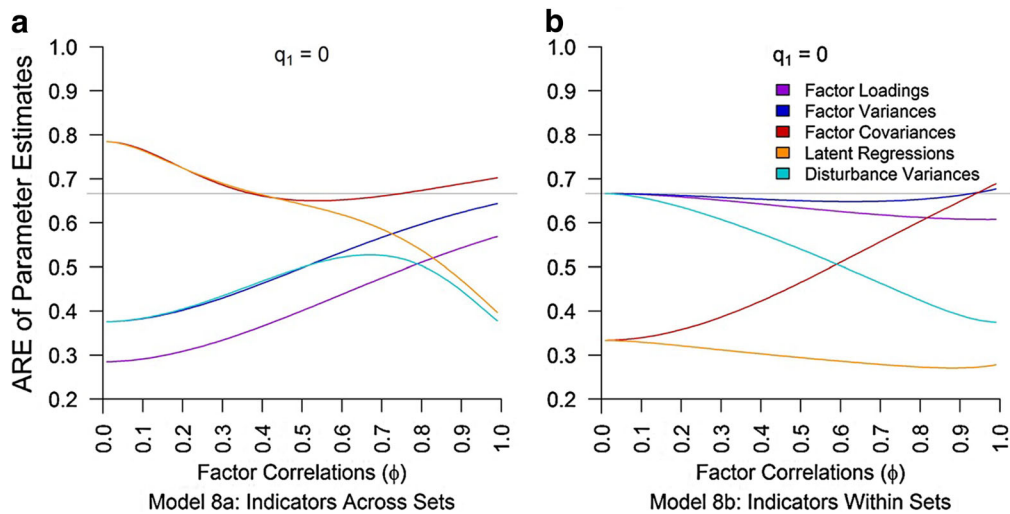


FIGURE 15.

Adjusted relative efficiency of parameter estimates in Study 4, Models 8a, and 8b. The Y axis displays adjusted relative efficiency.  $\phi$  is the value of all factor correlations, varied along the X axis. Factor loading values are all .7, and residual variances are all .51. The model is identified by fixing the first loading of each factor to .7 (its population value). The gray line at  $2/3$  reflects all parameters' ARE in the RN design.

## 7. Discussion

In four studies, we examined the effect of planned missing data designs (PM, RN, and a full range of INT designs) on the adjusted asymptotic relative efficiency of estimates of different types of parameters, under a variety of models and the full range of item correlations. Below, we summarize the results and offer a discussion of key findings.

### 7.1. Parameter Type

Some types of parameters are more efficiently estimated under PM designs than RN designs. In particular, parameters involving only exogenous observed variables—means, variances, within-set covariances, and even between-set covariances when variables are highly correlated—are more efficiently estimated under PM and INT designs than the RN design. In contrast, regression coefficients among observed variables are rarely more efficient under a PM design, and worse, they frequently display unacceptably low ARE. Study 2 found that between-set regression coefficients are frequently half as efficient in the PM design as in the RN design, suggesting that to have the same power as the RN design, sample size (and cost) would need to be doubled. Using an INT design, a modification of the PM design that administers all variables to a fraction of the sample ameliorates the situation by bringing ARE for regression coefficients closer to what they would be under an RN design; however, including as much as 30 % complete data might improve ARE just 15 %. Even if other good reasons to use a PM design exist, this dramatic drop in efficiency should make researchers pause before choosing to use a PM design if the model of interest is a regression or a path analysis model.

### 7.2. Assignment to Sets

It is possible to manipulate the assignment of variables to sets to improve ARE of regression coefficients: ARE improves substantially when predictors and dependent variables are included in

the same set. In addition, Study 3 shows that when either the dependent variables or the independent variables come from the X set, efficiency is improved even more dramatically. Studies 3 and 4 reveal that PM and INT designs may optimize the efficiency of certain parameters if the model of interest is a latent variable model. When indicators are divided across sets, factor covariances and latent regression coefficients are estimated more efficiently in the PM design than the RN design across most values of factor correlations. In contrast, measurement model parameters tend to be less efficiently estimated under the “indicators across sets” PM design than the RN design. When all indicators of a factor are assigned to the same set, the opposite pattern is found: measurement model parameters are estimated about as efficiently as in the RN design but factor covariances and latent regression coefficients are estimated less efficiently.

Including an X set also improves efficiency of many parameters relative to a PM or INT design without the X set. In particular, all regression paths connecting two variables, where one is in the X set, have acceptably high ARE relative to the RN design (i.e., ARE of these parameters are typically only about 5 % lower in the PM design than RN design). In latent variable models, including an X set brings ARE of all parameters to a minimum of 60 % (vs 30 % when the X set is not included). Covariances and regression paths between pairs of variables in different sets, where neither is in the X set, are not noticeably improved by the addition of an X set.

These results suggest that if the goal is to study the relations among constructs, then the best strategy is to measure constructs using scales (rather than single items), to split scale items across sets, to include an X set, and to use a latent variable model to capture the relations among the latent constructs. Measurement parameters will have lower efficiency than they would in an RN design, but this is less important unless the goal is to study the measurement properties of the items themselves (for example, if the scales have not been previously validated).

### 7.3. Correlation Strength

In a saturated model, the ARE of estimates of variances, covariances, and means increases with increasing correlations for PM designs. This increase in ARE can be as high as .45 for some parameters, such as between-set covariances. This effect is wholly due to increasing between-set correlations: as within-set correlations increase independently of between-set correlations, ARE of all these parameters actually decreases slightly. Therefore, the best conditions for estimating exogenous variables’ parameters occur when between-set correlations are high and within-set correlations are low. This situation can be achieved by placing variables that are likely to be highly correlated in different sets.

When the model is a regression model, ARE ratios of the corresponding parameters are less strongly affected by between- and within-set correlations. The efficiency of between-set regression coefficients does not change much when all correlations increase together. The efficiency of within-set regression coefficients, in contrast, decreases by more than 25 % as all correlations increase. This is an interesting finding in light of a widely reported assumption that stronger correlations lead to a lower rate of missing information in regression coefficients (e.g., Enders 2010; Graham et al., 1996).

In latent variable models, ARE of all parameters are highly affected by the strength of latent correlations. Parameters that are estimated inefficiently when correlations are low (e.g., factor loadings, factor variances, residual variances in Models 6, 7, and 8a) become more efficient with rising latent correlations. In contrast, latent regression coefficients are negatively affected by increasing factor correlations: they have the highest ARE when factors are uncorrelated, but it drops below that of the RN design when latent correlations reach about .5 (with no X-set; Model 7), or .8 (with X set; Model 7X). It is perhaps uncommon to have a latent variable model with factor correlations higher than .5, and thus latent covariances, correlations, and regressions will typically be more efficiently estimated with a PM design than an RN design.

#### 7.4. Cost Efficiency

One of the primary motivations to use a planned missing design is cost savings. Researchers may thus be interested in knowing to what extent it is possible to create a planned missingness design to reduce the price of data collection while holding power constant. There are two components to such a calculation: relative efficiency and relative cost. In this paper, we have been concerned with measuring the relative efficiency component. However, if relative cost of data collection is known, then it is possible to use relative efficiency to compute efficiency per research dollar.

By comparing PM, RN, and INT designs on a per-data-point basis, we implicitly assume a cost model in which the cost of implementing a design is a linear function of the number of data points collected. This is a plausible way to estimate the cost savings of a PM design relative to collecting complete data, but it will not always apply. If the real costs of data collection fit this model, then ARE can be used to assess cost savings directly. Alternatively, the cost of gathering data from one participant may be the same regardless of how many items she completes (in this case, the rationale for using a PM design would be based entirely on fatigue effects, rather than cost), or the cost of collecting complete data may be double that of collecting PM data if it requires a second testing occasion to complete the additional items.

To adjust ARE for a different ratio of cost efficiency, we can compute  $ARE^* = \left(\frac{2}{3}\right) \frac{ARE}{c(q_1+(1-q_1)RC)}$  (Models 1–3, 6, 7, 8a, 8b) or  $ARE^* = \left(\frac{3}{4}\right) \frac{ARE}{c(q_1+(1-q_1)RC)}$  (Models 4–5, 6x, 7x), where RC is the relative cost of collecting PM data on one participant to that of collecting complete data on one participant, and  $c$  and  $q_1$  are defined in Sect. 2.1.

For example, consider a design in which collecting complete data is twice as expensive, per participant, as collecting PM data, i.e.,  $RC = 1/2$ . For a regression coefficient in Model 5a with  $ARE = .70$  in the PM design ( $q_1 = 0, c = 1$ ) and  $ARE = .75$  in the RN design ( $q_1 = 1, c = 3/4$ ), we would get  $ARE^* = (3/2) ARE = 1.05$  for the PM design and  $ARE^* = ARE = .75$  for the RN design. That is, given a 2:1 cost ratio, the PM design would result in more efficient estimation of the regression parameter *per research dollar* than the RN design.

#### 7.5. Conclusions and Recommendations

Many practical reasons exist for using planned missing designs in psychological research, most of which have to do with the increased validity of measures collected from participants who are not overburdened. The goal of the present research was to examine whether PM designs may also be preferred for theoretical (statistical) reasons, as was implied by Graham et al. (2001). The results are mixed, and strongly dependent on the type of model estimated, the strength of the correlations among the variables, and the type of the PM design. Some of the practical reasons for preferring PM designs (e.g., minimizing fatigue) may be compelling enough to lead researchers to use a PM design even when efficiency compared to an RN design is low. Additionally, while some parameters suffer a large efficiency loss under PM designs, there are conditions that can minimize these losses. We encourage researchers to consider these issues carefully when creating a PM design for data collection. Our results also support the common recommendation that including an X-set is essential as this prevents ARE from dropping too low.

Designs that blended the PM and the RN designs by including some proportion of complete cases (INT designs) almost always displayed ARE somewhere between the PM and RN designs, and this increase in ARE was typically steepest for the lower proportion of complete cases. When collecting complete data is not expected to be overly burdensome, the most prudent design may be one of these INT designs: by including 50 % complete cases, estimates are protected from suffering too much efficiency loss. A potential benefit of this strategy is that it allows researchers to compare the results from those participants who completed a PM form to those who completed all items (e.g., Harel et al., 2011 did this and found interesting effect size differences among the

groups). Such an analysis can quantify the gains in validity or reliability that come from using a PM design, and may be very useful in informing future research. On the other hand, if participant burden is high in the complete case design, then having 50 % of the sample provides complete data, which essentially guarantees lower-quality data in this half of the sample.

Finally, our research question has been whether, and under what conditions, PM designs have a theoretical advantage over RN designs. This focus has naturally created limitations. First, our study has been that of relative efficiency, as captured by the increase in the size of the asymptotic variance due to the introduction of incomplete data. When testing if a parameter is different from zero, however, relative efficiency only implies a relative drop in power, not low power in absolute terms. Absolute power varies with effect size and sample size, and PM designs may have adequate power to test parameters that have exhibited low relative efficiency in our studies, depending on the true value of the parameter and the sample size. Figure 1 allows the reader to translate ARE into power. Second, our results are asymptotic. In real data, sampling variability may exert effects that cannot be seen here.

### Acknowledgments

We thank Kristopher Preacher and Alex Schoemann for comments on an earlier draft of this manuscript. This research was supported in part by a Banting Postdoctoral Fellowship from the Social Sciences and Humanities Research Council of Canada (SSHRC) to Mijke Rhemtulla, an SSHRC Grant to Victoria Savalei, and NSF Grant NSF0066969 to Todd Little (W. Wu, co-PI).

### Appendix A: Translating Relative Efficiency into Change in Power

Let  $\theta$  be the parameter of interest. We want to test the null  $H_0 : \theta \leq 0$  against the alternative  $H_1 : \theta > 0$ . Let  $\hat{\theta}_{CD}$  be the complete data estimator based on a given  $n$ . We assume that the sampling distribution of this estimator is approximately  $N\left(\theta, \frac{1}{n}SE_{\theta,CD}^2\right)$ . Similarly, let  $\hat{\theta}_{PM}$  be the incomplete data estimator (from a PM design) based on a given  $n$ , with the approximate sampling distribution  $N\left(\theta, \frac{1}{n}SE_{\theta,PM}^2\right)$ . Without loss of generality, and for simplicity, we assume  $n = 1$  below (none of the relevant equations depend on  $n$ ). Additionally, we assume for simplicity that standard errors are known, to avoid complicating the expressions with the “hat” notation. The expressions are then approximate and become exact asymptotically. Finally, the most serious but necessary simplifying assumption is that the variance of the estimator does not depend on the population value; i.e.,  $SE_{\theta,CD} = SE_{CD}$  and  $SE_{\theta,PM} = SE_{PM}$  for all values of  $\theta$ .

The Wald test statistic is given by  $z_{CD} = \frac{\hat{\theta}_{CD}}{SE_{CD}}$ , and its critical value for a one-tailed test at a given significance level  $\alpha$  is  $z_{CRIT} = \Phi^{-1}(1 - \alpha)$ . Suppose the true value is  $\theta_0$ . Then, power is given by

$$\begin{aligned} \pi_{\theta_0,CD} &= 1 - \Pr(z_{CD} \leq z_{CRIT} | \theta = \theta_0) = 1 - \Pr\left(\frac{\hat{\theta}_{CD}}{SE_{CD}} \leq z_{CRIT} | \theta = \theta_0\right) \\ &= 1 - \Pr\left(\frac{\hat{\theta}_{CD} - \theta_0}{SE_{CD}} \leq z_{CRIT} - \frac{\theta_0}{SE_{CD}} | \theta = \theta_0\right) = 1 - \Phi\left(z_{CRIT} - \frac{\theta_0}{SE_{CD}}\right). \end{aligned}$$

It follows that

$$1 - \pi_{\theta_0,CD} = \Phi\left(z_{CRIT} - \frac{\theta_0}{SE_{CD}}\right)$$

$$\Phi^{-1}(1 - \pi_{\theta_0, CD}) = z_{\text{CRIT}} - \frac{\theta_0}{\text{SE}_{CD}} = \Phi^{-1}(1 - \alpha) - \frac{\theta_0}{\text{SE}_{CD}}$$

$$\text{SE}_{CD} = \frac{\theta_0}{\Phi^{-1}(1 - \alpha) - \Phi^{-1}(1 - \pi_{\theta_0, CD})}. \quad (2)$$

Similarly, for incomplete data,

$$\text{SE}_{PM} = \frac{\theta_0}{\Phi^{-1}(1 - \alpha) - \Phi^{-1}(1 - \pi_{\theta_0, PM})}. \quad (3)$$

The square-root of the (adjusted) relative efficiency (RE or ARE; recall that these are identical for PM designs) is the ratio of (2) and (3), and plugging these expressions in, we obtain

$$\sqrt{\text{ARE}_{\theta}} = \frac{\text{SE}_{CD}}{\text{SE}_{PM}} = \frac{\Phi^{-1}(1 - \alpha) - \Phi^{-1}(1 - \pi_{\theta_0, PM})}{\Phi^{-1}(1 - \alpha) - \Phi^{-1}(1 - \pi_{\theta_0, CD})}.$$

After some algebraic manipulation, we obtain the final expression for the power of the incomplete data design as a function of relative efficiency and the power of the complete data design:

$$\left[ \Phi^{-1}(1 - \alpha) - \Phi^{-1}(1 - \pi_{\theta_0, CD}) \right] \sqrt{\text{ARE}_{\theta}} = \Phi^{-1}(1 - \alpha) - \Phi^{-1}(1 - \pi_{\theta_0, PM})$$

$$\Phi^{-1}(1 - \alpha) - \left[ \Phi^{-1}(1 - \alpha) - \Phi^{-1}(1 - \pi_{\theta_0, CD}) \right] \sqrt{\text{ARE}_{\theta}} = \Phi^{-1}(1 - \pi_{\theta_0, PM})$$

$$\pi_{\theta_0, PM} = 1 - \Phi \left\{ \Phi^{-1}(1 - \alpha) - \left[ \Phi^{-1}(1 - \alpha) - \Phi^{-1}(1 - \pi_{\theta_0, CD}) \right] \sqrt{\text{ARE}_{\theta}} \right\}.$$

Or equivalently,

$$\pi_{\theta_0, PM} = 1 - \Phi \left\{ \Phi^{-1}(1 - \alpha)(1 - \sqrt{\text{ARE}_{\theta}}) + \Phi^{-1}(1 - \pi_{\theta_0, CD}) \sqrt{\text{ARE}_{\theta}} \right\}.$$

#### References

- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. M. R. E. Schumacker (Ed.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah: Lawrence Erlbaum Associates Inc.
- Arminger, G., & Sobel, M. E. (1990). Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *Journal of the American Statistical Association*, *85*, 195–203.
- Bentler, P. M. (2007). *EQS 6 structural equations program manual*. Encino: Multivariate Software.
- Bentler, P. M., & Lee, S.-Y. (1978). Matrix derivatives with chain rule and rules for simple, Hadamard, and Kronecker products. *Journal of Mathematical Psychology*, *17*, 255–262.
- Bunting, B. P., Adamson, G., & Mulhall, P. K. (2002). A Monte Carlo Examination of an MTMM Model with planned incomplete data structures. *Structural Equation Modeling*, *9*, 369–389. doi:10.1207/S15328007SEM0903\_4.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
- Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In L. M. Collins & L. Seitz (Eds.), *Advances in data analysis for prevention intervention research: National Institute on Drug Abuse Research monograph series No. 142* (pp. 13–63). Washington, DC: National Institute on Drug Abuse.
- 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*, 197–218. doi:10.1207/s15327906mbr3102\_3.
- Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing data designs in the analysis of change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). Washington, DC: American Psychological Association. doi:10.1037/10409-011.
- Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323–343. doi:10.1037/1082-989X.11.4.323.

- Harel, O., Stratton, J., Aseltine, R. (2011). *Designed missingness to better estimate efficacy of behavioral studies*. Technical Report (11–15), The Department of Statistics, University of Connecticut.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.
- Magnus, J. R., & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics*. New York: Wiley.
- McArdle, J. J., & Woodcock, R. W. (1997). Expanding test-retest designs to include developmental time-lag components. *Psychological Methods*, 2, 403–435. doi:10.1037/1082-989X.2.4.403.
- Mistler, S. A., & Enders, C. K. (2012). Planned missing data designs for developmental research. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of Developmental Research Methods* (pp. 742–754). New York: Guilford Press.
- Mooijaart, A., & Bentler, P. M. (1991). Robustness of normal theory statistics in structural equation models. *Statistica Neerlandica*, 45, 159–170.
- Nel, D. G. (1980). On matrix differentiation in statistics. *South African Statistical Journal*, 14, 137–193.
- Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. *Paper presented at the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California.
- R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>. Accessed 1 Jan 2014.
- Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54–63. doi:10.1080/01621459.1995.10476488.
- Raykov, T., Marcoulides, G. A., & Patelis, T. (2013). Saturated versus identified models: A note on their distinction. *Educational and Psychological Measurement*, 73, 162–168.
- Revilla, M., & Saris, W. E. (2013). The split-ballot multitrait-multimethod approach: Implementation and problems. *Structural Equation Modeling*, 20, 27–46. doi:10.1080/10705511.2013.742379.
- Saris, W. E., Satorra, A., & Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design. *Sociological Methodology*, 34, 311–347. doi:10.1111/j.0081-1750.2004.00155.x.
- Savalei, V. (2010). Expected vs. observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods*, 15, 352–367. doi:10.1037/a0020143.
- Savalei, V., & Rhemtulla, M. (2011). Properties of local and global measures of fraction of missing information: Some explorations. In *Talk presented at the 76th Annual and the 17th International Meeting of the Psychometric Society*, Hong Kong.
- Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling*, 19, 477–494.
- Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge: Ballinger Publishing Company.
- Sirotnik, K., & Wellington, R. (1977). Incidence sampling: An integrated theory for “matrix sampling”. *Journal of Educational Measurement*, 14, 343–399. doi:10.1111/j.1745-3984.1977.tb00050.x.
- Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the National Health and Nutrition Examination Survey. *Survey Methodology*, 32, 217.
- Wacholder, S., Carroll, R. J., Pee, D., & Gail, M. H. (1994). The partial questionnaire design for case-control studies. *Statistics in Medicine*, 13, 623–634.
- Yuan, K., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, 30, 165–200. doi:10.1111/0081-1750.00078.

*Manuscript Received: 12 MAY 2013*

*Published Online Date: 13 SEP 2014*