Research & Methods
ISSN 1234-9224 Vol. 20 (1, 2011): 81–105
Institute of Philosophy and Sociology
Polish Academy of Sciences, Warsaw
www.ifispan.waw.pl
e-mail: publish@ifispan.waw.pl

Missing by Design: Planned Missing-Data Designs in Social Science

Artur Pokropek
Institute of Philosophy and Sociology Polish Academy of Sciences, Warsaw

This article presents research designs that employ modern statistical tools to optimize costs and precision of research along with some additional methodological advantages. In planned missing-data designs some parts of information about respondent are purposely not collected. This gives flexibility and opportunity to explore a broad range of solutions with considerably lower cost. Modern statistical tools for coping with missing-data, namely multiple imputation (MI) and maximum likelihood estimation with missing data (ML) are presented. Several missing-data designs are introduced and assessed by Monte Carlo simulation studies. Designs particularly useful in surveys, longitudinal analysis and measurement applications are showed and tested in terms of statistical power and bias reduction. Article shows advantages, opportunities and problems connected with missing-data designs and their application in social science researches.

Key words: missing-data; research design; multiple imputation; maximum likelihood; Monte Carlo simulations; statistical power.

INTRODUCTION

As one of the leading experts in missing data analysis (Graham 2009: 551) said, "Contrary to the old adage that the best solution to missing data is not to have them, there are times when building missing data into the overall measurement design is the best use of limited resources." In empirical science, researchers always want to measure things they are interested in as precisely as possible. In the social sciences, particularly in survey research, precision implies the need for sufficient sample size (to account for sampling error) and a suitably detailed measurement instrument (to account of measurement error). In the real world these requirements are often hard to achieve with complete data, i.e. all respondents from the large

sample are measured by detailed instrument. Large samples are very expensive to obtain. Long, detailed questionnaires with complex measuring instruments are inconvenient for respondents who, after too long period of questioning, may refuse to cooperate any further or do not agree to participate in panel research.

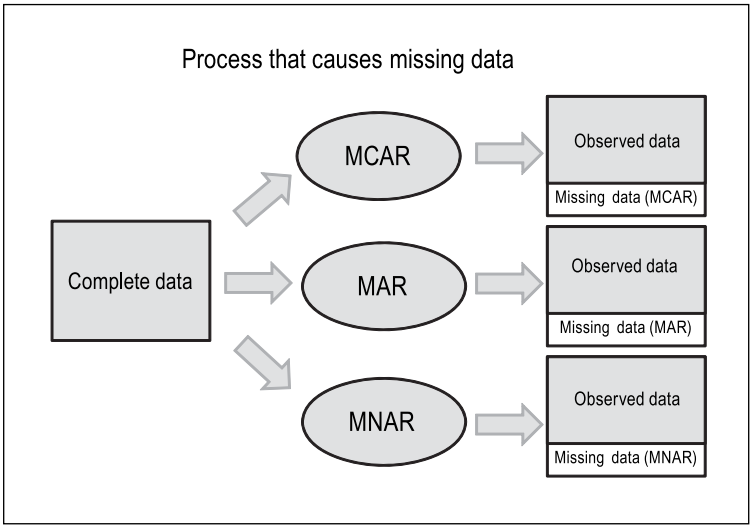
One way of coping with these problems is to keep a sufficient sample size and reduce the information from some respondents who did not quit as a result of intensely detailed survey items. This may be achieved by not giving all respondents the same set of questions; in other words, purposely omitting some of the respondents from some of the items. This procedure – along with appropriate statistical procedures – may substantially reduce costs and increase precision of research.

Using missing-data designs without appropriate statistical tools may do more harm than good; I begin by presenting a statistical approach to missing data analysis which missing-data designs require.

MISSING DATA THEORY

Rubin's *Interference and missing data* (1976), established the framework for inference from incomplete data. This framework has become the statistical basis for most popular missing data handling methods, namely maximum likelihood (ML) estimation with missing data and multiple imputation (MI) methods.

Figure 1 Process that causes missing data and its outcomes



In Rubins' framework missingness is regarded as an probabilistic phenomenon (Rubin 1976: 581) and is governed by the "process that causes missing data". The mechanism by which some data are recorded and others not is understood in terms of statistical device to describe patterns of missing values, specifically the relations between missing and complete data; it is only a statistical model that is not strictly connected with real-world processes that lay behind missing data (Schafer and Graham 2002, 7:150). A graphical representation of Rubin's framework is shown in Figure 1.

To understand Rubin's approach we need to assume the existence of an ideal complete dataset comprised of all variables specified by the research and has no missing values. Missing values appear during the collection of data (or more generally by process that causes missing data) and is driven by three different statistical mechanisms, the outcomes of which define type of missingness in the dataset, and each one produces a different kind of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

When M_y is defined as an indicator of missingness (M_y =1 if data are missing and M_y = 0 if not).[1] Data on *Y* are said to be missing completely at random (MCAR) when probability of missing data on *Y,* that is $P(M_Y)$, is unrelated to the value of *Y* itself or to the values of any other variables in the dataset. This may be formally described as:

$$P(M_Y \mid Y) = P(M_Y) \tag{1}$$

Data on *Y* are said to be missing at random (MAR) if the probability of missing data on *Y* is unrelated to the value of *Y*, after controlling other variables (*X*) in analysis, more formally:

$$P(M_Y \mid Y, X) = P(M_Y \mid X) \tag{2}$$

Which means that conditional probability of missing data on *Y* given both *Y* and *X* is equal to the probability of missing data on *Y* given alone *X* (Allison 2001:4).

The third type of data are missing not at random (MNAR) which means that probability of missingness on *Y* is related to the values of *Y* itself even after controlling for additional variables *X*. This situation is simply described by following equation:

$$P(M_Y \mid Y, X) \neq P(M_Y \mid X) \tag{3}$$

Table 1 Artificial data with three different missingness mechanisms

Complete data		Missingness mechanism			Y with missing data		
Y	X	MCAR	MAR	MNAR	MCAR	MAR	MNAR
46	23	1	0	0	-	46	46
46	55	1	1	0	-	-	46
47	36	1	0	0	-	47	47
47	39	1	0	0	-	47	47
48	42	1	1	0	-	-	48
48	53	0	1	0	48	-	48
50	26	1	0	0	-	50	50
50	33	0	0	0	50	50	50
54	36	1	0	0	-	54	54
55	59	1	1	0	-	-	55
55	34	0	0	1	55	55	-
55	56	1	1	1	-	-	-
56	22	0	0	1	56	56	
57	69	1	1	1	-	-	-
57	53	1	1	1	-	-	-
61	55	1	1	1	-	-	-
63	65	0	1	1	63	-	-
63	40	1	0	1	-	63	-
64	58	0	1	1	64	-	-
64	50	0	1	1	-	-	-
64	57	0	1	1	-	-	-
66	56	1	1	1	-	-	-
68	55	1	1	1	-	-	-
70	46	0	1	1	70	-	-
70	38	1	1	1	-	70	-
73	75	1	1	1	-	-	-
74	57	1	1	1	-	-	-
77	66	1	1	1	-	-	-
77	60	1	1	1	-	-	-
77	83	0	1	1	77	-	-

The example of three different mechanisms that causes missing data is presented in Table 1. In the first part of the table complete data on *Y* and *X* are presented (the correlation between those two variables is 0.60). In the second part ("Missingness

mechanism") indicators of missingness (M_y =1 if data are missing and M_y =0 if not) are presented. Those indicators are applied to variable *Y* presented in the first part of the table and they define three missingness mechanisms: MCAR, MAR and MNAR. The outcomes of three mechanisms are presented in the third part of the table ("*Y* with missing data") and are simply data that mimic three different missing data situations. In MCAR situation missing data are simply random draws of *Y*. In MAR situation *Y* becomes missing when the value of *X* is greater than 40 (missingness depends on *X*) and in MNAR *Y* becomes missing for the twenty highest values of itself (missingness depends on *Y*). All three situations produce 66,7% of missing data of *Y*.

After producing missingness one may wish to check whether the dataset really reflects desired situations. Having a complete dataset together with missing data indicators the task is quite simple. One of the ways to check this is to perform multiple regression:

$$Y_{com} = \beta_0 + \beta_1 M_y + \beta_2 X + e \tag{4}$$

where Y_{com} is variable *Y* with complete data on *Y*, M_y is indicator of missingness. Applying this model to three types of missingness and their definition in equations (1–3) we should expect that indicator for missingness should not be significant after controlling for *X* in MCAR and MAR situation. But we should be able to find significant relation between Y_{com} and M_y in MNAR situation. Results of such experiment are shown in Table 2. As one could suppose results are exactly along expectations; we shouldn't be surprised because generating missing data was strictly under control. Indicators of missing data after controlling for *X* are not significant for MCAR and MAR situation but are significant for the MNAR situation.

Table 2 Regression models to check missing data assumptions

Y(complete)	MCAR	MAR	MNAR
Missingness mechanism:			
MCAR Y	-0.78		
MAR Y		-3.44	
MNAR Y			13.43***
X	0.41***	0.50*	0.21*
Constans	39.51**	34.09**	54.17***

* p<0.05; ** p<0.01; *** p<0.001

Depending on what type of missing data researcher is currently working with, the distinction between MCAR, MAR and MNAR will define how the methods for coping with missing data will work. If data are MCAR, even simple methods like listwise deletion of variables will work and the main cost will be in reducing sample size and connected with that power of analysis. In MAR simple methods may produce substantial bias but others like maximum likelihood estimation with missing data (MI) or multiple imputations (MI) will work reasonably well. When data are MNAR serious bias is plausible regardless of missing data handling techniques.

Facing real datasets with missing values we rarely know whether data are MCAR, MAR or MNAR. The big question is if it is reasonable to question whether data are MCAR, MAR or MNAR, given that there is no way to test it in real situation. Serious concerns are not present only when the researcher controls missingness mechanism. This is the case of designs where missingness is planned by and the process that causes missingness is under control, as in the artificial example above. Because I am concerned with designs that control missingness, the assumption about missingness mechanisms becomes a decision of which mechanism is most suitable; the decision in nearly all cases is to have MCAR or MAR designs.

Only two methods of handling missing data both in MCAR and MAR scenarios give unbiased estimates of wide range of different parameters and their standard errors with reasonably small loss of power due to missingness: maximum likelihood estimation with missing data (MI) or multiple imputations (MI). Using an empirical example I briefly discuss those methods .

MAXIMUM LIKELIHOOD ESTIMATION WITH MISSING DATA

Roots of Maximum Likelihood (ML) technique go back to early 1920's and are connected with work of Sir Ronald Fisher, one of the founders of modern statistics. Although the base for ML techniques have been known for many years, only through development of computers and their computational speed is it possible to take all advantages of this very flexible approach. One of the main advantages of ML is that it allows us to incorporate observations with missing data into estimation process along with observations with complete data in a way that is natural for this framework. In classical estimation such approach is not possible – observations with missing variables must be ruled out from estimation or their missing values have to be replaced by some predictions (mean substitution, conditional imputation, and hot-deck technique). ML allows observations with missing data to contribute directly to final estimates; that only partial information is available for them is no problem for the ML estimation.

A short introduction to the general benefits of the ML aids in understanding how this technique may be used in handling missing data. Let us define dataset as a matrix **y**. We may assume that **y** is the outcome of the random process which can be characterized in terms of probability density function (p.d.f.) $p(\mathbf{y}|\boldsymbol{\theta})$ where data are conditioned on matrix of parameters $\boldsymbol{\theta}$ in the interior of the parameter space $\boldsymbol{\Omega}$. Given the parameters $\boldsymbol{\theta}$ and probability model the p.d.f. function describes the plausibility of observed data. The p.d.f function is the function of **y**. On the other hand the likelihood function describes plausibility of parameters $\boldsymbol{\theta}$ given the fixed data **y** and may be written as $L(\boldsymbol{\theta} \mid \mathbf{y})$. The likelihood is the function of $\boldsymbol{\theta}$ for fixed data. By definition likelihood function is defined as:

$$L(\boldsymbol{\theta} \mid \mathbf{y}) = k(\mathbf{y}) f(\mathbf{y} \mid \boldsymbol{\theta}) \propto f(\mathbf{y} \mid \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{5}$$

Where the function $k(\mathbf{y})$ does not depend on $\boldsymbol{\theta}$ but may depend on data. The likelihood function is neither a p.d.f. function nor any probability function. While true values of $\boldsymbol{\theta}$ are fixed, likelihood should be interpreted like Fisher proposed as a rationale measure of degree of belief or at least as a relative fit as some authors do (Enders 2010: 59).

The ML estimate $\hat{\boldsymbol{\theta}}$ is the most likely value of $\boldsymbol{\theta}$ given the data, more formally:

$$L(\hat{\boldsymbol{\theta}} \mid \mathbf{y}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \mathbf{y}) \tag{6}$$

Computational procedures used in ML estimation are designed for finding such values of $\boldsymbol{\theta}$ that maximize function $L(\boldsymbol{\theta} \mid \mathbf{y})$. Computational algorithms repeatedly assign different values of matrix $\boldsymbol{\theta}$ until the likelihood for whole data reaches maximum, less formally algorithms that search for a combination of parameters form matrix $\boldsymbol{\theta}$ that best fit to the data.

Conceptually applying ML to missing data is straightforward. In case of most statistical analysis the likelihood function is computed using individual data of particular observations. Assuming that responses of individuals are independent, the overall likelihood of $\boldsymbol{\theta}$ given data **y** for any dataset is the product of the N individual likelihood values:

$$L(\boldsymbol{\theta} \mid \mathbf{y}) = \prod_{i=1}^{N} f(\mathbf{y}_i \mid \boldsymbol{\theta}) \tag{7}$$

When missing data are present one thing which must be done is to split the formula into two parts of data – variables with and without missing data. Let us refer to a simple case of bivariate distribution. This example corresponds to the dataset presented in Table 1. In this example for the first m observations both x and

y are observed. For the rest of observations only data on *x* are available thus the likelihood function may be rewritten for two parts: first where observations are available for two variables and second where only values of *x* are known (Allison 2001:21):

$$L(\boldsymbol{\theta} \mid x, y) = \prod_{i=1}^{m} f(x_i, y_i \mid \boldsymbol{\theta}) \prod_{i=m+1}^{n} f(x_i \mid \boldsymbol{\theta}) \tag{8}$$

Finding proper estimates of **θ** (in case of bivariate distribution matrix of parameters consists of means and variances of the variables and covariance between them) becomes only a technical problem of how to find **θ** that maximizes the likelihood function in reasonably few iteration steps. This problem in most cases of missing data analysis is solved by applying so-called E-M algorithm (Expectation-Maximization) (Dempster, Laird, and Rubin 1977: 1–38). E-M algorithm starts from obtaining values of parameters in matrix **θ** – usually standard methods are used that disregard missing values. In E-step parameters of **θ** are used to compute regression which is used to predict the incomplete values from the observed variables. In M-step values of **θ** are updated using filled-in data from the E-step and standard methods of computing parameters. Then the E-step is repeated and updated parameters of **θ** are used to compute new regression and new predictions for missing values. After this the next M-step is conducted. Algorithm continues to repeat E- and M-steps until subsequent matrixes **θ** obtained in following steps do not differ in a significant way. When subsequent matrixes **θ** do not differ, the algorithm has converged on the maximum likelihood estimates.

MULTIPLE IMPUTATION: BAYESIAN FRAMEWORK FOR MISSING DATA

Multiple imputation (MI) was developed by Rubin (1987) and has become a flexible alternative to ML methods in missing data applications. MI is heavy grounded in Bayesian framework but conceptual principles of MI can be understood without relying on Bayesian statistics. The goal of multiple imputation is to create *m* imputed data sets (in most cases 5 to 10 datasets) in which missing data are replaced by unique imputations. Each imputed dataset is a plausiblly alternative version of the complete dataset.

The imputed values are random draws from conditional distribution that depend on observed data (Enders 2010: 192–194):

$$\mathbf{Y}_{mis}^{(t)} \sim P(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}; \boldsymbol{\theta}^{(t)}) \tag{9}$$

Where $\mathbf{Y}_{mis}^{(t)}$ represents the imputed values in *t* step. $\mathbf{Y}_{mis}$ and $\mathbf{Y}_{obs}$ are respectively missing portion and observed portion of the data. $\boldsymbol{\theta}^{(t)}$ is a matrix of parameters used in *t*-th imputation. What distinguishes MI from others missing data handling techniques is the Bayesian approach in which parameters are treated as random rather than fixed values. Treating parameters as random implies that in each imputation parameters used to predict missing data are different by random factor representing uncertainty about parameters.

Data augmentation relays on two steps algorithm: imputation step (I-step) and posterior step (P-step). At first, like in E-M algorithm, estimates of $\boldsymbol{\theta}$ are needed and usually are obtained ignoring missing data. In I-step the parameters from matrix $\boldsymbol{\theta}$ are used to build set of regression equations to predict missing data on incomplete variables from information from other variables. A normally distributed residual term from the regression equation is added to each predicted value so as not to underestimate the level of variability in the whole data set. As it may be derived from Bayesian framework, adding the residual term from the regression equation is not sufficient to keep variation that mimics the complete data variability. This may be achieved only if in different imputation, different mean vectors and covariance matrixes are used in regressions to predict missing data on incomplete variables. Consequently in the imputation process the error term is also added to $\boldsymbol{\theta}$ matrixes in process of data imputation producing matrix $\boldsymbol{\theta}^{(t)}$. In P-step missing values for missing data are simply predicted using information from $\boldsymbol{\theta}^{(t)}$. After the first cycle ends the algorithm goes back to I-step. Usually a couple of hundred cycles are conducted and the final imputed datasets is a random sample from all cycles (Rubin 1987).

After producing several datasets each of imputed data set is analyzed by completed data methods. This involves first fitting sets of models, each with one dataset and obtaining *m* sets of results. Then these analyses are aggregated using the Rubin rule, where the point estimator is simply the average of *m* estimators obtained from previous analysis:

$$\bar{\theta} = \frac{1}{m}\sum_{t=1}^{m} \hat{\theta}_t \tag{10}$$

where $\bar{\theta}_t$ is the parameter estimate from data set *t* and $\bar{\theta}$ is the average point estimator. Standard errors are average of standard errors obtained from previous analysis enlarged by variation between them (Little and Rubin 1987):

$$SE(\bar{\theta}) = \sqrt{V_w + V_B + V_B/m} \tag{11}$$

where V_w is the within-imputation variance. V_b is the between-imputation variance which quantifies the variability between imputed datasets. More formally:

$$V_w = \frac{1}{m}\sum_{t=1}^{m} SE_t^2 \qquad V_B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}_t - \bar{\theta})^2 \tag{12}$$

where SE^2 is the squared standard error from data *t*.

ESTIMATION WITH MISSING DATA: AN EXAMPLE

In Table 3 artificial datasets (from Table 1) are used to demonstrate the performance of missing data handling methods in estimation of mean, standard deviation (sd) and correlation. Three methods were used: classical one, where observations with missing data are excluded from analysis (i.e. listwise deletion LD), MI and ML. As in Table 1 three types of missing data on *Y* were used – MCAR, MAR and MNAR – to show the performance of different techniques of handling with missing data.

Table 3 Analysis with missing data results from artificial datasets

Data	Correlations with X	Mean	Sd
	Complete data		
Complete data X	1.00	49.90	15.01
Complete data Y	0.60	60.07	10.11
	Listwise deletion		
MCAR response Y	0.67	60.38	9.98
MAR response Y	0.36	53.80	7.71
MNAR response Y	0.21	49.10	3.18
	Multiple imputation		
MCAR response Y	0.60	61.69	10.12
MAR response Y	0.61	60.11	8.26
MNAR response Y	0.10	49.35	3.40
	Maximum likelihood		
MCAR response Y	0.57	60.92	9.10
MAR response Y	0.67	60.99	9.20
MNAR response Y	0.27	49.64	3.06

In the first part of the table „Complete data" we find information about variables and their associations when there is no missing data values. The mean of *X* is about

50 and the mean of *Y* is about 60. Standard deviations are respectively about 15 and 10. The correlation between *X* and *Y* equals 0.60.

The next three parts of Table 3 present results of three different estimation procedures in three conditions defined by three missing data types. Analysis with LD and MI were performed in Stata 10. For generating MI, the Stata procedure known as 'ICE' (Royston 2005: 527) was used. For ML estimation Mplus 6 (Muthen, Kaplan, and Hollis 1987: 431–462) was used. In MCAR situation all methods perform astonishingly well, keeping in mind that only 33% of data on *Y* is not missing. In all cases the correlation is about 0.6, estimated mean of *Y,* is about 60. In LD and MI standard deviation is just about 10 which exactly corresponds to the true value. In case of ML sd is about 9 which indicates a small underestimation of this parameter by ML technique[2].

MI and ML show their supremacy over LD in case where data are MAR. LD gives huge underestimation of correlation parameter by estimating it as 0.36. Mean and standard deviations are also noticeably underestimated by this technique. MI and ML perform reasonably well in MAR condition. Only estimation of sd in MI case is noticeably worse than it's true value but still estimation by MI is much closer than LD estimation.

Results from MNAR situations show that when the missingness mechanism is out of control, i.e. probability of missingness is connected with the value of variable itself, even controlling for variables and using the most sophisticated statistical procedures the situation is hopeless. In case of MNAR we face huge bias in estimation of standard deviation and mean in variables with missing values as well as in correlation.

This exercise show that MI and ML are techniques that outperform the classical method of missing data analysis, namely LD ; they are more accurate and flexible. One should not treat this analysis as proof of superiority. The intention was only to show an example of ML and MI performance in comparison to classical methods. Proofs showing that MI and ML are much more reliable, unbiased and analysis with them has much more statistical power than classical ones (not only LD but also pairwise deletion, single imputation, stochastic regression imputation, and hot deck imputation) are widely available in literature (Allison 2001: Enders 2010; Graham 2009: 549–576; Graham, Cumsille, and Elek-Fisk 2003: 87–114; Little and Rubin 1987). MI and ML are very accurate techniques and as both of them provide unbiased results that are asymptotically equivalent (Enders 2010: 189).

There are few hidden flaws in MI and ML. Both methods are computationally demanding and even with modern computers generating MI or estimating complex model by ML takes a long time. A few years ago both ML and MI techniques were restricted by multivariate normal assumption which indicates that variables which are not normal should not be applied to these techniques (i.e. binary variables,

count data, and nominal data). However, ML and MI estimates, obtained under the multivariate normal assumption, often have good properties even some variables with missing data that have the non-normal distribution (Allison 2001: 19, 38). Moreover transforming variables (for instance using logarithmic scale) may be very good solution in many cases. Assumption of normality is not demanded in modern solutions; only information about distribution of particular variables is required.

One of the serious disadvantages which may appear in MI or ML is misspecification of the model. Serious bias may arise when in MI procedure variable is omitted but used later in modeling (including interaction terms). Such a lack of variables in imputation phase attenuates associations in modeling phase. This may bring some problems as MI demands to predict in advance the variables and interactions terms which will be used in the modeling phase. "The advice has always been to include as many variables as possible when doing multiple imputation" (Rubin 1996: 473–489). Practically using all variables in the imputation phase is a common practice but including all interaction terms especially in large datasets is a less frequent routine. Even in medium datasets the number of interactions increases dramatically especially when not two-way interactions and tree-way interactions are taken into account. Number of variables to impute increases so dramatically that even a modern computer will have problems to conduct imputations properly and computational algorithms may fail when data matrix is too large.

In ML estimations a similar problem may occur. When doing ML all significant correlates of missingness – especially in MAR situation – should be included in the model even if they are not a primary interest. There are methods to include auxiliary variables in Structural Equation Modeling (SEM) framework (Muthen, Kaplan, and Hollis 1987: 431–462) with no harm to model of primary interest. But when very large number of auxiliary variables is introduced, serious problems may arise in estimation process.

PLANNED MISSING-DATA DESIGNS

One of the main goals of this article is to show different research designs that use the capabilities of modern estimation methods with missing data. Planned missing-data designs are designed for cost reduction and the following methodological reasons: not to expose respondents to too long questionnaires, not to harass respondents in panel designs too often, and to facilitate measurement instruments. Planned missing-data designs fulfil these tasks well but at the cost of reducing power of the statistical analysis.

Due to their complexity, there are no analytical formulas for estimating power in planned missing-data designs. The loss of power in such designs is generally not

strictly related to the decrease in sample size. Fortunately Monte Carlo simulations is a good alternative to analytical formulas (Rubinstein and Kroese 2008, 707). In Monte Carlo simulations hypothetical values of the population parameters are specified. Next, a large number of samples are drawn from the hypothetical population and to-be-tested model is estimated on each of them. Estimated parameters and theirs standard errors from each sample build an empirical sampling distribution for each parameter in the sampling model. If parameters in the hypothetical population differ from zero, the power for each parameter is simply a percent of the statistically significant result in the distribution of parameters' sampling distribution.

I examine several models according to loss of statistical power. I test different conditions and models, but in each Monte Carlo simulation 10 000 replications were used. For conducting simulations built-in routines of Mplus 6 computer programs were used. As MI and ML are asymptoticaly equivalent results from only one method are presented in following analyses, namely ML. Results from MI in case of simulations are merely the same and bring no additional information.

MISSING DESIGNS IN SURVEYS

Three-Form Design

One of the most popular of planned missing-data designs in methodological literature is so-called Three-Form Design. The design is presented in Table 4. The whole idea of it is to split questionnaire into four parts (blocks): X, A, B, C and produce three different versions of the questionnaire. The most important questions are placed in part X and are given to all respondents (in Table 4. "1" means that block of the questionnaire is present in particular version and "0" indicates that is not present). Versions of the Three-Form Design differ in such way that in each one of the item blocks (A, B,C) is missing. In version 1, respondents are answering questions from block X, B, C but not A, in version 2 blocks X, A and C are present but not B.

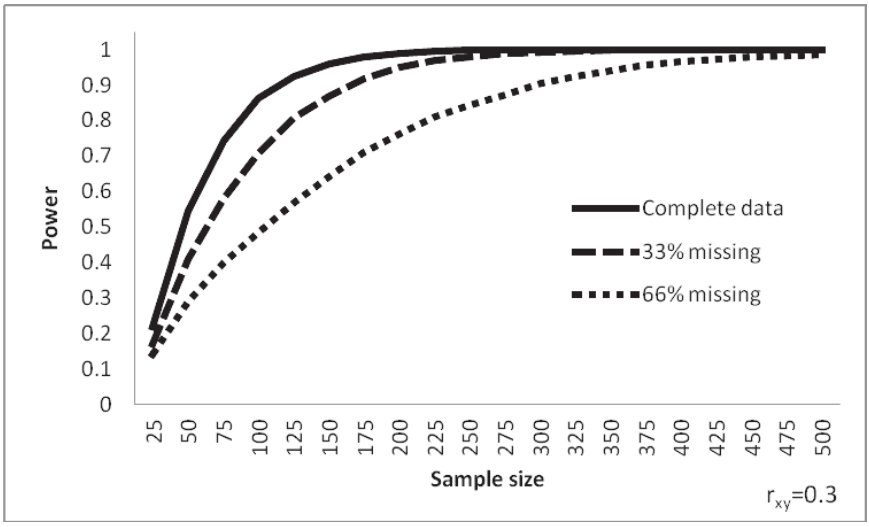
Table 4 Three-Form Design

Version:	Item blocks of questionnaire:			
	X	A	B	C
1	1	0	1	1
2	1	1	0	1
3	1	1	1	0

1 – present; 0 – not present

Presented design allows us to ask about 25% more questions than in classical designs with no extra time. For this benefit the cost is missing data, as each respondent will have 25% of missing data. Moreover, as in each different version different parts of the questionnaire are missing, in whole datasets multivariate analysis may have different rates of missing data. When one wants to use only variables from block X, there will be no missing data patterns. When running model both from block X and A (or B, C), 33% of missing data will appear. The highest rate of missing data will appear when one wants to estimate a model based on two items of three blocks: A, B and C (i.e. AB, AC, BC). In this situation about 66%/67% of missing-data will appear.

Figure 2 Monte Carlo analysis of the power in three-form design according to different sample size. Correlation coefficient in population equals 0.3



When the process that causes missing data is controlled, the only concern of the researcher is decreasing statistical power. The more missing data in analysis, the less powerful the analysis is. In Figure 2 results from Monte Carlo study of power (correlation) is presented. In simulation study a short questionnaire consisting of eight artificial variables was generated. First two variables belong to part X, next two to part A and so on. I Assume that the population correlations between all variables equals 0.3 (which corresponds to medium effect size). In Figure 2 results from 10 000 of Monte Carlo replications are shown. Each curve presents estimated power of correlation analysis depending on the sample size. The solid curve represents complete part of data from Three-Form Design, i.e. correlations

between variables in part X. This line outlines the maximum power which may be achieved with particular sample design. Dashed curve presents the average power of correlation based on variables from block X and the rest of blocks, i.e. the case with 33% of missing-data. The last curve (short-dashed) represents the average power of correlations between variables form blocks A, B and C.

The power of complete data reaches a reasonable level (more than 0.9) when the sample is greater than 100 and the power becomes almost perfect (near 1.0) when the sample size is greater than 200. In case of analysis with missing data the shape of the curves are similar but shift to the lower values of power. The differences are not very significant. The power of correlation analysis with 33% of missing data starts to be undistinguishable when the sample size reaches 225–250 respondents. The analysis with 66% of missing data reaches very high level of power when sample size is about 400, which is a standard effective sample size in nearly all survey research.

Figure 3 Monte Carlo analysis of the power in Three-Form Design according to different values of correlations in population. Sample size 400

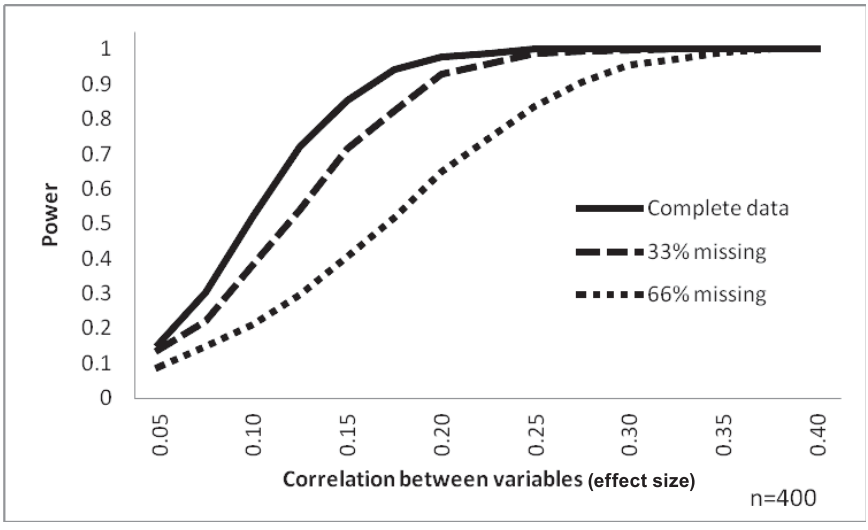


Figure 3 shows relation between power of the Three-Form Design and correlation of parameters to be estimated in a sample size typical for survey research: 400 observations. Obviously, when the size of the effect in population is small, statistical power to detect it is not very high. When the size of the effect becomes larger, probability of detecting it rises in different rates, depending on how missing data are present in analysis. The power for detecting correlation of 0.3

is high and do not vary substantially between three situations (complete data, 33% missing and 66% missing). This pattern changes when we are looking at smaller effects. With complete data, i.e. variables from block X in Three-Form Design, we are confidently able to detect (the power is about 90%) the correlation between 0.15 and 0.2. When 33% of missing data are present, capability of detecting true relations between variables does not change dramatically – correlation greater than 0.20 will be detected in more than 90% of cases. A not so bright picture emerges when we look at results of simulation studies concerning analysis with 66% of missing-data – power of the statistical tests will be achieved only for true correlations above 0.25/0.30.

Monte Carlo results presented in this section confirm usefulness of Three-Form Design. In standard sociological surveys, when effective sample size is about 400 and more, the power of detecting medium size effects is almost as large as in complete designs. For smaller size effects, when the effective sample size is below 400, some consideration about minimum level of accepted power in analyses should be taken into account in choosing this design. Three-Form Design may not be the best choice when small effects (below 0.2) are to be measured, particularly in the case of multivariate analysis, when 66% of missing data are present. When correlations between variables are expected to be small, those variables should be located in block X of questionnaire (or at least some of them should be placed there).

BIB7 – Youden Squares Design

Another interesting missing-data design that has many desirable characteristics is the 7-block Youden Squares Design, originally used in experimental biological research designs (Preece 1990: 65–75) and recently widely used in educational measurement (Aitkin and Aitkin 2011; Rutkowski et al. 2010: 142). This design is often referred to as the BIB7 design and is presented in Table 5. The design consists of seven versions and seven blocks of items. Each form contains three blocks. Interesting propriety of BIB7 is that each block appears once in each position in the design. Each block is answered by 43% of respondents and also appears once with each of the other blocks, so the multivariate analyses between variables from two different blocks are conducted on 14% of the data. BIB7 is applied when many variables must be collected in a single sample of respondents, as one version of questionnaire has less than half of all item pool administered in the conducted survey.

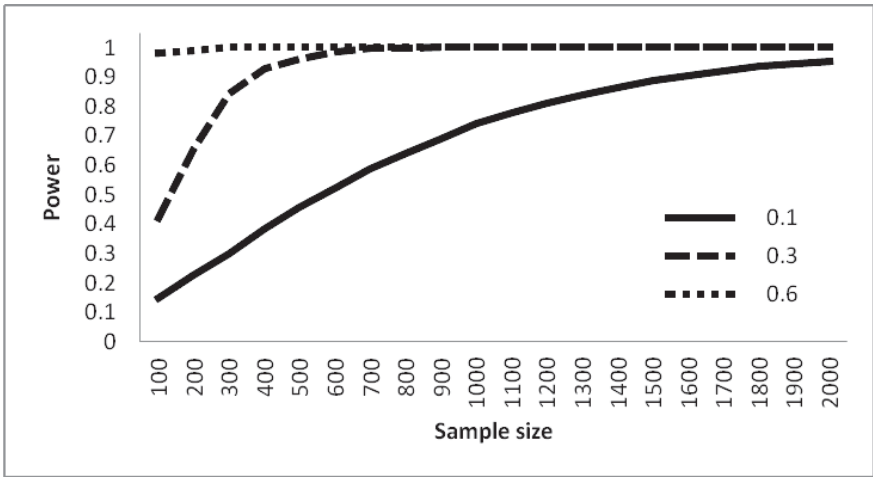
Table 5 BIB7 – Youden squares design

Version:	Placement of item blocks			Distribution of item blocks						
				A	B	C	D	E	F	G
1	A	B	D	1	1	0	1	0	0	0
2	B	C	E	0	1	1	0	1	0	0
3	C	D	F	0	0	1	1	0	1	0
4	D	E	G	0	0	0	1	1	0	1
5	E	F	A	1	0	0	0	1	1	0
6	F	G	B	0	1	0	0	0	1	1
7	G	A	C	1	0	1	0	0	0	1

1 – present; 0 – not present

As in all missing-data designs the main drawback of BIB7 is a loss of power. Figure 4 shows results from Monte Carlo power analysis in BIB7 design according to different sample size and effect size. Three effect sizes are considered here, corresponding to the values of correlations from two randomly chosen variables (from different item blocks) and set to be 0.1, 0.3 and 0.6. On the horizontal axis different sample size used for BIB7 design are shown.

Figure 4 Monte Carlo analysis of the power in BIB7 design according to different values of sample sizes and effect sizes



BIB7 with modern missing-data analysis techniques seems to be very suitable device in detecting medium (0.3) and large (0.6) effect sizes. With a sample size of 400 it gives reasonably large power to conduct statistical analysis for medium and large effects. When effects are expected to be very small, some serious considerations about sample size must be taken into account. When such small effect sizes are expected, also complete-data designs are not the ultimate solution. Going back to Figure 3, the power of complete data design for effect size 0.1 with sample size 400 is about 50%, corresponding power in BIB7 is about 40%. That means that in BIB7 we get 20% less power than in the complete design but BIB7 is able to administer more than two times question than in complete-design.

MISSING DESIGNS IN PANEL ANALYSIS

The implementation of missing-designs into longitudinal analysis seems to be particularly useful for methodological reasons. Bringing controlled missingness into panel designs allows not only to reduce the cost but also to reduce the number of waves per respondent, it minimizes participants' attrition over time due to pestering by research procedures (Graham, Taylor, and Cumsille 2001: 335–353).

In Table 6 I present a simple missing-data design in which the overall sample is divided into 6 groups tested in 5 panel waves. In classical panel design all groups will be participating in all waves. In presented design (Graham, Taylor, and Cumsille 2001: 335–353) only first group is present in all measurements (this group corresponds to block X in Three-Form Design) and reminding ones are not present in one panel wave. The design is constructed in such way that only one group does not participate in particular wave in the same time.

Table 6 Panel Missing Design (design 1)

Group:	Panel wave				
	1	2	3	4	5
1	1	1	1	1	1
2	1	1	1	1	0
3	1	1	1	0	1
4	1	1	0	1	1
5	1	0	1	1	1
6	0	1	1	1	1

1 – present; 0 – not present

Another design that may be useful in panel designs is Panel Chained Design. In this design as well as in previous longitudinal designs respondents are divided into 6 groups and 5 waves of panel are conducted. In this design respondents, except for the first and last group, participate only twice in research in two following waves. This design reduces dramatically the number of measurements, as only 33% of respondents participate in one wave of the panel. This carries a large reduction in expenditures and minimizes respondents' effort.

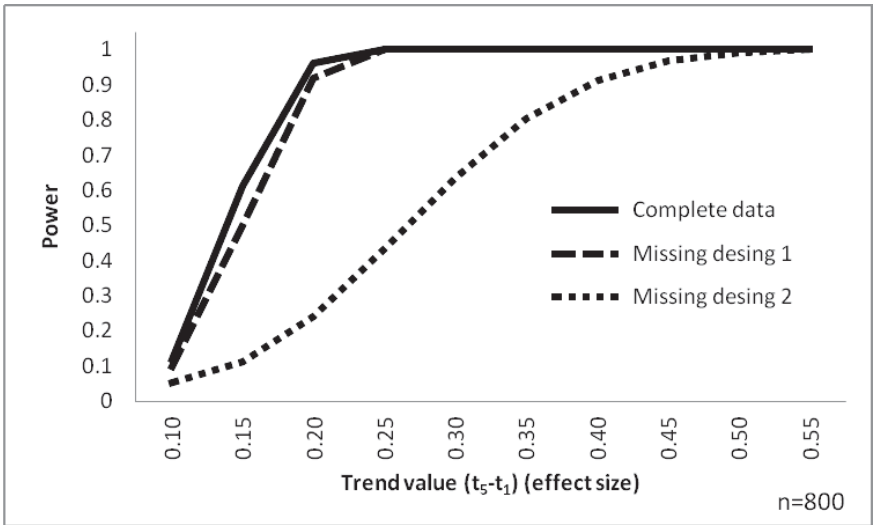
Table 7 Panel Chained Design (design 2)

Group:	Panel wave				
	1	2	3	4	5
1	1	0	0	0	0
2	1	1	0	0	0
3	0	1	1	0	0
4	0	0	1	1	0
5	0	0	0	1	1
6	0	0	0	0	1

1 – present; 0 – not present

As for others designs Monte Carlo simulations were conducted testing presented designs. In case of panel data they differ in such way that a more complex model were used, namely the Latent Growth Curve Model (Muthen and Muthen 2011:101) estimates growth. The population model was specified in such way that linear growth with the same rate between panel waves was imposed. Several scenarios were tested in which the growth between first and last wave was defined to be from 0.1 to 0.55 of stand deviation of first measurement. As growth is defined in terms of standard deviations it may be considered as an effect size indicator (like correlations in case of previous examples). For each scenario the Latent Growth Curve Model was estimated (10 000 times for each effect size and each design) and the parameter defining rate of growth was tested according to statistical power. The baseline sample size for this simulations is 800 as longitudinal researches often excide number of 400, which is most commonly used sample size in survey designs. Results from Monte Carlo studies referring to longitudinal analysis are presented in Figure 5.

Figure 5 Monte Carlo analysis of the power in longitudinal designs according to different values of effect sizes



The solid curve represents results from complete data design, the long dashed curve represents the Panel Missing Design (design 1) and the short dashed line represents Panel Chained Design (design 2). Using design 1 brings only minor changes to power of the study comparing to complete data design. Removing about 17% of respondents from each wave will bring noticeable cost reduction but not marked deterioration in power. The Chained Design may seem to be tempting, because huge reduction of sample size appears to bring remarkable reduction of power comparing to complete data design or design 1. The Chained Design with an effective sample size of 800 and 5 panel waves is able to detect (with acceptable certainty) only large effects of sizes greater than 0.4. To take advantage of Chained Design a large effect size must be expected or sufficiently large initial sample size must be chosen.

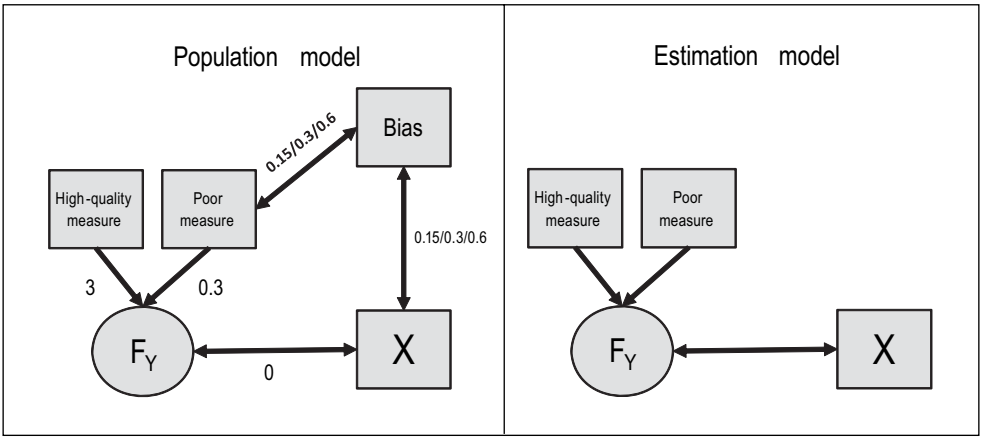
MEASUREMENT DESIGNS WITH MISSING DATA

When measurement is low quality, even the most sophisticated statistical analysis and most brilliant hypothesis will fail to save the research. Missing-data designs bring the opportunity to strengthen measurement and cope with problems carried by classical methods. Good measurement instruments that are complex or expensive often exhibit these two features. In many cases the researcher is restricted by budget conditions and faces a choice between sufficient sample size with a poor-quality measure and a small sample size with little power to bring significant results

but with high-quality measure. For instance it is easier to ask a respondent about health than to access the medical records. Missing-data design response to such dilemmas is quite straightforward. When one has two measurement instruments, one is high-quality (but expensive) and the second one are low-quality (but cheap), both should be used – give lower quality measure to all respondents and for others assign also the high-quality measure. The missing-data framework makes it possible to treat respondents without high-quality measurement as an observation with missing data on high-quality measure outcome and handle it as ordinary missing-data situation.

Figure 6 presents a simple example of using a missing-data design to improve measurement and evaluate the effectiveness of this approach by Monte Carlo simulations.[3] The aim is to show the relation between unobservable variable F_y and observable variable X. The true value (in population model) of correlation between those two variables is set to be 0. Variable F_y is measured by two instruments: high-quality measure and poor measure. The factor loading of high-quality measure (3) is 10 times greater than the poor measure (0.3). In addition poor measure is set to be biased by confounding correlation with other variable ("Bias") which is correlated also with X. Different scenarios with different amount of bias were tested setting correlations connected with bias to be 0.15, 0.3 and 0.6, which may be considered as a small bias, medium bias and large bias respectively.

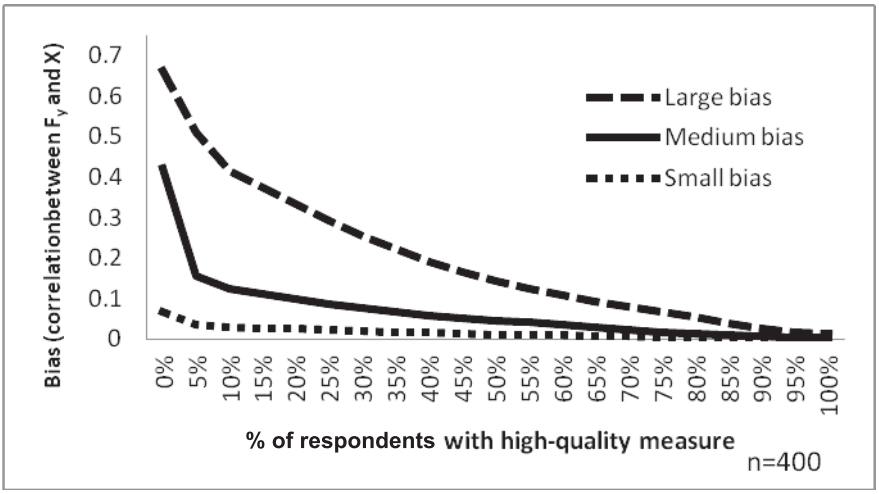
Figure 6 Diagram outlining a design of Monte Carlo simulation: measurement designs with missing data



Assessing the usefulness of the missing design relies on the result of the estimation where information about bias is not available (which is indicated on

the right side of Figure 6.) In estimation all respondents have non-missing data on poor-measurement but not all of them have results of high-quality measure. In Monte Carlo simulation different rate of respondents with high-quality measures were tested: from 0% to 100%. Otherwise, as in the previous Monte Carlo analysis presented in this paper, I present the amount of bias introduced by analysis. Bias is defined here as an estimated correlation between F_y and X. As the real correlation equals zero, any correlation exciding this number will be considered as bias. The results of Monte Carlo study on measurement instruments are presented in Figure 7.

Figure 7 Monte Carlo analysis of the power in measurement according to different rate of high-quality measure



When all respondents (of sample size 400) are measured by poor-quality as well high-quality measures even a large bias in poor-quality measure could not affect the overall results: correlations between F_y and X are simply 0. On the other hand, when there is no information from high-quality measure, the amount of bias is extremely high for large bias (more than 0.6), quite high for medium bias (about 0.4) and noticeable for small bias (about 0.1). Adding some respondents with two measures reduces bias. The most interesting thing about results from Figure 7 is that only small number of respondents with two measures allows for substantial bias reduction. Adding 5% to 10%, namely 20 to 40 of respondents with two kinds of measures may significantly reduce bias or, if it is small, may help to get rid of it completely.

MISSING AT RANDOM IN MISSING-DATA DESIGNS

All the examples and Monte Carlo simulation studies presented in this article refer to the MCAR situation, however there are no restrictions to limit missing-data designs to MCAR situation. If in longitudinal studies one wants to oversample younger respondents or in Three-Form Design one insists that some questions should be given more frequently to one group of the respondents than to others, there are a few restrictions. The process of generating missing data must be controlled by the researcher to mimic the MAR situation and what follows from this, variable or variables correlated with process that causes missing data must be included into imputation procedure (in case of MI) or must be implemented in estimation model (in case of ML).

Designs with MAR situation and analysis performed on these data might be conducted using weighting and/or re-weighting procedures (Guo and Fraser 2010). Weighting procedures differs from MI and ML in such way that information about process that causes missing data is incorporated into weights. Weighting procedures are relatively straightforward in use but are less efficient than MI or ML (Carpenter, Kenward, and Vansteelandt 2006: 571–584). In case of missing data analysis one should use MI or ML instead of weighting.

CONCLUSION

Missing-data designs are a tempting alternative to classical designs when all data are expected to be observed and missingness is considered as a problem. Missing-data designs allow us to reduce costs of research with low loss in precision and power of the analysis.

Simulation studies presented in this article confirm usefulness of the missing-data designs. In standard surveys when effective sample size is about 400 or more, the designs work extremely well with medium and large effects. However for smaller size effects, when the effective sample size is below 400, some consideration about minimum level of accepted power in analyses should be taken into account in choosing design.

The implementation of missing-designs into survey designs may reduce the time of interview more than 50% (BIB7 design) with no reduction in the number of questions. In longitudinal analysis missing-data design seems to be particularly useful for methodological reasons – controlled missingness in panel designs allows to not only reduce the cost but also to reduce respondents' effort to participate in the panel, which may lead to lower drop-out rate.

The great advantage of missing-data design is found in the measurement area. Monte Carlo studies show that adding 5% to 10% may significantly reduce or eliminate bias completely from the research at relatively low costs. Missing-data

designs should be considered in planning researches. While putting controlled missingness into a data does not solve all researchers' problems, it certainly brings some good opportunities.

NOTES

1 Rubin in different papers uses different notation for this indicator. In this paper I use most intuitive one in my opinion but different from Rubin's convention.
2 One should not take this as an evidence that ML underestimates sd in all situations. On the contrary this happens only when samples are relatively small.
3 This is an extended and different version of analyses known in the psychology literature (Graham et al. 2006: 323).

REFERENCES

Aitkin, Irit and Murray Aitkin. 2011. *Statistical Modeling of the National Assessment of Educational Progress*.: Not Avail.

Allison, Paul D. 2001. *Missing data*. Thousad Oaks: Sage Publications, Inc.

Carpenter, James R., Michael G. Kenward and Stijn Vansteelandt. 2006. 'A Comparison of Multiple Imputation and Inverse Probability Weighting for Analyses with Missing Data.' *Journal of the Royal Statistical Society, Series A* 169 (3): 571–584.

Dempster, Arthur P., Nan M. Laird and Donald B. Rubin. 1977. 'Maximum Likelihood from Incomplete Data Via the EM Algorithm.' *Journal of the Royal Statistical Society. Series B (Methodological)*: 1–38.

Enders, Craig K. 2010. *Applied Missing Data Analysis*. New York: The Guilford Press.

Graham, John W. 2009. 'Missing Data Analysis: Making it Work in the Real World.' *Annual Review of Psychology* 60: 549–576.

Graham, John W., Patricio E. Cumsille and Elvira Elek-Fisk. 2003. 'Methods for Handling Missing Data.' *Handbook of Psychology* 87–114.

Graham, John W., Bonnie J. Taylor and Patricio E. Cumsille. 2001. 'Planned Missing-Data Designs in Analysis of Change.' Pp. 335–353. In: *New Methods for the Analysis of Change*, edited by L.M. Collins and A.G. Sayer. Washington, DC: American Psychological Association.

Graham, John W., Bonnie J. Taylor, Allison E. Olchowski and Patricio E. Cumsille. 2006. 'Planned Missing Data Designs in Psychological Research.' *Psychological Methods* 11 (4): 323.

Guo, Shenyang and Mark W. Frase. 2010. *Propensity Score Analysis: Statistical Methods and Applications*. Los Angeles: Sage.

Little, Roderick J.A. and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Muthen, Bengt O., David Kaplan and Michael Hollis. 1987. 'On Structural Equation Modeling with Data that are not Missing Completely at Random.' *Psychometrika* 52 (3): 431–462.

Muthen, Linda K. and Bengt O. Muthen. 2011. *Mplus user's Guide. 6th*. Los Angeles: Muthen & Muthen.

Preece, Donald A. 1990. 'Fifty years of Youden Squares: A Review.' *Bulletin of the Institute of Mathematics and its Applications* 26 (4): 65–75.

Royston, Patrick. 2005. 'Multiple Imputation of Missing Values: Update of Ice.' *Stata Journal* 5 (4): 527.

Rubin, Donald B. 1976. 'Inference and Missing Data.' *Biometrika* 63 (3): 581.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New Jersey: Wiley.

Rubin, Donald B. 1996. 'Multiple Imputation after 18+ Years.' *Journal of the American Statistical Association*: 473–489.

Rubinstein, Reuven Y. and Dirk P. Kroese. 2008. *Simulation and the Monte Carlo Method*. New Jersey: Wiley.

Rutkowski, Leslie, Eugenio Gonzalez, Marc Joncas and Matthias von Davier. 2010. 'International Large-Scale Assessment Data.' *Educational Researcher* 39 (2): 142.

Schafer, Joseph L. and John W. Graham. 2002. 'Missing Data: Our View of the State of the Art.' *Psychological Methods* 7 (2): 147.

Artur Pokropek, Assistant Professor in the Institute of Philosophy and Sociology of the Polish Academy of Sciences in Warsaw. Member of Research Group on Interdisciplinary Studies on Education. He also works in the Institute of Educational Research (IBE), Warsaw, in a Department of Measurement. Head of methodological group in Polish Central Examination Board, where he is responsible for modeling and developing Educational Value Added in Polish educational system. His specialized in, multilevel modeling, contextual analysis, applied psychometrics and missing data handling methods. Aside of methodological and statistical field he works on such topics as: social structure, gender segregation in carrier expectation, determinates of school effectiveness and peer effects.

E-mail: artur.pokropek@gmail.com