Single missing data imputation in PLS-SEM

Ned Kock

December 2014

ScriptWarp Systems™
Laredo, Texas
USA

Single missing data imputation in PLS-SEM

Ned Kock

Full reference:

Kock, N. (2014). *Single missing data imputation in PLS-SEM*. Laredo, TX: ScriptWarp Systems.

Abstract

An important source of bias in structural equation modeling (SEM) employing the partial least squares method (PLS) is missing data. Deletion methods, such as listwise and pairwise deletion, have traditionally been used to deal with missing data. These methods are perceived as leading to selective loss of data and significant related biases. Missing data imputation methods, on the other hand, do not resort to deletion. We discuss five single missing data imputation methods in the context of PLS-SEM employing the PLS Mode A algorithm. Among these five methods, two hierarchical methods are new. The results of a Monte Carlo experiment suggest that Multiple Regression Imputation yielded the least biased mean path coefficient estimates, followed by Arithmetic Mean Imputation. With respect to mean loading estimates, Arithmetic Mean Imputation yielded the least biased results, followed by Stochastic Hierarchical Regression Imputation and Hierarchical Regression Imputation. Our study suggests that single missing data imputation methods perform better with PLS-SEM than expected based on past research on their performance with other multivariate analysis techniques such as multiple regression and covariance-based SEM. The methods are implemented as part of the software WarpPLS, starting in version 5.0.

KEYWORDS: Partial Least Squares; Structural Equation Modeling; Missing Data Imputation; Path Bias; Stochastic Regression; Monte Carlo Simulation

Introduction

The method of partial least squares (PLS) has been experiencing explosive growth in the context of structural equation modeling (SEM). PLS-SEM estimates latent variables through composites, which are exact linear combinations of the indicators assigned to the latent variables. Parameter estimation in models with composites tends to be biased (Kock, 2014; 2014b), with the magnitude of the biases decreasing with increases in sample size, the number of indicators used, and their loadings.

Another source of bias in PLS-SEM is missing data (Newman, 2014). Among patterns of missing data, particularly common is that known as "missing at random" (MAR), which is actually a misnomer. This pattern occurs when the probability of a missing value is related to other measured variables, but unrelated to the underlying values of the variable that are missing.

Researchers have traditionally used deletion methods, often listwise and pairwise deletion, to deal with missing data (Enders, 2010). A report by the American Psychological Association Task Force on Statistical Inference stated that these techniques are ''among the worst methods available for practical applications'' (Wilkinson, 1999, p. 598).

Missing data imputation methods provide an alternative to deletion methods. Through imputation missing data elements are replaced with well informed "guesses", obtained through various algorithms, leading to no reduction in sample size. We discuss five single missing data imputation methods in the context of PLS-SEM, with MAR data. Among these five methods, two are new. The performance of the methods is comparatively assessed though a Monte Carlo experiment.

The missing data imputation methods are implemented in version 5.0 of WarpPLS, which is under intensive internal testing and nearing beta release at the time of this writing. WarpPLS is a SEM software tool that is unique in that it is the first and only (at the time of this writing) to enable nonlinear analyses where best-fitting nonlinear functions are estimated for each pair of structurally linked variables in path models, and subsequently used (i.e., the nonlinear functions) to estimate path coefficients that take into account the nonlinearity.

Linear analyses are supported by WarpPLS as well. The discussion presented here builds on linear analyses. WarpPLS also provides a comprehensive set of model fit and quality indices that are compatible with both composite-based and factor-based SEM. Starting in version 5.0 of WarpPLS, factor-based SEM algorithms will be available in addition to composite-based algorithms (Kock, 2014). Factor-based SEM is also generally known as covariance-based SEM.

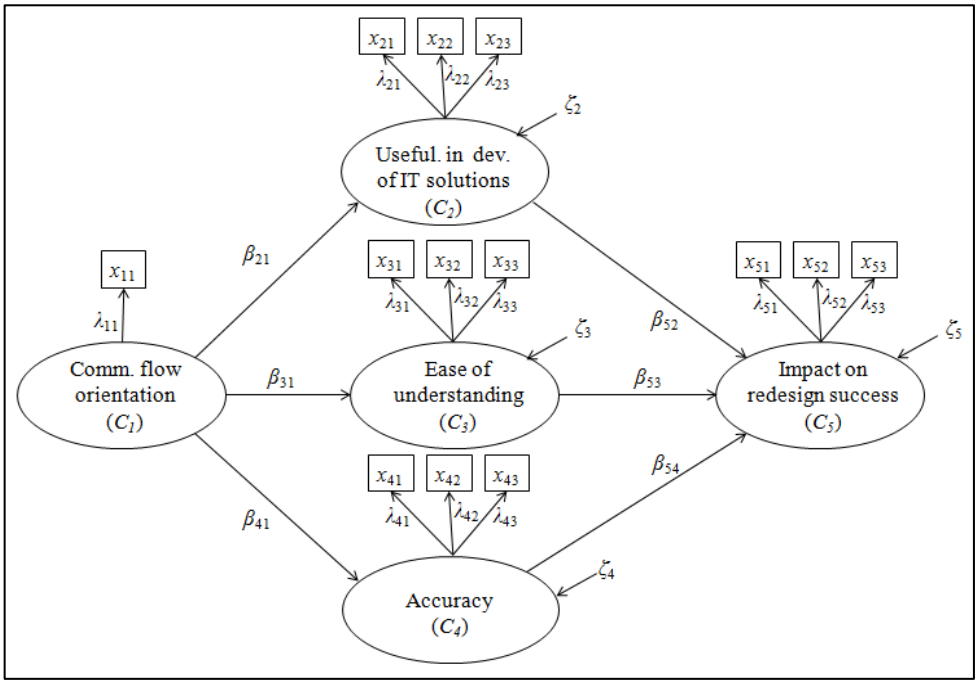
Illustrative model

We used an illustrative model to aid us in our presentation, as well as a basis for our Monte Carlo experiment and empirical illustration, which are discussed later. The illustrative model is depicted in Figure 1, and contains five latent variables, for which composites are estimated via PLS-SEM. The latent variables, which refer to theoretical constructs, are: communication flow orientation (C_1), usefulness in the development of information technology (IT) solutions (C_2), ease of understanding (C_3), accuracy (C_4), and impact on redesign success (C_5).

The mathematical symbols used in the model and in the following sections are adapted from the classic path analysis, covariance-based SEM, and PLS literatures: β_{ij} is the path coefficient for the link going from composite C_j to composite C_i, λ_{ij} is the loading for the *j*th indicator of

composite C_i, and ζ_i is the structural error associated with an endogenous composite C_i. With exception of communication flow orientation (C_1), a set of indicators x_{ij} is used to measure each composite C_i. When more than one indicator is used to measure a composite, each indicator is assumed to measure the composite with a certain degree of imprecision.

Figure 1. Illustrative model



Communication flow optimization theory (Danesh-Pajou, 2005; Kock, 2003) is the foundation on which the illustrative model is built. While the theory is not the focus of our investigation, it is useful for readers to know the theory's main prediction. The theory predicts that a greater focus on how communication takes place in business processes, in redesign efforts, is associated with better business process redesign outcomes.

Communication flow orientation (C_1) is the degree to which a business process modeling approach explicitly shows how communication interactions take place in a business process. This latent variable can be measured through a single indicator storing either 1 or 0, for a study contrasting two "opposite" modeling approaches, corresponding to a high or low communication flow orientation of a business process modeling approach used.

Usefulness in the development of IT solutions (C_2) is the degree to which a process modeling approach is useful in the development of a generic IT solution to automate the redesigned process. The need to automated redesigned processes with IT is almost universal in modern businesses. An example of question-statement that can be used for measurement of this latent variable is: "This process modeling approach is useful in the development of a generic IT solution to automate the redesigned process".

Ease of understanding (C_3) is the degree to which a process modeling approach is perceived to yield a process representation that is easy to understand. An example of question-statement that can be used for measurement of this latent variable is: "Processes modeled using this approach are easy to understand".

Accuracy (C_4) is the degree to which a process modeling approach is perceived to lead to an accurate representation of the process. An example of question-statement that can be used for measurement of this latent variable is: “This process modeling approach leads to accurate process representations”.

Finally, impact on redesign success (C_5) is the degree to which the process modeling technique used is perceived to lead to an actual improvement of the targeted business process. An example of question-statement that can be used for measurement of this latent variable is: “Using this process modeling approach is likely to contribute to the success of a process redesign project”.

Missing data imputation methods analyzed

In our analyses we focused on traditional single missing data imputation methods, plus two methods that we have developed. These new methods can be seen as “hierarchical” variations of two of the traditional methods. All of the missing data imputation methods are summarized below.

All variables are assumed to be standardized. This has no effect on the implementation of the methods; the methods can take as inputs unstandardized variables, store means and standard deviations for later unstandardization, standardize the variables, apply the various operations that define the methods, and finally unstandardize the variables again prior to generating the outputs.

Arithmetic Mean Imputation

Let x_i be a column vector denoting one of the k manifest variables used in a SEM model. The Arithmetic Mean Imputation (MEAN) method assigns values to each missing element $\dot{x}_{ir}$ according to (1), where N_m is the number of missing values in x_i, and $\bar{x}_i$ is the arithmetic mean of variable x_i.

$$\dot{x}_{ir} = \bar{x}_i, \quad r = 1 \dots N_m. \tag{1}$$

As its name implies, the Arithmetic Mean Imputation (MEAN) method replaces each missing element $\dot{x}_{ir}$ in a column of data i within a dataset, which refers to a manifest variable, with the average (or arithmetic mean) of that column. This method can be seen as the simplest of the imputation methods discussed here. While it can be employed by itself, this method also plays an ancillary role in other methods, as will be seen in the remainder of this section.

Multiple Regression Imputation

The Multiple Regression Imputation (MREGR) method assigns values to each missing element $\dot{x}_{ir}$ according to (2), where k is the number of manifest variables used in a model, N_m is the number of missing values in x_i, and each of the elements of the matrix of estimated regression coefficients $\hat{\beta}_{x_i x_j}$ is calculated through a multiple regression analysis with x_i as the criterion and x_j ($j = 1 \dots k, j \neq i$) as the predictors.

$$\dot{x}_{ir} = \sum_{j=1}^{k} \hat{\beta}_{x_i x_j} x_{jr}, \tag{2}$$

$$j = 1 \ldots k, j \neq i, r = 1 \ldots N_m.$$

In the Multiple Regression Imputation (MREGR) method each missing element $\dot{x}_{ir}$ is replaced with the corresponding expected value of x_i given all of the other variables x_j ($j = 1 \ldots k, j \neq i$) in the dataset. The regression coefficients $\hat{\beta}_{x_i x_j}$ for each variable x_i are obtained via a multiple regression analysis after an Arithmetic Mean Imputation (MEAN) is applied to the dataset.

An alternative to using Arithmetic Mean Imputation (MEAN), which tends to lead to an exacerbation of the biases and that is therefore not employed here, is to conduct the multiple regression analysis to obtain the regression coefficients $\hat{\beta}_{x_i x_j}$ after a listwise deletion. The use of deletion is particularly problematic here because the regression equation will typically have quite a few predictors, and thus a great deal of data may end up being lost after a listwise deletion.

Hierarchical Regression Imputation

This is one of the two new methods discussed here. The Hierarchical Regression Imputation (HREGR) method assigns values to each missing element $\dot{x}_{ir}$ according to (3), where k is the number of manifest variables used in a model, N_m is the number of missing values in x_i, and each of the elements of the matrix of estimated correlations $\hat{\Sigma}_{x_i x_j}$ is calculated after a pairwise deletion of missing elements is conducted for each pair of variables x_i and x_j. In this equation $max\left(\hat{\Sigma}_{x_i x_j}\right)$ is the maximum estimated correlation between the manifest variable x_i and any other manifest variable x_j for which a corresponding non-missing value x_{jr} exists.

$$\dot{x}_{ir} = max\left(\hat{\Sigma}_{x_i x_j}\right) x_{jr}, \tag{3}$$

$$j = 1 \ldots k, j \neq i, r = 1 \ldots N_m.$$

In the Hierarchical Regression Imputation (HREGR) method each missing element $\dot{x}_{ir}$ is replaced with the corresponding expected value of x_i given a variable x_j, stored in column j of the dataset, where x_j is the variable with the highest correlation with x_i after a pairwise deletion of missing elements.

Here a pairwise deletion is preferred over an Arithmetic Mean Imputation (MEAN) for the calculation of the correlations $\hat{\Sigma}_{x_i x_j}$ because it leads to less bias, as indicated by exploratory versions of this method that we developed and tested. In datasets with multiple variables and widespread missing data elements, pairwise deletions usually lead to much lesser amounts of data loss than listwise deletions. Nevertheless, the results of analyses conducted after pairwise deletions tend to be dependent on the pair-specific idiosyncrasies of missing data patterns.

Stochastic Multiple Regression Imputation

The Stochastic Multiple Regression Imputation (MSREG) method assigns values to each missing element $\dot{x}_{ir}$ according to (4), where k is the number of manifest variables used in a model, N_m is the number of missing values in x_i, and $Srandn(\quad)$ is a function that returns a different element of a standardized normally distributed random column vector each time it is invoked.

$$\dot{x}_{ir} = \sum_{j=1}^{k} \hat{\beta}_{x_i x_j} x_{jr} + \left(\sqrt{\left(1 - \sum_{j=1}^{k} \hat{\beta}_{x_i x_j} \hat{\Sigma}_{x_i x_j} \right)} \right) Srandn(\quad), \tag{4}$$

$$j = 1 \dots k, j \neq i, r = 1 \dots N_m.$$

The Stochastic Multiple Regression Imputation (MSREG) method is similar to the Multiple Regression Imputation (MREGR) method. The key difference is that in this stochastic variety, implemented via the equation above, normal random error is added to the new values due to the assumption that not doing so can create a downward bias in standard errors. Such a bias could lead to an exacerbation of type I errors. The random error elements yielded by $Srandn(\quad)$ are weighted so that they collectively account for all of the variance in x_i that is not explained by the predictors x_j ($j = 1 \dots k, j \neq i$).

While the above assumption regarding standard error bias may be a reasonable one with respect to standard multiple regression and covariance-based SEM, in PLS-SEM path coefficients tend to present downward biases even without missing data. Therefore a downward bias in standard errors may compensate for the related decrease in statistical power, due to the downward path coefficient bias, in turn countering an exacerbation in type II errors (and a reduction in power).

Stochastic Hierarchical Regression Imputation

This is the other of the two new methods discussed here. The Stochastic Hierarchical Regression Imputation (HSREG) method assigns values to each missing element $\dot{x}_{ir}$ according to (5), where k is the number of manifest variables used in a model, N_m is the number of missing values in x_i, and $Srandn(\quad)$ is a function that returns a different element of a standardized normally distributed random column vector each time it is invoked.

$$\dot{x}_{ir} = max\left(\hat{\Sigma}_{x_i x_j} \right) x_{jr} + \left(\sqrt{1 - max\left(\hat{\Sigma}_{x_i x_j} \right)^2} \right) Srandn(\quad), \tag{5}$$

$$j = 1 \dots k, j \neq i, r = 1 \dots N_m.$$

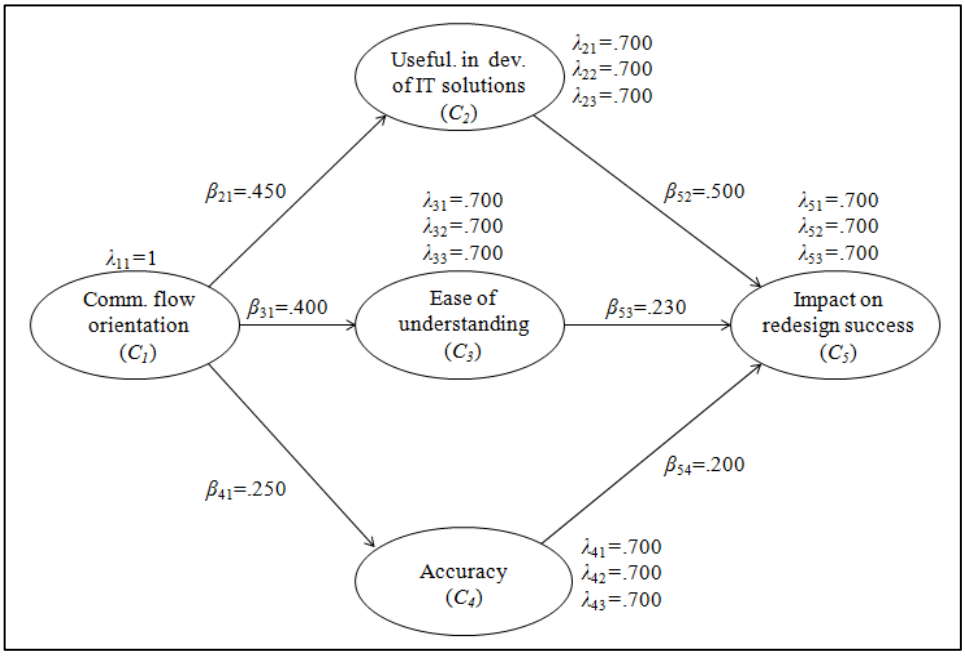
The Stochastic Hierarchical Regression Imputation (HSREG) method is similar to the Hierarchical Regression Imputation (HREGR) method. The key difference (analogously to the discussion above) in this stochastic variety is that normal random error is added to the new values due to the assumption that not doing so can create a downward bias in standard errors and

an overall deleterious effect on type I error rates. Again, while this assumption may find general application in standard multiple regression and covariance-based SEM, it may not readily apply to PLS-SEM.

Monte Carlo experiment

A Monte Carlo experiment based on the true population model shown in Figure 2 was conducted to assess the performance of the five missing data imputation methods discussed in the previous section. Performance was assessed in terms of path coefficient bias and standard error inflation. This Monte Carlo experiment was conducted as part of extensive internal tests of version 5.0 of WarpPLS.

Figure 2. True population model



When creating data for our Monte Carlo experiment we varied the following conditions: percentage of missing data (0%, 30%, 40%, and 50%), and sample size (100, 300, and 500). This led to a 4 x 3 factorial design, with 12 conditions. We created an analyzed 1,000 samples for each of these 12 conditions; a total of 12,000 samples.

The PLS Mode A algorithm with the path weighting scheme (Lohmöller, 1989) was used in the analyses. These are the most widely used algorithm (PLS Mode A) and inner model estimation scheme (path weighting) in the context of PLS-SEM. Results were obtained for analyses with no missing data (NMD), Arithmetic Mean Imputation (MEAN), Multiple Regression Imputation (MREGR), Hierarchical Regression Imputation (HREGR), Stochastic Multiple Regression Imputation (MSREG), and Stochastic Hierarchical Regression Imputation (HSREG).

A summarized set of results is shown in Table 1, where we restrict ourselves to $N = 300$ and 30% missing data (MAR). True path coefficients, mean path coefficient estimates, and standard errors of path coefficient estimates are shown next to one another. Full results, for all percentages of missing data and sample sizes included in the simulation, are available in

Appendix A. Since all loadings are the same in the true population model, loading-related estimates for only one indicator of the composites are shown. This avoids crowding and repetition, as the same pattern of results repeats itself in connection with all loadings.

Table 1. Summarized Monte Carlo experiment results ($N = 300$, 30% MAR data)

Missing data imputation scheme	NMD	MEAN	MREGR	HREGR	MSREG	HSREG
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450
CO>GT(AvgPath)	.390	.348	.367	.354	.333	.300
CO>GT(SEPath)	.075	.113	.110	.113	.138	.162
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.349	.312	.321	.313	.289	.262
CO>EU(SEPath)	.069	.101	.108	.106	.133	.151
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250
CO>AC(AvgPath)	.219	.198	.206	.195	.188	.161
CO>AC(SEPath)	.062	.078	.090	.083	.100	.108
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.381	.357	.359	.352	.334	.312
GT>SU(SEPath)	.127	.152	.156	.158	.179	.195
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.192	.183	.199	.178	.188	.163
EU>SU(SEPath)	.062	.072	.077	.078	.082	.089
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.165	.157	.176	.154	.166	.141
AC>SU(SEPath)	.058	.067	.073	.072	.077	.081
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.811	.691	.606	.649	.623	.652
GT3<GT(SELoad)	.113	.042	.120	.076	.115	.090

Notes: NMD = no missing data; MEAN = Arithmetic Mean Imputation; MREGR = Multiple Regression Imputation; HREGR = Hierarchical Regression Imputation; MSREG = Stochastic Multiple Regression Imputation; HSREG = Stochastic Hierarchical Regression Imputation; XX>YY = link from composite XX to YY; CO = communication flow orientation (C_1); GT = usefulness in the development of IT solutions (C_2); EU = ease of understanding (C_3); AC = accuracy (C_4); SU = impact on redesign success (C_5); TruePath = true path coefficient; AvgPath = mean path coefficient estimate; SEPath = standard error of path coefficient estimate; TrueLoad = true loading; AvgLoad = mean loading estimate; SELoad = standard error of loading estimate.

The mean path coefficient estimates that are shown underlined were obtained through the application of the PLS Mode A algorithm to datasets where no data was missing (NMD). Note that they generally underestimate the true path coefficients. This underestimation stems from the use of composites in PLS-SEM, discussed earlier, which leads to an attenuation of composite *correlations* (Nunnally & Bernstein, 1994). This correlation attenuation extends to the path coefficients (Kock, 2014), leading to the observed underestimation. The opposite effect is observed in connection with loadings, which tend to be overestimated in PLS-SEM.

Multiple Regression Imputation (MREGR) yielded the least biased mean path coefficient estimates, followed by Arithmetic Mean Imputation (MEAN). When we look at mean loading estimates, Arithmetic Mean Imputation (MEAN) yielded the least biased results, followed by Stochastic Hierarchical Regression Imputation (HSREG) and Hierarchical Regression Imputation (HREGR.

Compared with the no missing data condition (NMD), none of the methods induced a significant reduction in standard errors. This is noteworthy since prior results outside the context of PLS-SEM have tended to show a significant downward bias in standard errors, particularly for

non-stochastic varieties. Such downward bias in standard errors has led to concerns regarding an inflation in type I errors, and warnings against the use of single missing data imputation methods in general (Enders, 2010; Newman, 2014).

Empirical illustration

Table 2 summarizes the results of an empirical field study related to the illustrative and true population models discussed earlier. The field study in fact served as the basis for the development of the illustrative and true population models. Shown next to one another are estimated path coefficients (top part of the table), and loadings (bottom part of the table). All path coefficients and loadings are shown. Except for the column "NMD", all other columns show results with 30% missing data (MAR).

Table 2. Empirical study results

Missing data imputation scheme	NMD	MEAN	HREGR	HSREG	MREGR	MSREG
CO>GT	.485 [a]	.427 [a]	.472 [a]	.445 [a]	.462 [a]	.379 [a]
CO>EU	.362 [a]	.244 [a]	.282 [a]	.313 [a]	.248 [a]	.263 [a]
CO>AC	.269 [a]	.184 [b]	.209 [b]	.183 [b]	.195 [b]	.213 [b]
GT>SU	.506 [a]	.531 [a]	.536 [a]	.527 [a]	.532 [a]	.493 [a]
EU>SU	.217 [b]	.184 [b]	.204 [b]	.233 [b]	.187 [b]	.174 [c]
AC>SU	.194 [b]	.181 [b]	.150 [c]	.146 [c]	.173 [c]	.170 [c]
GT1<GT	.926	.854	.938	.883	.899	.900
GT2<GT	.880	.883	.919	.887	.897	.863
GT3<GT	.893	.878	.929	.885	.907	.855
EU1<EU	.796	.740	.815	.801	.786	.742
EU2<EU	.875	.831	.853	.816	.862	.827
EU3<EU	.910	.884	.909	.901	.903	.871
AC1<AC	.916	.926	.925	.918	.926	.926
AC2<AC	.868	.812	.863	.847	.840	.794
AC3<AC	.753	.674	.723	.634	.703	.677
SU1<SU	.937	.914	.950	.913	.934	.895
SU2<SU	.947	.934	.957	.916	.949	.919
SU3<SU	.932	.913	.944	.925	.933	.908

Notes: N = 156; [a] P < .001, [b] P < .01, [c] P < .05; PLS algorithm used = PLS Mode A; P values calculated via bootstrapping with 500 resamples; NMD = no missing data; MEAN = Arithmetic Mean Imputation; MREGR = Multiple Regression Imputation; HREGR = Hierarchical Regression Imputation; MSREG = Stochastic Multiple Regression Imputation; HSREG = Stochastic Hierarchical Regression Imputation; XX>YY = link from variable XX to YY; CO = communication flow orientation (C_1); GT = usefulness in the development of IT solutions (C_2); EU = ease of understanding (C_3); AC = accuracy (C_4); SU = impact on redesign success (C_5); XX1 … XXn = indicators associated with composite XX.

The data for this empirical study was collected from 156 individuals who participated in various business process redesign projects in organizations located in Northeastern U.S.A. The participants employed one of two business process modeling approaches. One of the modeling approaches focused primarily on the communication flow within business processes. The other focused primarily on the chronological flow of activities. Both approaches are illustrated in Appendix B. Appendix C has the questionnaire used for data collection.

Overall, all missing data imputation methods analyzed yielded estimates consistent with communication flow optimization theory (Kock, 2003). No method led to biases that were severe

enough, at 30% missing data, to generate non-significant *P* values. Given this, we could say that the empirical study results provide "real data" validation of all imputation methods, and to a certain extend qualified support for all of them. This is because the theory, which forms the underlying theoretical foundation for the model, has been validated before in multiple empirical studies employing different datasets and methods (Danesh-Pajou, 2005; Danesh-Pajou & Kock, 2005; Kock et al., 2008; 2009).

Discussion and conclusion

An important source of bias in PLS-SEM is missing data. Deletion methods, such as listwise and pairwise deletion, have traditionally been used to deal with missing data. While these methods are perceived as problematic because they can lead to reductions in sample size, particularly problematic are the possible biases that they can introduce. For example, missing data may be associated with groups of respondents who share some characteristics, and whose exclusion from datasets can significantly influence the strength of relationships among variables.

We discussed and compared five single missing data imputation methods in the context of PLS-SEM: Arithmetic Mean Imputation (MEAN), Multiple Regression Imputation (MREGR), Hierarchical Regression Imputation (HREGR), Stochastic Multiple Regression Imputation (MSREG), and Stochastic Hierarchical Regression Imputation (HSREG). Two of these methods are new – the hierarchical varieties (HREGR and HSREG). The relative performance of the methods was assessed though a Monte Carlo experiment.

The results from the Monte Carlo experiment suggest that Multiple Regression Imputation (MREGR) yielded the least biased mean path coefficient estimates, followed by Arithmetic Mean Imputation (MEAN). With respect to mean loading estimates, Arithmetic Mean Imputation (MEAN) yielded the least biased results, followed by Stochastic Hierarchical Regression Imputation (HSREG) and Hierarchical Regression Imputation (HREGR).

None of the methods induced a significant reduction in standard errors when compared with the no missing data condition (NMD). This is at odds with past results outside the context of PLS-SEM, which tended to show a significant downward bias in standard errors, particularly for non-stochastic imputation methods. This observed downward bias in standard errors has led to concerns regarding type I error inflation, and admonitions against the use of single missing data imputation methods in general. Our results suggest that PLS-SEM may be a fertile ground for the application of single missing data imputation methods, although more research is needed to shed light as to whether this is truly the case and why.

Users of WarpPLS, starting in version 5.0, will be able to test the methods for themselves. We hope that the discussion presented here will provide enough details for implementations of the methods in numerical programming environments such as R and GNU Octave. As these are developed and tested under various conditions, we welcome comments, suggestions, and corrections.

Acknowledgments

The author is the developer of the software WarpPLS, which has over 7,000 users in more than 33 different countries at the time of this writing, and moderator of the PLS-SEM e-mail distribution list. He is grateful to those users, and to the members of the PLS-SEM e-mail distribution list, for questions, comments, and discussions on topics related to SEM and to the use of WarpPLS.

References

Danesh-Pajou, A. (2005). *IT-enabled process redesign: Using communication flow optimization theory in an information intensive environment*. Doctoral Dissertation. Philadelphia, PA: Temple University.

Danesh-Pajou, A., & Kock, N. (2005). An experimental study of process representation approaches and their impact on perceived modeling quality and redesign success. *Business Process Management Journal*, 11(6), 724-735.

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

Kock, N. (2003). Communication-focused business process redesign: Assessing a communication flow optimization model through an action research study at a defense contractor. *IEEE Transactions on Professional Communication*, 46(1), 35-54.

Kock, N. (2014). *A note on how to conduct a factor-based PLS-SEM analysis*. Laredo, TX: ScriptWarp Systems.

Kock, N. (2014b). *One-tailed or two-tailed P values in PLS-SEM?* Laredo, TX: ScriptWarp Systems.

Kock, N., Danesh-Pajou, A., & Komiak, P. (2008). A discussion and test of a communication flow optimization approach for business process redesign. *Knowledge and Process Management*, 15(1), 72-85.

Kock, N., Verville, J., Danesh-Pajou, A., & DeLuca, D. (2009). Communication flow orientation in business process modeling and its effect on redesign success: Results from a field study. *Decision Support Systems*, 46(2), 562-575.

Lohmöller, J.-B. (1989). *Latent variable path modeling with partial least squares*. Heidelberg, Germany: Physica-Verlag.

Newman, D.A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, 17(4), 372-411.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American psychologist*, 54(8), 594-604.

Appendix A: Full Monte Carlo experiment results

The full Monte Carlo experiment results are provided in the tables below. Notes: NMD = no missing data; MEAN = Arithmetic Mean Imputation; MREGR = Multiple Regression Imputation; HREGR = Hierarchical Regression Imputation; MSREG = Stochastic Multiple Regression Imputation; HSREG = Stochastic Hierarchical Regression Imputation; XX>YY = link from composite XX to YY; CO = communication flow orientation (C_1); GT = usefulness in the development of IT solutions (C_2); EU = ease of understanding (C_3); AC = accuracy (C_4); SU = impact on redesign success (C_5); TruePath = true path coefficient; AvgPath = mean path coefficient estimate; SEPath = standard error of estimate; TrueLoad = true loading; AvgLoad = mean loading estimate; SELoad = standard error of estimate.

Sample size	**100**	**100**	**100**	**100**	**100**	**100**
Percentage of missing data	**0%**	**30%**	**30%**	**30%**	**30%**	**30%**
Missing data imputation scheme	**NMD**	**MEAN**	**HREGR**	**HSREG**	**MREGR**	**MSREG**
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450
CO>GT(AvgPath)	.394	.354	.364	.308	.362	.327
CO>GT(SEPath)	.094	.129	.133	.175	.148	.171
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.355	.323	.326	.280	.335	.308
CO>EU(SEPath)	.096	.120	.130	.161	.145	.156
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250
CO>AC(AvgPath)	.227	.205	.205	.172	.214	.196
CO>AC(SEPath)	.093	.111	.124	.140	.148	.153
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.384	.355	.353	.319	.351	.328
GT>SU(SEPath)	.141	.170	.178	.206	.188	.206
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.193	.188	.187	.172	.207	.196
EU>SU(SEPath)	.094	.103	.112	.121	.121	.129
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.172	.165	.167	.150	.193	.183
AC>SU(SEPath)	.091	.107	.114	.123	.130	.134
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.810	.687	.645	.644	.593	.603
GT3<GT(SELoad)	.118	.072	.105	.128	.156	.165

Sample size	**100**	**100**	**100**	**100**	**100**	**100**
Percentage of missing data	**0%**	**40%**	**40%**	**40%**	**40%**	**40%**
Missing data imputation scheme	**NMD**	**MEAN**	**HREGR**	**HSREG**	**MREGR**	**MSREG**
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450
CO>GT(AvgPath)	.394	.309	.315	.247	.307	.264
CO>GT(SEPath)	.094	.188	.193	.240	.223	.251
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.355	.280	.283	.225	.275	.240
CO>EU(SEPath)	.096	.185	.194	.226	.219	.239
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250

Sample size	100	100	100	100	100	100
Percentage of missing data	0%	40%	40%	40%	40%	40%
Missing data imputation scheme	NMD	MEAN	HREGR	HSREG	MREGR	MSREG
CO>AC(AvgPath)	.227	.186	.182	.145	.189	.165
CO>AC(SEPath)	.093	.170	.188	.185	.208	.211
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.384	.320	.324	.272	.311	.280
GT>SU(SEPath)	.141	.222	.227	.263	.246	.270
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.193	.191	.189	.163	.189	.178
EU>SU(SEPath)	.094	.144	.157	.163	.186	.195
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.172	.177	.177	.146	.186	.164
AC>SU(SEPath)	.091	.157	.172	.170	.204	.208
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.810	.479	.440	.444	.395	.398
GT3<GT(SELoad)	.118	.261	.295	.306	.347	.359

Sample size	100	100	100	100	100	100
Percentage of missing data	0%	50%	50%	50%	50%	50%
Missing data imputation scheme	NMD	MEAN	HREGR	HSREG	MREGR	MSREG
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450
CO>GT(AvgPath)	.394	.241	.248	.170	.227	.183
CO>GT(SEPath)	.094	.272	.287	.327	.323	.345
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.355	.215	.211	.145	.190	.159
CO>EU(SEPath)	.096	.263	.284	.308	.323	.327
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250
CO>AC(AvgPath)	.227	.146	.151	.110	.136	.113
CO>AC(SEPath)	.093	.227	.242	.228	.276	.270
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.384	.267	.263	.208	.238	.207
GT>SU(SEPath)	.141	.292	.303	.337	.351	.359
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.193	.172	.168	.137	.163	.139
EU>SU(SEPath)	.094	.212	.239	.213	.264	.259
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.172	.152	.149	.118	.153	.135
AC>SU(SEPath)	.091	.219	.242	.213	.270	.263
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.810	.284	.250	.263	.217	.214
GT3<GT(SELoad)	.118	.451	.480	.483	.511	.526

Sample size	300	300	300	300	300	300
Percentage of missing data	0%	30%	30%	30%	30%	30%
Missing data imputation scheme	NMD	MEAN	HREGR	HSREG	MREGR	MSREG
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450

Sample size	**300**	**300**	**300**	**300**	**300**	**300**
Percentage of missing data	**0%**	**30%**	**30%**	**30%**	**30%**	**30%**
Missing data imputation scheme	**NMD**	**MEAN**	**HREGR**	**HSREG**	**MREGR**	**MSREG**
CO>GT(AvgPath)	.390	.348	.354	.300	.367	.333
CO>GT(SEPath)	.075	.113	.113	.162	.110	.138
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.349	.312	.313	.262	.321	.289
CO>EU(SEPath)	.069	.101	.106	.151	.108	.133
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250
CO>AC(AvgPath)	.219	.198	.195	.161	.206	.188
CO>AC(SEPath)	.062	.078	.083	.108	.090	.100
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.381	.357	.352	.312	.359	.334
GT>SU(SEPath)	.127	.152	.158	.195	.156	.179
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.192	.183	.178	.163	.199	.188
EU>SU(SEPath)	.062	.072	.078	.089	.077	.082
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.165	.157	.154	.141	.176	.166
AC>SU(SEPath)	.058	.067	.072	.081	.073	.077
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.811	.691	.649	.652	.606	.623
GT3<GT(SELoad)	.113	.042	.076	.090	.120	.115

Sample size	**300**	**300**	**300**	**300**	**300**	**300**
Percentage of missing data	**0%**	**40%**	**40%**	**40%**	**40%**	**40%**
Missing data imputation scheme	**NMD**	**MEAN**	**HREGR**	**HSREG**	**MREGR**	**MSREG**
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450
CO>GT(AvgPath)	.390	.309	.311	.240	.308	.264
CO>GT(SEPath)	.075	.160	.165	.224	.173	.209
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.349	.273	.274	.211	.271	.234
CO>EU(SEPath)	.069	.147	.152	.204	.162	.191
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250
CO>AC(AvgPath)	.219	.176	.174	.132	.178	.156
CO>AC(SEPath)	.062	.113	.116	.142	.129	.138
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.381	.323	.320	.264	.314	.282
GT>SU(SEPath)	.127	.191	.196	.246	.207	.235
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.192	.186	.180	.157	.201	.184
EU>SU(SEPath)	.062	.087	.094	.101	.096	.099
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.165	.161	.161	.138	.180	.163
AC>SU(SEPath)	.058	.083	.085	.097	.099	.103
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.811	.496	.461	.475	.423	.440
GT3<GT(SELoad)	.113	.221	.256	.253	.296	.286

Sample size	**300**	**300**	**300**	**300**	**300**	**300**
Percentage of missing data	**0%**	**50%**	**50%**	**50%**	**50%**	**50%**
Missing data imputation scheme	**NMD**	**MEAN**	**HREGR**	**HSREG**	**MREGR**	**MSREG**
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450
CO>GT(AvgPath)	.390	.243	.252	.176	.243	.193
CO>GT(SEPath)	.075	.229	.226	.288	.246	.284
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.349	.217	.223	.152	.213	.172
CO>EU(SEPath)	.069	.209	.208	.264	.230	.260
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250
CO>AC(AvgPath)	.219	.145	.150	.099	.143	.112
CO>AC(SEPath)	.062	.150	.154	.179	.180	.194
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.381	.271	.273	.212	.264	.227
GT>SU(SEPath)	.127	.246	.249	.300	.263	.295
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.192	.183	.185	.143	.194	.168
EU>SU(SEPath)	.062	.104	.114	.130	.134	.138
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.165	.160	.159	.126	.171	.151
AC>SU(SEPath)	.058	.112	.123	.124	.141	.137
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.811	.329	.296	.311	.256	.268
GT3<GT(SELoad)	.113	.386	.417	.412	.456	.453

Sample size	**500**	**500**	**500**	**500**	**500**	**500**
Percentage of missing data	**0%**	**30%**	**30%**	**30%**	**30%**	**30%**
Missing data imputation scheme	**NMD**	**MEAN**	**HREGR**	**HSREG**	**MREGR**	**MSREG**
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450
CO>GT(AvgPath)	.389	.346	.352	.296	.363	.328
CO>GT(SEPath)	.070	.110	.109	.162	.104	.135
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.343	.308	.309	.258	.317	.286
CO>EU(SEPath)	.067	.100	.102	.149	.102	.129
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250
CO>AC(AvgPath)	.219	.197	.192	.159	.204	.183
CO>AC(SEPath)	.052	.070	.077	.103	.077	.090
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.380	.354	.348	.309	.358	.333
GT>SU(SEPath)	.124	.151	.157	.196	.151	.175
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.189	.180	.176	.160	.198	.184
EU>SU(SEPath)	.055	.064	.070	.083	.065	.073
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.164	.154	.151	.137	.174	.164
AC>SU(SEPath)	.054	.063	.067	.077	.061	.067
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.811	.692	.652	.654	.609	.627
GT3<GT(SELoad)	.113	.035	.069	.082	.113	.106

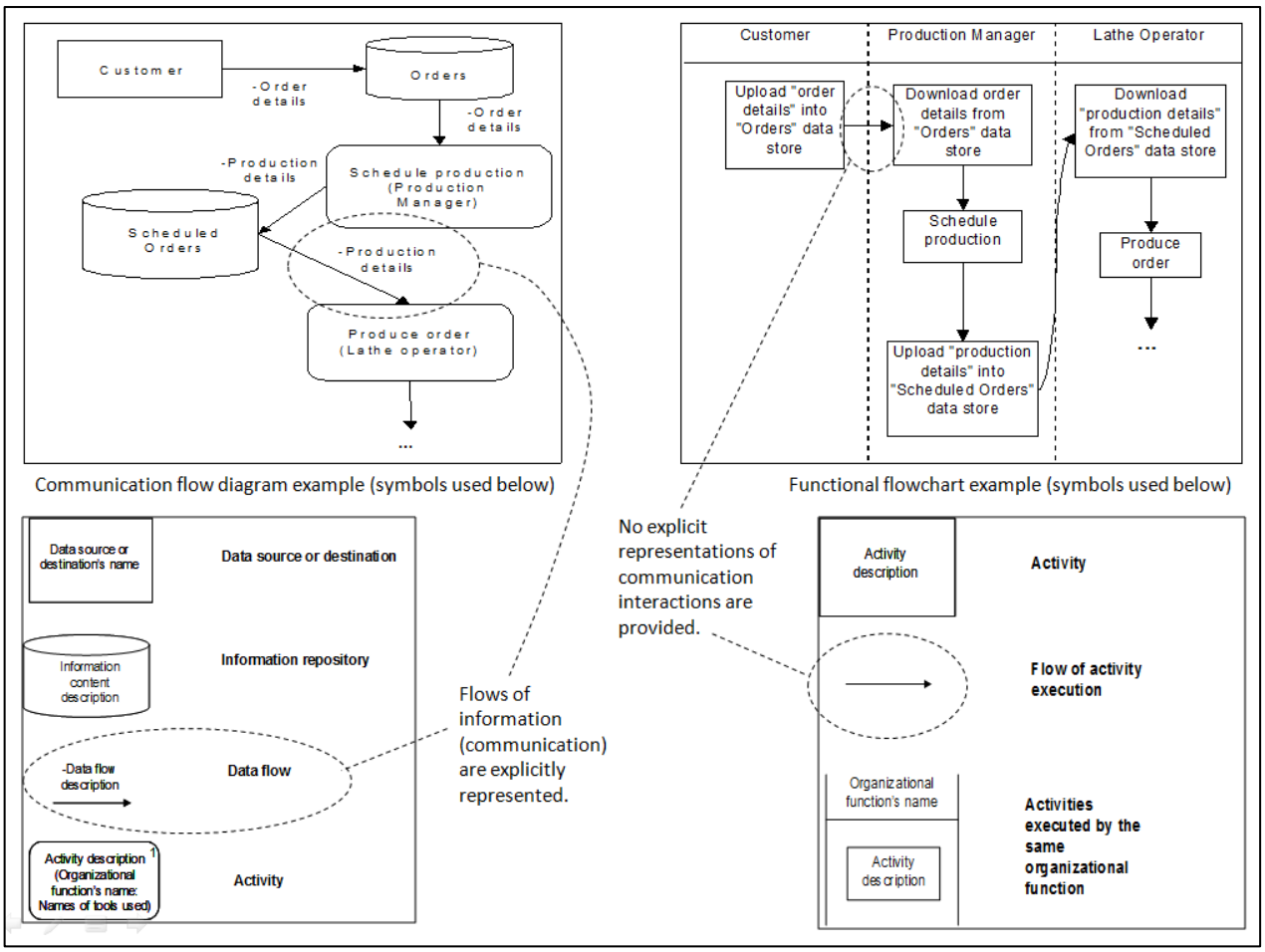
Sample size	**500**	**500**	**500**	**500**	**500**	**500**
Percentage of missing data	**0%**	**40%**	**40%**	**40%**	**40%**	**40%**
Missing data imputation scheme	**NMD**	**MEAN**	**HREGR**	**HSREG**	**MREGR**	**MSREG**
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450
CO>GT(AvgPath)	.389	.307	.308	.236	.307	.265
CO>GT(SEPath)	.070	.155	.158	.223	.164	.201
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.343	.270	.267	.205	.267	.230
CO>EU(SEPath)	.067	.145	.151	.205	.157	.188
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250
CO>AC(AvgPath)	.219	.174	.171	.129	.175	.151
CO>AC(SEPath)	.052	.098	.104	.135	.109	.125
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.380	.321	.315	.260	.312	.280
GT>SU(SEPath)	.124	.187	.194	.246	.200	.230
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.189	.181	.178	.152	.194	.177
EU>SU(SEPath)	.055	.078	.082	.097	.084	.090
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.164	.161	.157	.134	.178	.163
AC>SU(SEPath)	.054	.072	.076	.088	.078	.082
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.811	.501	.468	.486	.433	.455
GT3<GT(SELoad)	.113	.213	.245	.237	.281	.267

Sample size	**500**	**500**	**500**	**500**	**500**	**500**
Percentage of missing data	**0%**	**50%**	**50%**	**50%**	**50%**	**50%**
Missing data imputation scheme	**NMD**	**MEAN**	**HREGR**	**HSREG**	**MREGR**	**MSREG**
CO>GT(TruePath)	.450	.450	.450	.450	.450	.450
CO>GT(AvgPath)	.389	.245	.250	.171	.238	.193
CO>GT(SEPath)	.070	.218	.218	.288	.236	.274
CO>EU(TruePath)	.400	.400	.400	.400	.400	.400
CO>EU(AvgPath)	.343	.213	.216	.150	.209	.168
CO>EU(SEPath)	.067	.205	.206	.260	.218	.251
CO>AC(TruePath)	.250	.250	.250	.250	.250	.250
CO>AC(AvgPath)	.219	.143	.144	.098	.140	.113
CO>AC(SEPath)	.052	.133	.137	.168	.154	.168
GT>SU(TruePath)	.500	.500	.500	.500	.500	.500
GT>SU(AvgPath)	.380	.270	.270	.206	.263	.227
GT>SU(SEPath)	.124	.240	.243	.301	.254	.285
EU>SU(TruePath)	.230	.230	.230	.230	.230	.230
EU>SU(AvgPath)	.189	.172	.170	.134	.183	.158
EU>SU(SEPath)	.055	.098	.103	.119	.105	.115
AC>SU(TruePath)	.200	.200	.200	.200	.200	.200
AC>SU(AvgPath)	.164	.157	.158	.127	.175	.151
AC>SU(SEPath)	.054	.090	.095	.103	.104	.109
GT3<GT(TrueLoad)	.700	.700	.700	.700	.700	.700
GT3<GT(AvgLoad)	.811	.339	.307	.322	.267	.285

Sample size	**500**	**500**	**500**	**500**	**500**	**500**
Percentage of missing data	**0%**	**50%**	**50%**	**50%**	**50%**	**50%**
Missing data imputation scheme	**NMD**	**MEAN**	**HREGR**	**HSREG**	**MREGR**	**MSREG**
GT3<GT(SELoad)	.113	.373	.403	.395	.443	.431

Appendix B: Business process modeling approaches used

The figure below illustrates the two types of representations used in the business process redesign projects. In the context of our data analyses example, the one on the left was coded as 1, and the one on the right as 0. They correspond to high and low communication flow orientations, respectively, of the business process modeling approach used.



Appendix C: Questionnaire used in empirical study

The question-statements below were used for latent variable measurement in the illustrative study. Except for communication flow orientation (C_1), all question-statements were answered on 7-point Likert-type scales.

Communication flow orientation (C_1)

- C_{11}: Coded as either 1 or 0, corresponding to high or low communication flow orientation of the business process modeling approach used.

Usefulness in the development of IT solutions (C_2)

- C_{21}: This process modeling approach is useful in the development of a generic IT solution to automate the redesigned process.
- C_{22}: Creating a generic IT solution to enable the redesigned process is easy based on this process modeling approach.
- C_{23}: Graphical process representations using this approach facilitate the generation of a generic IT solution to automate the redesigned process.

Ease of understanding (C_3)

- C_{31}: Processes modeled using this approach are easy to understand.
- C_{32}: Graphical representations of processes using this approach are clear.
- C_{33}: This process modeling approach leads to graphical models that are easy to understand.

Accuracy (C_4)

- C_{41}: This process modeling approach leads to accurate process representations.
- C_{42}: Models created using this approach are correct representations of a process.
- C_{43}: Graphical representations using this approach clearly reflect the real process.

Impact on redesign success (C_5)

- C_{51}: Using this process modeling approach is likely to contribute to the success of a process redesign project.
- C_{52}: Success chances are improved if this process modeling approach is used.
- C_{53}: Using the graphical process representations in this approach is likely to make process redesign projects more successful.