Statistics Surveys
Vol. 16 (2022) 182–209
ISSN: 1935-7516
https://doi.org/10.1214/22-SS137

General-purpose imputation of planned missing data in social surveys: Different strategies and their effect on correlations*

Julian B. Axenfeld

Mannheim Centre for European Social Research,
A5, 6, 68159 Mannheim, Germany

e-mail: julian.axenfeld@uni-mannheim.de
and

Christian Bruch

GESIS Leibniz Institute for the Social Sciences
B6, 4-5, 68159 Mannheim, Germany

Mannheim Centre for European Social Research,
University of Mannheim,
A5, 6, 68159 Mannheim, Germany

e-mail: christian.bruch@gesis.org
and

Christof Wolf

GESIS Leibniz Institute for the Social Sciences
B6, 4-5, 68159 Mannheim, Germany

Mannheim Centre for European Social Research,
University of Mannheim,
A5, 6, 68159 Mannheim, Germany

e-mail: christof.wolf@gesis.org

Abstract: Planned missing survey data, for example stemming from split questionnaire designs are becoming increasingly common in survey research, making imputation indispensable to obtain reasonably analyzable data. However, these data can be difficult to impute due to low correlations, many predictors, and limited sample sizes to support imputation models. This paper presents findings from a Monte Carlo simulation, in which we investigate the accuracy of correlations after multiple imputation using different imputation methods and predictor set specifications based on data from the German Internet Panel (GIP). The results show that strategies that simplify the imputation exercise (such as predictive mean matching with dimensionality reduction or restricted predictor sets, linear regression

* This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) [project numbers: BL 1148/1-1, BR 5869/1-1, WO 739/20-1]. The Monte Carlo simulations were run on the High Performance Computing facilities of the state of Baden-Württemberg (bwHPC). This paper uses data from the German Internet Panel (GIP) funded by the DFG through the Collaborative Research Center (SFB) 884 "Political Economy of Reforms" (SFB 884) [Project-ID: 139943784].

models, or the multivariate normal model without transformation) perform well, while especially generalized linear models for categorical data, classification trees, and imputation models with many predictor variables lead to strong biases.

MSC2020 subject classifications: Primary 62D10; secondary 65C05, 62P25.
Keywords and phrases: Bias, Monte Carlo simulation, multiple imputation, imputation methods, split questionnaire design.

Received February 2022.

1. Introduction

Long questionnaires pose a serious threat to the quality of survey data, triggering low response rates and poor response quality [22, 41]. Recently, survey projects such as the PISA 2012 context questionnaire [40, pp. 48-58] or the European Values Study [34] have attempted to overcome this problem with methods such as the split questionnaire design (SQD) [43]. In an SQD survey, a long questionnaire is split into different overlapping, shorter questionnaires. Consequently, respondents receive only a part of the full questionnaire while all bivariate combinations of variables and their covariances are observed. Obviously, this results in a large amount of planned missing data (i.e., data that intentionally remain unobserved). As a result, dropping the incomplete cases from the analysis (listwise deletion) is usually unfeasible with SQD data, since in SQDs fully observed cases are rare or nonexistent. Therefore, SQD surveys require appropriate methods to deal with the intentionally unobserved data.

Multiple imputation (MI) [47] is one of the state-of-the-art methods for handling missing data. Based on an imputation model, MI replaces missing values with multiple potential values drawn from the joint distribution of the data. Given an adequately specified imputation model, data imputed via MI can be analyzed through standard statistical techniques. Yet, from a practical perspective the responsibility of imputing SQD data cannot easily be shifted to the data user, as only a minority of users are experts for imputation. Furthermore, it can be argued that in the interest of transparent, replicable and cumulative research it would be beneficial if researchers were able to work with the same imputed data. This means that it could be beneficial if the data is published with imputed data for general research purposes, giving data users with different substantive interests a reliable basis for their analysis. However, as we argue in the following paragraphs, more research is needed to determine which imputation strategies can adequately handle such data scenarios in practice.

A general-purpose imputation of SQD data faces the following challenges: First, imputation models ideally should cover all variable relations studied in an analysis model. If variable relations that are omitted in the imputation are included in an analysis model, they will be biased towards zero unless the true relationship is equal to zero [7]. For our scenario of a general-purpose imputation this means using all available variables as predictors, because they may be included in a researcher's substantive analysis model. However, to impute large numbers of variables with large predictor sets, large samples are needed. This often will not be the case for SQD.

Second, because analyses of SQD data largely rely on imputed data, the selection of the imputation strategy is crucial, for even minor misspecifications in the imputation model could significantly damage the estimates.

Third, noisy data and especially low correlations are common features of social surveys, even though the exact conditions may vary depending on a survey's content and measurement scales. This complicates the definition of accurate imputation models since SQD data typically will contain only limited information that can be utilized for imputation.

In sum, an adequate imputation strategy must deal with potentially huge predictor sets but limited sample sizes, comparatively little information input, and the threat to distort relations in the overall data. Axenfeld et al. [5], for example, observe in a Monte Carlo simulation that especially relationships between variables (more so than univariate distributions) can turn out considerably biased in imputed SQD data. In another real-data simulation of an SQD, Bahrami et al. [6] report regression coefficients with complete and imputed SQD data, also revealing systematic biases in most coefficients. Hence, it is necessary to evaluate which simplifying assumptions must be made in the imputation regarding both the predictor set and the imputation method.

To answer our question how planned missing data from an SQD survey can be imputed as a service for the research community independent of a specific purpose of analysis, we evaluate different imputation strategies (methods and predictor set specifications) in their ability to reproduce relations in the data. To this end, we present findings from a Monte Carlo study simulating planned missing data from an SQD based on real survey data that we subsequently impute.

This paper proceeds as follows: In Section 2, we discuss the theory on planned missing data and MI as well as different imputation methods. Section 3 explains our data and method. In Section 4, we describe our results for the different strategies, first for strong and then for weak relationships between variables. Section 5 concludes with a discussion of the implications and limitations of this study.

2. Imputation of planned missing survey data

2.1. Planned missing data

Planned missing data occur when items are intentionally removed from questionnaires for specific groups of (usually) randomly selected respondents to shorten questionnaires and reduce respondent burden. In a simple planned missing data design each respondent is assigned to a predetermined number of items randomly selected from the complete questionnaire [38, 53]. The split questionnaire design [23, 43] is a modification of this procedure and involves allocating items to distinct split modules and subsequently randomly assigning each respondent to a subset of two or more split modules. In addition, a core module with particularly important items can be assigned to all participants to avoid planned missing data on these items.

SQDs result in a fixed share of planned missing data corresponding to the modules omitted by design. For example, with a questionnaire split into five modules of equal length, assigning three modules to each respondent produces 40% planned missing data. As a result, researchers wanting to analyze variables from different split modules will oftentimes end up with an empty dataset.

In consequence, Raghunathan and Grizzle [43] and Graham et al. [23] propose completing the missing data via MI (see also [1,6,28,41,44,56]). However, as discussed in the previous section, this may be challenging in practice: Large proportions of the data have to be imputed, making the quality of results particularly susceptible to misspecifications of the imputation models. A further challenge is the large number of variables in the predictor set of the imputation models in relation to the relatively small sample sizes. Furthermore, predominantly low correlations may also mean that the uncertainty of imputed values remains high, and many potential predictors do not improve the imputation but only add complexity to the model.

2.2. Imputation

The past decades have produced developments that allow for properly dealing with missing data by replacing them with several plausible values through multiple imputation [47,58]. To understand MI, suppose we have a variable $\mathbf{Y}$ that contains both observed values and planned missing values identified by vector $\mathbf{Z} = \{0; 1\}$, where 1 indicates that a value is observed and 0 that it is missing. Our scenario assumes that all missing data $\mathbf{Y}|(\mathbf{Z} = 0)$ is planned as described above and thus missing completely at random (MCAR). MI aims to replace $\mathbf{Y}|(\mathbf{Z} = 0)$ with m potential values that are plausible given a matrix of predictor variables $\mathbf{X}$ [58, pp. 19-20]. To this end, we rely on an imputation model that estimates the conditional probability distribution of $\mathbf{Y}$ given $\mathbf{X}$ using an adequate imputation method, accounting for all variable relationships as well as noise in the data and parameter uncertainty [58, pp. 65-68]. Multiple imputed values are drawn randomly from this conditional distribution for each missing value, generating m independently imputed datasets [58, p. 67]. With a properly specified model, the imputed data should reproduce the relationships between variables as well as uncertainty about these relationships and about the true unobserved values [47, pp. 12-16].

To analyze imputed data, estimates can be calculated separately for each of the m datasets with standard methods for complete data [47, p. 12]. Subsequently, these estimates are combined into a single estimate using Rubin's Rules [47, 58, pp. 145-147], yielding one combined estimator for each estimated parameter.

2.3. Predictors included in imputation models

An important decision in MI is what to include in the set of predictor variables **X**. The general recommendation is to include at least all variables that will be analyzed in a model together with the imputed variable unless their true relationship is zero [7]. Because we are interested in imputing data as a service to other researchers, we do not know which models will be applied to the data. In this situation, including all variables as predictors of the missing variable, and thereby using as much information as possible, may theoretically be the best option.

However, including all variables is often not feasible in practice [39, 58, pp. 167-170, 259-271, 70]. Each additional variable included in **X** makes the task of modeling the distribution of **Y** conditional on **X** more complex. At some point, the sample would not be sufficient anymore to support a reliable estimation of this conditional distribution. Therefore, common recommendations are to use at most 15 to 25 [58] or 30 to 40 [25] variables in imputation models. This is particularly important because otherwise, unattainably huge increases in sample sizes would be necessary.

In case predictor sets need to be restricted during the imputation of planned missing data, we argue that predictors should cover at least all variables that are substantively correlated with **Y**. These variables are essential to reduce the uncertainty of the imputations [59], as they contribute to the variance of the imputed variable. Under MCAR, imputation models excluding variables that are not correlated with **Y** may also be the most reasonable choice regarding their potential use in analysis models because there is no relationship to be preserved by the imputation.

In this study we consider both restricted and unrestricted predictor set specifications.

2.4. Imputation methods

To model the conditional distribution of **Y** through **X**, we need an adequate imputation method. In the following, we discuss several established methods, which differ both in their distributional assumptions regarding the imputed variable and in how its relationship to the predictor variables **X** is modeled.

2.4.1. Linear regression models (LRM)

First, linear regression can be used for MI [47, pp. 166-167, 58, pp. 67-74] if **Y** is continuous. However, since social research often treats ordinal variables as continuous, especially if the number of categories is high (see Wu and Leung [71] for a broader discussion and simulation), researchers might also consider LRM as a method to impute ordinal planned missing survey data.

To impute data using Bayesian LRM [58, p. 67], a linear model of $\mathbf{Y}$ conditional on $\mathbf{X}$ is specified:

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of Bayesian estimates of the regression coefficients for the predictor variables in $\mathbf{X}$ and ϵ represents the residuals. Accordingly, the posterior distribution $P(\mathbf{Y}|\mathbf{X})|(\mathbf{Z} = 1)$ can be estimated, from which imputations are randomly drawn. In an alternative frequentist setting, imputations can be calculated by adding an error drawn from the normal distribution of errors to a bootstrapped point estimate of $\mathbf{Y}$ [58, p. 67].

This procedure is associated with strong model assumptions. First, residuals are assumed to be normally distributed. With primarily categorical survey data, the normality assumption is likely violated. If this assumption does not hold, some authors recommend transformation techniques to approximate normality [26,31] while others show that outcomes can be biased with transformed variables as well (see for example von Hippel [64]).

Furthermore, linear regression does not account for restrictions such as discrete scales or logical bounds [33], potentially leading to implausible imputations [70, 58, p. 78, 64]. For example, if $\mathbf{Y}$ is an ordinal, Likert scale–based variable defined for integers from 0 to 10, non-integer and potentially even negative imputations would be obtained. Although the analysis results are not necessarily negatively affected by implausible imputations [2, 63, 64], imputed data with lots of implausible values may be considered inappropriate for publication, and standard analysis methods for categorical variables would most likely fail with data imputed by LRM.

In addition, all predictors are included as linear terms. This requires their actual relationship with $\mathbf{Y}$ to be exclusively linear as well. If there are any additional relationships in the data, say quadratic or interaction effects, these must be explicitly specified in the model [50,63].

While possibly oversimplifying the relationship between predictors and imputed variables, LRM have the clear advantage of only needing one parameter (the regression coefficient) to describe the relationship of a predictor with an outcome. This relatively simple imputation task facilitates the estimation of many relationships considering the practical problems with the imputation described above. In contrast, methods that attempt to address categorical data specifically or model non-linear relationships require more parameters for the same set of variables.

2.4.2. Categorical regression models (CRMs)

To circumvent some of the theoretical disadvantages of LRM, we might consider using categorical regression models (CRMs) [16, 47, pp. 169-170, 60]) from the general class of generalized linear models (GLMs). To accommodate the estimation of non-normal outcomes such as categorical variables, LRMs are generalized through

$$\mathbf{Y} = g(\boldsymbol{X\beta})\,, \tag{2}$$

where g stands for a link function that depends on the assumed distribution of $\mathbf{Y}$. A simple example for a CRM is logistic regression for estimating the probability of $\mathbf{Y} = 1$ in binary variables, where g stands for the logit function

$$Pr(\mathbf{Y} = 1) = \frac{e^{\boldsymbol{X\beta}}}{1 + e^{\boldsymbol{X\beta}}}\,. \tag{3}$$

In this way, the non-normal distribution of categorical outcomes can be accounted for. As a result, CRMs with a correct specification of $\mathbf{Y}$'s discrete distribution allow for directly drawing imputations that stick to empirically possible values. However, we still assume that all effects of $\mathbf{X}$ on the transformed $\mathbf{Y}$ variable will be linear, so non-linear relationships must be explicitly modeled, like with LRM. Similarly, we assume error terms to follow a predefined distribution, meaning that the imputation quality could be impaired if these restrictive distributional assumptions do not hold.

CRMs can also cause new problems if the sample size is small, since modeling categories instead of the variables themselves increases the complexity of the imputation model. Accordingly, van Buuren [58, p. 91] notes that the "imputation of categorical data is more difficult than continuous data". As a rule of thumb, at least about ten cases per predictor category times outcome category are required for CRM to produce stable estimates [57, p. 87, 58, p. 91]. As categorical predictors are usually represented as dummy variables, this means thousands of respondents would be required to impute a variable with ten categories only using one predictor with equally ten categories. Correspondingly, in a similar context White et al. [70] report that they have found particularly structures with several nominal variables "challenging to work with" when imputing them by multinomial logistic regression. Furthermore, Wu et al. [72] observe that LRMs outperform CRMs in various scenarios with binary and ordinal variables.

2.4.3. Predictive mean matching (PMM)

Another common method used in MI is predictive mean matching (PMM) [32, 46]. PMM is a two-stage method: First, a regression is applied to the data. However, instead of drawing imputations directly, predicted values $\hat{\mathbf{Y}}$ are calculated and a real observed $\mathbf{Y}$ value is drawn from a set of donors with similar $\hat{\mathbf{Y}}$. Extensions of this method add bootstrapping, propensity score matching as a special case for categorical variables, and an alternative to draw imputations weighted by distance instead of randomly from the donor set [30, 54].

This solves several problems of conventional regression methods. First, imputations do not take impossible values, as all imputed values are taken from real observations on other cases. Second, although all effects are still expected to be linear, evidence shows that PMM is quite robust against violations of this assumption [30, 37, 58, pp. 77-79]. However, model misspecifications can still result in biases [30,37,50]. For example, interaction effects must be specified explicitly [50]. Moreover, when missing cases do not have enough potential donors nearby, PMM falls back to more distant donors to draw imputed values, which may also result in bias [29].

2.4.4. Partial least squares PMM (PLS-PMM)

Although PMM relaxes some assumptions on the imputation, large numbers of potential predictors could still be a problem. Robitzsch and Grund [45] implement partial least squares (PLS) regression [20,35] as a two-step method to reduce the dimensionality of the predictor space before imputing the data. In a first step, PLS regression is used to extract a predetermined number of k components of $\mathbf{X}$ that describe the maximum possible covariance of $\mathbf{X}$ and $\mathbf{Y}$ [20]. These PLS components are uncorrelated latent variables optimized to predict $\mathbf{Y}$ and ordered by decreasing importance for predicting $\mathbf{Y}$. In the second step, missing values are imputed (by default, with PMM) using the k components as predictor set rather than the original data.

Such an approach suggests unique advantages over other methods. First, by using comparatively few PLS components for the imputation rather than many original predictors $\mathbf{X}$, the number of parameters in the model is reduced. At the same time, most of the information on $\mathbf{Y}$ is preserved, as the PLS components were extracted from $\mathbf{X}$ specifically to predict $\mathbf{Y}$. Second, substituting the original variables $\mathbf{X}$ by their (uncorrelated) PLS components also removes potential multicollinearity (although due to the rather small correlations, multicollinearity should be low). Third, by using PMM to draw imputations based on the PLS components, only empirically possible values are imputed. Thus, PLS-PMM might help preserve information considering that the data context supposedly requires restricting the number of parameters because of the limited case numbers and large amounts of missing data to deal with.

However, PLS-PMM may also introduce new difficulties, particularly due to potential information loss caused by dimensionality reduction. Extracting only k PLS components from $\mathbf{X}$ means that some other information in $\mathbf{X}$ will be ignored in the imputation. If this ignored part of $\mathbf{X}$ still contains additional information on $\mathbf{Y}$, corresponding relationships would be to some extent lost. In consequence, k should ideally be set such that all relevant information on the covariance between $\mathbf{X}$ and $\mathbf{Y}$ is included in the imputation, that is, a potential $k + 1$-th component must not provide any substantial further information on $\mathbf{Y}$. Furthermore, PLS-PMM still assumes that all relationships in the data are linear. Thus, non-linear terms such as interactions must be explicitly specified in the PLS model.

2.4.5. Classification and regression trees (CART)

Finally, we could also decide to drop all assumptions about distributions and relationships in the data, choosing an algorithm that attempts to learn about these features. Classification and regression trees (CART), as described by Breiman et al. [17], have shown to be a relatively simple method for this purpose [18,21]. Other tree-based algorithms such as random forests work similarly, but often go beyond CART by combining estimates of various trees (see, for example, Shah et al. [52]), making them quite computationally demanding.

CART creates a decision tree predicting $\mathbf{Y}$ by repeatedly partitioning the data into two subregions along the values of the predictor variables. After having started with an unconditional estimate of $\mathbf{Y}$ (i.e., the mean or mode, depending on whether $\mathbf{Y}$ is continuous or categorical), a cut-off point on a variable in $\mathbf{X}$ is chosen and $\mathbf{Y}$ is estimated separately below and above the cut-off point (i.e., with two mean or mode values). In doing so, as many possible cut-off points as possible are tested and the one that optimizes the goodness of fit is chosen. For example, for categorical $\mathbf{Y}$ this means the cut-off point that reduces entropy the most is accepted. After that, the same procedure starts again separately within both subregions, leading to the data being cut into four subregions in total. This procedure is repeated again and again, creating smaller and smaller subregions, and stops only when (a) an external stopping criterion is reached, (b) the goodness of fit cannot be further improved, or (c) there are not enough data left for another cut, thereby eventually reaching a terminal node. To impute a missing value, an observed value can be randomly drawn from one of the observed cases in the same terminal node [58, p. 86].

CART's main advantage is that it accounts for all kinds of relationships (including interactions) automatically without the need to specify a functional form. Furthermore, it generates plausible imputations by drawing observations from the same terminal node. Thus, CART seems ideal for a general-purpose imputation, as it provides imputations that make intuitive sense and is agnostic to the functional form of data users' eventual analysis models. Some evidence also suggests that CART outperforms CRM and PMM especially in reproducing complex relationships [3, 18, 21]. Slade and Naylor [51] observe a similar performance of CART and correctly specified PMM.

However, large predictor sets might create particularly severe problems for CART. Remember that CART stops partitioning a subregion of the data when not enough data are available to support another split. As one imputed value must be randomly drawn from a pool of several potential donors in the terminal node, several (say, five) cases must be left in each terminal node. However, if this node size limit is reached before all relevant predictor variables are accounted for, the remaining ones are implicitly omitted from the imputation.

For example, suppose we have $1,600$ observed cases on $\mathbf{Y}$. On average, each repeated cut divides the average case numbers remaining in each subregion by two. For simplicity, suppose that these two subregions are always equally large. Consequently, we would reach terminal nodes after only eight successive cuts, with $1,600/2^8 = 6.25$ cases per subregion. Thus, including more than eight predictor variables would mean that some are necessarily omitted in the imputation. Furthermore, even eight predictors would only work in the unlikely case that one binary cut per predictor variable suffices to represent all its relationship with $\mathbf{Y}$. For instance, Doove et al. [21] observe particular problems with reproducing linear main effects, arguing that such structures likely require several consecutive cuts per variable. Effectively, we might thus end up with only a few predictors sufficiently utilized by CART.

CART could thus run into problems even with relatively large samples: Assume we quadruple the sample in our example survey, yielding $6,400$ observed cases. Even this would only allow for two more cuts on average (ten cuts in total). Thus, we may face a *curse of dimensionality* problem [8], in which adding more predictors requires an exponential growth in case numbers. In consequence, CART implicitly assumes that only a few predictors in $\mathbf{X}$ really determine $\mathbf{Y}$ and all other predictors are negligible.

In this context, generally low but non-zero correlations as commonly found in survey data could even exacerbate such problems. First, CART might face difficulties in identifying optimal cut-off points due to high uncertainty in the data. Furthermore, in a data context in which predictive information on $\mathbf{Y}$ is not primarily stored in a few strong correlations but in many different weak correlations, much information on $\mathbf{Y}$ may be lost in the imputation when the selected imputation method limits the number of predictors so strictly.

2.5. Imputing multivariate missing data

With planned missing data as produced by an SQD, missing data is usually obtained not on one but on many variables. This means that, when imputing a variable $\mathbf{Y}$ with missing values, there will also be missing values in $\mathbf{X}$. To deal with such multivariate missing data, one can apply the previously discussed imputation methods for each variable consecutively via fully conditional specification (FCS; sometimes also referred to as multiple imputation by chained equations) or alternatively, use joint modelling (JM) as a holistic method instead of integrating univariate imputation methods.

JM is the classical application of MI described by Rubin [47]. It entails modeling the joint distribution of multivariate missing data in a single multivariate model [58, pp. 112, 115-119]. This requires an explicit assumption about the true distribution that applies to all variables in the imputation model. Usually, a multivariate normal distribution is assumed, and variables violating normality are often transformed [26, 49]. This normality assumption must hold for all (transformed) variables in the model alike. After estimating the multivariate distribution parameters, imputations can be drawn directly from the distribution.

FCS has been developed more recently [16, 58, 60] and divides the multivariate imputation task into multiple univariate imputation tasks that are processed one after the other. In doing so, an implicit joint distribution is approximated without having to specify it explicitly. To this end, an imputation model with relevant predictors is defined for each variable to be imputed, describing the conditional distribution of this variable. Predictors can either be fully observed or contain missing values that are imputed themselves. Furthermore, an imputation method (such as CART, PMM, etc.) is also specified for each variable to be imputed.

The FCS algorithm [58, pp. 120-121] iterates over all conditional distributions to impute the missing values. This means imputation models for each imputed variable are repeatedly run one after the other, eventually imputing the whole data. The first run starts with random draws from $\mathbf{Y}|(\mathbf{Z} = 1)$. Then, the first variable with missing values is imputed on the basis of the predictors, which rely on observed data completed by the random starting values. In doing so, the initial random imputations on this variable are replaced. Then the second variable is imputed, followed by the third, and so on, until all initial random imputations are replaced. Subsequently, the procedure starts again with the previously imputed values, imputing the first, second, third, etc. variable. This is repeated for a number of iterations to reach convergence, each time replacing the imputations from the former iteration. When a predictor variable has imputed values itself, imputation models always use its latest imputed version throughout the iterations.

JM and FCS are different in some respects. JM has a more bottom-up theoretical justification and is computationally faster, while FCS offers much more flexibility [58, pp. 130-131]: distributions must only be defined univariately for the imputed variables instead of an overarching multivariate distribution. This allows for using different imputation methods (for example, accounting for different levels of measurement) as well as different predictor sets for each imputed variable. In this study, we test both JM and FCS strategies, but due to the gains in flexibility, we mostly rely on FCS.

3. Data and methods

To test the different imputation strategies for their ability to reproduce relationships in planned missing data, we apply a Monte Carlo simulation based on real survey data. This section describes the preparation of the data, simulation setup, and measures.

3.1. Data

We use data from two survey waves of the German Internet Panel (GIP), a probability-based online panel of the general population in Germany [11–14,19]. The dataset includes 61 variables with items on the respondents' sociodemographic information and sampling cohort, organization membership, Big Five personality traits, lobbying in EU politics, domestic and party politics (this is the same dataset as used in Axenfeld et al. [5]).

Because our focus is on the evaluation of strategies to impute planned missing data stemming from split questionnaire designs, we removed all non-planned missing data (nonresponse) from the dataset. This is necessary to ensure that the reported effects of imputing planned missing data are not confounded by imputations for other missing data. To deal with unit nonresponse, we restricted our sample to respondents who took part in both waves of the GIP (dropping $1,390$ out of $5,411$ cases). Next, we had to deal with item nonresponse. Some item nonresponse could be matched with responses from earlier waves [9, 10]. The remaining item nonresponse (on average 167 values or 4% per item) was imputed with single imputations in *R* [42] via *mice* [61],[1] using PMM including all variables with Spearman correlations stronger than $|0.05|$. This procedure had negligible effects on correlations and marginal distributions in this dataset (see [5, Figure A.1]).

In a next step, we recode variables with rare events to allow for an appropriate imputation. This is because the simulation procedure reduces available sample sizes considerably in all simulation runs, and hence the number of available observations per category is much lower in the simulated SQD datasets than in the population. Thus, categories containing fewer than 100 cases (2.5%) are combined into somewhat broader categories to provide the imputation with sufficient case numbers.

Our final dataset, which we will refer to as population dataset, contains $4,061$ cases and 61 items. All variables are categorical and contain no missing values. From the 11 sociodemographic and sampling cohort variables, 1 variable is dichotomous, 7 are nominal with 3 to 12 categories, and 3 are ordinal with 5 to 12 categories. These are treated as core variables, which are complete and hence do not have to be imputed. Of the remaining 50 variables, 44 are ordinal with 3 to 11 categories and 6 are dichotomous. These 50 variables are imputed during the simulation.

3.2. Simulation of planned missing data

To assess the performance of different imputation strategies with planned missing data, we simulate the implementation of a split questionnaire design in our population data. To this end, we assume that the sociodemographic items and the sampling cohort constitute a core module. The remaining 50 items would be allocated randomly to five split modules with ten items each. Each respondent then receives the core module and three out of five randomly assigned split modules. This results in a 33% reduction in questionnaire length, with approximately 40% (2/5 modules) randomly missing data on each split item and no missing data on the core items.

Our simulation study picks up this scenario, repeating to simulate SQDs in 1,007 simulation runs using the bwHPC high-performance computing infrastructure.[2] In each simulation run, this entails the following tasks:

1. drawing a random sample from the population data;
2. randomly allocating items to modules;
3. randomly assigning modules to respondents;
4. setting values for modules not assigned to missing, mimicking an SQD;

[1] Other *R* packages used for this paper (if not cited elsewhere) are: DescTools [55], doMPI [66], foreach [36], ggplot2 [67], haven [69], MASS [62], Rmpi [73], tidyr [68].

[2] The exact number of 1,007 simulation runs was used for computational reasons, as the simulation ran parallelized on one processor for each run, and we had access to 1,008 processor cores (one of them is consumed by setting up the simulation.

5. applying MI to the simulated planned missing data for each imputation strategy, and
6. estimating Spearman correlations on the MI data to be compared against their population benchmarks for each imputation strategy.

3.3. Imputation strategies

In each simulation run, we test different imputation methods implemented in R. We implement JM via *Amelia* [26], a technique that draws from a multivariate normal distribution modeled using the expectation–maximization algorithm. With this method, we have the option to (correctly) declare our variables as ordinal, which will make *Amelia* transform the initial continuous imputations into discrete categories. However, forcing continuous values into integer imputations can compromise the accuracy of estimates [2, 27], so Honaker et al. [26, p. 16] suggest letting *Amelia* impute continuous values without ordinal transformation, if feasible. However, this produces implausible imputations, which may be a problem if the data is to be published. In consequence, we include both *Amelia* with transformed (JM-T) and with untransformed imputations (JM-U) in our simulation.

Moreover, we use some FCS imputation methods implemented in *mice* [61]: the *mice* default (CRM, here: logistic regression and ordinal logistic regression), *norm* (Bayesian LRM), *pmm*, and *cart*. Furthermore, we use *pls* (PLS-PMM) from the *miceadds* package [45], which includes 20 PLS components in the imputation. For these FCS techniques we draw values after 10 iterations, because an initial test simulation suggested that more iterations could not improve our estimates.

As a benchmark for poor imputations we include *sample* (also included in *mice*), an unconditional hot deck sampling replacing missing values with randomly selected observed values, to assess in how far the other methods outperform a purely random replacement of missing values.

In the basic design, predictor sets include all variables in the data. Additionally, two refinements with fewer predictors are implemented for all eligible imputation methods. These two options exclude predictor variables with Spearman correlations either weaker than $|0.1|$ (option 1) or weaker than $|0.2|$ (option 2) to the imputed variable and are applied to LRM, CRM, PMM, and CART. *Amelia*, as a JM technique, does not allow for excluding different predictor variables per imputed variable, and PLS applies a dimensionality reduction before imputation, generally including all variables in $\mathbf{X}$.

The correct specification of m to adequately represent the distribution of potential values for a missing value is subject to a lively debate. Sometimes, $m = 5$ may suffice (see, for example, Schafer and Olsen [48]), but depending on the data and analysis purpose, m must often be considerably larger [15, 24, 65]. In our study, we create $m = 20$ imputed datasets for each imputation strategy because an initial test simulation suggested that results do not improve with more imputations.

3.4. Measures

We compare different imputation strategies regarding how well they reproduce bivariate relationships based on Spearman correlations. For each pair of variables i, j (with $i \neq j$) in split modules, Spearman correlations $\rho_{i,j}$ are calculated as benchmarks based on the population data. With the imputed SQD data, Spearman correlations $\hat{\rho}_{i,j,s}^{imputed}$ are estimated for the same variable pairs in each simulation run s. This entails that Spearman correlations are estimated separately in each imputed dataset and subsequently pooled through applying *Fisher's Z* transformation on the correlations, calculating the mean and transforming it back into a correlation [58, p. 146].

The correlations turn out generally low in the population data, as is typically the case with many surveys. Of the 1,225 correlations, 85 (7%) are stronger than |0.2| with a maximum value of 0.70, 140 (11%) are stronger than |0.1| but at most |0.2|, 248 (20%) are stronger than |0.05| but at most |0.1|, and 752 (61%) are weaker than or equal to |0.05|. Thus, many variables are hardly correlated, whereas few have relatively strong correlations.

In case $\hat{\rho}_{i,j,s}^{imputed}$ estimates $\rho_{i,j}$ validly, we should observe that random differences between the MI estimate and its population benchmark average out over many simulation runs. Therefore, we compute the (raw) Monte Carlo bias $Bias^{MC}$ of the average MI estimate $\hat{\rho}_{i,j}^{imputed}$ over all simulation runs S,

$$Bias^{MC}(\hat{\rho}_{i,j}^{imputed}) = \frac{1}{S}\sum_{s=1}^{S}\hat{\rho}_{i,j,s}^{imputed} - \rho_{i,j}\ , \tag{4}$$

representing the average difference between MI estimates and the true correlation benchmark. To obtain a more intuitive measure of bias, we can calculate the percentage Monte Carlo bias by dividing the raw bias by the true correlation $\rho_{i,j}$ and multiplying it by 100:

$$\%Bias^{MC}(\hat{\rho}_{i,j}^{imputed}) = \frac{Bias^{MC}(\hat{\rho}_{i,j}^{imputed})}{\rho_{i,j}} \times 100\,. \tag{5}$$

The percentage bias indicates by how much percent the MI correlation is underestimated or overestimated.

Percentage biases have the disadvantage that they are only meaningful for correlations that are clearly different from zero: A $\rho_{i,j}$ near zero in the denominator of Equation 5 can lead to exceedingly large relative deviations even when the actual difference between estimate and benchmark is negligible. Furthermore, a $\rho_{i,j}$ exactly equal to zero means a denominator equal to zero, making $\%Bias^{MC}(\hat{\rho}_{i,j}^{imputed})$ impossible to calculate. In consequence, a reliable estimation of the percentage bias is only feasible for correlations clearly different from zero. This is especially relevant given that, as described before, correlations in our population dataset tend to be weak. Accordingly, percentage biases work poorly for the many very small correlations, for which we observe percentage biases up to 84,606% with deviations that are often negligible in absolute size (as small absolute deviations may be divided by much smaller correlations close to zero). Thus, to analyze very small correlations in a meaningful way we resort to the raw bias as defined in Equation 4, which does not share this problem. Observing that extremely large percentage biases as just mentioned appear exclusively in correlations below $|0.05|$, we therefore use the percentage bias for the 473 correlations stronger than $|0.05|$ and the raw bias for the 752 correlations equal or weaker than $|0.05|$.

4. Results

We now discuss the performance of the implemented imputation strategies as measured by percentage and raw biases in Spearman correlations. First, we describe the results for item pairs that have strong or moderate relationships in our population data. In doing so, we concentrate on the relationships which have the most to lose in terms of substantive relationships when the imputation fails. In this part, we also include different predictor set specifications. Subsequently, for the sake of completeness, we also show the results for item pairs with weak or null relationships.

4.1. Item pairs with moderate or strong relationships

Figure 1 displays the average percentage biases in Spearman correlations for the 85 item pairs with moderate or strong relationships (stronger than $|0.2|$ in the population data), broken down by imputation method and predictor set specification. Each point displayed in a row represents the average bias over the $1,007$ simulation runs for one specific variable pair. The boxplots condense the information given by these point clouds that depict the average biases for the different variable pairs into an aggregate image of how the Monte Carlo biases are distributed for each strategy. In addition, the corresponding quantile distributions are shown in an appendix (Table A1).

First, the random imputations with unconditional hot-deck sampling lead to biases that concentrate at about -65%. Consequently, this is the approximate average bias we could expect from a method that completely fails to incorporate relationships in the imputation.

With LRM, biases are relatively small, with the central 50% (i.e., the area from the first through the third quartile) of biases ranging from -6.8% to -2.6%. Some outliers appear at both tails up to or slightly exceeding $\pm 20\%$. Although most biases are negative, many are close to zero. Excluding predictors correlated less than $|0.1|$ with the imputed variable (option 1) results in a shift to the right, suggesting weaker biases: Here, the central 50% of biases range from -4.0% to $+0.6\%$. Further removing predictors correlated less than $|0.2|$ with the imputed variable (option 2) yields no additional improvement (the central 50% range from -4.2% to $+0.7\%$).

CRM tends to produce strong biases. With an unrestricted predictor set, the central 50% of biases range from -50.7% to -21.9%. We observe no biases closer

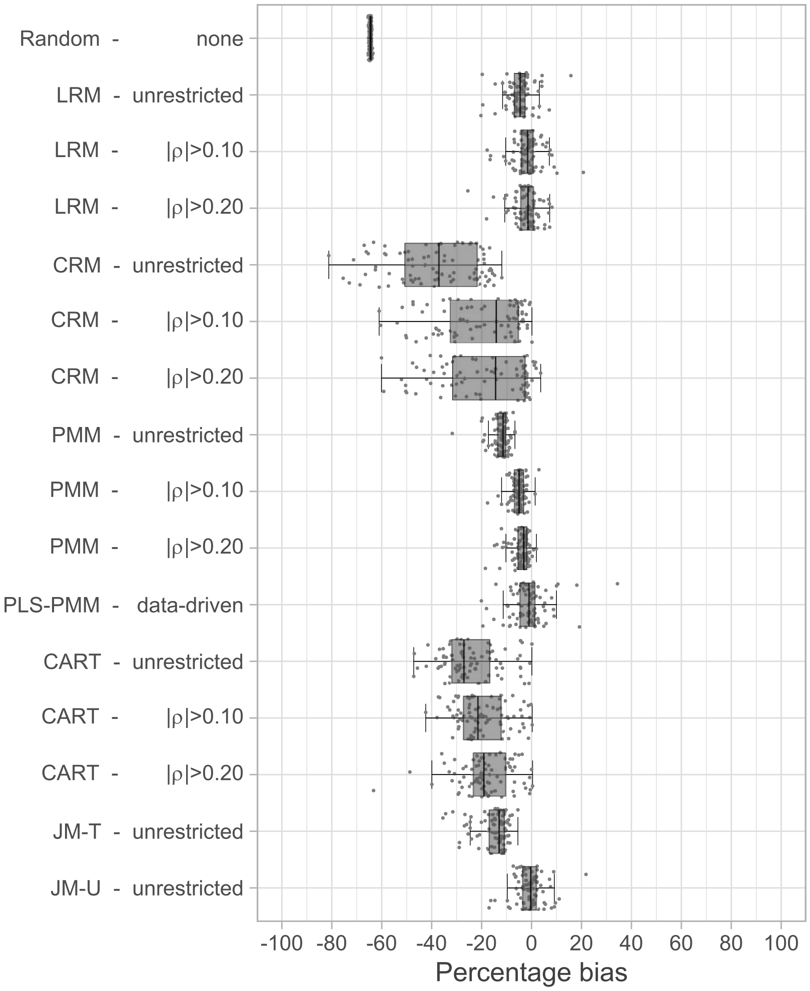

Fig 1. *Average percentage Monte Carlo biases of Spearman correlations for 85 item pairs with moderate or strong relationships (true correlations stronger than* $|0.2|$*), by imputation method and predictor set specification.*

Note: Random = unconditional hot-deck sampling; LRM = linear regression model; CRM = categorical regression model; PMM = predictive mean matching; PLS-PMM = predictive mean matching on partial least squares components; CART = classification and regression trees; JM-T = joint modeling with transformed imputations; JM-U = joint modeling with untransformed imputations.
Unrestricted = with all variables in the predictor set; $|\rho| > 0.1/0.2$ = with only predictors with $|\rho| > 0.1$ / $|\rho| > 0.2$ in the predictor set; data-driven = 20 PLS components; none = no predictors.

to zero than −10% but some biases stronger than −65%. Thus, all correlations appear biased, with some even further from the truth than randomly imputed values. Again, we observe some predictor set effects on biases shifting the distribution of biases to the right: The central 50% of biases range from −32.5% to −5.6% (option 1) or from −31.6% to −2.7% (option 2). With both options, biases also have a smaller tendency towards extreme values, with minimum values at about −60%. Thus, CRM performs poorly with unrestricted predictor sets and improves a little when we remove weak predictors, but even severely restricted predictor sets cannot eliminate the biases, which are still mostly much stronger than −10%.

PMM performs better than CRM but, at least with unrestricted predictor sets, shows still moderate biases, with the central 50% ranging from −13.5% to −10.5% and no biases closer to zero than −6%. Only two biases exceed −20%, yet one extreme outlier has a bias of −31.7%. These biases can be reduced considerably by excluding weak predictors from the imputation models: The central 50% of biases then range from −6.8% to −3.4% with option 1 or even from −5.4% to −2.0% with option 2. Furthermore, both option 1 and option 2 make the extreme outlier disappear, with the strongest biases being less pronounced than −20% in both cases. Thus, we can obtain relatively accurate estimates with PMM, almost catching up with JM-U when using restricted predictor sets.

With PLS-PMM most biases are even smaller, with the central 50% ranging from −4.6% to +1.4%. Concurrently, we observe outliers mostly up to about ±20% and one at +34.4%.

CART leads to relatively strong biases, although they are less pronounced than with CRM: With unrestricted predictor sets, the central 50% of biases range from −31.7% to −16.8%, with the strongest bias being −47.2%. Furthermore, only few correlations are almost unbiased, with maximum values of +0.1%. Again, removing weak predictors from the imputation models yields an improvement. However, the central 50% of biases still range from −27.2% to −12.3% with option 1 and from −23.2% to −10.5% with option 2. However, with option 2, we also observe two extreme biases with a minimum value of −63.3%. Thus, despite some improvements with restricted predictor sets, CART in general performs poorly.

JM performs much better than CRM and CART but still leads to moderate biases when normal imputations are transformed to ordinal values (JM-T): The central 50% of biases range from −16.9% to −11.2%. We also observe outliers with some biases stronger than −30%. There are no biases closer to zero than −5%, so correlations appear quite universally biased. However, JM-U (i.e., declaring the variables (incorrectly) as continuous) considerably reduces biases: The central 50% of biases range from −3.4% to +1.7%, with the most extreme outliers at about ±20%. Thus, despite some remaining biases, JM with untransformed imputations overall performs well.

To sum up, strong correlations over $|0.2|$ are best reproduced by PLS-PMM and JM-U when the entire set of variables is considered in the imputation. PMM and LRM approach their level of accuracy with predictor sets restricted to stronger correlations. While CRM and CART perform exceptionally poorly, PMM with unrestricted predictor sets and JM-T also produce systematically biased results.

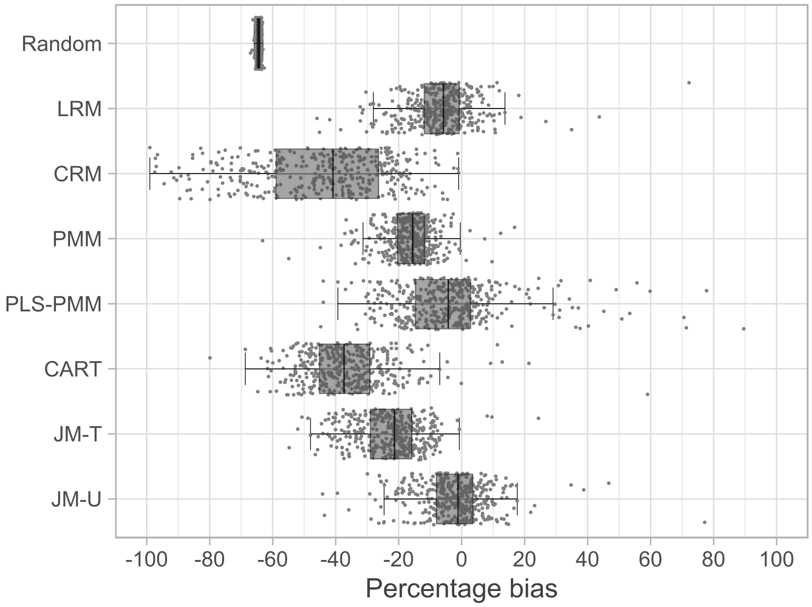

Fig 2. *Average percentage Monte Carlo biases of Spearman correlations for 388 item pairs with weak relationships (true correlations weaker than* |0.2| *but stronger than* |0.05|*), by imputation method.*

See Note Figure 1.

4.2. Item pairs with weak or null relationships

Figure 2 displays the average percentage biases for the 388 item pairs that had weak relationships in the population (between |0.05| and |0.2|), again for different imputation methods. Alternatively, quantile distributions are given in Table A2. Restrictions of the predictor set are not presented here, as they exclude (some of) the relationships under study from the imputation and thus produce biased estimates per se. Apart from this, the information displayed in the graph is equivalent to Figure 1, with point clouds and boxplots showing the distributions of biases for each strategy.

In general, Figure 2 reproduces most patterns observed for strong relationships. With random imputations, we still observe biases concentrating at about −65%. Furthermore, JM-U, LRM and PLS-PMM yield the least biased estimates, followed by PMM and JM-T, while CART and CRM have the strongest biases among all methods (except for random imputations).

However, percentage biases tend to be more pronounced for these weak relationships than for the stronger relationships discussed in the previous Section 4.1. CART and JM-T are particularly affected, with distributions visibly shifted away from zero. Biases with the other strategies also appear slightly shifted to the negative, but primarily scatter more compared to strong relationships,

Table 1

Quantile distribution of absolute raw average Monte Carlo biases of Spearman correlations for 752 item pairs with relationships close to zero (true correlations weaker than $|0.05|$*), by imputation method.*

	Min.	5%	25%	50%	75%	95%	Max.
Random	0.000	0.001	0.006	0.012	0.021	0.03	0.033
LRM	0.000	0.000	0.001	0.003	0.006	0.014	0.039
CRM	0.000	0.001	0.004	0.009	0.017	0.035	0.056
PMM	0.000	0.000	0.002	0.004	0.007	0.013	0.027
PLS-PMM	0.000	0.000	0.002	0.004	0.008	0.018	0.045
CART	0.000	0.001	0.004	0.009	0.015	0.024	0.039
JM-T	0.000	0.000	0.002	0.005	0.009	0.016	0.024
JM-U	0.000	0.000	0.001	0.003	0.006	0.014	0.041

See Note Figure 1.

causing an increased prevalence of extreme biases. Correspondingly, biases considerably larger than zero (i.e., positive percentage biases) occur with CRM, JM-T, JM-U, and PLS-PMM, each with maximum values of about +60% or more. With CRM, some biases also fall out of the display range defined between −100% and +100%: Ten correlations have biases exceeding −100% with a minimum value of −119.8%. PLS-PMM also has one bias out of display range (+106.6%).

Table 1 displays the quantile distribution of the absolute values of raw average biases for the 752 relationships close to zero (weaker than $|0.05|$ in the population) for the different imputation methods. Due to the small true relationship strength, their raw biases are mostly small as well. We observe that random imputations lead to biases between 0.000 and 0.033. In contrast, biases with other imputation methods are mostly smaller, but all methods except JM-T and PMM have maximum values larger than those obtained with random imputations. Apart from that, patterns with this kind of relationship again largely reproduce the findings above: CRM and CART have comparatively large biases concentrating around 0.009. At the other extreme we again have JM-U and LRM producing biases of only 0.003 at the median, while JM-T, PMM and PLS-PMM show biases concentrating at around 0.004 and 0.005.

5. Discussion

As we described in the introduction, a general-purpose imputation of planned missing data resulting from using a split questionnaire design holds special challenges. They stem primarily from the large amount of missing data to be imputed on many variables using many partially missing predictors, combined with survey-typical features such as comparatively small sample sizes and low correlations. Using a Monte Carlo simulation, we tested the accuracy of several imputation strategies with real survey data. In doing so, we first analyzed correlations stronger than $|0.2|$ in the population data, and then turned to the weaker correlations. Overall, the relative performance of imputation methods is similar in both cases.

Surprisingly, LRM performed exceptionally well, with mostly low biases in Spearman correlations even with unrestricted predictor sets. This finding stands in sharp contrast to statistical intuition suggesting that methods should account for the variables' levels of measurement, which raises the question of why LRM performed so well. First, our data context characterized by low correlations and high uncertainty, limited case numbers, and many potential predictors may have promoted the use of simple methods that need comparably few data to efficiently estimate relationships between all variables. Here, linear regression can excel because it estimates only one coefficient per predictor. Thus, LRM's benefits due to simplicity might have outweighed its disadvantages, such as assuming an incorrect level of measurement and strict linearity in relationships. Second, although our data are not continuous, they are at least binary or ordinal. Presumably, the performance of LRM would quickly drop if we shifted our focus to non-ordered categorical data. Third, LRM might perform well with reproducing the correlations covered by our study but still fail with other types of relationships or estimates. Perhaps strongly non-linear relationships were absent in our data, which would give LRM an advantage over competing methods. Furthermore, we must bear in mind that LRM will inevitably destroy discrete distributions of categorical variables, leading to implausible imputations. Hence, an LRM general-purpose imputation would heavily restrict data users in their analyses. For example, frequency counts or classification models such as logistic regression would most likely fail. Consequently, we might be tempted to round imputations to discrete values, but this practice has shown to cause bias (for example, see Horton et al. [27]). Moreover, the assumption of normally distributed error terms is unlikely to hold with LRM on categorical data.

CRM consistently showed a dissatisfactory performance under all the predictor set specifications we studied. Some biases were even stronger than with random imputations drawn without any predictor variables. This confirms earlier findings reporting inaccuracies with similar methods (e.g., White et al. [70] and Wu et al. [72]).

PMM was found to perform much better than CRM, even though unrestricted predictor sets still lead to moderate biases. We showed that these biases were significantly reduced by simplifying the imputation model. This could be done either by removing predictors that are only weakly correlated with the imputed variable or through dimensionality reduction (PLS-PMM), suggesting that an adequately specified imputation via PMM might work well.

CART performed poorly with all predictor set specifications, although better than CRM. This finding is especially noteworthy considering that there is evidence suggesting that CART may outperform other imputation methods, such as PMM [21]. We suspect this is primarily due to the complex imputation exercise of our planned missing data context, which is characterized by a limited number of cases and many relevant but predominantly weakly correlated variables. However, as CART has also been previously reported to be challenged specifically by predicting linear relationships [21], future research could examine whether CART plays more to its strengths with non-continuous relationships. Furthermore, future research might investigate whether other, more sophisticated decision tree techniques (such as random forests) could provide an improvement over CART that is sufficient to impute large amounts of planned missing survey data from SQDs.

Joint modeling via *Amelia* showed moderate biases when we correctly specified the measurement level as ordinal (JM-T), resulting in imputations transformed into discrete categories. When we instead specified the level of measurement as continuous (JM-U), we mostly got rid of these biases, similarly as with FCS via LRM, for example. This is no coincidence, as "FCS using all linear regressions is identical to imputation under the multivariate normal model" [58, p. 130]. However, this means that both also share many disadvantages, especially as, in contrast to JM-T, they lead to implausible imputations not matching the discrete distributions and bounds of categorical variables.

For the imputation methods we analyzed, removing weak predictors leads to more accurate estimates. However, this also involves a strong theoretical assumption: Either the true relationship of imputed variable and predictor must be zero or both variables must eventually not be analyzed together. In contrast, an analysis-specific imputation could explicitly select predictors by whether they will be used in an analysis model. Thus, an analysis-specific imputation could be expected to yield a better estimation accuracy if neither part of the aforementioned assumption holds.

PLS-PMM with a dimensionality reduction of the predictor space could show a way out of this dilemma. This method allows to include all variables in the imputation with a performance comparable to solutions with restricted predictor sets. Furthermore, PMM is in general more robust against violations of the normality assumption than LRM (e.g., [30]) and maintains the discrete scale of the variables. In principle, with PLS-PMM we could also include non-linear terms and interaction effects as predictors if they are highly correlated with the imputed variable, enabling data users to explore phenomena beyond linear effects with their analysis models. Finally, PMM automatically generates plausible imputations, preserving categorical variables. For a general-purpose imputation, this is a significant advantage over methods such as JM-U and LRM, which performed comparably well but produce implausible continuous imputations and thus might not be considered optimal to impute data from a SQD for general usage. Thus, PLS-PMM appears as the currently most promising approach for a general-purpose imputation of data from an SQD, being able to yield both plausible values and produce only little bias in bivariate relationships in the data.

Future research should explore how the current implementation of PLS-PMM can be refined to produce valid general-purpose imputations of SQD data. For example, one challenge is to find more theoretically or empirically justified methods to set the number of PLS components used for imputation.

Moreover, in this study we focused on biases of Spearman correlations because they have previously been found to be particularly adversely affected when imputing data from an SQD [5], constituting a good target to measure the performance of imputation strategies. However, further tests could focus more on precision and coverage, as well as additional targets, such as regression coefficients.

Another aspect is how nonresponse by respondents interacts with the imputation of SQD data, which we explicitly did not study here. This may be relevant not only as nonresponse by respondents will increase the proportion of missing values, but also because the resulting missing data might not be MCAR.

Future research should also test whether our findings hold under different data contexts and parameter settings. On the one hand, data with a higher number of strong correlations or considerably larger sample sizes could hypothetically yield better results. On the other hand, challenges could grow with surveys having more items (increasing the number of potential predictors) or primarily relying on nominal response scales (reducing the options regarding adequate imputation methods). Continuing to focus particularly on the practical issues of imputing planned missing survey data from SQDs will be crucial to ensure the future usability and validity of data and the research stemming from these designs.

Appendix. Quantile distributions for the information displayed in Figures 1 and 2.

Table A1

Quantile distribution of average percentage Monte Carlo biases of Spearman correlations for 85 item pairs with moderate or strong relationships (true correlations stronger than |0.2|), by imputation method and predictor set specification.

Method	Predictor set	Min.	5%	25%	50%	75%	95%	Max.
Random	None	−65.5	−65.1	−64.8	−64.6	−64.4	−63.9	−63.5
LRM	unrestricted	−20.3	−14.0	−6.8	−4.5	−2.6	3.9	15.8
LRM	$\lvert\rho\rvert > 0.10$	−17.8	−10.1	−4.0	−1.5	0.6	7.4	20.7
LRM	$\lvert\rho\rvert > 0.20$	−25.5	−10.1	−4.2	−1.4	0.7	6.5	8.2
CRM	unrestricted	−81.2	−68.4	−50.7	−37.1	−21.9	−16.1	−11.9
CRM	$\lvert\rho\rvert > 0.10$	−61.0	−50.3	−32.5	−14.3	−5.6	−1.8	0.2
CRM	$\lvert\rho\rvert > 0.20$	−60.1	−50.3	−31.6	−14.5	−2.7	0.5	3.7
PMM	unrestricted	−31.7	−19.1	−13.5	−11.6	−10.5	−8.2	−6.6
PMM	$\lvert\rho\rvert > 0.10$	−17.6	−9.6	−6.8	−5.0	−3.4	−1.7	3.0
PMM	$\lvert\rho\rvert > 0.20$	−18.5	−12.0	−5.4	−3.2	−2.0	−0.7	2.0
PLS-PMM	data-driven	−20.1	−13.7	−4.5	−1.0	1.4	9.8	34.4
CART	unrestricted	−47.2	−40.9	−31.7	−27.0	−16.8	−3.9	0.1
CART	$\lvert\rho\rvert > 0.10$	−42.4	−33.4	−27.2	−21.4	−12.3	−1.6	0.3
CART	$\lvert\rho\rvert > 0.20$	−63.3	−34.5	−23.2	−19.1	−10.5	−1.1	0.4
JM-T	unrestricted	−35.5	−28.5	−16.9	−13.2	−11.2	−8.8	−5.5
JM-U	unrestricted	−17.0	−9.1	−3.4	−0.3	1.7	8.7	21.8

See Note Figure 1.

Acknowledgments

The authors wish to thank Azim Selvi for his assistance and Hannah Laumann for language editing. The authors gratefully acknowledge support by the state

Table A2

Quantile distribution of average percentage Monte Carlo biases of Spearman correlations for 388 item pairs with weak relationships (true correlations weaker than |0.2| but stronger than |0.05|), by imputation method.

	Min.	5%	25%	50%	75%	95%	Max.
Random	−67.0	−65.4	−64.9	−64.5	−64.1	−63.6	−62.7
LRM	−45.9	−24.6	−11.9	−5.9	−1.0	7.7	72.2
CRM	−119.8	−89.7	−60.3	−41.2	−27.1	−14.3	−1.0
PMM	−63.2	−29.9	−20.2	−15.6	−12.0	−5.6	16.8
PLS-PMM	−44.8	−27.7	−14.6	−4.2	2.9	31.1	106.6
CART	−79.9	−55.5	−45.1	−37.4	−29.3	−16.2	59.0
JM-T	−54.8	−40.6	−28.9	−21.4	−15.9	−8.5	24.4
JM-U	−44.2	−21.4	−7.9	−1.3	3.3	13.3	77.2

See Note Figure 1.

of Baden-Württemberg through bwHPC for providing high-performance computing facilities for the Monte Carlo simulation.

Supplementary Material

Code and Data Availability

(doi: 10.1214/22-SS137SUPP; .zip). All computer code used for this paper and a documentation are available as supplementary material for replication purposes [4]. The data can be accessed via the GESIS Leibniz Institute for the Social Sciences:
https://doi.org/10.4232/1.12607 (wave 1),
https://doi.org/10.4232/1.12619 (wave 13),
https://doi.org/10.4232/1.13390 (wave 37),
https://doi.org/10.4232/1.13391 (wave 38).

References

[1] Adigüzel, F. and Wedel, M. (2008). Split questionnaire design for massive surveys. *Journal of Marketing Research* **45** 608–617.

[2] Allison, P. D. (2005). Imputation of Categorical Variables with PROC MI. In *Proceedings of the SAS Users Group International (SUGI)* **30** 113–30. SAS Institute, Cary.

[3] Akande, O, Li, F. and Reiter, J. (2017). An Empirical Comparison of Multiple Imputation Methods for Categorical Data. *The American Statistician* **71** 162–170. MR3668704

[4] Axenfeld, J. B., Bruch, C. and Wolf, C. (2022). *Code and Data Availability*. Supplement to "General-purpose imputation of planned missing data in social surveys: Different strategies and their effect on correlations."

[5] Axenfeld, J. B., Blom, A.G., Bruch, C. and Wolf, C. (2022). Split Questionnaire Designs for Online Surveys: The Impact of Module Construction on Imputation Quality. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smab055

[6] Bahrami, S., Assmann, C., Meinfelder, F. and Rässler, S. (2014). A split questionnaire survey design for data with block structure correlation matrix. In *Improving Survey Methods: Lessons from Recent Research*, (U. Engel, B. Jann, P. Lynn, A. Scherpenzeel and P. Sturgis, eds.) 368–380. Routledge, New York.

[7] Bartlett, J. W., Seaman, S. R., White, I. R. and Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research* **24** 462–487. MR3372102

[8] Bellman, R. E. (1961). *Adaptive control processes: a guided tour*. Princeton University Press, Princeton. MR0134403

[9] Blom, A. G., Bossert, D., Funke, F., Gebhard, F., Holthausen, A. and Krieger, U.; SFB 884 "Political Economy of Reforms" Universität Mannheim (2016). *German Internet Panel, Wave 1 - Core Study (September 2012)*. GESIS Data Archive, Cologne. ZA5866 Data file Version 2.0.0. https://doi.org/10.4232/1.12607.

[10] Blom, A. G., Bossert, D., Gebhard, F., Funke, F., Holthausen, A. and Krieger, U.; SFB 884 "Political Economy of Reforms" Universität Mannheim (2016). *German Internet Panel, Wave 13 - Core Study (September 2014)*. GESIS Data Archive, Cologne. ZA5924 Data file Version 2.0.0. https://doi.org/10.4232/1.12619.

[11] Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T. and Wenz, A.; SFB 884 "Political Economy of Reforms", Universität Mannheim (2019). *German Internet Panel, Wave 37 - Core Study (September 2018)*. GESIS Data Archive, Cologne. ZA6957 Data file Version 1.0.0. https://doi.org/10.4232/1.13390.

[12] Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, R. and Wenz, A.; SFB 884 "Political Economy of Reforms", Universität Mannheim (2019). *German Internet Panel, Wave 38 (November 2018)*. GESIS Data Archive, Cologne. ZA6958 Data file Version 1.0.0. https://doi.org/10.4232/1.13391.

[13] Blom, A. G., Gathmann, C. and Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. *Field Methods* **27** 391–408.

[14] Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U. and Bossert, D. (2017). Does the recruitment of offline households increase the sample representativeness of probability-based online panels? Evidence from the German Internet Panel. *Social Science Computer Review* **35** 498–520.

[15] Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal* **15** 651–675. MR2530371

[16] Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Erasmus University Rotterdam, Rotterdam.

[17] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J.

(1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey. MR0726392

[18] Burgette, L. F. and Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, **172** 1070–1076.

[19] Cornesse, C., Felderer, B., Fikel, M., Krieger, U. and Blom, A. G. (2021). Recruiting a probability-based online panel via postal mail: experimental evidence. *Social Science Computer Review*. doi:10.1177/08944393211006059

[20] De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18** 251–263.

[21] Doove, L. L., van Buuren, S. and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis* **72** 92–104. MR3139350

[22] Galesic, M. and Bosnjak, M (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly* **73** 349–360.

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

[24] Graham, J. W., Olchowski, A. E. and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, **8** 206–213.

[25] Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, **54** 561–581.

[26] Honaker, J., King, G. and Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software* **45** 1–47.

[27] Horton, N. J., Lipsitz, S. R. and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician* **57** 229–232. MR2016255

[28] Imbriano, P. M. and Raghunathan, T. E. (2020). Three-Form Split Questionnaire Design for Panel Surveys. *Journal of Official Statistics* **36** 827–854.

[29] Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology* **14** 3–15.

[30] Koller-Meinfelder, F. (2009). *Analysis of incomplete survey data-multiple imputation via Bayesian bootstrap predictive mean matching*. University of Bamberg, Bamberg.

[31] Lee, K. J. and Carlin, J. B. (2010). Multiple imputation in the presence of non-normal data. *Statistics in Medicine* **171** 624–632. MR3594613

[32] Little, R. J. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics* **6** 287–296.

[33] Long, J. S. (1997). *Regression models for categorical and limited*

dependent variables. Sage, Thousand Oaks.

[34] Luijkx, R., Jónsdóttir, G. A., Gummer, T., Ernst Stähli, M., Fredriksen, M., Reeskens, T., Ketola, K., Brislinger, E., Christmann, P., Gunnarsson, S. Þ., Bragi, Á., Hjaltason, D. J., Lomazzi, V., Maineri, A. M., Milbert, P., Ochsner, M., Pollien, A., Sapin, M., Solanes, I., Verhoeven, S. and Wolf, C. (2021). The European Values Study 2017: On the way to the future using mixed-modes. *European Sociological Review* **37** 330–346.

[35] Mevik, B.-H. and Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software* **18**(2) 1–24.

[36] Microsoft and Weston, S. (2020). *foreach: Provides Foreach Looping Construct*. R package version 1.5.0.

[37] Morris, T. P., White, I. R. and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology* **14** 1–13.

[38] Munger, G. F. and Loyd, B. H. (1988). The use of multiple matrix sampling for survey research. *The Journal of Experimental Education* **56** 187–191.

[39] Nicoletti, C. and Peracchi, F. (2006). The effects of income imputation on microanalyses: evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169** 625–646. MR2236924

[40] OECD (2014). *PISA 2012 Technical Report*. OECD, Paris.

[41] Peytchev, A. and Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods* **11** 361–368.

[42] R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

[43] Raghunathan, T. E. and Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association* **90** 54–63.

[44] Rässler, S., Koller, F. and Mäenpää, C. (2002). A split questionnaire survey design applied to German media and consumer surveys. In *Friedrich-Alexander University Erlangen-Nuremberg, Chair of Statistics and Econometrics Discussion Papers* [online], available at https://www.statistik.rw.fau.de/files/2016/03/d0042b.pdf.

[45] Robitzsch, A. and Grund, S. (2021). *miceadds: Some Additional Multiple Imputation Functions, Especially for 'mice'*. R package version 3.11-6.

[46] Rubin, D. B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics* **4** 87–94.

[47] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York. MR0899519

[48] Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research* **33** 545–571.

[49] Schafer, J. L. (1999). *NORM users guide (version 2)*. The Methodology Center, The Pennsylvania State University, University Park.

[50] Seaman, S. R., Bartlett, J. W. and White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology* **12** 1–13.

[51] Slade, E. and Naylor, M. G. (2020). A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Statistics in Medicine* **39** 1156–1166. MR4075852

[52] Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology* **179** 764–774.

[53] Shoemaker, D. M. (1973). *Principles and Procedures of Multiple Matrix Sampling*. Ballinger, Cambridge, MA.

[54] Siddique, J. and Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine* **27** 83–102. MR2416864

[55] Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., Arppe, A., Baddeley, A., Barton, K., Bolker, B., Borchers, H. W., Caeiro, F., Champely, S., Chessel, D., Chhay, L., Cooper, N., Cummins, C., Dewey, M., Doran, H. C., Dray, S., Dupont, C., Eddelbuettel, D., Ekstrom, C., Elff, M., Enos, J., Farebrother, R. W., Fox, J., Francois, R., Friendly, M., Galili, T., Gamer, M., Gastwirth, J. L., Gegzna, V., Gel, Y. R., Graber, S., Gross, J., Grothendieck, G., Harrell Jr, F. E., Heiberger, R., Hoehle, M., Hoffmann, C. W., Hojsgaard, S., Hothorn, T., Huerzeler, M., Hui, W. W., Hurd, P., Hyndman, R. J., Jackson, C., Kohl, M., Korpela, M., Kuhn, M., Labes, D., Leisch, F., Lemon, J., Li, D., Maechler, M., Magnusson, A., Mainwaring, B., Malter, D., Marsaglia, G., Marsaglia, J., Matei, A., Meyer, D., Miao, W., Millo, G., Min, Y., Mitchell, D., Mueller, F., Naepflin, M., Navarro, D., Nilsson, H., Nordhausen, K., Ogle, D., Ooi, H., Parsons, N., Pavoine, S., Plate, T., Prendergast, L., Rapold, R., Revelle, W., Rinker, T., Ripley, B. D., Rodriguez, C., Russell, N., Sabbe, N., Scherer, R., Seshan, V. E., Smithson, M., Snow, G., Soetaert, K., Stahel, W. A., Stephenson, A., Stevenson, M, Stubner, R., Templ, M., Temple Lang, D., Therneau, T., Tille, Y., Torgo, L., Trapletti, A., Ulrich, J., Ushey, K., VanDerWal, J., Venables, B., Verzani, J., Villacorta Iglesias, P. J., Warnes, G. R., Wellek, S., Wickham, H., Wilcox, R. R., Wolf, P., Wollschlaeger, D., Wood, J., Wu, Y., Yee, T. and Zeileis, A. (2020). *DescTools: Tools for descriptive statistics*. R package version 0.99.36.

[56] Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J. and Johnson, C. L. (2006). An evaluation of matrix sampling methods

using data from the National Health and Nutrition Examination Survey. *Survey Methodology* **32** 217–231.

[57] Van Belle, G. (2002). *Statistical Rules of Thumb*. John Wiley & Sons, New York. MR1886359

[58] Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC press, Boca Raton, 2nd Edition.

[59] Van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18** 681–694.

[60] Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76** 1049–1064. MR2307507

[61] Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **45**(3) 1–67.

[62] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York. MR1337030

[63] Von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology* **39** 265–291.

[64] Von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research* **42** 105–138. MR3190726

[65] Von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research* **49** 699–718. MR4123147

[66] Weston, S. (2017). *doMPI: foreach parallel adaptor for the Rmpi package*. R package version 0.2.2.

[67] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.

[68] Wickham, H. and Henry, L. (2019). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.8.3.

[69] Wickham, H. and Miller, E. (2019). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.1.1.

[70] White, I. R., Royston, P. and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* **30** 377–399. MR2758870

[71] Wu, H. and Leung, S.O. (2017). Can Likert scales be treated as interval scales?—A simulation study. *Journal of Social Service Research* **43** 527–532.

[72] Wu, W., Jia, F. and Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research* **50** 484–503.

[73] Yu, H. (2002). Rmpi: Parallel statistical computing in R. *R News* **2**(2) 10–14.