Imputation of Missing Data From Split Questionnaire Designs in Social Surveys

Julian B. Axenfeld

Imputation of Missing Data From Split Questionnaire Designs in Social Surveys

Julian B. Axenfeld

Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Sozialwissenschaften der Universität Mannheim

vorgelegt von
Julian B. Axenfeld

Dekan der Fakultät für Sozialwissenschaften:
Prof. Dr. Michael Diehl
1. Betreuerin:
Prof. Annelies G. Blom, Ph.D.
2. Betreuer:
Prof. Dr. Christof Wolf
Erstgutachter:
Prof. Dr. Florian Keusch
Zweitgutachter:
Prof. Joseph W. Sakshaug, Ph.D.
Tag der Verteidigung:
24.07.2023

List of papers

I Split Questionnaire Designs for Online Surveys: The Impact of Module Construction on Imputation Quality

II General Purpose Imputation of Planned Missing Data in Social Surveys: Different Strategies and Their Effect on Correlations

III The Performance of Multiple Imputation in Social Surveys With Missing Data From Planned Missingness and Item Nonresponse

IV Effects of General Purpose Imputations in Planned Missing Survey Data on the Estimation of a Multiple Regression Model: A Case Study

Acknowledgments

This dissertation would not have been possible without the help and support by countless people over the previous years. I would like to express my immense gratitude towards each and everyone of them. Attempting to acknowledge everyone may be a futile endeavor. However, as planned missingness is the very topic of this dissertation, I feel that I can take the risk of producing incomplete data and thank some people explicitly.

I am indebted to my supervisors and co-authors, Annelies Blom and Christof Wolf, for their invaluable guidance and unwavering support throughout all aspects of this dissertation. Their insights and inspirations have played an instrumental role in shaping this work. I would also like to express my heartfelt gratitude to my colleague and co-author Christian Bruch, whose valuable advice in numerous in-depth discussions, even on seemingly trivial details, have significantly strengthened this work.

Furthermore, I would like to extend my appreciation to Barbara Felderer, Jessica Herzing, Jan Karem Höhne, Tobias Rettig, Marina Ungefucht, and all my other former colleagues at the German Internet Panel for their helpful advice and support throughout the process of preparing this dissertation. I would also like to acknowledge support by the state of Baden-Württemberg through bwHPC for providing the necessary high-performance computing resources for this research.

Finally, my deepest thanks go towards my family and friends. I am grateful for their understanding and patience especially during all those times when we have seen each other far too infrequently. I want to thank my parents, Karin Heiß and Hans Georg Axenfeld, who have always supported and encouraged me throughout this journey, and my brother, Laurin Axenfeld, who also deserves special appreciation for his incredibly helpful feedback on this work. Last but not least, I would like to say a special thank you my wife, Luisa, for always being by my side, for all her advice, and for helping me see the positive, even in times when the world seemed full of insurmountable obstacles to me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Survey research in the social sciences is facing a conundrum: Response rates have persistently declined for decades and across different countries (de Heer and de Leeuw, 2002; de Leeuw et al., 2018). Not only does this development continuously increase the fieldwork efforts necessary to obtain sufficient case numbers, making surveys more expensive. It also raises doubts about validity of inferences drawn about relevant target populations due to nonresponse bias. Consequently, ensuring acceptable response rates constitutes a significant challenge of survey methodology.

Managing the length of a survey is a critical factor in optimizing response rates. On the one hand, researchers may desire data that thoroughly cover all potentially relevant aspects of the survey topic. On the other hand, this may lead to a lengthy, burdensome survey that may discourage respondents, resulting in low response rates (Dillman et al., 1993). Beyond that, response quality may suffer, with increased breakoff and measurement error (Galesic and Bosnjak, 2009; Peytchev and Peytcheva, 2017).

This is becoming even more relevant given the recent continual shift from traditional survey modes to self-administered online surveys. On the one hand, online surveys are relatively inexpensive compared to other survey modes (e.g., Lozar Manfreda et al., 2008), which helps contain the ever-increasing costs of conducting a survey. On the other hand, online surveys have narrow limits in questionnaire length due to a greater susceptibility for breakoffs (Peytchev, 2009; Tourangeau et al., 2013, p. 52). Therefore, limiting survey length is considered especially impor-

tant for online surveys. Thus, by moving online, one may be forced to cut down on the number of questions asked in a survey, potentially resulting in the cancellation of important research projects due to limited resources.

1.1 Planned Missingness and Split Questionnaire Designs

One idea proposed by previous research to resolve this issue is planned missingness (e.g., Shoemaker, 1973; Raghunathan and Grizzle, 1995; Graham et al., 1996), where each respondent receives only a subset of all questions rather than the entire questionnaire. This results in shorter questionnaires for individual respondents but also generates considerable amounts of planned missing data. Several approaches to implement planned missingness in surveys have been developed building on this notion. With *multiple matrix sampling* (Good, 1969, 1970; Shoemaker, 1973; Munger and Loyd, 1988), each respondent is assigned a predefined number of questions from the entire questionnaire via simple random sampling. This can be complemented by a so-called core module that is presented to each respondent (e.g., Munger and Loyd, 1988), containing items that are deemed essential and therefore need to be observed completely. Although effectively reducing questionnaire length, this procedure may yield data in which some pairs of variables have no overlapping observations, making it impossible to study their relationships.

The *split questionnaire design* (SQD; Raghunathan and Grizzle, 1995) and similarly the *3-form design* (Graham et al., 1996) are advancements of multiple matrix sampling that solve this problem.[1] Here, questions are allocated to one of several modules (also called components). Then, a subset of two or more modules is randomly assigned to each respondent. This modularization procedure limits the number of different questionnaire forms and also ensures that there are pairwise complete observations available for at least each bivariate relation of variables. Figure 1.1 illustrates this procedure in a fictional example survey with 24 items (displayed in the columns) and 20 respondents (displayed in the rows).[2] In this example, each item is allocated to one of six modules. To ensure similar questionnaire length for each questionnaire form, all modules contain the same number of items (four).

[1] Although this dissertation's focus is on SQDs, note that there also are further, somewhat different developments such as two-method measurement designs (Graham et al., 2006).

[2] Note that this is a small-scale example to ease the reader's understanding. In practice, SQD surveys may cover much more items and respondents.

Module	Core				1				2				3				4				5			
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Respondent 1	•	•	•	•					•	•	•	•	•	•	•	•					•	•	•	•
Respondent 2	•	•	•	•	•	•	•	•					•	•	•	•					•	•	•	•
Respondent 3	•	•	•	•	•	•	•	•					•	•	•	•	•	•	•	•				
Respondent 4	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•								
Respondent 5	•	•	•	•									•	•	•	•	•	•	•	•	•	•	•	•
Respondent 6	•	•	•	•	•	•	•	•	•	•	•	•									•	•	•	•
Respondent 7	•	•	•	•	•	•	•	•	•	•	•	•					•	•	•	•				
Respondent 8	•	•	•	•					•	•	•	•					•	•	•	•	•	•	•	•
Respondent 9	•	•	•	•	•	•	•	•									•	•	•	•	•	•	•	•
Respondent 10	•	•	•	•					•	•	•	•	•	•	•	•	•	•	•	•				
Respondent 11	•	•	•	•									•	•	•	•	•	•	•	•	•	•	•	•
Respondent 12	•	•	•	•	•	•	•	•									•	•	•	•	•	•	•	•
Respondent 13	•	•	•	•	•	•	•	•					•	•	•	•	•	•	•	•				
Respondent 14	•	•	•	•	•	•	•	•					•	•	•	•					•	•	•	•
Respondent 15	•	•	•	•	•	•	•	•	•	•	•	•					•	•	•	•				
Respondent 16	•	•	•	•	•	•	•	•	•	•	•	•									•	•	•	•
Respondent 17	•	•	•	•					•	•	•	•	•	•	•	•	•	•	•	•				
Respondent 18	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•								
Respondent 19	•	•	•	•					•	•	•	•	•	•	•	•					•	•	•	•
Respondent 20	•	•	•	•					•	•	•	•					•	•	•	•	•	•	•	•

Figure 1.1: Illustration of a split questionnaire design in a fictional example survey with 24 items and 20 respondents. Bullet points indicate an item is presented to a respondent.

The left-most module in Figure 1.1 is a core module, implying that all respondent receive the items from this module. These might cover, for example, important sociodemographic characteristics or central outcome variables of this data collection project. The other five modules are split modules, with each respondent receiving a randomly selected subset of these modules. In this example, each respondent receives three split modules plus the core module, resulting in 2 out of 6 modules (or 8 out of 24 items) missing by design for each respondent. Hence, items in split modules here have 40% planned missing data. Overall (i.e., including the core module), 33% of all values in the data are missing by design, meaning that here the questionnaire length is approximately reduced by a third.

1.2 Imputation of Planned Missing Data

Enormous amounts of planned missing data as displayed in Figure 1.1 can affect the analyzability of the data. In this example, two items from different split modules have only 36% of the entire sample pairwise observed. This effect is especially troublesome with multivariate analyses that include variables from multiple different split modules. In this case, the case numbers available for the analysis may quickly drop to zero.

To help with this issue, Raghunathan and Grizzle (1995) propose using multiple imputation (Rubin, 1987; van Buuren, 2018) to complete the planned missing data. Multiple imputation is a statistical technique that replaces each missing value in a dataset by multiple values that are statistically plausible given the information available in the observed data. To impute a target variable, an imputation model is estimated based on a set of predictor variables. In doing so, multiple imputation aims to preserve marginal distributions, the relations between variables, and the uncertainty of the missing data. The resulting imputed datasets can then be analyzed separately, with the resulting multiple estimates being pooled thereafter.

However, the imputation of planned missing social survey data comes with numerous challenges. First, the estimation of imputation models relies on correlations between the imputed variable and predictor variables, but correlations between variables in social survey data are often weak. Second, survey data is often categorical rather than continuous, but categorical data is considered more difficult to impute than continuous data (van Buuren, 2018, p. 91). Third, large proportions of missing data need to be imputed. This implies that even small inaccuracies in the imputation model can have significant effects on estimates after imputation. Furthermore, large proportions of missing data also challenge the estimation of imputation models, as this means they have to rely on very limited amounts of observed data.

Moreover, the imputation scenario itself could be a challenge. In an ideal case, planned missing survey data may be imputed right away by the data-collecting research institute for general research purposes in order to provide imputed data to individual data users. This would take away burden from data users who might lack the resources and training necessary to conduct multiple imputation by themselves. Furthermore, the data-collecting institute may have the statistical expertise, fieldwork knowledge, and the computational resources to set up a suitable imputation procedure adequately taking into account the particular features of the data and of the data collection process. This approach might also be the most efficient in terms of financial costs, working hours, and energy consumption.

Yet, a general purpose imputation strategy makes the task of imputing planned missing survey data even more difficult: In statistical theory, imputation models must (at the very least) always include all analysis variables to preserve their relation with the imputed variable (Meng, 1994). However, with a general purpose imputation strategy, the analysis models typically are unknown. As this may mean that in principle, all relations between variables could end up in an analysis model, imputation models may need to cover all variables as predictors. This intensifies

issues with the estimation of imputation models: The imputation needs to deal with a lot of predictors that need to be included simultaneously but also with very limited observed data.

So far, there is little evidence on how well the multiple imputation of planned missing data might work with actual social survey data under these circumstances. Moreover, there is an imperative need for more research on different strategies how to implement split questionnaire designs in social surveys to determine when and how multiple imputation can be successfully applied with these data. This need for research involves the whole process of implementing an SQD, starting with the design of questionnaires up to the imputation and provision of the resulting data.

First, existing research often claims that it is essential to distribute highly correlated variables across different modules in order to ensure an adequate quality of imputations (e.g., Raghunathan and Grizzle, 1995). However, there is little evidence as to what extent this makes a difference for estimates with real social survey data, which often lack particularly strong correlations.

Second, despite existing research on the performance of different imputation procedures in general (e.g., Akande et al., 2017; Collins et al., 2001; Slade and Naylor, 2020), we currently lack evidence on which imputation procedures may perform satisfactorily specifically in imputing planned missing survey data. This is especially the case for the enormous, far-reaching task of imputing data for general research purposes.

Third, beyond the planned missing data that is usually missing completely at random, surveys typically also exhibit some degree of potentially non-random item nonresponse by the survey participants which needs to be imputed as well to mitigate nonresponse bias. However, there currently are no studies on how the combined presence of planned missingness and item nonresponse affect imputation performance, whether the imputation still manages to adjust for nonresponse bias under these circumstances, and how item nonresponse should be taken into account in the design of split questionnaires.

Finally, there currently is no systematic investigation of how estimates for a multivariate model may be affected by a general purpose imputation strategy compared to an analysis-specific imputation strategy. However, such research is critical in order to find best-practice procedures for implementing SQDs as well as to evaluate the value of SQDs for social survey practice.

1.3 Contribution of This Dissertation

This dissertation contributes to the research on SQDs and their imputation, addressing the overarching research question under which conditions accurate estimates can be obtained in a social survey with an SQD in practice. Using a series of Monte Carlo simulation studies based on real social survey data from the German Internet Panel (GIP; Blom et al., 2015; Cornesse et al., 2021) and the European Social Survey (ESS; European Social Survey, 2018a,b,c), I examine the quality of univariate, bivariate, and multivariate estimates after imputation. In doing so, I manipulate various features of the SQD survey and of the imputation procedure, dealing with the research gaps discussed above.

This work improves the research community's understanding of how multiple imputation in surveys may perform in practice for SQD scenarios under realistic conditions and with real data. Through this work, I identify appropriate strategies for designing split questionnaires and imputing the resulting data. Moreover, I outline the conditions necessary to enable acceptably accurate estimates in practice given state-of-the-art imputation routines. The following paragraphs provide a more detailed synopsis of the four papers resulting from this research.

1.4 Synopsis of Papers in This Dissertation

1.4.1 Paper I: Split Questionnaire Designs for Online Surveys: The Impact of Module Construction on Imputation Quality

The first paper investigates the impact of module construction on the quality of univariate and bivariate estimates after imputation in SQDs. In doing so, different perspectives on module construction are taken into account. On the one hand, questionnaire developers may want to design a questionnaire that appears coherent and easy to understand to respondents, and may therefore like to avoid frequent changes in topics. In consequence, their preferred modularization strategy may be to construct modules each containing only a single topic. On the other hand, Raghunathan and Grizzle (1995) argue that highly correlated items should be systematically allocated to different modules to ensure a good quality of imputations. Yet, in practice highly correlated variables can usually be found mostly within the same survey

topic. Following this logic, constructing modules covering all survey topics would be a preferable option.

In a Monte Carlo simulation using data from the GIP waves 37 and 38 (Blom et al., 2019a,b), SQDs are simulated by deleting values from the complete sample data. Three module construction strategies are examined: Randomly constructed modules, modules each covering a single survey topic, and modules covering diverse (all) survey topics. After multiple imputation of the simulated missing values, univariate frequencies and bivariate Spearman correlations are estimated between all variables in the data and compared to estimates based on the complete sample data.The main finding is that while random and diverse topics modules perform very similarly, single topic modules lead to mostly less accurate estimates. Nonetheless, each of the modularization strategies results in some estimates—particularly correlations—being severely biased.

1.4.2 Paper II: General Purpose Imputation of Planned Missing Social Survey Data: Different Strategies and Their Effect on Correlations

Following up on the finding that especially correlation estimates from SQD data after imputation can turn out severely biased, in the second paper a wide range of different imputation procedures are reviewed and tested in their accuracy regarding Spearman correlation estimates. This entails both different imputation methods and different predictor set specifications are being tested. With respect to the latter, two strategies are tested that systematically exclude predictor variables with near-zero correlations to the imputed variable from the imputation model. This builds on the assumption that not reproducing near-null relationships would not harm estimates after imputation too much (see also the concept of semi-compatibility, Bartlett et al., 2015). Furthermore, partial least squares predictive mean matching is tested as well as a more sophisticated technique to simplify predictor sets (Robitzsch et al., 2016; Robitzsch and Grund, 2021). This technique uses partial least squares regression to reduce the dimensionality of the predictor space before imputing the data via predictive mean matching.

Again, a Monte Carlo simulation is applied using the same dataset as in Paper I, simulating SQDs with random modules. Two major findings stand out. First, several established imputation methods can result in strong biases in correlation estimates when imputing SQD data, especially generalized linear models for categori-

cal data and classification trees. Second, combining predictive mean matching with restricted predictor sets or partial least squares regression can help to reduce biases in correlation estimates. These findings also highlight the challenge with imputing multinomial variables, for which no satisfying solution appears to be available yet.

1.4.3 Paper III: The Performance of Multiple Imputation in Social Surveys With Missing Data From Planned Missingness and Item Nonresponse

So far, previous research (including the two preceding papers) considered only one single source of missing data at once. In an SQD survey, however, different sources of missing data entailing very different challenges can be expected to emerge. The primary challenge of planned missing data, as discussed above, lies in its typically large quantity in relation to the amount of observed data. In contrast, the primary challenge of nonresponse by survey participants is that its missingness may not emerge randomly (as is typically the case with planned missing data). Both kinds of missingness combined may result in a large amount of missing data with a potentially non-random, heterogeneous missingness mechanism: One part is missing completely at random, while another part follows a distinct unknown mechanism. This might interfere with the imputation in several ways: First, the larger and partially uncontrollable amounts of missing data from both sources may further aggravate issues with many values to be imputed but little data to support an imputation model. Second, it needs to be investigated whether the imputation model can still account for a heterogeneous missingness mechanism (as induced by the combined presence of planned missingness and item nonresponse) in spite of relatively little observed data and a large set of predictor variables.

In consequence, this paper provides another Monte Carlo simulation based on the GIP dataset in which the accuracy of univariate frequency estimates and bivariate Spearman correlation estimates is examined under a wide range of different scenarios. These cover the amount and mechanism of item nonresponse and the amount of planned missing data. The simulation of item nonresponse mimics the item nonresponse observed in the GIP as modeled through elastic net logistic regressions. The results show that besides the item nonresponse potentially being non-ignorable, the major challenge of item nonresponse in an SQD survey is that it can increase the already large proportions of missing data from the SQD to such a degree that estimates can turn out severely biased after imputation.

1.4.4 Paper IV: Effects of General Purpose Imputations in Planned Missing Survey Data on the Estimation of a Multiple Regression Model: A Case Study

The final paper of this dissertation examines the effects of a general purpose imputation strategy as compared to an analysis-specific imputation strategy based on a case study of a multiple regression model from the social sciences literature (Safi, 2010). Here, I investigate two aspects in which general purpose imputation may differ crucially from analysis-specific imputation: First, a general purpose imputation must preserve all relations in the data and therefore needs to include a lot of predictor variables, while an analysis-specific imputation may have the flexibility to restrict the predictor set quite heavily. Second, a general purpose imputation model would usually be based on the entire survey sample, while an analysis-specific imputation model can be fitted to the specific analysis subsample (assuming they are not the same).

I apply another Monte Carlo simulation study based on data from the ESS, in which I again simulate SQDs with random modules by deleting observed values from a complete dataset and subsequently estimate a regression model of general life satisfaction on a wide range of regressors. In this context, I test the effect of (a) using the gross survey sample vs. the net analysis sample to impute the data and (b) defining imputation models using only the analysis variables vs. using varying numbers of additional correlated and uncorrelated predictor variables. This analysis shows that given adequate dimensionality reduction through partial least squares regression, even adding additional uncorrelated predictor variables (i.e., variables that are worthless for predicting the imputed variable) may have no adverse effects on regression coefficients or standard errors. However, partly considerable biases can occur when the analysis model sample and the imputation model sample are not the same.

1.5 Lessons Learned and the Way Forward

In this dissertation, I investigate the performance of SQDs and multiple imputation with real social survey data, covering the entire process from designing the questionnaires and planning the amount of missingness to imputing and providing the data for data users. This reveals a multitude of important insights with implications for future research.

Imputing the planned missing data under real-data conditions (that is, enormous amounts of missing data, predominantly low correlations and often categorical data) turns out to be a challenge that can potentially be a substantial source of bias for univariate and bivariate as well as multivariate estimates. These difficulties intensify with a general purpose imputation strategy, which aims to preserve all relations between variables (rather than only some specific relations of interest) and therefore often needs to include a vast number of variables into the imputation models. Fortunately, however, suitable strategies for setting up an SQD and imputing the resulting data can largely (though not completely) eliminate biases. These include:

- not using single-topic modules,
- avoiding too large amounts of missing data on each item,
- considering to allocate items with nominal levels of measurement to a core module,
- using appropriate imputation procedures, such as partial least squares predictive mean matching, and
- transparently communicating for which analyses the general-purpose imputed data could be used and for which analyses an analysis-specific imputation is required (this is the case especially for modeling non-continuous effects or analysis models based on a subset of the entire survey sample).

Yet, even if researchers follow these recommendations, the imputation may still remain some source of bias. Therefore, this work suggests that when designing a survey, researchers should carefully weigh their expectations about the benefits of an SQD in terms of response quality and cost savings against potential inaccuracies from the imputation.

Beyond that, the remaining issues with imputing planned missing survey data identified in this dissertation highlight the need for further research regarding the design of SQDs and the imputation of the resulting data. For instance, future research might further develop existing imputation procedures so that multinomial data can be accounted for more appropriately. In addition, future research may also focus on other aspects of SQDs that were out of scope for this thesis. These include, for example, the effects of different modularization techniques on actual response behavior. As such, empirical experimental research on SQDs may be needed in the future, as simulation studies are not suitable to investigate real behavioral effects on respondents.

References

Akande, O, Li, F., & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician 71*(2), 162-170.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462-487.

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

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

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. *Field Methods*, *27*(4), 391-408.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330-351.

Cornesse, C., Felderer, B., Fikel, M., Krieger, U., & Blom, A. G. (2021). Recruiting a probability-based online panel via postal mail: Experimental evidence. *Social Science Computer Review*, *40*(5), 1259-1284.

de Heer, W., & de Leeuw, E. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.), *Survey nonresponse* (pp. 41-54). Wiley.

de Leeuw, E., Hox, J., & Luiten, A. (2018). International nonresponse trends across countries and years: An analysis of 36 years of Labour Force Survey data. *Survey Insights: Methods from the Field*. https://surveyinsights.org/?p=10452

Dillman, D. A., Sinclair, M. D., & Clark, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for

occupant-addressed census mail surveys. *Public Opinion Quarterly*, *57*(3), 289-304.

European Social Survey (2018a). *ESS-1 2002 documentation report. Edition 6.6*. European Social Survey Data Archive, Sikt - Norwegian Agency for Shared Services in Education and Research, Norway for ESS ERIC. https://doi.org/10.21338/NSD-ESS1-2002

European Social Survey (2018b). *ESS-2 2004 documentation report. Edition 3.7*. European Social Survey Data Archive, Sikt - Norwegian Agency for Shared Services in Education and Research, Norway for ESS ERIC. https://doi.org/10.21338/NSD-ESS2-2004

European Social Survey (2018c). *ESS-3 2006 documentation report. Edition 3.7*. European Social Survey Data Archive, Sikt - Norwegian Agency for Shared Services in Education and Research, Norway for ESS ERIC. https://doi.org/10.21338/NSD-ESS3-2006

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360.

Good, I. J. (1969). Split questionnaires. *The American Statistician*, *23*(4), 53-54.

Good, I. J. (1970). Split questionnaires II. *The American Statistician*, *24*(2), 36-37.

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

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological methods*, *11*(4), 323-343.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, *50*(1), 79-104.

Meng, X.-L. (1994). Multiple imputation with uncongenial sources of input. *Statistical Science*, *9*(4), 538-558.

Munger, G. F. & Loyd, B. H. (1988). The use of multiple matrix sampling for survey research. *The Journal of Experimental Education*, *56*(4), 187-191.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74-97.

Peytchev, A. & Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods*, *11*(4), 361-368.

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

Robitzsch, A. & Grund, S. (2021). *miceadds: Some additional multiple imputation functions, especially for 'mice'*. R package version 3.10-28. https://CRAN.R-project.org/package=miceadds

Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In Breit, S. & Schreiner, C. (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der Österreichischen Bildungsstandardüberprüfung [Methodological foundation of standard achievement testing]* (pp. 259-293). facultas.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Safi, M. (2010). Immigrants' life satisfaction in Europe: Between assimilation and discrimination. *European Sociological Review*, *26*(2), 159-176.

Shoemaker, D. M. (1973). *Principles and Procedures of Multiple Matrix Sampling*. Ballinger.

Slade, E. & Naylor, M. G. (2020). A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Statistics in Medicine*, *39*(8), 1156-1166.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford University Press.

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Chapter 2

Split Questionnaire Designs for Online Surveys: The Impact of Module Construction on Imputation Quality

Abstract

Established face-to-face surveys encounter increasing pressures to move online. Such a mode-switch is accompanied with methodological challenges, including the need to shorten the questionnaire that each respondent receives. Split questionnaire designs (SQDs) randomly assign respondents to different fractions of the full questionnaire (modules) and, subsequently, impute the data that are missing by design. Thereby, SQDs reduce the questionnaire length for each respondent. Although some researchers have studied the theoretical implications of SQDs, we still know little about their performance with real data, especially regarding potential approaches to constructing questionnaire modules. In a Monte Carlo study with real survey data, we simulate SQDs in three module-building approaches: random, same topic, and diverse topics. We find that SQDs introduce bias and variability in univariate and, especially, in bivariate distributions, particularly when modules

This paper is joint work with Annelies Blom, Christian Bruch, and Christof Wolf. A similar version of this paper has been published in:
Axenfeld, J. B., Blom, A. G., Bruch, C., & Wolf, C. (2022). Split questionnaire designs for online surveys: The impact ofmodule construction on imputation quality. *Journal of Survey Statistics and Methodology*, *10*(5), 1236–1262.

are constructed with items of the same topic. However, single topic modules yield better estimates for correlations between variables of the same topic.

2.1 Introduction

Surveys are an indispensable source of evidence in the social sciences. Many large-scale face-to-face surveys like the General Social Survey (Smith et al., 2019) or the British Social Attitudes survey (Curtice et al., 2019) stimulate scientific discourse with high-quality data. However, face-to-face surveys are increasingly under pressure due to decreasing response rates (de Leeuw et al., 2018) and increasing costs (e.g., Roberts et al., 2014; Calinescu et al., 2013).

With close to universal internet coverage in Western countries (International Telecommunication Union, 2019), online surveys have become a viable alternative to face-to-face data collection in recent years. At considerably lower cost (e.g., Bianchi et al., 2017; Olson et al., 2021) several large-scale probability-based online surveys have been established across the world (e.g., the KnowledgePanel (Ipsos, 2021) in the US, the LISS Panel in the Netherlands (Knoef and de Vos, 2009), and the German Internet Panel (GIP; Blom et al., 2015)).

Consequently, survey projects face pressures to switch to the less expensive online mode (e.g., Jäckle et al., 2015; Bianchi et al., 2017). However, there is one major obstacle to moving face-to-face surveys online: Online surveys are typically much shorter than those conducted face-to-face, because researchers worry about higher breakoff rates (Galesic, 2006; Mavletova and Couper, 2015; Peytchev, 2009; Revilla, 2017; Tourangeau et al., 2013, p. 52), lower response quality, and higher measurement error (Galesic and Bosnjak, 2009; Peytchev and Peytcheva, 2017) in lengthy online questionnaires. When asking directly, the median online survey respondent reports that they would like to answer surveys of 25 minutes at maximum (Revilla and Höhne, 2020). Many established face-to-face surveys, however, are with approximately one hour considerably longer (Curtice et al., 2019, p. 257) and, thus, would have to be shortened when moved online.

Split questionnaire designs (SQDs) may provide a solution to such obstacles. It allocates the items of a given questionnaire to different modules and randomly assigns respondents to a subset of these modules. Data for the questions not presented to a respondent are missing by design and can subsequently be imputed to allow for applying conventional analysis techniques (Raghunathan and Grizzle, 1995).

While SQDs theoretically provide an attractive solution to shortening online questionnaires, little is still known about their practical implications. Importantly, low variable correlations in real social survey data driven by multi-topic questionnaires and non-exact measurement may lead to biases and inefficiencies in the imputation process: Imputation models rely fundamentally on information on the unobserved data stored in the observed data. Due to generally low correlations, however, observed data cannot contribute much information. Moreover, with SQDs large proportions of the data are imputed, implying that poor imputations could severely affect substantive analyses on the data. Consequently, preserving as much of the scarce information as possible for the imputation is a major challenge for SQD surveys. Otherwise, imputation models might fail to reproduce distributions and relationships in the data, implying potentially inefficient and biased estimates.

In this paper, we therefore shed light on an important practical aspect of SQDs: The construction of the questionnaire modules and its impact on the quality of the imputed data (i.e., biases and variability of frequency and correlation estimates). For a realistic examination of modularization strategies, this study relies on real (non-synthetic) survey data to account for real-data challenges (e.g., low correlations or skewed distributions). We test three modularization methods: random modules (RM), where the questions are randomly allocated to modules, single topic modules (STM), where each module contains only one questionnaire topic, and diverse topics modules (DTM), where the various topics of a questionnaire are spread across several modules. We present findings from a Monte Carlo simulation that examines how RM, STM and DTM affect imputation quality in real survey data.

2.2 Administration of Split Questionnaire Designs

2.2.1 Split Questionnaire Design (SQD)

SQD is a planned missing data method developed by Raghunathan and Grizzle (1995) as an extension of matrix sampling (e.g., Shoemaker, 1973; Munger and Loyd, 1988). Items are bundled to mutually exclusive packages called modules (e.g., Raghunathan and Grizzle, 1995; Peytchev and Peytcheva, 2017). There may be one core module containing especially important items that are administered to all respondents (e.g., Raghunathan and Grizzle, 1995). Additionally, respondents are randomly assigned to a subset of the remaining modules.

Constructing modules instead of sampling items directly is an important aspect of SQD, guaranteeing sufficient pairwise observations for each pair of items (Raghunathan and Grizzle, 1995; Rässler et al., 2002). To this end, every split questionnaire must contain two split modules at minimum, and all possible combinations of split modules must be allowed to appear (Raghunathan and Grizzle, 1995). This general procedure is the same independent of the modularization strategy.

SQDs produce so much missing data that often too few observed cases are available for conventional complete case analyses. As a solution, Raghunathan and Grizzle (1995) suggest multiple imputation (MI; Rubin, 1987) to impute values missing by design.

2.2.2 Multiple Imputation (MI)

MI is a method for completing incomplete data matrices with plausible values to enable analyses on the full data (for a detailed overview, see Rubin, 1987; van Buuren, 2018). MI replaces missing values with values drawn from a posterior probability density distribution. This distribution is obtained by an imputation model relying on a set of predictor variables. Values are drawn multiple times to account for the uncertainty of the missing values, generating multiple datasets with different imputed values. Data analyses are carried out on each dataset separately and estimates are subsequently pooled using Rubin's Rules (Rubin, 1987).

The challenge of MI lies in the reproduction of distributions and relationships that would be observed in a complete dataset. In general, this challenge is best met when the missing information is limited (Madley-Dowd et al., 2019) and correlations between imputed and predictor variables are strong. However, correlations in surveys are typically weak, and SQDs produce lots of missing data. The aim of choosing a modularization strategy for SQDs is thus to maximize the information that predictors provide on the variables to be imputed (Raghunathan and Grizzle, 1995). In practice this means that relatively highly correlated variables need to be allocated to different modules to prevent them from being missing together.

2.2.3 Modularization Techniques

The module construction strategy may decisively shape the resulting SQD. First, as described above, the imputation requires retaining as much information as possible, i.e., correlated variables should be distributed across modules.

Second, however, certain items should not be separated (Raghunathan and Grizzle, 1995; Rässler et al., 2002). For example, this can be motivated by the need to maintain question filtering (see for instance Bishop et al., 1983; Kreuter et al., 2011, for question-filter effects on data quality), prevent differential order effects (e.g., McFarland, 1981; Silber et al., 2016), or limit frequent topic switches that may raise respondent burden.

Finally, module construction must be feasible in real survey settings. Thus, all information used during modularization must be available or obtainable before data collection. Exact variable correlations, for example, are not available a priori; instead we have to rely on previous surveys or collect this information during a pilot study.

Thus, guidance on modularization will depend on how various perspectives are weighted. Similar to Gonzalez and Eltinge (2007), we classify such different techniques into three general strategies: RM, STM, and DTM. Figure 2.1 illustrates these three strategies with a small example questionnaire.

Random modules (RM)

The upper part of Figure 2.1 shows one potential outcome when modules are constructed randomly in an example questionnaire. The questionnaire is a set Q of questions described by the index $q = 1, 2, \ldots, Z$, where Z is the total number of questions in the questionnaire (in this example, $Z = 9$). All questions in Q belong to mutually exclusive topics with each topic described by the index $h = 1, 2, \ldots, L$, where L is the total number of topics (here, $L = 3$). For RM, we want to randomly allocate all questions to a fixed number M of split modules, which are mutually exclusive and described by the set W with the index $w = 1, 2, \ldots, M$ denoting a certain module. The number of modules M can in principle be set to any value $2 < M \leq Z$ (in the example, we chose $M = 3$) so that each respondent can receive at least 2 modules.

Furthermore, we suppose modules should be balanced in size so that all respondents receive questionnaires of similar length (Rässler et al., 2002; Thomas et al., 2006). Therefore, we determine uniform module sizes $B_w = Z/M$ if $Z/M \in \mathbb{N}$. If $Z/M \notin \mathbb{N}$, we create two different subsets of modules by randomly drawing a subset V from the set of modules W that contains a number of $M\left(Z/M - \lfloor Z/M \rfloor\right)$ modules. For these two subsets, we define different module

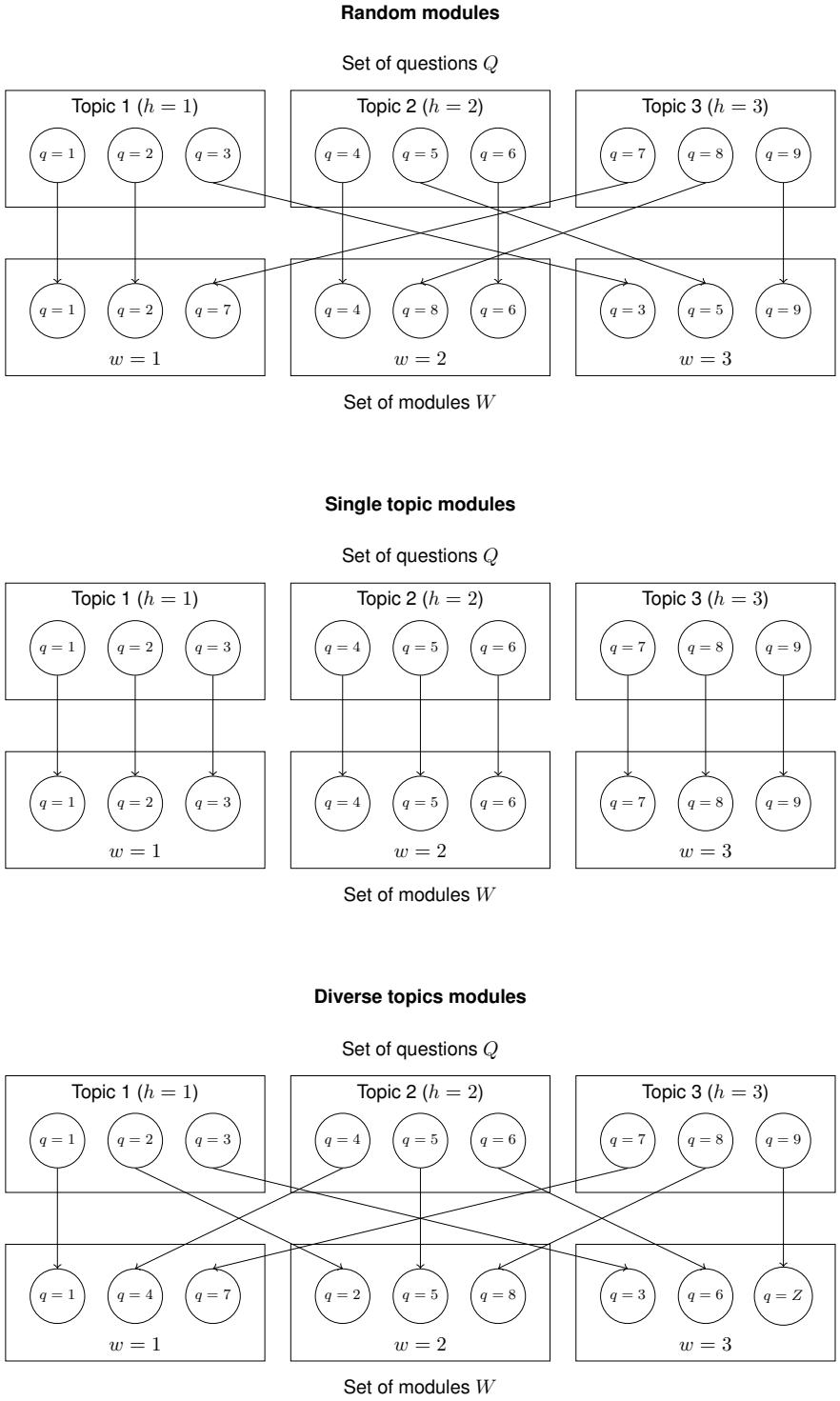

Figure 2.1: Illustration of modularization strategies.

sizes:

$$B_w = \begin{cases} \lceil Z/M \rceil & if \quad w \in V \\ \lfloor Z/M \rfloor & if \quad w \notin V \end{cases} . \tag{2.1}$$

This means each module w will receive a number of items B_w defined by either the ceiling or floor value of the total number of items Z over the total number of modules M, depending on whether the module was or was not in subset V. Then, we randomly assign all questions in Q to the modules with sizes B_w, with each question q having a probability of B_w/Z to be allocated to a module w before the assignment of questions starts.

RM considers no survey information other than the number of questions Z and the predetermined number of modules M. Consequently, imputation quality may suffer, because correlated items are not systematically distributed across modules optimally and could possibly amass within the same module by chance. From a practitioner's perspective, RM might not be optimal either, as question sequences are ignored and hence, meaningful and consistent questionnaires cannot be guaranteed using RM.

Single topic modules (STM)

STM's procedure is illustrated in the middle part of Figure 2.1, again with $Z = 9$ questions, $L = 3$ topics, and $M = 3$ modules. This is a fully deterministic process, where all items of one topic h are allocated to the same single module w. However, if one topic module contains considerably more (or more burdensome) questions than the other topic modules, the large single topic module may be additionally split to achieve balanced module lengths.

The key benefit of STM is that it avoids potential disruptions in the questionnaire structure. STM therefore seems to be the strategy of choice for many survey practitioners, who seek to obtain questionnaires that appear meaningful and consistent to respondents regarding its topics. Consequently, STM has many real-life applications such as in the 2017 European Values Study (Luijkx et al., 2017) and the 2012 PISA study (OECD, 2014, chap. 3).

However, STM may hinder imputation, because most variables on the same topic may deliver the highest correlations but are clustered within rather than distributed across modules. Hence, while RM may trigger adverse scenarios for MI by chance, STM will cause them by design.

Diverse topics modules (DTM)

Finally, DTM purposefully assigns the most highly correlated variables to different modules to optimize subsequent imputation. DTM constitutes a diverse group of techniques that optimize SQDs (examples can be found in Rässler et al., 2002; Thomas et al., 2006; Adigüzel and Wedel, 2008; Chipperfield and Steel, 2009, 2011; Chipperfield et al., 2018; Imbriano, 2018). From an imputation perspective DTM is attractive, because it maximizes the information available for the MI. However, it contains a conundrum: To determine which variables are highly correlated, the data must be available a priori, i.e., before fieldwork. Although some surveys can draw on data from a pilot study, typically these correlations are unknown during modularization. Therefore, this study uses a DTM approach proposed similarly by Bahrami et al. (2014), which assumes that variables correlate more strongly when they originate from questions on the same topic. This implies that all items from a topic h should be evenly distributed over all M modules, such that highly correlated variables will most likely end up in different modules. Since here the topics serve only to identify potentially highly correlated items, practitioners could also consider alternative ways to group highly correlated items other than topics (e.g., prior theoretical knowledge).

The bottom part of Figure 2.1 illustrates a potential outcome of this DTM approach. The procedure is a stratified random assignment, in which the topics described by the index $h = 1, 2, \ldots, L$ serve as strata. Hence, RM is applied separately within each topic h.

We first determine how many questions from a given topic h should end up in each of the modules. This number of questions $B_{w,h}$ is defined by $B_{w,h} = A_h / M$ if $A_h / M \in \mathbb{N}$ in a topic h, where A_h is the number of questions in a topic h (in the example, $A_h = 3$). In Figure 2.1, $B_{w,h} = 1$ for each w and h, so in each topic h one question is allocated to each module w.

Otherwise, if $A_h / M \notin \mathbb{N}$ in a topic h, we create two different subsets of modules by randomly drawing a subset U_h from the set of modules W that contains a number of $M\,(A_h / M - \lfloor A_h / M \rfloor)$ modules. For these two subsets, we define different topic-specific module sizes:

$$B_{w,h} = \begin{cases} \lceil A_h / M \rceil & if \quad w \in U_h \\ \lfloor A_h / M \rfloor & if \quad w \notin U_h \end{cases} \tag{2.2}$$

Thus, from a given topic h, each module w will receive a number of items defined by either the ceiling or floor value of the number of items in the topic A_h over

the number of modules M, depending on whether module w was or was not in U_h. Subsequently, we randomly assign $B_{w,h}$ questions from a topic h to each module w. We apply this procedure to each topic h, yielding modules constructed by stratified random assignment.

Compared to RM, the stratification in DTM can make module sizes vary slightly more. In our study, module sizes turn out constant (always equal to 10). However, practitioners may consider rejecting module structures with sizes that vary too much.

Whereas RM may lead to an underrepresentation of some topics in some modules (in Figure 2.1 for example, module 1 contains no question from topic 2), DTM obtained by stratified random assignment may eliminate the "unluckier" outcomes of RM while requiring only heuristic information on the correlation structure.

2.2.4 Prior Research

Prior research into SQD imputation with real data can be grouped into two categories: Monte Carlo simulations investigating imputation quality with one specific modularization strategy (Bahrami et al., 2014; Raghunathan and Grizzle, 1995; Thomas et al., 2006) and case studies that explore different modularization strategies (Adigüzel and Wedel, 2008; Imbriano and Raghunathan, 2020; Rässler et al., 2002).

From existing simulation studies we learn that "little is lost" regarding means and standard errors (Raghunathan and Grizzle, 1995). Thomas et al. (2006) report only small biases in means and regression coefficients but considerable precision losses in simulated SQDs compared to complete surveys. Bahrami et al. (2014) observe a small attenuation in most of their regression coefficients. As their MI estimates are overall still mostly in line with complete data estimates, they evaluate their design favorably in general.

Furthermore, three single-case (non-Monte Carlo) studies compare different modularization strategies: Adigüzel and Wedel (2008) suggest that data-driven solutions could retain more information than ad-hoc solutions. Additionally, Rässler et al. (2002) briefly report a poorer imputation performance when split modules consist of highly correlated items. Imbriano and Raghunathan (2020) compare different SQDs in a longitudinal health survey context, manipulating whether respondents receive repeatedly the same topics or different topics each wave (whereby correlations of one variable across waves are usually high). They find that univariate and regres-

sion estimates are reproduced best when respondents receive different items each wave (i.e., when highly correlated variables are separated).

To our knowledge, our study is the first to combine the application of Monte Carlo simulations with examining different modularization strategies (RM, STM, and DTM) using real survey data. Furthermore, it also goes beyond most existing real-data evidence through investigating bivariate in addition to univariate measures (e.g., Adigüzel and Wedel, 2008; Raghunathan and Grizzle, 1995, study 1).

2.3 Data and Methods

2.3.1 Data

Our study uses real data from an existing survey: the German Internet Panel (GIP), a probability-based online panel of the German population (for details on recruitment and response rates, see Blom et al., 2015, 2017; Cornesse et al., 2021). The GIP is particularly suited, because it has a reasonably large number of cases (5,411) and a multi-topic structure. The latter arises from independent research teams in various areas of economics, political science, sociology, and data science feeding questionnaires into the GIP to answer their respective research questions.

We used 61 variables from GIP waves 37 and 38 (Blom et al., 2019a,b). Table 2.1 depicts the topics and number of variables selected and indicates whether the variables were used in the core or split modules. The table also shows to which module the variables were allocated with STM. All variables are discrete, most of them ordinal or dichotomous, and seven variables in the core are nominal. Additional information on the wording of survey questions, field-time periods and response rates is provided in Tables 2.A1 and 2.A2.

To pursue our research question of examining different modularization strategies, we rely on imputed data of the planned missing SQD data. In order not to confound the effects of this type of missing data with regular missing data, we removed all unit and item nonresponse from the dataset. Consequently, participants who did not respond to either wave 37 or 38 were excluded from the GIP dataset. Furthermore, where possible, missing observations were matched to responses from earlier waves (Blom et al., 2016a,b). Finally, the remaining item nonresponse was replaced with single imputations using predictive mean matching (PMM) as implemented in the *mice* package in R[1] (van Buuren and Groothuis-Oudshoorn, 2011; R

[1]Other R packages used for this paper are: DescTools (Signorell et al., 2020), doMPI (Weston, 2017), dplyr (Wickham et al., 2019), faux (DeBruine, 2020), foreach (Microsoft and Weston, 2020),

Table 2.1: Variables used in Monte Carlo simulation

Topic	# variables	SQD constituent	Origin	STM allocation
Sociodemographics	10	core	wave 37	core
Sampling cohort	1	core	wave 37	core
Organization membership	10	split	wave 37	module 1
Big Five personality traits	10	split	wave 37	module 2
Lobbying in EU politics	10	split	wave 38	module 3
Domestic and party politics	20	split	wave 38	modules 4&5

Core Team, 2020), using all variables as predictors that have Spearman correlations of $|0.05|$ or stronger. The effects of this procedure on univariate frequencies and correlations appear negligible, as both turn out extremely similar when calculated without imputation (with pairwise deletion) and with imputation (for details, see Figure 2.A1).

Finally, rarely observed categories with fewer than 100 cases were combined into broader categories to avoid obtaining empty categories in the simulation. This yielded a completely observed dataset with 4,061 cases as the population for our simulation.

2.3.2 Variable Correlations Within and Between Topics

To consider the variable correlations in the data set, we calculate a Spearman correlation matrix for the 50 split variables (see Figure 2.A2 for an illustration). Absolute values of correlations range from 0.000 to 0.702 with 81.6% smaller than 0.10. We further evaluate average absolute correlations within and between topics using Fisher's-Z transformation: Different-topic variable pairs tend to have weaker correlations than same-topic variable pairs with an average correlation of 0.046 compared to 0.162 (average correlations within topics are between 0.107 and 0.258). 45.3% of within-topic correlations and 89.8% of between-topic correlations are below 0.10.

Finally, we take a glimpse at the correlations of variables of different modules. The absolute Spearman correlations between variables of different modules are on average 0.049 with STM, 0.070 with RM, and 0.072 with DTM.

ggcorrplot (Kassambara, 2019), MASS (Venables and Ripley, 2002), Matrix (Bates and Maechler, 2019), Rmpi (Yu, 2002), tidyverse (Wickham et al., 2019).

2.3.3 Simulation of SQDs

We applied a Monte Carlo simulation, repeating modularization and imputation on different samples over 1,007 simulation runs.[2] Accordingly, we randomly drew 1,007 samples with each 2,000 respondents from our GIP population data. Unlike single simulations, this procedure produces findings beyond anecdotal evidence by ruling out random differences. The following paragraphs describe the steps taken in each simulation run.

Generating module structures

To generate module structures, we implemented RM, STM and DTM as described above in R. With each modularization technique, we create five split modules with 10 items each. This results in three module structures tested in each simulation run. While the arrangement of variables with RM and DTM differs across simulation runs due to their stochastic procedure, STMs are predefined (see Table 2.1) and thus do not vary.

Creating reduced datasets

To generate SQD datasets, we randomly assigned three out of five split modules plus the core module to each respondent in the sample. All possible combinations of split modules had equal chances to appear (although empirical frequencies of occurrence may vary randomly). All values from unassigned modules were deleted from the sample data, generating reduced datasets with 67% of the original size.

Completing the reduced data

For all three strategies and in each simulation run, we applied MI with the *mice* package in R with 40 imputations drawn after 15 iterations to complete the reduced data. Like Rässler et al. (2002), we used PMM as imputation method, because a small-scale test with one simulation run and RM showed enormous shifts in univariate distributions and correlation sizes with the *mice* default methods (logistic regression for binary variables, proportional odds logistic regression for ordinal variables) but not with PMM (see Figure 2.B1 for details). This complies with prior research revealing difficulties with imputation using categorical regression methods

[2] This number of simulation runs $(1,007)$ was favored over $1,000$ because and we had access to $1,008$ processor cores (one core per simulation run, except for one consumed by setting up the simulation).

(van Buuren, 2018, p. 91; White et al., 2011; Wu et al., 2015) and recommending PMM at least as a fallback option (van Buuren, 2018, p. 166; Koller-Meinfelder, 2009, pp. 48-68).

Small-scale tests also showed that restricting imputation models to predictor variables with Spearman correlations stronger than $|0.10|$ in the non-imputed SQD data may lead to improved imputations. Thereby, imputation models include on average between 2 and 22 predictors (median: 11). If no predictors are included in a simulation run, we resort to unconditional hot-deck sampling. Also considering that general recommendations are to include at most 15-25 (van Buuren, 2018) or 30-40 (Honaker and King, 2010) predictors, we proceeded with this approach. The excluded variables' correlations with the imputed variable are thereby assumed to be zero. Hence, their strength may be underestimated after imputation, but these underestimations should be small because the correlations are close to zero. Results from an additional simulation that instead includes all variables as predictors can be found in Figures 2.B2 and 2.B3, with substantively identical findings for the relative performance of modularization strategies. Overall, these unrestricted predictor sets yield much larger biases especially in univariate estimates. Bivariate estimates also have a tendency towards more extreme biases. At the same time, many of the biases that are very small with unrestricted predictor sets are slightly larger with restricted predictor sets, because restricting predictor sets in this way implies slight biases in very weak correlations.

Estimating distribution parameters

We examine how well univariate and bivariate distributions in the complete sample data can be reproduced with the imputed data. In consequence, distribution parameters were estimated in each simulation run with the complete sample dataset and with all imputed datasets. For each modularization strategy, the resulting estimates were pooled using Rubin's Rules. Consequently, for each parameter and in each simulation run, we have one pooled estimate per strategy and, as a benchmark, one estimate for the complete sample data.

To cover univariate distributions, we estimated relative univariate frequencies. All split items in our simulation are available as categorical variables. The index c describes a single category of any of these variables. We calculated relative univariate frequencies for each variable category c in each simulation run s based on the complete sample data ($\widehat{\pi}_{c,s}^{\text{complete}}$) and imputed data ($\widehat{\pi}_{c,s}^{\text{imputed}}$).

For bivariate distributions we used Spearman correlations. We first generated dummy variables for all categories of the seven nominal type variables in the core module, increasing the total number of variables to 99. Then, Spearman correlations $\widehat{\rho}_{i,j,s}^{\text{complete}}$ for the complete sample data and $\widehat{\rho}_{i,j,s}^{\text{imputed}}$ for the imputed data were estimated in each simulation run s for each relevant unique pair of variables i, j. We excluded all variable pairs that did not include at least one split module, that is, imputed, variable.

2.3.4 Measures

The basis of our analyses is the deviation $\widehat{\Delta}$ of imputed data estimates from complete-data estimates in each simulation run s.[3] For a frequency $\widehat{\pi}_{c,s}$ of a category c or correlation $\widehat{\rho}_{i,j,s}$ of a variable pair i, j each simulation run s entails the following operation:

$$\widehat{\Delta}\left(\widehat{\pi}_{c,s}\right) = \widehat{\pi}_{c,s}^{\text{imputed}} - \widehat{\pi}_{c,s}^{\text{complete}} \tag{2.3}$$

$$\widehat{\Delta}\left(\widehat{\rho}_{i,j,s}\right) = \widehat{\rho}_{i,j,s}^{\text{imputed}} - \widehat{\rho}_{i,j,s}^{\text{complete}} \tag{2.4}$$

A positive value on $\widehat{\Delta}(\widehat{\pi}_{c,s})$ or $\widehat{\Delta}\left(\widehat{\rho}_{i,j,s}\right)$ means that the corresponding estimate has been overestimated, whereas a negative value indicates an underestimation.

Bias

If a given estimate is Monte Carlo unbiased, we expect the average of its deviations $\widehat{\overline{\Delta}}$ over all simulation runs to be zero. In contrast, a positive (negative) average suggests that the estimate is systematically overestimated (underestimated).

The Monte Carlo bias of a frequency estimate $\widehat{\pi}_c$ for a category c is obtained through the average over its deviations in all $S = 1,007$ simulation runs:

$$\widehat{\overline{\Delta}}\left(\widehat{\pi}_c\right) = \frac{1}{S}\sum_{s=1}^{S}\widehat{\Delta}\left(\widehat{\pi}_{c,s}\right) \tag{2.5}$$

The Monte Carlo bias of a correlation estimate $\widehat{\rho}_{i,j}$ for variables i and j is:

$$\widehat{\overline{\Delta}}\left(\widehat{\rho}_{i,j}\right) = \frac{1}{S}\sum_{s=1}^{S}\widehat{\Delta}\left(\widehat{\rho}_{i,j,s}\right) \tag{2.6}$$

[3] Dividing $\widehat{\Delta}$ by the complete-data benchmark would yield percentage deviations. This study, however, does not consider such a measure because it turned out unstable for the many correlations near zero, as this implies dividing by numbers very close or equal to zero.

Variability

Another important aspect of the quality of an estimate is its precision. In practice, this means that ideally standard errors are relatively small. The Monte Carlo simulation allows to approximate the variance of a given point estimate through taking the estimate's variance over all simulation runs (e.g., Münnich and Rässler, 2005; Mashregi et al., 2014; Bruch, 2016). Because the point estimator of interest is the deviation from the complete-sample estimate, we use the variance of these deviations in Equations 2.3 and 2.4 instead of the variance of the frequency or correlation estimates themselves. (In doing so, we focus more on the variance caused by the SQD, but standard errors of the frequencies and correlation estimates as approximated through the simulation (see Figures 2.C1 and 2.C2) yield equivalent findings.) Thus, for a frequency $\widehat{\pi}$ of a category c we measure the variability of deviations across all simulation runs from the average deviation through the standard deviation of deviations (SDD) $\widehat{\sigma}\left\{\widehat{\Delta}(\widehat{\pi}_c)\right\}$:

$$\widehat{\sigma}\left\{\widehat{\Delta}(\widehat{\pi}_c)\right\} = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\{\widehat{\Delta}\left(\widehat{\pi}_{c,s}\right) - \widehat{\overline{\Delta}}\left(\widehat{\pi}_c\right)\}^2}. \tag{2.7}$$

Correspondingly, $\widehat{\sigma}\left\{\widehat{\Delta}(\widehat{\rho}_{i,j})\right\}$ is the SDD for a correlation $\widehat{\rho}$ of two variables i and j:

$$\widehat{\sigma}\left\{\widehat{\Delta}(\widehat{\rho}_{i,j})\right\} = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\{\widehat{\Delta}\left(\widehat{\rho}_{i,j,s}\right) - \widehat{\overline{\Delta}}\left(\widehat{\rho}_{i,j}\right)\}^2} \tag{2.8}$$

An SDD equal to zero means that imputed and complete data produce identical estimates in each simulation run net of systematic bias, while larger SDDs correspond to more uncertain estimates. Hence, a modularization technique that obtains small biases and SDDs will yield high imputation quality. However, since RM and DTM rely on a stochastic procedure, this additional source of randomness may increase the estimates' variability.

2.3.5 Evaluation Strategy

As we generate a huge number of imputation quality measures (297 for frequencies and 3,675 for correlations), we need to condense the information displayed in our results. Therefore, we produce one summary graph each for univariate and bivariate biases and SDDs. We combine this evaluation of general patterns with additional

analyses on specific sets of variable pairs to gain more insight into potential differences between variable pairs.

We focus on two aspects: First, we provide additional analyses restricted to variable pairs that were used in all their respective imputation models throughout the simulation, because whether a variable is included in an imputation model may decisively determine if its correlation to the imputed variable can be estimated correctly.

Second, we perform separate analyses for correlations based on within-topic and different-topic variable pairs. Depending on the modularization strategy, this difference has important consequences: For instance, consider a correlation of two variables within the same topic. With STM, the two variables are always in the same module, implying all cases are either pairwise observed or unobserved. Therefore, the imputation can rely on many commonly observed values, but we must impute both variables for all other cases. With DTM, however, the variables tend to end up in different modules. Consequently, there are relatively few pairwise observed cases, but many cases where only one of both variables must be imputed. Thus, two variables may have systematically different bivariate missing data patterns depending on the modularization strategy.

2.4 Results

2.4.1 Univariate Frequencies

Figure 2.2 displays the distribution of average Monte Carlo biases of univariate frequencies for the imputed data for RM (first boxplot), STM (second boxplot), and DTM (third boxplot). The rug plots in the second section of Figure 2.2 show the complete distribution of biases for the three strategies (same order). Each data point represents the average bias of one variable category over all simulation runs.

Many biases concentrate closely around zero. With RM and DTM 80% of biases range from -0.002 to $+0.002$. However, some frequencies have stronger biases: The largest biases are -0.006 and $+0.006$ with RM and -0.005 and $+0.005$ with DTM. Biases are larger with STM, where 80% of biases range from -0.004 to $+0.003$ with outliers of up to ± 0.014.

Figure 2.3 summarizes the sizes of SDDs for the imputed frequencies with boxplots and rugs in the same fashion as for biases. Again, each data point represents the SDD of a certain category's frequency. Although small SDDs would be prefer-

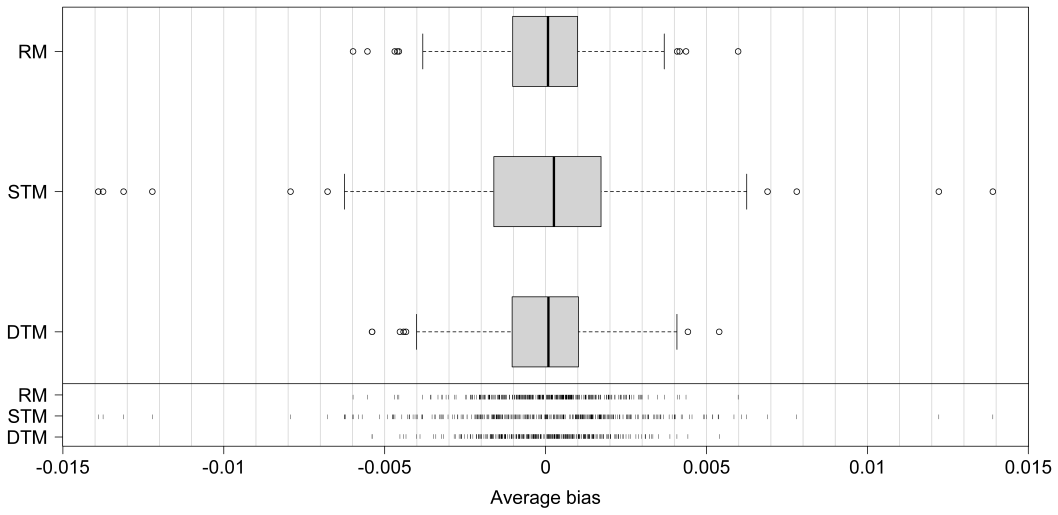

Figure 2.2: Average biases for 297 univariate frequencies according to equation 5, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).
Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

able, unlike average biases they cannot be expected to approach zero. Like with the biases, the differences between RM and DTM are negligible. At the same time, SDDs with STM tend to be somewhat larger than with RM and DTM. For example, the largest SDD with STM is 0.011, while it is 0.010 with RM and DTM.

2.4.2 Bivariate Correlations

Figure 2.4 displays the distribution of average Monte Carlo biases of bivariate correlations for the imputed data for RM (first boxplot), STM (second boxplot), and DTM (third boxplot). The rug plots show the complete distribution of biases for the three strategies (same order). Each data point represents an average bias for one variable pair over all simulation runs.

With both RM and DTM 50% of average biases range from -0.006 to $+0.006$, 90% from -0.017 to $+0.017$, and the most extreme bias is 0.082. Note that these are absolute measures, thus some correlations are highly biased. The outlier with a value of 0.082, for example, belongs to a correlation that is -0.065 in the complete data and on average, $+0.017$ in the imputed data. Hence, it is overestimated by 126%, entailing a sign change. The second-most extreme bias is -0.081 (with RM) with a correlation of 0.206 in the complete data and on average, 0.125 in the imputed data, suggesting it was underestimated by 39%. Furthermore, the rug plots also show some average biases in the area closely around zero. STM has

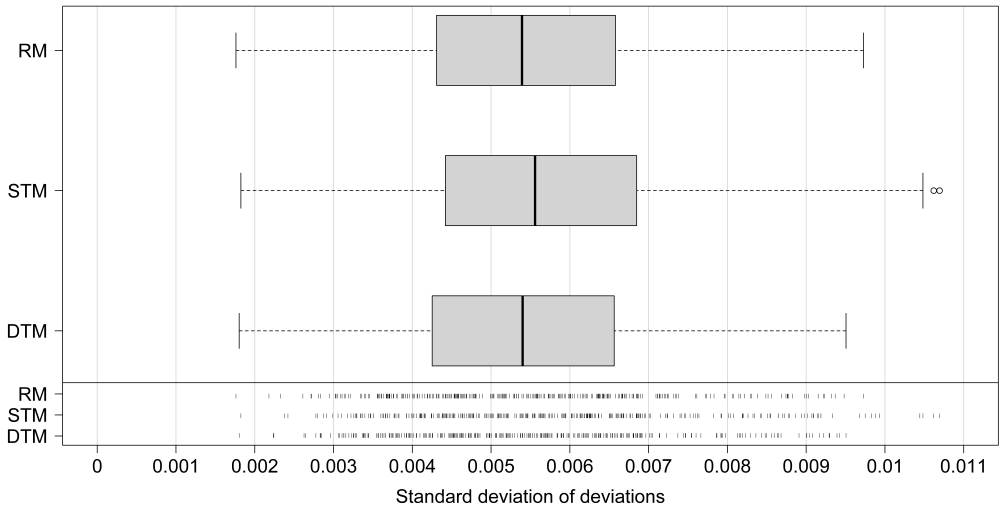

Figure 2.3: Standard deviations of deviations (SDDs) of 297 univariate frequencies according to equation 7, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

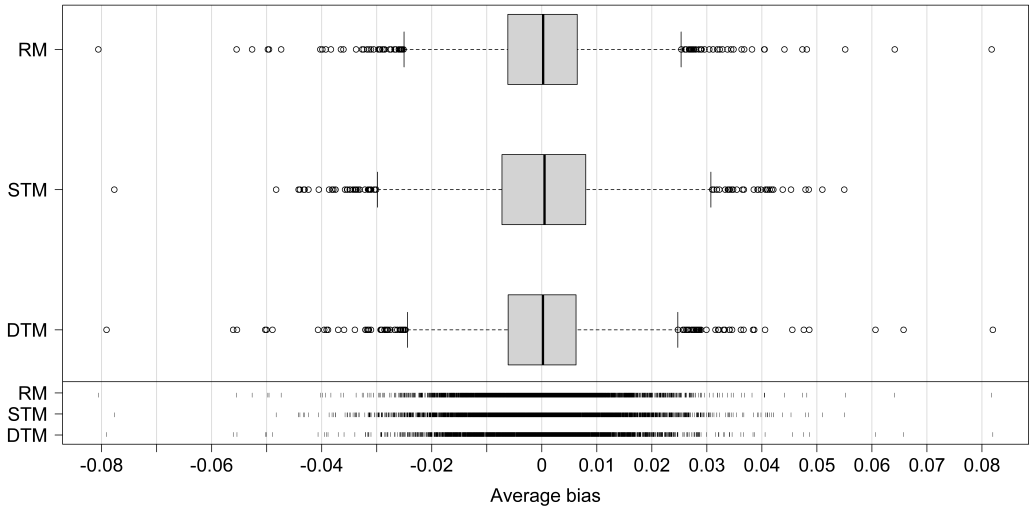

Figure 2.4: Average biases of 3,675 bivariate correlations according to equation 6, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

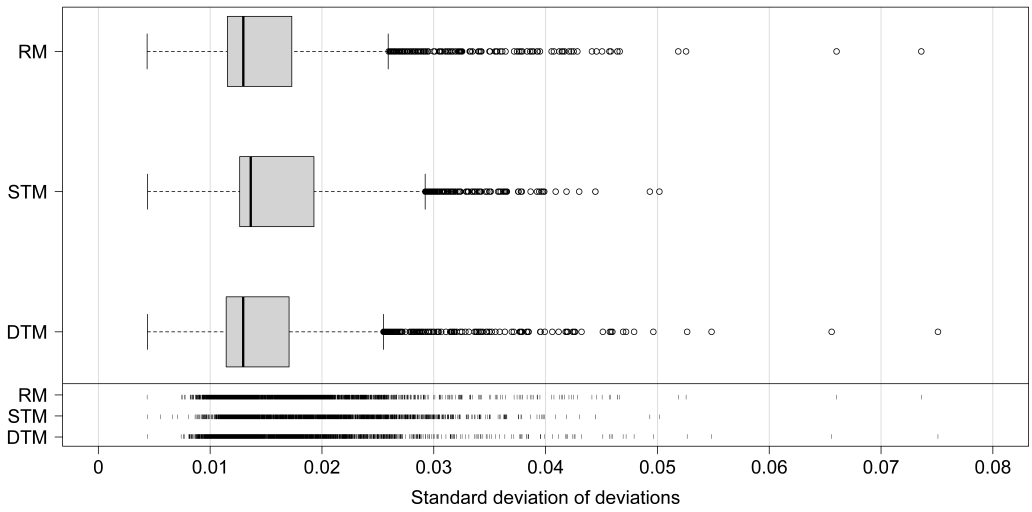

Figure 2.5: Standard deviations of deviations (SDDs) of 3,675 bivariate correlations according to equation 8, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).
Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

a different pattern: 50% range from -0.007 to $+0.008$ and 90% from -0.020 to $+0.020$. Furthermore, STM produces fewer extreme outliers larger than ± 0.050 (three correlations) than RM (six correlations) and DTM (eight correlations).

Figure 2.5 summarizes the SDDs for Spearman correlations. STM tends to produce larger SDDs than RM and DTM, with boxes visibly shifted to the right. Again, however, STM yields fewer extreme outliers: The largest SDD with STM is 0.050, while the largest SDDs with RM and DTM are 0.074 and 0.075.

Analysis by topic

To further investigate effects of the modularization on biases in bivariate correlations, Figure 2.6 shows the distributions of average biases, separately for correlations between variables of different topics (on the left) and correlations between variables of the same topic (on the right).

For different-topic correlations 50% of average biases with RM and DTM are between -0.008 and $+0.009$. Biases with STM are larger with 50% between -0.010 and $+0.013$. The strongest biases are 0.037 with RM and DTM and 0.048 with STM.

For within-topic correlations 50% of average biases with RM and DTM are between -0.015 and $+0.005$ and 50% of biases with STM between -0.009 and $+0.007$. STM leads to fewer extreme biases of larger than ± 0.050 (two with STM,

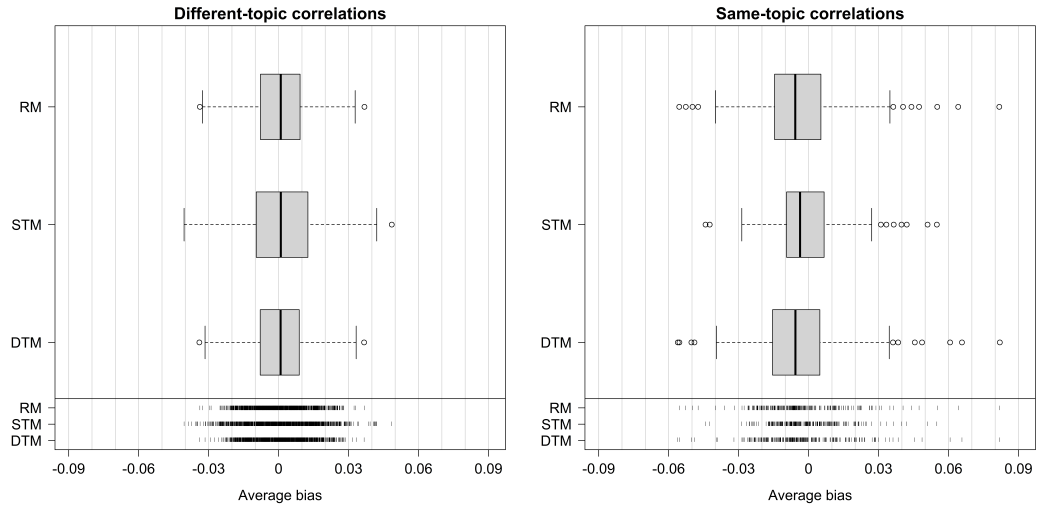

Figure 2.6: Average biases of 3,675 bivariate correlations according to equation 6, separating correlations of variables of different vs. same topics, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).
Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

five with RM, and six with DTM). Correspondingly, the strongest biases with RM and DTM are 0.082 but only 0.055 with STM.

In addition, within-topic correlations seem to be underestimated: With RM, 66.7% of within-topic correlations have biases smaller than zero, 60.0% with STM and 68.0% with DTM.

Figure 2.7 shows the sizes of SDDs for different-topic and within-topic correlations. For different-topic correlations, small SDDs are again less common with STM than with RM or DTM: With RM and DTM, the majority of SDDs are smaller than 0.020, while with STM, the majority of SDDs are larger than 0.020. For same-topic correlations, however, STM tends to produce smaller SDDs.

Subset by representation in the imputation models

Figure 2.8 displays average biases exclusively for variable pairs included in each imputation model throughout the simulation. Note that this subset covers only a small fraction (72 correlations) of all correlations. These correlations are generally stronger, as imputation models only included correlations stronger than 0.10. Even in this subset, biases are still different from zero. This underscores the challenges of SQDs for the imputation. Again, correlations in both graphs tend to be underestimated. For different-topic correlations, all correlations are underestimated and

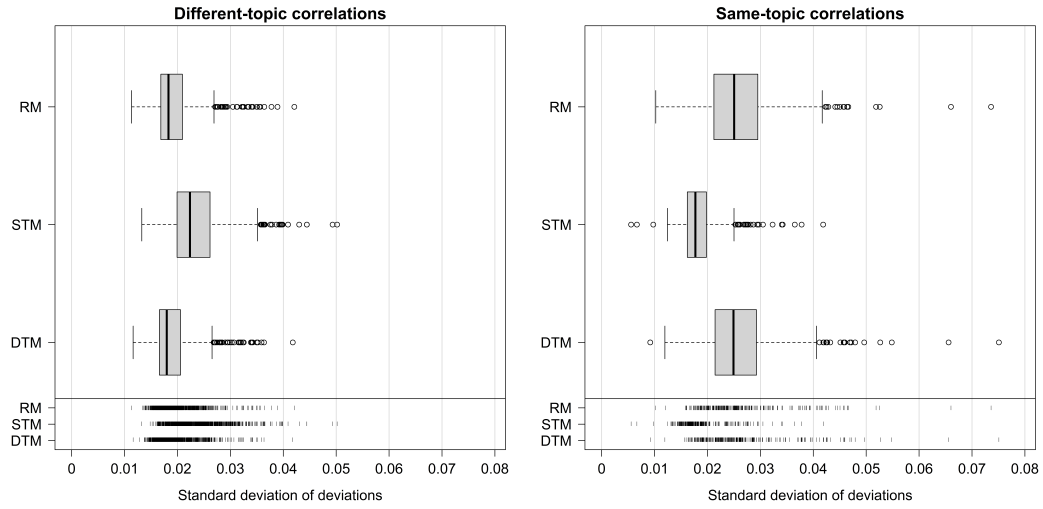

Figure 2.7: Standard deviations of deviations (SDDs) of 3,675 bivariate correlations according to equation 8, separating correlations of variables in different vs. same topics, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).
Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

73.2% (RM and DTM) and 71.4% (STM) of same-topic correlations are underestimated.

50% of biases of different-topic correlations are between -0.019 and -0.014 with RM and DTM (STM: -0.026 and -0.013). The most extreme biases are -0.025 (RM), -0.036 (STM) and -0.023 (DTM). For same-topic correlations 50% of the biases are between -0.021 and $+0.005$ with RM, -0.012 and $+0.002$ with STM, and -0.021 and $+0.004$ with DTM. The most extreme biases are $+0.055$ (RM), -0.027 (STM), and $+0.061$ (DTM).

SDDs are displayed in Figure 2.9. STM clearly produces larger SDDs for different-topic correlations ranging from 0.026 to 0.033 whereas SDDs with RM range from 0.023 to 0.026 and SDDs with DTM from 0.022 to 0.026. For within-topic correlations STM leads to smaller SDDs than RM and DTM ranging from 0.012 to 0.025, while SDDs with RM range from 0.019 to 0.042 and with DTM from 0.018 to 0.043.

2.4.3 Alternative Correlation Structures

In contrast to our expectations, DTM and RM generally performed similarly. Potentially, the lack of high correlations even within topics may have prevented such an effect. To test this hypothesis, we applied two additional simulations (using the

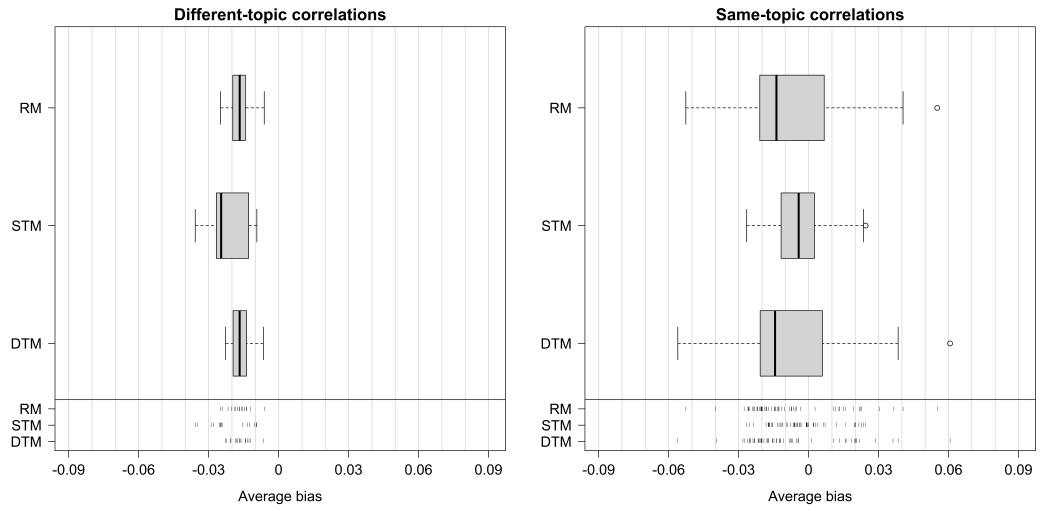

Figure 2.8: Average biases of 72 bivariate correlations according to Equation 6 for correlations represented in every imputation model throughout the simulation, separately for correlations of variables of different vs. same topics, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).
Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

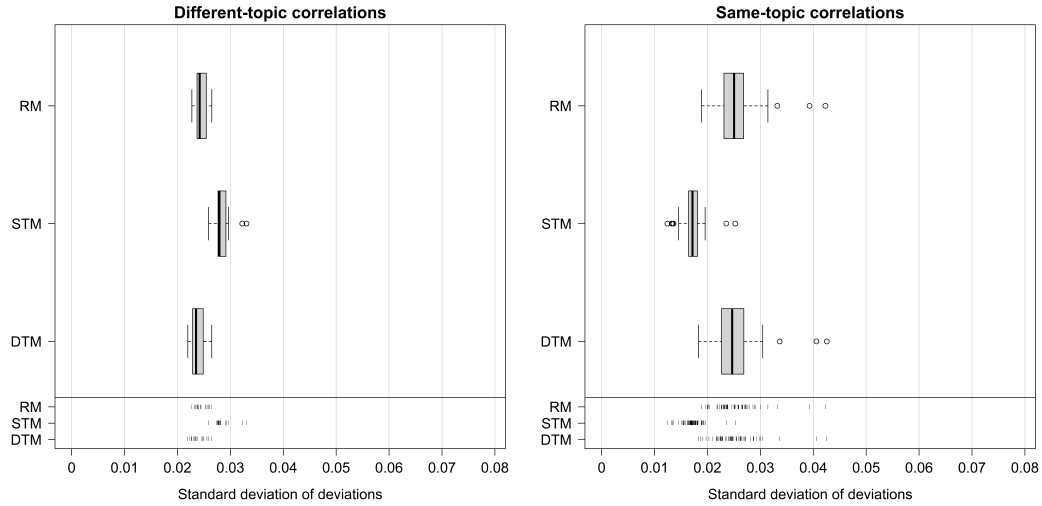

Figure 2.9: Standard deviations of deviations (SDDs) of 72 bivariate correlations according to equation 8, for correlations represented in every imputation model throughout the simulation, separately for correlations of variables in different vs. same topics, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).
Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

same procedure as with the main simulation) with synthetic data. Here, we maintained the univariate distributions found in the GIP dataset but manipulated correlation structures to assess whether DTM outperforms RM when there is one highly correlated predictor within the same topic for each imputed variable (see appendix D for a description of the data-generating process). Scenario 1 (control condition) largely adopts the original correlation structure but with maximum correlations of $|0.20|$. Scenario 2 is the same except for one same-topic correlation per imputed variable increased towards ± 0.90.

Results (see appendix Figures 2.D1 through 2.D4) indeed show somewhat small"-er biases and SDDs with DTM than with RM for scenario 2, while STM performs exceptionally poorly. However, even in this extreme scenario DTM's advantage over RM remains quite small. Scenario 1 largely replicates the findings from the main simulation study, with STM producing somewhat larger biases and SDDs than RM and DTM, which perform similarly.

2.5 Summary

In this paper, we simulated the impact of different modularization strategies on imputation quality in an SQD. By using real data from a probability-based online survey, our goal was to test approaches to implementing SQDs under realistic conditions, characterized by a large number of variables with many missing cases to be imputed using a wide range of relatively weakly correlated predictor variables that are partially missing themselves.

The evidence suggests that univariate frequencies tend to be slightly biased. More concerning are our results concerning bivariate relationships captured by correlations. Although some biases are small, others are comparatively large. This observation holds for all examined modularization strategies, among within-topic correlations and different-topic correlations as well as for correlations included in all imputation models.

Thereby, correlations tend to be attenuated. Most correlations that are positive in the population data have biases smaller than zero (RM: 81.0 %; STM: 81.0%; DTM: 81.5%). However, most correlations that are *negative* in the population data have biases larger than zero (RM: 84.2%; STM: 86.1%; DTM: 83.9%). (Note that overestimating a truly negative correlation implies a loss in correlation strength.)

Overall, we find that STM leads to larger biases and variability in estimates than RM and DTM. This effect is most pronounced for frequencies but holds for

correlations in the overall pattern as well. However, STM performs better than RM and DTM for same-topic correlations, suggesting that correlations with more pairwise observed cases (here: correlations based on variables in the same module) can be estimated with higher quality.

2.6 Conclusions

We draw several conclusions: First, modularization strategies affect imputation quality. Overall, STM produced estimates with larger biases and variability compared to RM and DTM. Thus, from a statistical perspective, modules should be designed heterogeneously regarding topics. This concurs with the notion that strongly correlated items should not be allocated to the same module (Raghunathan and Grizzle, 1995; Rässler et al., 2002). Though STM may be a solution when analyses are conducted within one topic only and thus do not require imputation.

Second, results for RM and DTM hardly differed. As suggested by the additional synthetic data simulations, DTM might outperform RM in different data scenarios if for instance, one correlation per imputed variable within the same topic was considerably increased. However, even these effects were small, potentially because the probability for some highly correlated variable pair to end up in the same module is already quite small with RM.

However, DTM might also have insufficiently exploited the correlation structure. To test this, we applied the modified cluster analysis technique for modularization developed by Rässler et al. (2002) on our (original) population data, a method that minimizes correlations within modules. The resulting average between-module correlation was 0.073 (compared to 0.072 with DTM and 0.070 with RM). Thus, the added value of such data-driven methods may be limited for settings with low variable correlations.

Third, differences between modularization strategies were detectable, but average biases and variability seem to differ more between estimates for different categories or variable pairs than between modularization strategies. This suggests independent of modularization strategy, items in split modules should be designed well-suited for imputation. Additionally, modularization strategy might also affect response quality, as for example, topic switches would be more frequent with DTM than with STM. Thus, we encourage future research into response effects to complement our findings.

Finally, imputation remains a great challenge for SQD data. Especially relationships between variables are not fully retained. This finding is compatible with Bahrami et al. (2014), who report small downwards slants in regression estimates. Perhaps, further restricting the number of predictors in the imputation models may help more, but the more the model is restricted, the larger will be the risk of underestimating relevant relationships. Thus, future research should further investigate on how SQD data can be imputed in real-data contexts.

This study has some limitations. First, our findings may be sensitive to changes in the data context. For example, surveys with more items could aggravate problems with the complexity of imputation models.

Second, alternative imputation strategies could change the results. Although we do not expect differences in the relative performance of modularization strategies, future research should explore how different imputation strategies generally affect imputation quality for SQDs.

Third, our research should be extended to testing the performance of multivariate models. This was beyond the scope of this paper. However, the biases in bivariate correlations revealed by our simulation suggest that multivariate coefficients may also be biased. Therefore, future research would benefit the state of the art by running simulations of SQD on real data with models commonly found in the social science literature.

Fourth, our analyses ignored item nonresponse in the data caused by respondent behavior. Again, for our purposes, this was out of scope. However, we look forward to future research that investigates how missingness by SQD and item nonresponse differentially affect analyses and may be best imputed.

Fifth, simulating reduced data (rather than implementing an SQD in a real survey) does not allow to examine response behavior with different SQDs. Again, we encourage future research on this.

We anticipate that with the continued growth in online surveys, the pressure to shorten questionnaires with SQD will increase, too. Our study, however, demonstrates the challenges to the imputation of SQD data. We show that the choice of modularization strategy may alleviate some of these challenges. Moreover, our findings stress the need for further exploration of how existing SQD procedures may be enhanced to fit the reality of social data and thereby ensure high data quality for future surveys.

References

Adigüzel, F., & Wedel, M. (2008). Split questionnaire design for massive surveys. *Journal of Marketing Research*, *45*(5), 608-617.

Bahrami, S., Aßmann, C., Meinfelder, F., & Rässler, S. (2014). A split questionnaire survey design for data with block structure correlation matrix. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving survey methods: Lessons from recent research* (pp. 368-380). Routledge.

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods.* R Package version 1.2-18. https://CRAN.R-project.org/package=Matrix

Bianchi, A., Biffignandi, S., & Lynn, P. (2017). Web-face-to-face mixed-mode design in a longitudinal survey: Effects on participation rates, sample composition, and costs. *Journal of Official Statistics*, *33*(2), 385-408.

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1983). Effects of filter questions in public opinion surveys. *Public Opinion Quarterly*, *47*(4), 528-546.

Blom, A. G., Bossert, D., Funke, F., Gebhard, F., Holthausen, A., & Krieger, U.; SFB 884 "Political Economy of Reforms" Universität Mannheim (2016a). *German Internet Panel, wave 1 - core study (September 2012)*. GESIS Data Archive, ZA5866 Data file Version 2.0.0. https://doi.org/ 10.4232/1.12607.

Blom, A. G., Bossert, D., Gebhard, F., Funke, F., Holthausen, A., & Krieger, U.; SFB 884 "Political Economy of Reforms" Universität Mannheim (2016b). *German Internet Panel, wave 13 - core study (September 2014)*. GESIS Data Archive, ZA5924 Data file Version 2.0.0. https://doi.org/10.4232/1.12619.

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

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

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. *Field Methods*, *27*(4), 391-408.

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., & 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*(4), 498-520.

Bruch, C. (2016). *Varianzschätzung unter Imputation und bei komplexen Stichprobendesigns* [Doctoral dissertation]. University of Trier.

Calinescu, M., Bhulai, S., & Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, *226*(1), 115-121.

Chipperfield, J. O., Barr, M. L., & Steel, D. G. (2018). Split questionnaire designs: Collecting only the data that you need through MCAR and MAR designs. *Journal of Applied Statistics*, *45*(8), 1465-1475.

Chipperfield, J. O. & Steel, D. G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, *25*(2), 227-244.

Chipperfield, J. O. & Steel, D. G. (2011). Efficiency of split questionnaire surveys. *Journal of Statistical Planning and Inference*, *141*(5), 1925-1932.

Cornesse, C., Felderer, B., Fikel, M., Krieger, U., & Blom, A. G. (2022). Recruiting a probability-based online panel via postal mail: Experimental evidence. *Social Science Computer Review*, *40*(5), 1259-1284.

Curtice, J., Clery, E., Perry, J., Phillips M., & Rahim, N. (Eds.) (2019). *British Social Attitudes: The 36th report*. National Centre for Social Research.

de Leeuw, E., Hox, J., & Luiten, A. (2018). International nonresponse trends across countries and years: An analysis of 36 years of Labour Force Survey data. *Survey Insights: Methods from the Field.* https://surveyinsights.org/?p=10452.

DeBruine, L. (2020). *faux: Simulation for factorial designs*. R package version 0.0.1.5.

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, *22*(2), 313-328.

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360.

Gonzalez, J. M. & Eltinge, J. L. (2007). Multiple matrix sampling: A review. In *JSM proceedings, survey research methods section* (pp. 3069-3075). American Statistical Association.

Honaker, J. & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561-581.

Imbriano, P. (2018). *Methods for improving efficiency of planned missing data designs* [Doctoral dissertation]. University of Michigan, Ann Arbor.

Imbriano, P. M. & Raghunathan, T. E. (2020). Three-form Split questionnaire design for panel surveys. *Journal of Official Statistics*, *36*(4), 827-854.

International Telecommunication Union (2019). *World telecommunication/ICT indicators database* (23rd ed.). ITUPublications.

Ipsos (2021). *KnowledgePanel*. https://www.ipsos.com/en-us/solutions/public-affairs/knowledgepanel

Jäckle, A., Lynn, P., & Burton, J. (2015). Going online with a face-to-face household panel: Effects of a mixed mode design on item and unit non-response. *Survey Research Methods*, *9*(1), 57-70.

Kassambara, A. (2019). *ggcorrplot: Visualization of a correlation matrix using 'ggplot2'*. R package version 0.1.3. https://CRAN.R-project.org/package=ggcorrplot

Knoef, M. & de Vos, K. (2009). *The representativeness of LISS, an online probability panel*. CentERdata. https://www.lissdata.nl/sites/default/files/bestanden/paper_knoef_devos_-website.pdf

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

Kreuter, F., McCulloch, S., Presser, S., & Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods & Research*, *40*(1), 88-104.

Luijkx, R., Jónsdóttir, G. A., Gummer, T., Ernst Stähli, M., Fredriksen, M., Reeskens, T., Ketola, K., Brislinger, E., Christmann, P., Gunnarsson, S. Þ., Hjaltason, Á. B., Joye, D., Lomazzi, V., Maineri, A. M., Milbert, P., Ochsner, M., Ólafsdóttir, S., Pollien, A., Sapin, M., ... Wolf, C. (2021). The European Values Study 2017: On the way to the future using mixed-modes. *European Sociological Review*, *37*(2), 330-346.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, *110*, 63-73.

Mashreghi, Z., Léger, C., & Haziza, D. (2014). Bootstrap methods for imputed data from regression, ratio and hot-deck imputation. *Canadian Journal of Statistics*, *42*(1), 142-167.

Mavletova, A. & Couper, M. P. (2015). A meta-analysis of breakoff rates in mobile web surveys. In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), *Mobile research methods: Opportunities and challenges of mobile research methodologies* (pp. 81-98). Ubiquity Press.

McFarland, S. G. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, *45*(2), 208-215.

Microsoft & Weston, S. (2020). *foreach: Provides foreach looping construct*. R package version 1.5.0. https://CRAN.R-project.org/package=foreach

Munger, G. F. & Loyd, B. H. (1988). The use of multiple matrix sampling for survey research. *The Journal of Experimental Education*, *56*(4), 187-191.

Münnich, R. & Rässler, S. (2005). PRIMA: A new multiple imputation procedure for binary variables. *Journal of Official Statistics*, *21*(2), 325-341.

Nicoletti, C. & 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*, *169*(3), 625-646.

OECD (2014). *PISA 2012 technical report*. OECD.

Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N. A., McCarthy, J. S., O'Brien, E., Opsomer, J. P., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z. T., Turakhia, C., & Wagner, J. (2021). Transitions from

telephone surveys to self-administered and mixed-mode surveys: AAPOR task force report. *Journal of Survey Statistics and Methodology*, *9*(3), 381-411.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74-97.

Peytchev, A. & Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods*, *11*(4), 361-368.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

Rässler, S., Koller, F. & 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*. https://www.statistik.rw.fau.de/files/2016/03/d0042b.pdf

Revilla, M. (2017). Analyzing survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. *methods, data, analyses*, *11*(2), 135-162.

Revilla, M. & Höhne, J. K. (2020). How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research*, *62*(5), 538-545.

Roberts, C., Vandenplas, C., & Ernst Stähli, M. (2014). Evaluating the impact of response enhancement methods on the risk of nonresponse bias and survey costs. *Survey Research Methods*, *8*(2), 67-80.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Shoemaker, D. M. (1973). *Principles and Procedures of Multiple Matrix Sampling*. Ballinger.

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., . . . Zeileis, A. (2020). *DescTools: Tools for descriptive statistics*. R package version 0.99.36. https://CRAN.R-project.org/package=DescTools

Silber, H., Höhne, J. K., & Schlosser, S. (2016). Question order experiments in the German-European context. *Survey Methods: Insights from the Field*. https://surveyinsights.org/?p=7645

Smith, T. W., Davern, M., Freese, J., & Morgan, S. L. (2019). *General Social Surveys, 1972-2018, cumulative codebook*. NORC.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the national health and nutrition examination survey. *Survey Methodology*, *32*(2), 217-231.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford University Press.

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer.

Weston, S. (2017). *doMPI: foreach parallel adaptor for the Rmpi package*. R package version 0.2.2. https://CRAN.R-project.org/package=doMPI

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *dplyr: A grammar of data manipulation*. R package version 1.0.6. https://cran.r-project.org/package=dplyr

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377-399.

Wu, W., Jia F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research*, *50*(5), 484-503.

Yu, Hao (2002). Rmpi: Parallel statistical computing in R. *R News*, *2*(2), 10-14.

Appendix

A. Additional Information on the Data Used for the Simulation

Table 2.A1: Dates and AAPOR response rates of GIP waves used in the simulation.

Wave	Field-time period	Response rate face-to-face recruitment 1	Cumulative panel registration rate 2	Wave completion rate 3	Cumulative response rate at wave 4
1	Sept/01/2012 – Sept/30/2012	52.1%	18.5%	92.5%	17.1%
13	July/01/2014 – July/31/2014	52.1%	18.5%	86.5%	17.8%
37	Sept/01/2018 – Sept/30/2018	52.1% (2012 sample)	18.5% (2012 sample)	75.5%	17.4%
		47.5% (2014 sample)	21.0% (2014 sample)		
38	Nov/01/2018 – Nov/30/2018	52.1% (2012 sample)	18.5% (2012 sample)	54.0%	12.0%
		47.5% (2014 sample)	21.0% (2014 sample)		
			24.1% (2018 sample)		

Note: The GIP is based on a probability sample of the general population aged 16-75 at recruitment. It was recruited face-to-face with area probability samples in 2012 and 2014 (see Blom et al., 2015, for a description). In 2018, it was recruited postally by means of a sample of individuals from the population registers (see Cornesse et al., 2021, for a description). All proportions are rounded to one decimal place but calculated with unrounded figures.

[1] The response rate face-to-face recruitment is calculated with AAPOR RR2 for face-to-face surveys.

[2] The cumulative panel registration rate is all individuals registered for the online panel over all eligible members in the gross sample (in the gross area probability sample for the 2012 and 2014 recruitments and in the gross register sample for the 2018 recruitment). This is thus also equivalent to AAPOR RR2 for face-to-face surveys. Overall and weighted for the differential gross sample sizes of the three recruitment rounds, the cumulative panel registration rate is 21.5%.

[3] The basis for the calculation of the wave completion rates is all individuals registered for the online panel. The GIP tracks completion for members of its three recruitment rounds.

*[4] The cumulative response rate is the cumulative panel registration rate * the wave completion rate.*

Table 2.A2: Wording of GIP items used in the simulation.

Variable	Item	German	English
gender_18	gender	**Fragetext:**	**Question text:**
		Geschlecht	Gender
		Antwortkategorien:	**Answer categories:**
		1. männlich	1. male
		2. weiblich	2. female
year_of_-	year of birth	**Fragetext:**	**Question text:**
birth	categories	Geburtsjahr	Year of birth
_cat_18		**Antwortkategorien:**	**Answer categories:**
		1. 1935-1939	1. 1935-1939
		...	...
		13. 1995-1999	13. 1995-1999
		14. 2000 und später	14. 2000 and later

Variable	Item	German	English
educ_school_18	highest educational degree	**Fragetext:** Welches ist Ihr höchster allgemeinbildender Schulabschluss? **Antwortkategorien (7, nominal):** • Noch Schüler-/in • Schule beendet ohne Abschluss • Volks-/Hauptschulabschluss bzw. Polytechnische Oberschule mit Abschluss 8. oder 9. Klasse • Mittlere Reife, Realschulabschluss bzw. Polytechnische Oberschule mit Abschluss 10. Klasse • Fachhochschulreife (Abschluss einer Fachoberschule etc.)	**Question text:** Which is the highest school degree that you have obtained? **Answer categories (7, nominal):** • Still at school • Left school without degree • Volks-/Hauptschulabschluss or Polytechnic Secondary School leaving after 8^{th} or 9^{th} grade (lower secondary education degree) • Mittlere Reife, Realschulabschluss (intermediate secondary education degree, leaving after 10^{th} grade) • Fachhochschulreife (high secondary education degree, leaving after 11^{th} or 12^{th} grade)

Variable	Item	German	English
		• Abitur bzw. Erweiterte Oberschule mit Abschluss 12. Klasse (Hochschulreife) • Anderen Schulabschluss: Bitte tragen Sie Ihren Schulabschluss ein [Antwortfeld]	• Abitur or Extended Secondary School leaving after 12th grade (highest secondary education degree; universal higher education entrance qualification) • Other degree, Please enter [answer field]

Variable	Item	German	English
educ_job_18	highest professional qualification	**Fragetext:** Welchen höchsten beruflichen Ausbildungs- oder (Fach-) Hochschulabschluss haben Sie? **Antwortkategorien (12, nominal):** • Noch in beruflicher Ausbildung (Berufsvorbereitungsjahr, Auszubildende/-r, Praktikant/-in, Student/-in) • Schüler/-in und besuche eine berufsorientierte Aufbau-, Fachschule o. ä. • Keinen beruflichen Abschluss und bin nicht in beruflicher Ausbildung • Beruflich-betriebliche Berufsausbildung (Lehre) abgeschlossen • Beruflich-schulische Ausbildung (Berufsfachschule, Handelsschule, Vorbereitungsdienst für den mittleren Dienst in der öffentlichen Verwaltung) abgeschlossen	**Question text:** Which is your highest vocational qualification? **Answer categories (12, nominal):** • Still in vocational training (vocational preparation year, apprentice, intern, student) • High shool student and attending a job-oriented school, college or similar • No professional qualification and not in vocational training • Completed vocational training (apprenticeship) • Completed vocational training (vocational school, commercial school, preparation for medium level civil service) • Completed training at a technical school of the GDR

Variable	Item	German	English
		• Ausbildung an einer Fachschule der DDR abgeschlossen • Ausbildung an einer Fach-, Meister-, Technikerschule, Berufs- oder Fachakademie abgeschlossen • Bachelor an (Fach-)Hochschule abgeschlossen • Fachhochschulabschluss (z. B. Diplom, Master) • Universitätsabschluss (z. B. Diplom, Magister, Staatsexamen, Master) • Promotion • Ein anderer beruflicher Abschluss, und zwar (bitte eintragen) [Antwortfeld]	• Completed training at a specialist, master, technical school, vocational or technical college • Completed Bachelor's degree at a university or a university of applied sciences • Degree from a university of applied sciences (e.g. Diploma, Master) • University degree (e.g., Diploma, Magister, State examination, Master) • Doctoral degree • Another professional qualification, namely (please enter) [answer field]

Variable	Item	German	English
marital_status_18	marital status	**Fragetext:** Welchen Familienstand haben Sie? **Antwortkategorien (9, nominal):** • Verheiratet und leben mit Ihrem/Ihrer Ehepartner/-in zusammen • Verheiratet und leben getrennt • Verwitwet • Geschieden • Ledig • Eingetragene Lebenspartnerschaft, zusammenlebend • Eingetragene Lebenspartnerschaft, getrennt lebend • Eingetragene/-r Lebenspartner/-in verstorben • Eingetragene Lebenspartnerschaft aufgehoben	**Question text:** What is your marital status? **Answer categories (9, nominal):** • Married and cohabiting • Married and not cohabiting • Widowed • Divorced • Single • Registered civil partnership, cohabiting • Registered civil partnership, not cohabiting • Registered civil partner deceased • Registered civil partnership annulled

Variable	Item	German	English
number_- hh_ members_18	number of household members	**Fragetext:** Wie viele Personen leben ständig in Ihrem Haushalt, Sie selbst eingeschlossen? **Antwortkategorien:** 1. 1 ... 5. 5 und mehr	**Question text:** How many people live permanently in your household, including yourself? **Answer categories:** 1. 1 ... 5. 5 and more

Variable	Item	German	English
occupation_-18	employment status	**Fragetext:** Und welche der Beschreibungen trifft am besten auf Ihre aktuelle berufliche Situation zu? Was würden Sie als Ihre Haupttätigkeit bezeichnen? **Antwortkategorien (17, nominal):** • Vollzeiterwerbstätig • Teilzeiterwerbstätig • Altersteilzeit (unabhängig davon, ob in der Arbeits- oder Freistellungsphase befindlich) • Geringfügig erwerbstätig, 400-Euro-Job, Minijob • "Ein-Euro-Job" (bei Bezug von Arbeitslosengeld II) • Gelegentlich oder unregelmäßig beschäftigt • In einer beruflichen Ausbildung/Lehre	**Question text:** And which of the following descriptions best applies to your current job situation? Which best describes your main occupation? **Answer categories (17, nominal):** • Working full-time • Working part-time • Partial retirement • Marginally employed, 400-Euro-Job, Minijob • "One Euro Job" (when receiving unemployment benefits) • Occasionally or irregularly employed • In an apprenticeship

Variable	Item	German	English
		• In Umschulung • Freiwilliger Wehrdienst, Bundesfreiwilligendienst • Freiwilliges Soziales/Ökologisches/Kulturelles Jahr • Mutterschafts-, Erziehungsurlaub, Elternzeit oder sonstige Beurlaubung (Altersteilzeit oben angegeben) • Schüler/-in an einer allgemeinbildenden Schule • Student/-in • Rentner/-in • Pensionär/-in, im Vorruhestand • Arbeitslos • Dauerhaft erwerbsunfähig • Hausfrau/Hausmann	• In retraining • Voluntary military service, voluntary federal service • Voluntary social/ecological/cultural year • Maternity leave, parental leave or other leave of absence (indicate partial retirement above) • High school student • University student • Pensioner • Pensioner (from civil service), in early retirement • Unemployed • Permanently unable to work • Housewife/houseman

Variable	Item	German	English
german_citizenship_18	german citizenship	**Fragetext:** Haben Sie die deutsche Staatsangehörigkeit? **Antwortkategorien (3, nominal):** • Ja, nur die deutsche Staatsangehörigkeit • Ja, die deutsche Staatsangehörigkeit • Nein, habe eine andere Staatsangehörigkeit	**Question text:** Do you hold German citizenship? **Answer categories (3, nominal):** • Yes, only German citizenship • Yes, German citizenship • No, hold a different citizenship
internet_usage_18	private internet usage	**Fragetext:** Wie oft nutzen Sie das Internet, das World Wide Web oder E-Mail für private Zwecke, egal ob zu Hause oder am Arbeitsplatz? **Antwortkategorien:** 1. Nutze ich nie 2. Weniger als einmal im Monat 3. Einmal im Monat 4. Mehrmals im Monat 5. Einmal in der Woche 6. Mehrmals in der Woche 7. Täglich	**Question text:** How often do you use the internet, the World Wide Web or email for personal purposes at home or at work? **Answer categories:** 1. Never 2. Less than once a month 3. Once a month 4. Several times a month 5. Once a week 6. Several times a week 7. Daily

Variable	Item	German	English
state_18	residence state	**Fragetext:** In welchem Bundesland haben Sie Ihren Hauptwohnsitz? **Antwortkategorien (12, nominal):** • Schleswig-Holstein / Hamburg • Niedersachsen / Bremen • Nordrhein-Westfalen • Hesse • Rheinland-Pfalz / Saarland • Baden-Württemberg • Bayern • Berlin / Brandenburg • Mecklenburg-Vorpommern • Sachsen • Sachsen-Anhalt • Thüringen	**Question text:** In which state do you have your main residence? **Answer categories (12, nominal):** • Schleswig-Holstein / Hamburg • Lower Saxony / Bremen • North Rhine-Westphalia • Hessen • Rhineland-Palatinate / Saarland • Baden-Württemberg • Bavaria • Berlin / Brandenburg • Mecklenburg Western Pomerania • Saxony • Saxony-Anhalt • Thuringia

Variable	Item	German	English
sample	year of recruitment	[keine Frage] Kategorien (3, nominal): • rekrutiert in 2012 • rekrutiert in 2014 • rekrutiert in 2018	[no question] Categories (3, nominal): • recruited 2012 • recruited 2014 • recruited 2018
AA37027	social activity: culture	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Sport-, Musik- oder Kulturverein, sonstige Hobbyvereinigung **Antwortkategorien:** 1. Ich bin kein Mitglied. 2. Ich bin passives Mitglied. 3. Ich bin aktives Mitglied.	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Sports, music or cultural association, other hobby association **Answer categories:** 1. I am not a member. 2. I am a passive member. 3. I am an active member.

Variable	Item	German	English
AA37028	social activity: environment	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Menschenrechts-, Umwelt- oder Tierschutzverein **Antwortkategorien:** 1. Ich bin kein Mitglied. 2. Ich bin passives Mitglied. 3. Ich bin aktives Mitglied.	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Human rights, environmental or animal protection association **Answer categories:** 1. I am not a member. 2. I am a passive member. 3. I am an active member.
AA37029	social activity: social cause	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Wohltätigkeitsverein oder Hilfsorganisation **Antwortkategorien:** 1. Ich bin kein Mitglied. 2. Ich bin passives Mitglied. 3. Ich bin aktives Mitglied.	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Charity or aid organization **Answer categories:** 1. I am not a member. 2. I am a passive member. 3. I am an active member.

Variable	Item	German	English
AA37030	social activity: religion	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Religiöse oder kirchliche Organisation **Antwortkategorien:** 1. Ich bin kein Mitglied. 2. Ich bin passives Mitglied. 3. Ich bin aktives Mitglied.	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Religious or church organization **Answer categories:** 1. I am not a member. 2. I am a passive member. 3. I am an active member.
AA37031	social activity: youth organization	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Jugend-, Eltern- oder Seniorenvereinigung **Antwortkategorien:** 1. Ich bin kein Mitglied. 2. Ich bin passives Mitglied. 3. Ich bin aktives Mitglied.	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Youth, parent or elderly association **Answer categories:** 1. I am not a member. 2. I am a passive member. 3. I am an active member.

Variable	Item	German	English
AA37032	social activity: civil action	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Bürgerinitiative **Antwortkategorien:** 1. Ich bin kein Mitglied. 2. Ich bin passives Mitglied. 3. Ich bin aktives Mitglied.	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Citizens' initiative **Answer categories:** 1. I am not a member. 2. I am a passive member. 3. I am an active member.
AA37033	social activity: professional	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Berufsverband **Antwortkategorien:** 1. Ich bin kein Mitglied. 2. Ich bin passives Mitglied. 3. Ich bin aktives Mitglied.	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Professional association **Answer categories:** 1. I am not a member. 2. I am a passive member. 3. I am an active member.

Variable	Item	German	English
AA37034	social activity: union	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Gewerkschaft **Antwortkategorien:** 1. Ich bin kein Mitglied. 2. Ich bin passives Mitglied. 3. Ich bin aktives Mitglied.	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Trade union **Answer categories:** 1. I am not a member. 2. I am a passive member. 3. I am an active member.
AA37035	social activity: political party	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Politische Partei **Antwortkategorien:** 1. Ich bin kein Mitglied. 2. Ich bin passives Mitglied. 3. Ich bin aktives Mitglied.	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Political party **Answer categories:** 1. I am not a member. 2. I am a passive member. 3. I am an active member.

Variable	Item	German	English
AA37036	social activity: open1	**Fragetext:** Im Folgenden werden verschiedene Organisationen und Vereine aufgelistet. Gehen Sie diese bitte durch und geben Sie an, was zurzeit auf Sie zutrifft. Sonstige/-r, und zwar: Antwortfeld: [Antwortfeld]	**Question text:** In the following several organizations and associations are listed. Please go through them and specify what currently applies to you. Other, namely: Answer field: [answer field]
AA37044	Big5 reserved	**Fragetext:** Nun kommen einige allgemeine Aussagen, die zur Beschreibung von Personen verwendet werden können. Diese Aussagen können auf Sie persönlich mehr oder weniger zutreffen. Bitte geben Sie bei jeder Aussage an, inwieweit die Aussage auf Sie selbst zutrifft. Ich bin eher zurückhaltend, reserviert. **Antwortkategorien:** 1. trifft überhaupt nicht zu 2. trifft eher nicht zu 3. weder noch 4. eher zutreffend 5. trifft voll und ganz zu	**Question text:** Now we present some general statements that can be used to describe people. These statements may more or less apply to you personally. Please indicate for every statement to what extent it applies to you personally. I am rather reserved. **Answer categories:** 1. does not apply at all 2. does not apply 3. neither nor 4. rather applies 5. totally applies

Variable	Item	German	English
AA37045	Big5 trusting	**Fragetext:** Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen. **Antwortkategorien:** 1. trifft überhaupt nicht zu 2. trifft eher nicht zu 3. weder noch 4. eher zutreffend 5. trifft voll und ganz zu	**Question text:** I trust others easily, believe in the good in people. **Answer categories:** 1. does not apply at all 2. does not apply 3. neither nor 4. rather applies 5. totally applies
AA37046	Big5 lazy	**Fragetext:** Ich bin bequem, neige zur Faulheit. **Antwortkategorien:** 1. trifft überhaupt nicht zu 2. trifft eher nicht zu 3. weder noch 4. eher zutreffend 5. trifft voll und ganz zu	**Question text:** I am easy-going, tend to be lazy. **Answer categories:** 1. does not apply at all 2. does not apply 3. neither nor 4. rather applies 5. totally applies

Variable	Item	German	English
AA37047	Big5 relaxed	**Fragetext:** Ich bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen. **Antwortkategorien:** 1. trifft überhaupt nicht zu 2. trifft eher nicht zu 3. weder noch 4. eher zutreffend 5. trifft voll und ganz zu	**Question text:** I am relaxed, I am above such things as stress. **Answer categories:** 1. does not apply at all 2. does not apply 3. neither nor 4. rather applies 5. totally applies
AA37048	Big5 few artistic interests	**Fragetext:** Ich habe nur wenig künstlerisches Interesse. **Antwortkategorien:** 1. trifft überhaupt nicht zu 2. trifft eher nicht zu 3. weder noch 4. eher zutreffend 5. trifft voll und ganz zu	**Question text:** I only have limited artistic interest. **Answer categories:** 1. does not apply at all 2. does not apply 3. neither nor 4. rather applies 5. totally applies

Variable	Item	German	English
AA37049	Big5 sociable	**Fragetext:**	**Question text:**
		Ich gehe aus mir heraus, bin gesellig.	I am outgoing, sociable.
		Antwortkategorien:	**Answer categories:**
		1. trifft überhaupt nicht zu	1. does not apply at all
		2. trifft eher nicht zu	2. does not apply
		3. weder noch	3. neither nor
		4. eher zutreffend	4. rather applies
		5. trifft voll und ganz zu	5. totally applies
AA37050	Big5 find fault with others	**Fragetext:**	**Question text:**
		Ich neige dazu, andere zu kritisieren.	I tend to criticize others.
		Antwortkategorien:	**Answer categories:**
		1. trifft überhaupt nicht zu	1. does not apply at all
		2. trifft eher nicht zu	2. does not apply
		3. weder noch	3. neither nor
		4. eher zutreffend	4. rather applies
		5. trifft voll und ganz zu	5. totally applies

Variable	Item	German	English
AA37051	Big5 thourough	**Fragetext:**	**Question text:**
		Ich erledige Aufgaben gründlich.	I complete tasks thoroughly.
		Antwortkategorien:	**Answer categories:**
		1. trifft überhaupt nicht zu	1. does not apply at all
		2. trifft eher nicht zu	2. does not apply
		3. weder noch	3. neither nor
		4. eher zutreffend	4. rather applies
		5. trifft voll und ganz zu	5. totally applies
AA37052	Big 5 nervous	**Fragetext:**	**Question text:**
		Ich werde leicht nervös und unsicher.	I get nervous and insecure quickly.
		Antwortkategorien:	**Answer categories:**
		1. trifft überhaupt nicht zu	1. does not apply at all
		2. trifft eher nicht zu	2. does not apply
		3. weder noch	3. neither nor
		4. eher zutreffend	4. rather applies
		5. trifft voll und ganz zu	5. totally applies

Variable	Item	German	English
AA37053	Big5 active imagination	**Fragetext:** Ich habe eine aktive Vorstellungskraft, bin fantasievoll. **Antwortkategorien:** 1. trifft überhaupt nicht zu 2. trifft eher nicht zu 3. weder noch 4. eher zutreffend 5. trifft voll und ganz zu	**Question text:** I have an active imagination, I am vsionary. **Answer categories:** 1. does not apply at all 2. does not apply 3. neither nor 4. rather applies 5. totally applies
BG38001	lobbying EU	**Fragetext:** Inwieweit beeinflusst Lobbyismus Ihrer Meinung nach die Politik der Europäischen Union? **Antwortkategorien:** 0. 0 Überhaupt nicht ... 4. 4 Sehr stark	**Question text:** In your opinion, to what extent does lobbyism affect the politics of the European Union? **Answer categories:** 0. 0 Not at all ... 4. 4 Very strongly

Variable	Item	German	English
BG38002	lobbying Germany	**Fragetext:** Im Vergleich zum Einfluss von Lobbyismus auf die Politik der Europäischen Union, inwieweit beeinflusst Lobbyismus Ihrer Meinung nach die deutsche Politik? **Antwortkategorien:** 0. Weniger 1. Eher weniger 2. In gleichem Maße 3. Eher mehr 4. Mehr	**Question text:** When comparing the influence of lobbyism on the politics of the European Union, to what extent does lobbyism influence German politics? **Answer categories:** 0. Less 1. A little less 2. To the same extent 3. A little more 4. More
BG38006_a	EU more_-industry	**Fragetext:** Welche der folgenden Akteure profitieren Ihrer Meinung nach am stärksten vom Lobbyismus auf EU-Ebene? Industrie **Antwortkategorien:** 0. Kategorie nicht gewählt 1. Kategorie gewählt	**Question text:** In your opinion, which of the following actors profits most from lobbying at the EU level? Industry **Answer categories:** 0. item not checked 1. item checked

Variable	Item	German	English
BG38006_b	EU more_-citizens	**Fragetext:** Welche der folgenden Akteure profitieren Ihrer Meinung nach am stärksten vom Lobbyismus auf EU-Ebene? EU-Bürger **Antwortkategorien:** 0. Kategorie nicht gewählt 1. Kategorie gewählt	**Question text:** In your opinion, which of the following profits most from lobbying at the EU level? EU citizens **Answer categories:** 0. item not checked 1. item checked
BG38006_c	EU more_-charity	**Fragetext:** Welche der folgenden Akteure profitieren Ihrer Meinung nach am stärksten vom Lobbyismus auf EU-Ebene? Wohltätigkeitsorganisationen **Antwortkategorien:** 0. Kategorie nicht gewählt 1. Kategorie gewählt	**Question text:** In your opinion, which of the following profits most from lobbying at the EU level? Charitable organizations **Answer categories:** 0. item not checked 1. item checked

Variable	Item	German	English
BG38006_d	EU more_-politician	**Fragetext:** Welche der folgenden Akteure profitieren Ihrer Meinung nach am stärksten vom Lobbyismus auf EU-Ebene? Politiker **Antwortkategorien:** 0. Kategorie nicht gewählt 1. Kategorie gewählt	**Question text:** In your opinion, which of the following profits most from lobbying at the EU level? Politicians **Answer categories:** 0. item not checked 1. item checked
BG38006_e	EU more_-other	**Fragetext:** Welche der folgenden Akteure profitieren Ihrer Meinung nach am stärksten vom Lobbyismus auf EU-Ebene? Sonstige, und zwar: **Antwortkategorien:** 0. Kategorie nicht gewählt 1. Kategorie gewählt	**Question text:** In your opinion, which of the following profits most from lobbying at the EU level? Other, namely: **Answer categories:** 0. item not checked 1. item checked

Variable	Item	German	English
BG38006_f	EU don't_know	**Fragetext:** Welche der folgenden Akteure profitieren Ihrer Meinung nach am stärksten vom Lobbyismus auf EU-Ebene? Weiß ich nicht **Antwortkategorien:** 0. Kategorie nicht gewählt 1. Kategorie gewählt	**Question text:** In your opinion, which of the following profits most from lobbying at the EU level? I don't know **Answer categories:** 0. item not checked 1. item checked
BG38007	lobbying climate	**Fragetext:** Inwieweit beeinflusst Lobbyismus Ihrer Meinung nach die Klimapolitik der Europäischen Union? **Antwortkategorien:** 0. 0 Überhaupt nicht ... 4. 4 Sehr stark	**Question text:** In your opinion, to what extent does lobbying affect climate politics of the European Union? **Answer categories:** 0. 0 Not at all ... 4. 4 Very strongly

Variable	Item	German	English
BG38009	lobbying climate protection	**Fragetext:** Inwieweit beeinflusst Lobbyismus das Maß an Klimaschutz der Europäischen Union? **Antwortkategorien:** 0. Viel mehr Klimaschutz 1. Eher mehr 2. Weder mehr noch weniger 3. Eher weniger 4. Viel weniger Klimaschutz	**Question text:** To what extent does lobbying affect the climate protection policies of the European Union? **Answer categories:** 0. Much more climate protection 1. A little more climate protection 2. Neither more nor less 3. A little less climate protection 4. Much less climate protection
CE38153	satisfaction_-party_-cducsu	**Fragetext:** Wie unzufrieden oder zufrieden sind Sie mit den Leistungen der CDU/CSU (Christlich Demokratische Union Deutschlands/Christlich-Soziale Union) im Bundestag? **Antwortkategorien:** 1. 1 völlig unzufrieden ... 11. 11 völlig zufrieden	**Question text:** How dissatisfied or satisfied are you with the performance of CDU/CSU (Christian Democratic Union of Germany/Christian Social Union) in the Bundestag? **Answer categories:** 1. 1 totally dissatisfied ... 11. 11 totally satisfied

Variable	Item	German	English
CE38154	satisfaction_-party_spd	**Fragetext:** Wie unzufrieden oder zufrieden sind Sie mit den Leistungen der SPD (Sozialdemokratische Partei Deutschlands) im Bundestag? **Antwortkategorien:** 1. 1 völlig unzufrieden ... 11. 11 völlig zufrieden	**Question text:** How dissatisfied or satisfied are you with the performance of SPD (Social Democratic Party of Germany) in the Bundestag? **Answer categories:** 1. 1 totally dissatisfied ... 11. 11 totally satisfied
CE38155	satisfaction_-party_-gruene	**Fragetext:** Wie unzufrieden oder zufrieden sind Sie mit den Leistungen der Partei Bündnis 90/Die Grünen im Bundestag? **Antwortkategorien:** 1. 1 völlig unzufrieden ... 11. 11 völlig zufrieden	**Question text:** How dissatisfied or satisfied are you with the performance of Alliance 90/The Greens in the Bundestag? **Answer categories:** 1. 1 totally dissatisfied ... 11. 11 totally satisfied

Variable	Item	German	English
CE38156	satisfaction_-party_linke	**Fragetext:** Wie unzufrieden oder zufrieden sind Sie mit den Leistungen der Partei Die Linke im Bundestag? **Antwortkategorien:** 1. 1 völlig unzufrieden ... 11. 11 völlig zufrieden	**Question text:** How dissatisfied or satisfied are you with the performance of The Left in the Bundestag? **Answer categories:** 1. 1 totally dissatisfied ... 11. 11 totally satisfied
CE38312	satisfaction_-party_afd	**Fragetext:** Wie unzufrieden oder zufrieden sind Sie mit den Leistungen der AFD (Alternative für Deutschland) im Bundestag? **Antwortkategorien:** 1. 1 völlig unzufrieden ... 11. 11 völlig zufrieden	**Question text:** How dissatisfied or satisfied are you with the performance of AfD (Alternative for Germany) in the Bundestag? **Answer categories:** 1. 1 totally dissatisfied ... 11. 11 totally satisfied

Variable	Item	German	English
CE38347	probability voting cdu/csu	**Fragetext:** Es gibt eine Reihe von politischen Parteien in Deutschland. Jede davon würde gern Ihre Stimme bekommen. Wie wahrscheinlich ist es auf einer Skala von 1 (sehr unwahrscheinlich) bis 11 (sehr wahrscheinlich), dass Sie die Partei CDU/CSU (Christlich Demokratische Union Deutschlands/Christlich-Soziale Union) jemals wählen werden? **Antwortkategorien:** 1. 1 sehr unwahrscheinlich ... 11. 11 sehr wahrscheinlich	**Question text:** There are a number of political parties in Germany. Each of them would like to gain your vote. On a scale from 1 (very unlikely) to 11 (very likely), how likely is it that you will ever vote for the party CDU/CSU (Christian Democratic Union of Germany/Christian Social Union)? **Answer categories:** 1. 1 very unlikely ... 11. 11 very likely

Variable	Item	German	English
CE38348	probability voting spd	**Fragetext:** Es gibt eine Reihe von politischen Parteien in Deutschland. Jede davon würde gern Ihre Stimme bekommen. Wie wahrscheinlich ist es auf einer Skala von 1 (sehr unwahrscheinlich) bis 11 (sehr wahrscheinlich), dass Sie die Partei SPD (Sozialdemokratische Partei Deutschlands) jemals wählen werden? **Antwortkategorien:** 1. 1 sehr unwahrscheinlich ... 11. 11 sehr wahrscheinlich	**Question text:** There are a number of political parties in Germany. Each of them would like to gain your vote. On a scale from 1 (very unlikely) to 11 (very likely), how likely is it that you will ever vote for the party SPD (Social Democratic Party of Germany)? **Answer categories:** 1. 1 very unlikely ... 11. 11 very likely

Variable	Item	German	English
CE38350	probability voting gruene	**Fragetext:** Es gibt eine Reihe von politischen Parteien in Deutschland. Jede davon würde gern Ihre Stimme bekommen. Wie wahrscheinlich ist es auf einer Skala von 1 (sehr unwahrscheinlich) bis 11 (sehr wahrscheinlich), dass Sie die Partei Bündnis 90/Die Grünen jemals wählen werden? **Antwortkategorien:** 1. 1 sehr unwahrscheinlich ... 11. 11 sehr wahrscheinlich	**Question text:** There are a number of political parties in Germany. Each of them would like to gain your vote. On a scale from 1 (very unlikely) to 11 (very likely), how likely is it that you will ever vote for the party Alliance 90/The Greens? **Answer categories:** 1. 1 very unlikely ... 11. 11 very likely

Variable	Item	German	English
CE38351	probability voting linke	**Fragetext:** Es gibt eine Reihe von politischen Parteien in Deutschland. Jede davon würde gern Ihre Stimme bekommen. Wie wahrscheinlich ist es auf einer Skala von 1 (sehr unwahrscheinlich) bis 11 (sehr wahrscheinlich), dass Sie die Partei Die Linke jemals wählen werden? **Antwortkategorien:** 1. 1 sehr unwahrscheinlich ... 11. 11 sehr wahrscheinlich	**Question text:** There are a number of political parties in Germany. Each of them would like to gain your vote. On a scale from 1 (very unlikely) to 11 (very likely), how likely is it that you will ever vote for the party The Left? **Answer categories:** 1. 1 very unlikely ... 11. 11 very likely

Variable	Item	German	English
CE38352	probability voting afd	**Fragetext:** Es gibt eine Reihe von politischen Parteien in Deutschland. Jede davon würde gern Ihre Stimme bekommen. Wie wahrscheinlich ist es auf einer Skala von 1 (sehr unwahrscheinlich) bis 11 (sehr wahrscheinlich), dass Sie die Partei AfD (Alternative für Deutschland) jemals wählen werden? **Antwortkategorien:** 1. 1 sehr unwahrscheinlich ... 11. 11 sehr wahrscheinlich	**Question text:** There are a number of political parties in Germany. Each of them would like to gain your vote. On a scale from 1 (very unlikely) to 11 (very likely), how likely is it that you will ever vote for the party AfD (Alternative for Germany)? **Answer categories:** 1. 1 very unlikely ... 11. 11 very likely
CE38056	unity_gov	**Fragetext:** Innerhalb einer Bundesregierung werden manchmal verschiedene Standpunkte vertreten. Nehmen Sie die Bundesregierung als zerstritten oder als geschlossen wahr? **Antwortkategorien:** 1. 1 sehr zerstritten ... 11. 11 sehr geschlossen	**Question text:** Within the federal government different positions are sometimes represented. Do you perceive the federal government as divided or united? **Answer categories:** 1. 1 very divided ... 11. 11 very united

Variable	Item	German	English
CE38250	unity_- party_cdu	**Fragetext:** Innerhalb einer Partei werden manchmal verschiedene Standpunkte vertreten. Nehmen Sie die CDU (Christlich Demokratische Union Deutschlands) als zerstritten oder als geschlossen wahr? **Antwortkategorien:** 1. 1 sehr zerstritten ... 11. 11 sehr geschlossen	**Question text:** Within a party different positions are sometimes represented. Do you perceive CDU (Christian Democratic Union of Germany) as divided or united? **Answer categories:** 1. 1 very divided ... 11. 11 very united
CE38252	unity_- party_csu	**Fragetext:** Innerhalb einer Partei werden manchmal verschiedene Standpunkte vertreten. Nehmen Sie die CSU (Christlich-Soziale Union) als zerstritten oder als geschlossen wahr? **Antwortkategorien:** 1. 1 sehr zerstritten ... 11. 11 sehr geschlossen	**Question text:** Within parties different positions are sometimes represented. Do you perceive CSU (Christian Social Union) as divided or united? **Answer categories:** 1. 1 very divided ... 11. 11 very united

Variable	Item	German	English
CE38254	unity_- party_spd	**Fragetext:** Innerhalb einer Partei werden manchmal verschiedene Standpunkte vertreten. Nehmen Sie die SPD (Sozialdemokratische Partei Deutschlands) als zerstritten oder als geschlossen wahr? **Antwortkategorien:** 1. 1 sehr zerstritten ... 11. 11 sehr geschlossen	**Question text:** Within parties different positions are sometimes represented. Do you perceive SPD (Social Democratic Party of Germany) as divided or united? **Answer categories:** 1. 1 very divided ... 11. 11 very united
CE38256	unity_- party_ fdp	**Fragetext:** Innerhalb einer Partei werden manchmal verschiedene Standpunkte vertreten. Nehmen Sie die FDP (Freie Demokratische Partei) als zerstritten oder als geschlossen wahr? **Antwortkategorien:** 1. 1 sehr zerstritten ... 11. 11 sehr geschlossen	**Question text:** Within parties different positions are sometimes represented. Do you perceive FDP (Free Democratic Party) as divided or united? **Answer categories:** 1. 1 very divided ... 11. 11 very united

Variable	Item	German	English
CE38258	unity_- party_- gruene	**Fragetext:** Innerhalb einer Partei werden manchmal verschiedene Standpunkte vertreten. Nehmen Sie Bündnis 90/Die Grünen als zerstritten oder als geschlossen wahr? **Antwortkategorien:** 1. 1 sehr zerstritten ... 11. 11 sehr geschlossen	**Question text:** Within parties different positions are sometimes represented. Do you perceive Alliance 90/The Greens as divided or united? **Answer categories:** 1. 1 very divided ... 11. 11 very united
CE38260	unity_- party_ linke	**Fragetext:** Innerhalb einer Partei werden manchmal verschiedene Standpunkte vertreten. Nehmen Sie Die Linke als zerstritten oder als geschlossen wahr? **Antwortkategorien:** 1. 1 sehr zerstritten ... 11. 11 sehr geschlossen	**Question text:** Within parties different positions are sometimes represented. Do you perceive The Left as divided or united? **Answer categories:** 1. 1 very divided ... 11. 11 very united

Variable	Item	German	English
CE38262	unity_- party_afd	**Fragetext:** Innerhalb einer Partei werden manchmal verschiedene Standpunkte vertreten. Nehmen Sie die AfD (Alternative für Deutschland) als zerstritten oder als geschlossen wahr? **Antwortkategorien:** 1. 1 sehr zerstritten ... 11. 11 sehr geschlossen	**Question text:** Within parties different positions are sometimes represented. Do you perceive AfD (Alternative for Germany) as divided or united? **Answer categories:** 1. 1 very divided ... 11. 11 very united
CE38280	vagueness CDU/CSU	**Fragetext:** Vor Wahlen machen Parteien in der Regel Aussagen darüber, welche Reformen sie nach den Wahlen umsetzen wollen. Diese Aussagen können sehr vage oder auch sehr genau sein. Für wie vage oder genau halten Sie die Aussagen von CDU/CSU (Christlich Demokratische Union Deutschlands/Christlich-Soziale Union)? **Antwortkategorien:** 1. 1 sehr vage ... 11.11 sehr genau	**Question text:** Before elections, parties usually make statements about the policies they want to implement after the election. These statements can be very vague or very precise. How vague or precise do you perceive the statements of CDU/CSU (Christian Democratic Union of Germany/Christian Social Union)? **Answer categories:** 1. 1 very vague ... 11. 11 very precise

Variable	Item	German	English
CE38329	turkey_-salience	**Fragetext:** Wie wichtig sind die Beitrittsverhandlungen der Türkei mit der EU für Sie? **Antwortkategorien:** 1. 1 sehr wichtig ... 7. 7 überhaupt nicht wichtig	**Question text:** How important are the accession negotiations between Turkey and the EU for you? **Answer categories:** 1. 1 very important ... 7. 7 not important at all

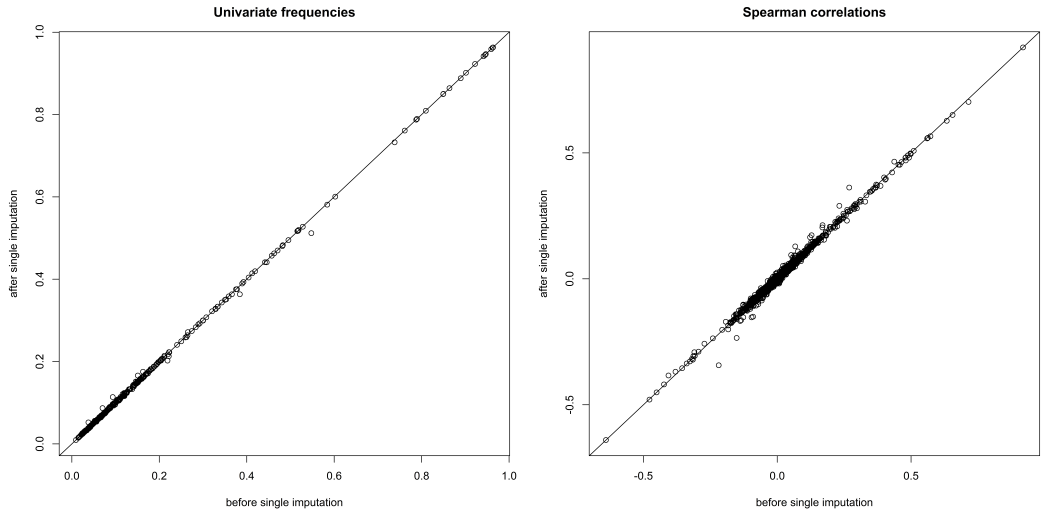

Figure 2.A1: Univariate frequencies and Spearman correlations before single imputation (on the horizontal axis, based on pairwise deletion) versus after single imputation (on the vertical axis) of item nonresponse in the population data.

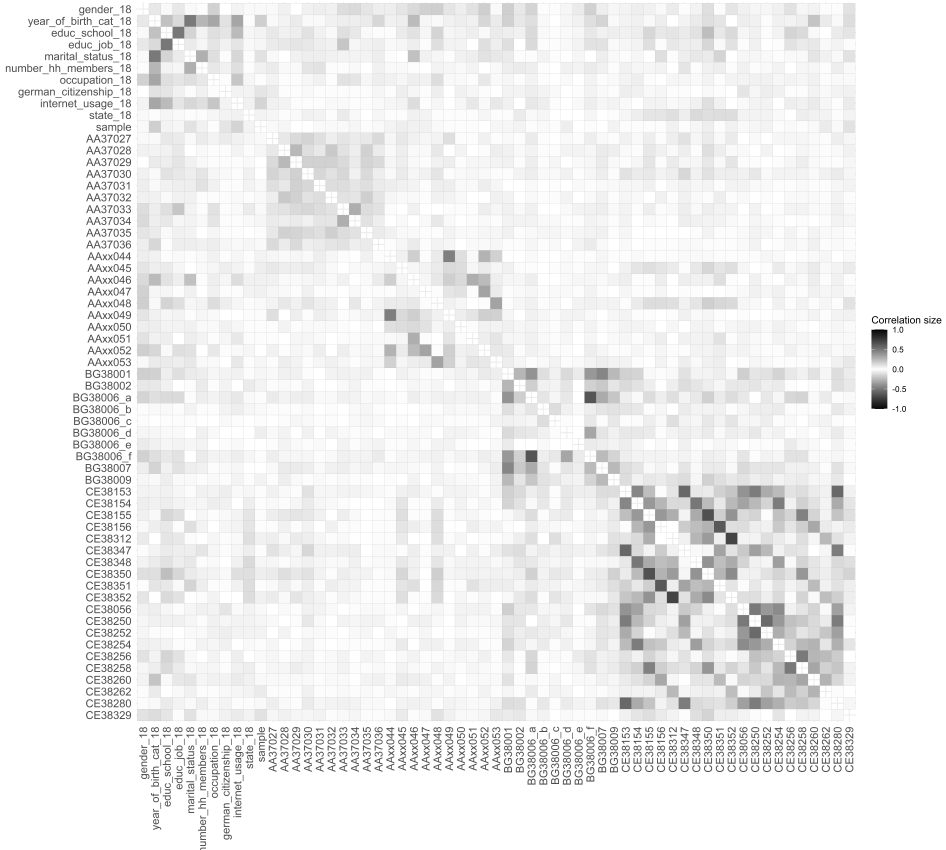

Figure 2.A2: Spearman correlations of GIP items used in the simulation.

Note: highest education degree, highest professional qualification, marital status, employment status, residence state, *and* year of recruitment *are nominal and thus their correlations allow only for limited interpretation.*

B. Alternative Imputation Strategies

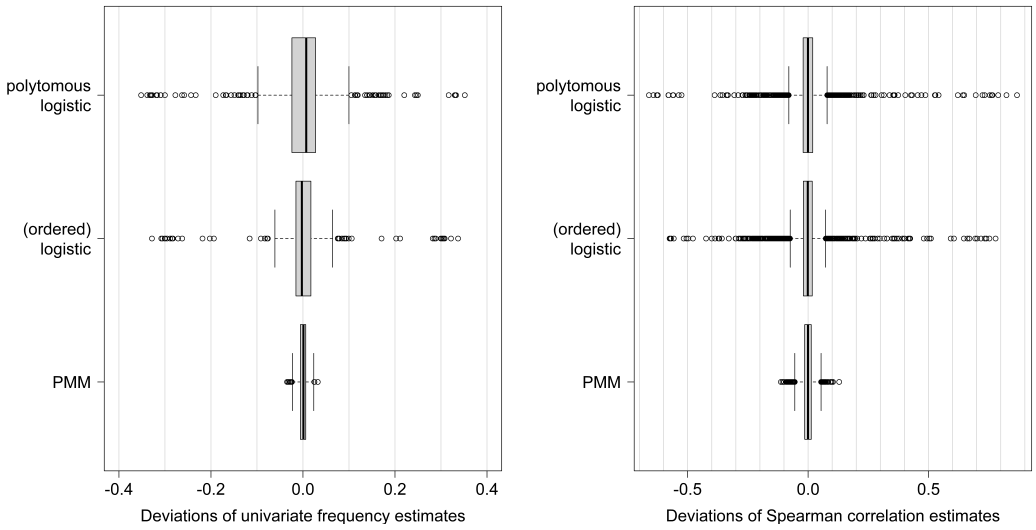

Figure 2.B1: Deviations of MI estimates from complete-sample estimates (Equations 2.3 and 2.4) by imputation method: (proportional odds) logistic regression vs. predictive mean matching.

Note: Based on a single test simulation run with n = 2,000 and random modularization.

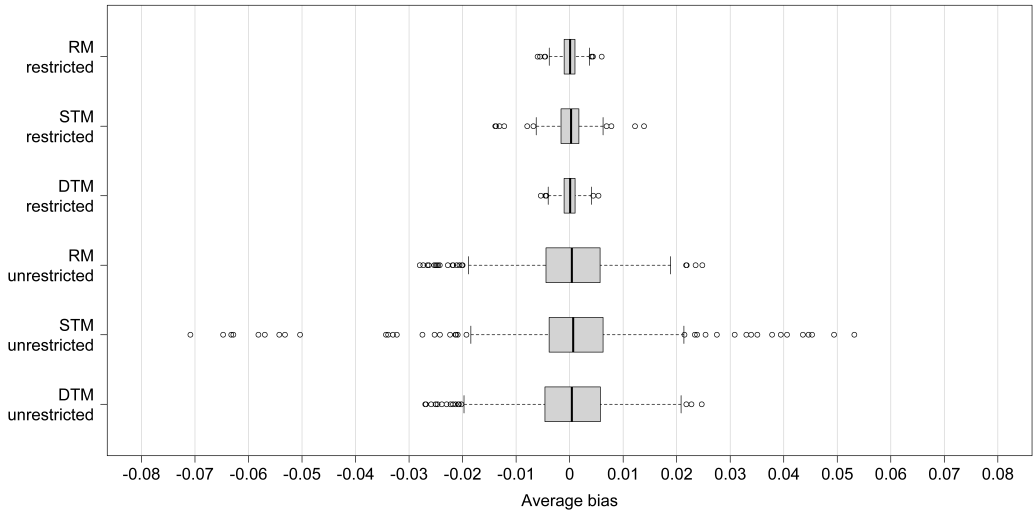

Figure 2.B2: Average biases for 297 univariate frequencies according to equation 5 with imputation model predictor sets including only variables with correlations stronger than |0.10| vs. all variables, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

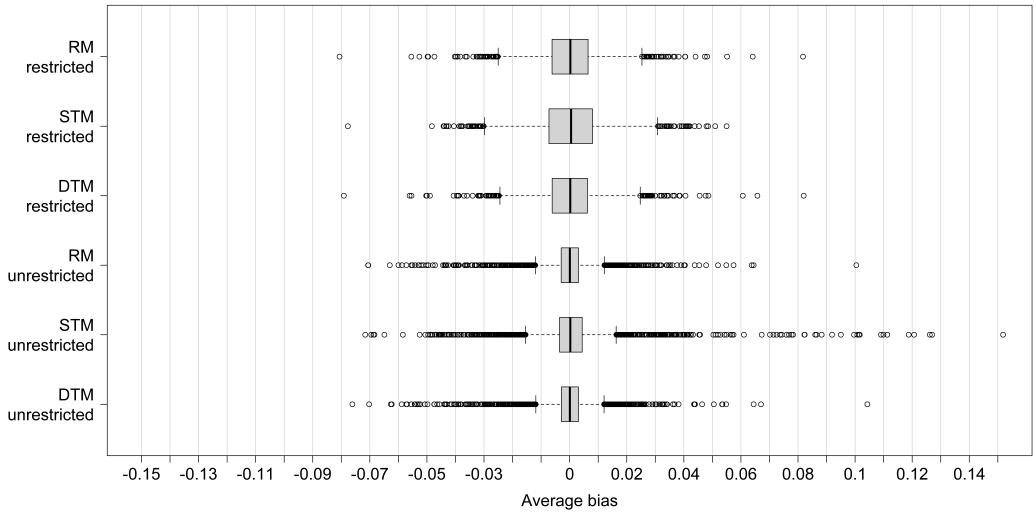

Figure 2.B3: Average biases for 3,675 bivariate correlations according to equation 6 with imputation model predictor sets including only variables with correlations stronger than |0.10| vs. all variables, by modularization technique. Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

C. Standard Errors of Estimates in the Simulation

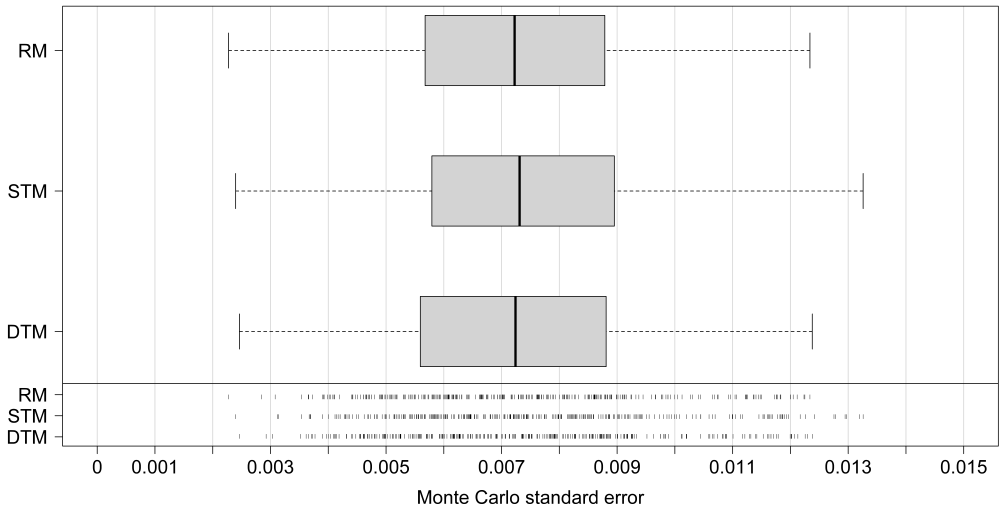

Figure 2.C1: Standard errors of 297 univariate frequencies, by imputation model predictor set and modularization technique. Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

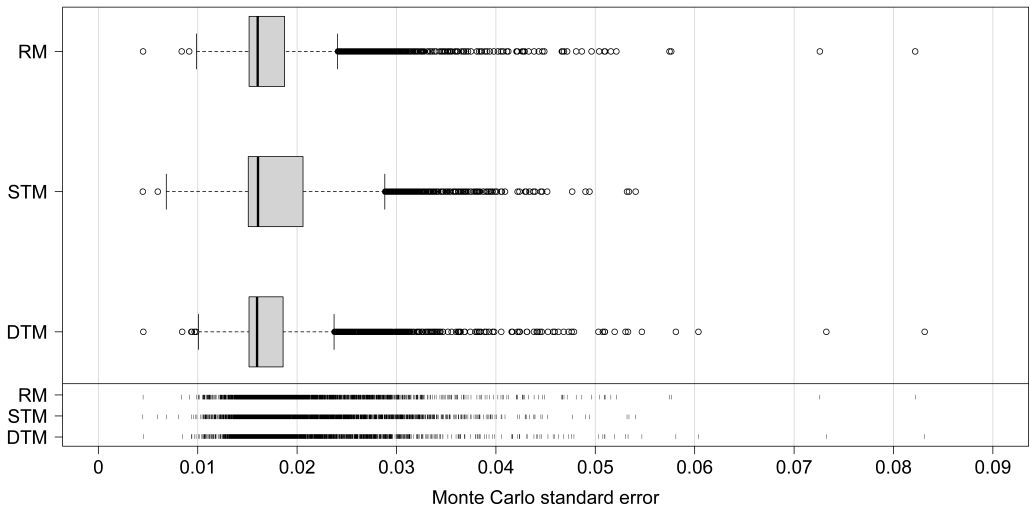

Figure 2.C2: Standard errors of 3,675 bivariate correlations, by imputation model predictor set and modularization technique. Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

D. Generation of and Results for Two Synthetic GIP Datasets With and Without High Correlations Within Topics

For additional simulations on the question how higher correlations within topics may affect imputation quality with the three modularization strategies, we generated two synthetic datasets based on the GIP, one with correlations no larger than 0.20 (scenario 1) and another with some correlations considerably increased within the same topic (scenario 2), according to the following data-generating procedure:

1. Estimate the Pearson correlation matrix for the GIP dataset as well as univariate distributions for all variables in the GIP.
2. Apply the Fisher's Z transformation on the correlation matrix.
3. Modify the transformed correlation matrix:
    a. Replace correlations larger than 0.20 (or smaller than -0.20) by 0.20 (-0.20) to generate data without high correlations.
    b. Only with scenario 2: Replace one same-topic correlation per variable by 0.90 (-0.90 if the correlation was negative before). Respective variable pairs are selected by their order in the data: Starting with the 11^{th} variable (representing the first split item in the data), highly correlated variable pairs are variables 11 and 12, variables 13 and 14, etc. until variable 61.
4. Apply the inverse Fisher's Z transformation on the modified correlation matrix.
5. Find the nearest positive-definite correlation matrix for the modified correlation matrix using the *nearPD* (Bates and Maechler, 2019) algorithm in R.
6. Generate a standard multivariate normal dataset based on the modified correlation matrix with 100,000 cases using *rnorm_multi* (DeBruine, 2020) in R.
7. Make the data categorical: For each variable, order the continuous values by size and assign a categorical value to it based on the quantile distribution of the respective categorical variable in the real observed GIP dataset. Through this procedure, the univariate distributions from the real GIP dataset are preserved in the synthetic data. At the same time, correlations decrease to some extent due to the concomitant information loss.

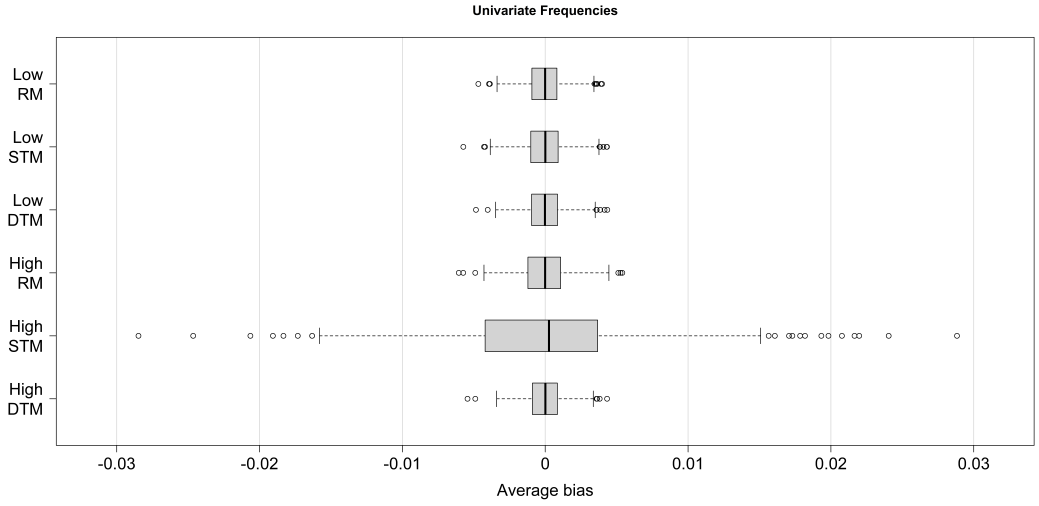

Figure 2.D1: Average biases for each 297 univariate frequencies according to equation 5 in two simulation studies on synthetic data with low correlations within the same topic (scenario "Low") and high correlations within the same topic (scenario "High"), by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on two Monte Carlo simulations with each 1,007 runs on 2,000 cases (40% missing data) each.

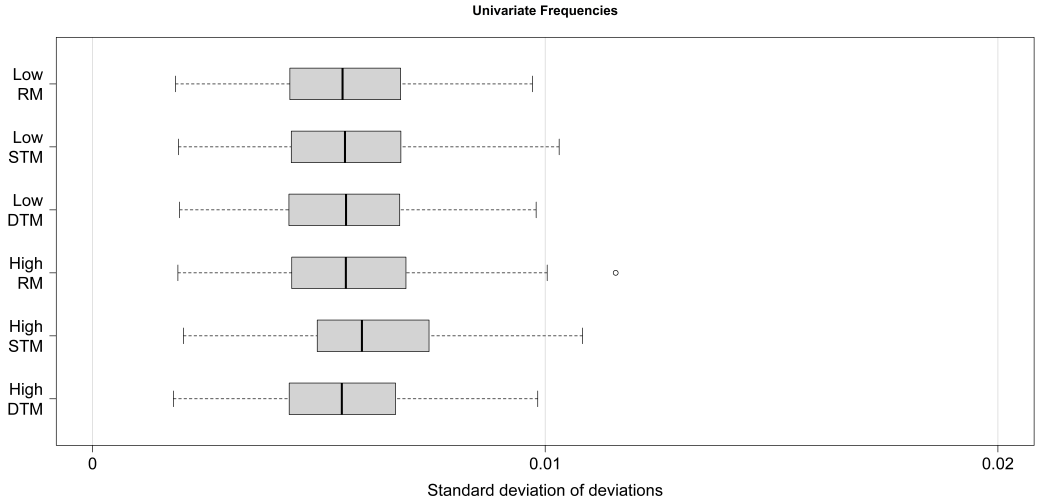

Figure 2.D2: Standard deviations of deviations for each 297 univariate frequencies according to equation 7 in two simulation studies on synthetic data with low correlations within the same topic (scenario "Low") and high correlations within the same topic (scenario "High"), by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on two Monte Carlo simulations with each 1,007 runs on 2,000 cases (40% missing data) each.

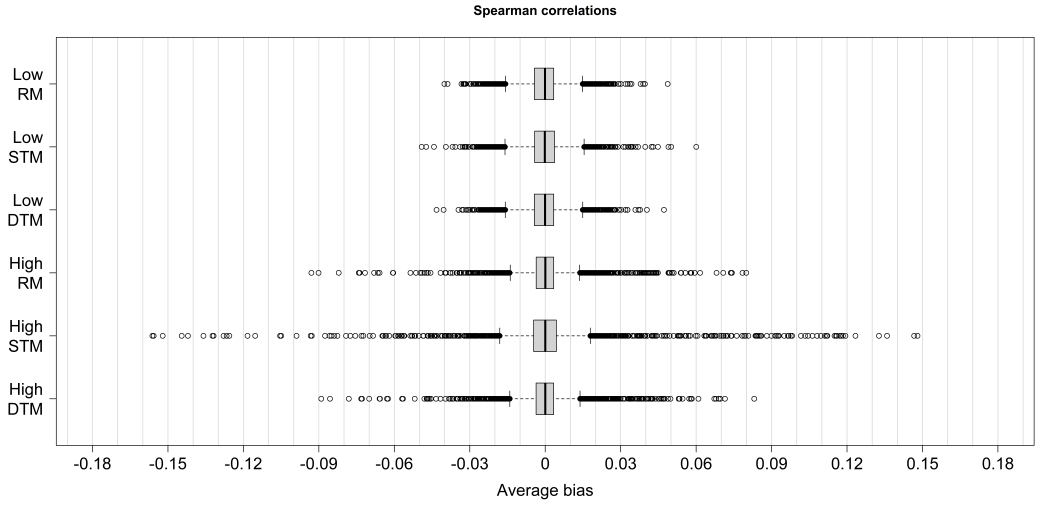

Figure 2.D3: Average biases for each 3,675 bivariate correlations according to equation 6 in two simulation studies on synthetic data with low correlations within the same topic (scenario "Low") and high correlations within the same topic (scenario "High"), by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on two Monte Carlo simulations with each 1,007 runs on 2,000 cases (40% missing data) each.

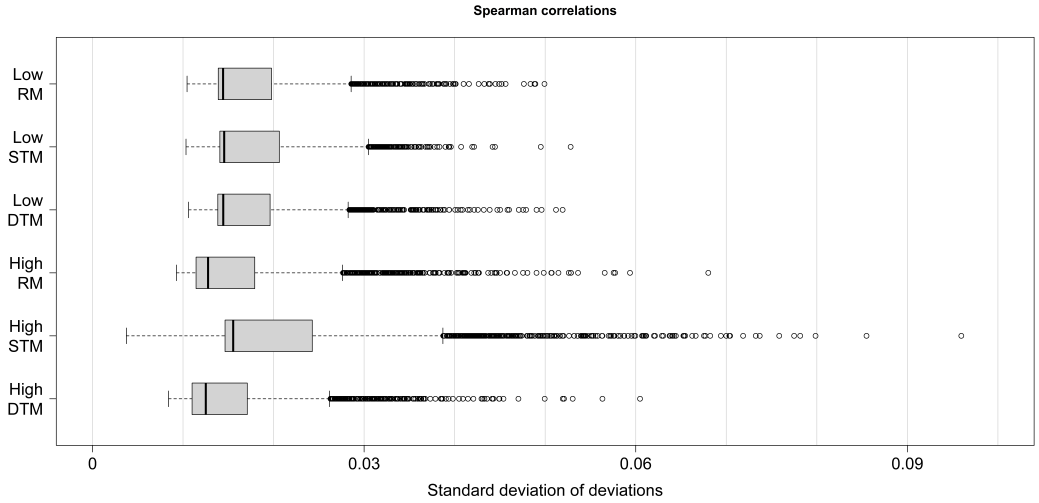

Figure 2.D4: Standard deviations of deviations (SDDs) for each 3,675 bivariate correlations according to equation 8 in two simulation studies on synthetic data with low correlations within the same topic (scenario "Low") and high correlations within the same topic (scenario "High"), by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM).

Note: Based on two Monte Carlo simulations with each 1,007 runs on 2,000 cases (40% missing data) each.

Chapter 3

General Purpose Imputation of Planned Missing Social Survey Data: Different Strategies and Their Effect on Correlations

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 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.

This paper is joint work with Christian Bruch, and Christof Wolf. A similar version of this paper has been published in:
Axenfeld, J. B., Bruch, C. & Wolf, C. (2022). General-purpose imputation of planned missing social survey data: Different strategies and their effect on correlations. *Statistics Surveys*, 16: 182-209.

3.1 Introduction

Long questionnaires pose a serious threat to the quality of survey data, triggering low response rates and poor response quality (Galesic and Bosnjak, 2009; Peytchev and Peytcheva, 2017). Recently, survey projects such as the PISA 2012 context questionnaire (OECD, 2014, pp. 48-58) or the European Values Study (Luijkx et al., 2021) have attempted to overcome this problem with methods such as the split questionnaire design (SQD) (Raghunathan and Grizzle, 1995). 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; Rubin, 1987) 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 (Bartlett et al., 2015). 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. Chapter 2 of this dissertation shows 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. (2014) 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 3.2, we discuss the theory on planned missing data and MI as well as different imputation methods. Section 3.3 explains our data and method. In Section 3.4, we describe our results for the different strategies, first for strong and then for weak relationships between variables. Section 3.5 concludes with a discussion of the implications and limitations of this study.

3.2 Imputation of Planned Missing Survey Data

3.2.1 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 (Munger and Loyd, 1988; Shoemaker, 1973). The split questionnaire design (Raghunathan and Grizzle, 1995; Graham et al., 1996) 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 (1995) and Graham et al. (1996) propose completing the missing data via MI (see also Adigüzel and Wedel, 2008; Bahrami et al., 2014; Imbriano and Raghunathan, 2020; Peytchev and Peytcheva, 2017; Rässler et al., 2002; Thomas et al., 2006). 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.

3.2.2 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 (Rubin, 1987; van Buuren, 2018). To understand MI, suppose we have a

variable Y that contains both observed values and planned missing values identified by vector $Z = \{0; 1\}$, where 1 indicates that a value is observed and 0 that it is missing. Our scenario assumes that all missing data $Y|(Z = 0)$ is planned as described above and thus missing completely at random (MCAR). MI aims to replace $Y|(Z = 0)$ with *m* potential values that are plausible given a matrix of predictor variables X (van Buuren, 2018, pp. 19-20). To this end, we rely on an imputation model that estimates the conditional probability distribution of Y given X using an adequate imputation method, accounting for all variable relationships as well as noise in the data and parameter uncertainty (van Buuren, 2018, pp. 65-68). Multiple imputed values are drawn randomly from this conditional distribution for each missing value, generating *m* independently imputed datasets (van Buuren, 2018, 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 (Rubin, 1987, pp. 12-16).

To analyze imputed data, estimates can be calculated separately for each of the *m* datasets with standard methods for complete data (Rubin, 1987, p. 12). Subsequently, these estimates are combined into a single estimate using Rubin's Rules (Rubin, 1987; van Buuren, 2018, pp. 145-147), yielding one combined estimator for each estimated parameter.

3.2.3 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 (Bartlett et al., 2015). 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 (Nicoletti and Peracchi, 2006; van Buuren, 2018, pp. 167-170, 259-271; White et al., 2011). 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 (van Buuren, 2018) or 30 to 40 (Honaker and King, , 2010) 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 (van Buuren et al., 1999), 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.

3.2.4 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.

Linear regression models (LRM)

First, linear regression can be used for MI (Rubin, 1987, pp. 166-167; van Buuren, 2018, 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, 2017, 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 (van Buuren, 2018, p. 67), a linear model of Y conditional on X is specified:

$$Y = X\beta + \epsilon\,, \tag{3.1}$$

where $\boldsymbol{\beta}$ is a vector of Bayesian estimates of the regression coefficients for the predictor variables in X and $\boldsymbol{\epsilon}$ represents the residuals. Accordingly, the posterior distribution $P(Y|X)|(Z = 1)$ 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 Y (van Buuren, 2018, 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 (Honaker et al., 2011; Lee and Carlin, 2010) while others show that outcomes can be biased with transformed variables as well (see for example von Hippel, 2013).

Furthermore, linear regression does not account for restrictions such as discrete scales or logical bounds (Long, 1997), potentially leading to implausible imputations (White et al., 2011; van Buuren, 2018, p. 78; von Hippel, 2013). For example, if 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 (Allison, 2005; von Hippel, 2009, 2013), 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 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 (Seaman et al., 2012; von Hippel, 2009).

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.

Categorical regression models (CRMs)

To circumvent some of the theoretical disadvantages of LRM, we might consider using categorical regression models (CRMs; Brand, 1999; Rubin, 1987, pp. 169-170; van Buuren et al., 2006) 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

$$Y = g(X\beta)\,, \tag{3.2}$$

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

$$Pr(Y = 1) = \frac{e^{X\beta}}{1 + e^{X\beta}}\,. \tag{3.3}$$

In this way, the non-normal distribution of categorical outcomes can be accounted for. As a result, CRMs with a correct specification of Y's discrete distribution allow for directly drawing imputations that stick to empirically possible values. However, we still assume that all effects of X on the transformed 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 (2018, 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 (van Belle, 2002, p. 87; van Buuren, 2018, 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. (2011) 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. (2015) observe that LRMs outperform CRMs in various scenarios with binary and ordinal variables.

Predictive mean matching (PMM)

Another common method used in MI is predictive mean matching (PMM; Little, 1988; Rubin, 1986). PMM is a two-stage method: First, a regression is applied to the data. However, instead of drawing imputations directly, predicted values $\hat{Y}$ are calculated and a real observed Y value is drawn from a set of donors with similar $\hat{Y}$. 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 (Koller-Meinfelder, 2009; Siddique and Belin, 2008).

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 (Koller-Meinfelder, 2009; Morris et al., 2014; van Buuren, 2018, pp. 77-79). However, model misspecifications can still result in biases (Koller-Meinfelder, 2009; Morris et al., 2014; Seaman et al., 2012). For example, interaction effects must be specified explicitly (Seaman et al., 2012). Moreover, when missing cases do not have enough potential donors nearby, PMM resorts to more distant donors to draw imputed values, which may also result in bias (Kleinke, 2018).

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 (2021) implement partial least squares (PLS) regression (de Jong, 1993; Mevik and Wehrens, 2007) 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 X that describe the maximum possible covariance of X and Y (de Jong, 1993). These PLS components are uncorrelated latent variables optimized to predict Y and ordered by decreasing importance for predicting 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 X, the number of parameters in the model is reduced. At the same time, most of the information on Y is preserved, as the PLS components were extracted from X specifically to predict Y. Second, substituting the original variables 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 X means that some other information in X will be ignored in the imputation. If this ignored part of X still contains additional information on 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 X and Y is included in the imputation, that is, a potential $k + 1$-th component must not provide any substantial further information on 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.

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. (1984), have shown to be a relatively simple method for this purpose (Burgette and Reiter, 2010; Doove et al., 2014). 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., 2014), making them quite computationally demanding.

CART creates a decision tree predicting Y by repeatedly partitioning the data into two subregions along the values of the predictor variables. After having started with an unconditional estimate of Y (i.e., the mean or mode, depending on whether Y is continuous or categorical), a cut-off point on a variable in X is chosen and 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 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 (van Buuren, 2018, 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 (Akande et al., 2017; Burgette and Reiter, 2010; Doove et al., 2014). Slade and Naylor (2020) 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 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 Y. For instance, Doove et al. (2014) 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 (Bellman, 1961), in which adding more predictors requires an exponential growth in case numbers. In consequence, CART implicitly assumes that only a few predictors in X really determine 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. Fur-

thermore, in a data context in which predictive information on Y is not primarily stored in a few strong correlations but in many different weak correlations, much information on Y may be lost in the imputation when the selected imputation method limits the number of predictors so strictly.

3.2.5 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 Y with missing values, there will also be missing values in 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 modeling (JM) as a holistic method instead of integrating univariate imputation methods.

JM is the classical application of MI described by Rubin (1987). It entails modeling the joint distribution of multivariate missing data in a single multivariate model (van Buuren, 2018, 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 (Honaker et al., 2011; Schafer, 1999). 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 (Brand, 1999; van Buuren et al., 2006; van Buuren, 2018) 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 (van Buuren, 2018, 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 $Y|(Z = 1)$. 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 (van Buuren, 2018, 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.

3.3 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.

3.3.1 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 (Blom et al., 2015, 2017, 2019a,b; Cornesse et al., 2021). 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 Chapter 2).

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 (Blom et al., 2016a,b). The remaining item nonresponse (on average 167 values or 4% per item) was imputed with single imputations in *R* (R Core Team, 2021) via *mice* (van Buuren and Groothuis-Oudshoorn, 2011),[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 Figure 2.A1).

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.

3.3.2 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

[1] Other *R* packages used for this paper (if not cited elsewhere) are: DescTools (Signorell et al., 2020), doMPI (Weston, 2017), foreach (Microsoft and Weston, 2020), ggplot2 (Wickham, 2016), haven (Wickham and Miller, 2019), MASS (Venables and Ripley, 2002), Rmpi (Yu, 2002), tidyr (Wickham and Henry, 2019).

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;
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.

3.3.3 Imputation Strategies

In each simulation run, we test different imputation methods implemented in R. We implement JM via *Amelia* (Honaker et al., 2011), 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 (Allison, 2005; Horton et al., 2003), so Honaker et al. (2011, 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* (van Buuren and Groothuis-Oudshoorn, 2011): the *mice* default (CRM, here: logistic regression and ordinal logistic regression), *norm* (Bayesian LRM), *pmm*, and *cart*.

[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.

Furthermore, we use *pls* (PLS-PMM) from the *miceadds* package (Robitzsch and Grund, 2021), 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 $|0.10|$ (option 1) or weaker than $|0.20|$ (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 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, 1998), but depending on the data and analysis purpose, *m* must often be considerably larger (Bodner, 2008; Graham et al., 2007; von Hippel, 2020). 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.

3.3.4 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 (van Buuren, 2018, 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

$|0.20|$ with a maximum value of 0.70, 140 (11%) are stronger than $|0.10|$ but at most $|0.20|$, 248 (20%) are stronger than $|0.05|$ but at most $|0.10|$, 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,

$$Bias^{MC}(\hat{\rho}_{i,j}^{imputed}) = \frac{1}{S}\sum_{s=1}^{S}\hat{\rho}_{i,j,s}^{imputed} - \rho_{i,j}\,, \tag{3.4}$$

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 bias by dividing the raw bias by the true correlation $\rho_{i,j}$ and multiplying it by 100:

$$\%Bias^{MC}(\hat{\rho}_{i,j}^{imputed}) = \frac{Bias^{MC}(\hat{\rho}_{i,j}^{imputed})}{\rho_{i,j}} \times 100\,. \tag{3.5}$$

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 3.5 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 3.4, 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 $|0.05|$.

3.4 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.

3.4.1 Item Pairs With Moderate or Strong Relationships

Figure 3.1 displays the average percentage biases in Spearman correlations for the 85 item pairs with moderate or strong relationships (stronger than $|0.20|$ 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 3.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 $|0.10|$ 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 $|0.20|$ 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 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.

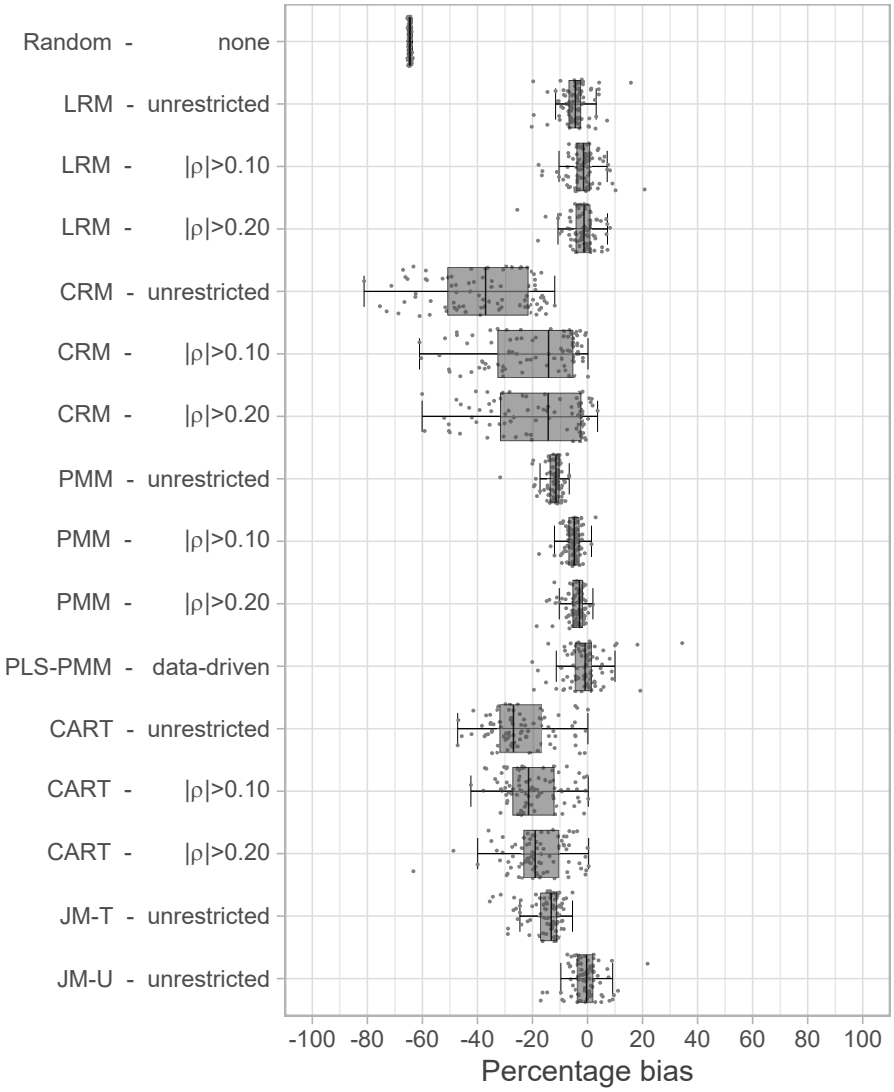

Figure 3.1: Average percentage Monte Carlo biases of Spearman correlations for 85 item pairs with moderate or strong relationships (true correlations stronger than $|0.2|$), by imputation method and predictor set specification.

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; $|\rho| > 0.10$/0.20 = with only predictors with $|\rho| > 0.10$ / $|\rho| > 0.20$ in the predictor set; data-driven = 20 PLS components; none = no predictors.
Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40% missing data) each.

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 $\pm 20\%$ 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 $\pm 20\%$.

Thus, despite some remaining biases, JM with untransformed imputations overall performs well.

To sum up, strong correlations over $|0.20|$ 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.

3.4.2 Item Pairs With Weak or Null Relationships

Figure 3.2 displays the average percentage biases for the 388 item pairs that had weak relationships in the population (between $|0.05|$ and $|0.20|$), again for different imputation methods. Alternatively, quantile distributions are given in Table 3.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 3.1, with point clouds and boxplots showing the distributions of biases for each strategy.

In general, Figure 3.2 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 3.4.1. 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, 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 3.1 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,

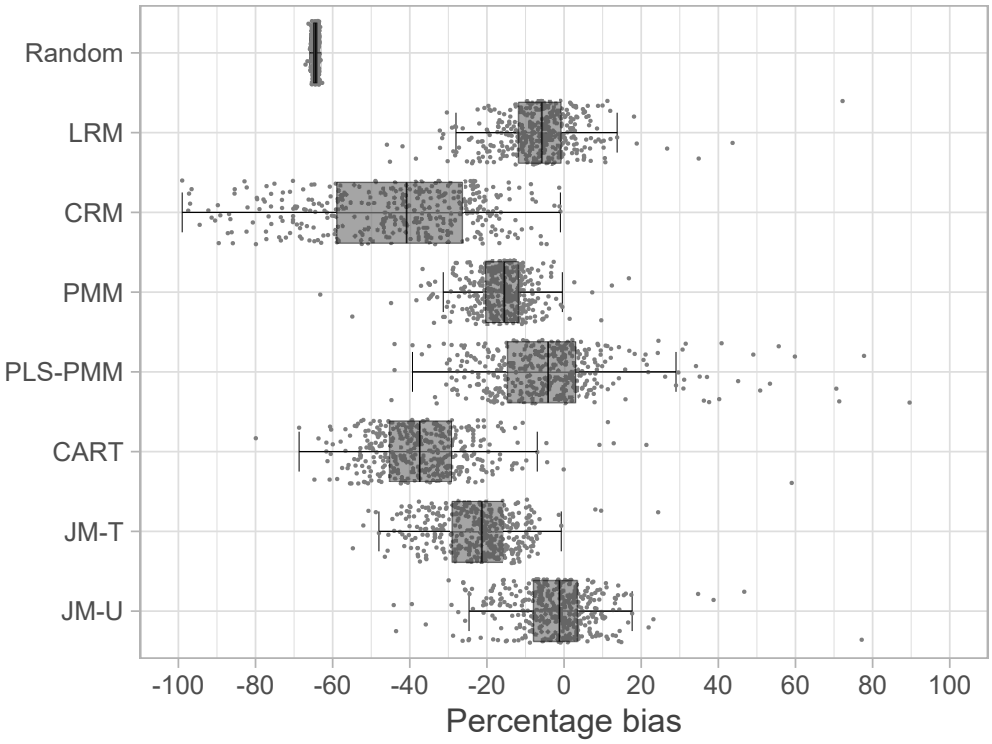

Figure 3.2: Average percentage Monte Carlo biases of Spearman correlations for 388 item pairs with weak relationships (true correlations weaker than $|0.20|$ but stronger than $|0.05|$), by imputation method.

See Note Figure 3.1.

Table 3.1: 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 3.1.

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.

3.5 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 $|0.20|$ 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., 2003). 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., 2011; Wu et al., 2015).

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

(Doove et al., 2014). 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 (Doove et al., 2014), 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" (van Buuren, 2018, 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., Koller-Meinfelder, 2009) 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 (see Chapter 2), 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.

References

Adigüzel, F. & Wedel, M. (2008). Split questionnaire design for massive surveys. *Journal of Marketing Research*, *45*(5), 608-617.

Allison, P. D. (2005). Imputation of categorical variables with PROC MI. In *Proceedings of the SAS Users Group International (SUGI)* (Vol. 30, pp. 113-130). SAS Institute.

Akande, O., Li, F. & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, *71*(2), 162-170.

Bahrami, S., Aßmann, C., Meinfelder, F., & Rässler, S. (2014). A split questionnaire survey design for data with block structure correlation matrix. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.) *Improving survey methods: Lessons from recent research* (pp. 368-380). Routledge.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462-487.

Bellman, R. E. (1961). *Adaptive control processes: A guided tour*. Princeton University Press.

Blom, A. G., Bossert, D., Funke, F., Gebhard, F., Holthausen, A., & Krieger, U.; SFB 884 “Political Economy of Reforms” Universität Mannheim (2016a). *German Internet Panel, wave 1 - core study (September 2012)*. GESIS Data Archive, ZA5866 Data file Version 2.0.0. https://doi.org/ 10.4232/1.12607.

Blom, A. G., Bossert, D., Gebhard, F., Funke, F., Holthausen, A., & Krieger, U.; SFB 884 “Political Economy of Reforms” Universität Mannheim (2016b). *German Internet Panel, wave 13 - core study (September 2014)*. GESIS Data Archive, ZA5924 Data file Version 2.0.0. https://doi.org/10.4232/1.12619.

Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., & Wenz, A.; SFB 884 “Political Economy of Reforms”, Universität Mannheim (2019a). *German Internet Panel, wave 37 - core study (September 2018)*. GESIS Data Archive, ZA6957 Data file Version 1.0.0. https://doi.org/10.4232/1.13390.

Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., & Wenz, A.; SFB 884 “Political Economy of Reforms”, Universität Mannheim (2019b). *German Internet Panel, wave 38 (November 2018)*. GESIS Data Archive, ZA6958 Data file Version 1.0.0. https://doi.org/10.4232/1.13391.

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. *Field Methods*, *27*(4), 391-408.

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., & 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*(4), 498-520.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 651-675.

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Burgette, L. F. & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070-1076.

Cornesse, C., Felderer, B., Fikel, M., Krieger, U., & Blom, A. G. (2021). Recruiting a probability-based online panel via postal mail: Experimental evidence. *Social Science Computer Review*, *40*(5), 1259-1284.

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251-263.

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

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360.

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

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

Honaker, J. & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561-581.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1-47.

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, *57*(4), 229-232.

Imbriano, P. M. & Raghunathan, T. E. (2020). Three-form split questionnaire design for panel Surveys. *Journal of Official Statistics*, *36*(4), 827-854.

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology*, *14*(1), 3-15.

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

Lee, K. J. & Carlin, J. B. (2010). Multiple imputation in the presence of non-normal data. *Statistics in Medicine*, *36*(4), 624-632.

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

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Sage.

Luijkx, R., Jónsdóttir, G. A., Gummer, T., Ernst Stähli, M., Fredriksen, M., Reeskens, T., Ketola, K., Brislinger, E., Christmann, P., Gunnarsson, S. Þ., Hjaltason, Á. B., Joye, D., Lomazzi, V., Maineri, A. M., Milbert, P., Ochsner, M., Ólafsdóttir, S., Pollien, A., Sapin, M., ... Wolf, C. (2021). The European Values Study 2017: On the way to the future using mixed-modes. *European Sociological Review*, *37*(2), 330-346.

Mevik, B.-H. & Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, *18*(2), 1-24.

Microsoft & Weston, S. (2020). *foreach: Provides foreach looping construct*. R package version 1.5.0. https://CRAN.R-project.org/package=foreach

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(75).

Munger, G. F. & Loyd, B. H. (1988). The use of multiple matrix sampling for survey research. *The Journal of Experimental Education*, *56*(4), 187-191.

Nicoletti, C. & 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*, *169*(3), 625-646.

OECD (2014). *PISA 2012 technical report*. OECD.

Peytchev, A. & Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods*, *11*(4), 361-368.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

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

Rässler, S., Koller, F., & 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*. https://www.statistik.rw.fau.de/files/2016/03/d0042b.pdf

Robitzsch, A. & Grund, S. (2021). *miceadds: Some additional multiple imputation functions, especially for 'mice'*. R package version 3.11-6. https://CRAN.R-project.org/package=miceadds

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87-94.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Schafer, J. L. (1999). *NORM users guide (version 2)*. The Methodology Center, The Pennsylvania State University. https://www.methodology.psu.edu/files/2019/09/NORM.pdf

Schafer, J. L. & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545-571.

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology*, *12*(46).

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & 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*(6), 764-774.

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

Siddique, J. & Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine*, *27*(1), 83-102.

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., ... Zeileis, A. (2020). *DescTools: Tools for descriptive statistics*. R package version 0.99.36.

Slade, E. & Naylor, M. G. (2020). A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Statistics in Medicine*, *39*(8), 1156-1166.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the national health and nutrition examination survey. *Survey Methodology*, *32*(2), 217-231.

van Belle, G. (2002). *Statistical rules of thumb*. John Wiley & Sons.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC press.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*(6), 681-694.

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049-1064.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer.

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*(1), 265-291.

von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research*, *42*(1), 105-138.

von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*, *49*(3), 699-718.

Weston, S. (2017). *doMPI: foreach parallel adaptor for the Rmpi package*. R package version 0.2.2. https://CRAN.R-project.org/package=doMPI

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377-399.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wickham, H. & Henry, L. (2019). *tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. R package version 0.8.3. https://CRAN.R-project.org/package=tidyr

Wickham, H. & Miller, E. (2019). *haven: Import and export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.1.1. https://CRAN.R-project.org/package=haven

Wu, H. & Leung, S.O. (2017). Can Likert scales be treated as interval scales?—A simulation study. *Journal of Social Service Research*, *43*(4), 527-532.

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research*, *50*(5), 484-503.

Yu, H. (2002). Rmpi: Parallel statistical computing in R. *R News*, *2*(2), 10-14.

Appendix

Quantile Distributions for the Information Displayed in Figures 3.1 and 3.2

Table 3.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 $|0.20|$), 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	$\|\rho\| > 0.10$	-17.8	-10.1	-4.0	-1.5	0.6	7.4	20.7
LRM	$\|\rho\| > 0.20$	-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	$\|\rho\| > 0.10$	-61.0	-50.3	-32.5	-14.3	-5.6	-1.8	0.2
CRM	$\|\rho\| > 0.20$	-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	$\|\rho\| > 0.10$	-17.6	-9.6	-6.8	-5.0	-3.4	-1.7	3.0
PMM	$\|\rho\| > 0.20$	-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	$\|\rho\| > 0.10$	-42.4	-33.4	-27.2	-21.4	-12.3	-1.6	0.3
CART	$\|\rho\| > 0.20$	-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 3.1.

Table 3.A2: Quantile distribution of average percentage Monte Carlo biases of Spearman correlations for 388 item pairs with weak relationships (true correlations weaker than $|0.20|$ 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 3.1.

Chapter 4

The Performance of Multiple Imputation in Social Surveys With Missing Data From Planned Missingness and Item Nonresponse

Abstract

Designs using planned missingness, such as the split questionnaire design, are becoming more and more important in social survey research. To ensure an acceptable questionnaire length, these approaches typically entail large amounts of planned missing data, which can be imputed after data collection. However, social surveys typically also include other types of missingness such as item nonresponse by survey participants, which need to be imputed as well. This entails a complex imputation task with amounts of missing data larger than initially planned and a potentially non-random, heterogeneous mechanism. Yet, it remains to be studied whether accurate multiple imputation estimates can be obtained in practice with planned missingness and item nonresponse.

To deal with this research gap, we apply a Monte Carlo simulation study using real social survey data. In this study, we simulate missing data based on item nonresponse with different mechanisms and proportions of item nonresponse as well as different proportions of planned missing data. We find that item nonresponse

This paper is joint work and based on a previously unpublished paper by Axenfeld, J. B., Bruch, C., Wolf, C., & Blom, A. G.

can jeopardize the quality of estimates after multiple imputation especially when the total amount of missing data from both sources is high or when there is a considerable proportion of item nonresponse that is missing not at random. Therefore, from an imputation perspective, survey designers should incorporate their expectations about item nonresponse on each variable when designing surveys with planned missing data.

4.1 Introduction

Survey designs using planned missingness are recently receiving a lot of attention in social survey research. This is marked by a growing body of research, particularly focusing on how to design the planned missingness patterns in such surveys (see for instance Chapter 2 of this dissertation; Adigüzel and Wedel, 2008; Bahrami et al., 2014; Imbriano and Raghunathan, 2020; Thomas et al., 2006). Increasingly, designs with planned missingness are also being applied in large-scale social surveys, such as the European Values Study 2017 (Luijkx et al., 2021) or the PISA 2012 context questionnaire (OECD, 2014, pp. 48-58). Examples of planned missingness designs are multiple matrix sampling (Shoemaker, 1973; Munger and Loyd, 1988), two-method measurement designs (Graham et al., 2006), and the 3-form design (Graham et al., 1996) or (similarly) the split questionnaire design (SQD; Raghunathan and Grizzle, 1995).

The SQD entails leaving out items for each respondent based on a random procedure. This usually serves to shorten questionnaires for individual respondents, considering that lengthy questionnaires can lead to reduced response rates, high breakoff, and increased measurement error (Galesic and Bosnjak, 2009; Peytchev and Peytcheva, 2017). This especially applies to self-administered online surveys (Callegaro et al., 2015; de Leeuw, 2008), which increasingly tend to compete with traditional face-to-face surveys.

The resulting planned missing data (PMD) is usually considered *missing completely at random* (MCAR). Yet, as all cases and most variables would be incomplete, simple pairwise deletion may often result in insufficient net sample sizes. Thus, as proposed by Raghunathan and Grizzle (1995), missing data from SQDs may need to be imputed.

Meanwhile, additional sources of missing data are typically present as well in SQD surveys. In particular, item nonresponse (INR) by survey participants is a com-

mon issue.[1] Unlike unit nonresponse, INR has been found to be little responsive to variations in survey length (Galesic and Bosnjak, 2009). Thus, we may expect that INR constitutes a similar challenge to SQD and conventional surveys alike. This also includes the potential for nonresponse bias, which would require appropriate treatment (see, for example, Durrant, 2009; Frick and Grabka, 2005; Rässler and Riphahn, 2006) through statistical techniques such as multiple imputation (MI; Rubin, 1987; van Buuren, 2018). Yet, INR is often not considered explicitly in research on imputing SQD survey data.

A realistic scenario of imputing SQD survey data has to take different types of missingness into account: PMD by the design and INR by the participants. These two types of missingness combined may cause an adverse scenario for the imputation: First, both types of missingness may in combination sum up to a very large overall proportion of missing data. On the one hand, this is because a considerable reduction in questionnaire length requires an equivalent amount of PMD. On the other hand, INR can unexpectedly cause considerable amounts of missingness because participants' response behavior is not under the control of the survey designer. Second, INR by participants may occur non-randomly, potentially causing nonresponse bias. In consequence, imputation models need to account for a potentially heterogeneous, non-random missingness mechanism for a potentially very large amount of missing data. This is important also because the resulting low case numbers available for the imputation model might hamper its capacity to account for the variables relevant for the response mechanism. Consequently, both types of missingness combined in a survey might adversely affect estimates after imputing the data. All this implies that future implementations of SQDs in social surveys may depend crucially on appropriate research telling if and under which conditions accurate estimates can be obtained. Existing research on imputing SQD survey data does not provide such inference.

We contribute to this research gap by investigating how the simultaneous occurrence of PMD and INR in social surveys affects estimates after imputation. In doing so, we seek to determine to what extent SQDs might still constitute a useful tool for social surveys when additional INR is factored in. We also examine if the imputation is able to deal with bias introduced by INR in such a situation.

In this paper, we use a Monte Carlo simulation study based on real social survey data. We vary the proportion of PMD, the proportion of INR, and the mechanism

[1] Note that our definition of INR in the following does not include planned missingness, i.e. we restrict the definition to cases where participants fail to deliver a response to a question assigned to them.

producing the INR. We investigate the accuracy of univariate frequency and bivariate correlation estimates after imputation in the different scenarios.

4.2 Theory

Assume we have a survey with $1, 2, \ldots, i, \ldots, n$ respondents and $1, 2, \ldots, j, \ldots, k$ variables yielding an $n \times k$ data matrix X with observations on a variable j identified by the vector $\vec{x}_j = \{x_{1,j}, x_{2,j}, \ldots, x_{i,j}, ..., x_{n,j}\}$. Some values in X are missing, with Z being the missingness indicator matrix of the same dimensionality as X identifying missing observations by 1 and available observations by 0.

4.2.1 Missingness Mechanisms

Missing data can have different effects on the analysis of survey data depending on the missingness mechanism. There are three types of missingness mechanisms (Rubin, 1976; Little and Rubin, 2020): missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

In the MCAR condition all observations have the same probability of being missing independent of any observed or relevant unobserved data. Consequently, the missingness does not introduce bias to analyses of the data. Hence, such data can in principle be analyzed using only the complete cases. However, this strategy may yield small case numbers if there is a relevant share of missing data. Thus, MCAR may not directly introduce bias but can cause difficulties through the consequential loss of cases for the analysis.

If the missing data are MAR, the missingness Z may depend on any observed data $X|(Z = 0)$ but not on the missing data $X|(Z = 1)$. In this situation, dropping incomplete cases from the analysis may result in biased estimates. Yet, we may still obtain unbiased and approximately efficient estimates through appropriate methods such as MI (Rubin, 1987), which model the missingness mechanism for $\vec{x}_j$ based on the information in the other variables, X_{-j}.

Under MNAR, by contrast, Z depends on the missing data $X|(Z = 1)$ itself or other unobserved parameters even after conditioning on $X|(Z = 0)$. This applies especially if the missing data in a variable j depends on $\vec{x}_j$, i.e., the concerned variable itself. In this situation, conventional imputation procedures relying only on conditioning on $X|(Z = 0)$ may be invalid (van Buuren, 2018). It obviously is not possible to condition on $X|(Z = 1)$ either, since this information is missing.

This can be resolved through specialized MNAR imputation procedures that introduce external information on the selection into $Z = 1$ to the imputation model. For instance, imputed values can be shifted upwards or downwards to match a known distribution (pattern-mixture models; Glynn et al., 1986) or prespecified response weights can be used (selection models; Heckman, 1976; for a detailed discussion of both methods, see Little, 2009). Another approach by Carpenter et al. (2007) proposes weighting the multiple estimates for a parameter produced by MI in order to correct for a MNAR mechanism. However, for social survey data such external information is often not available. Therefore, research practice often relies on more pragmatic approaches. When the MAR assumption is questionable, it is often suggested to use imputation procedures for MAR mechanisms but include as much information predictive of the missingness as possible into the imputation model to reduce bias in estimates (Collins et al., 2001; van Buuren, 2018, p. 165). Consequently, our study assumes that no external information on the missing data is available.

4.2.2 Planned Missing Data (PMD)

Our study supposes that missing data in X stems from two sources: INR by participants and PMD from an SQD.

PMD emerge by intentionally administering only parts of the complete questionnaire to each respondent (in the following described by the PMD indicator matrix Z^{ψ} (identifying planned-missing observations by 1 and not-planned-missing observations by 0) with PMD on a variable j identified by $\vec{z}_j^{\psi} = z_{1,j}^{\psi}, z_{2,j}^{\psi}, ..., z_{i,j}^{\psi}, ..., z_{n,j}^{\psi}$). The SQD, specifically, proceeds by allocating all items to modules. One of these modules may be a so-called core module, which is assigned to all respondents. Of the remaining modules (subsequently called split modules), a subset of two or more modules is assigned randomly to each respondent. In consequence, respondents receive only the items from the modules assigned to them. Due to the random assignment, the PMD are usually MCAR.

SQDs may yield large amounts of missing data for each respondent and on all variables excluding the core. This is because a meaningful reduction in questionnaire length presupposes a large amount of questions remaining unasked: Reducing the number of items presented to each respondent by 50%, for example, requires overall 50% PMD. This also leaves all cases and all split module variables observed incompletely. As a result, analysis strategies relying only on the complete cases may end up with an insufficient number of cases or no cases at all. In consequence,

Raghunathan and Grizzle (1995) propose imputing PMD to obtain analyzable data from SQDs.

4.2.3 Item Nonresponse (INR)

INR in surveys occurs when a sample unit participates in the survey but does not answer a specific item. In the following, we let the INR indicator matrix Z^{ω} (identifying observations missing through INR by 1 and observations not missing through INR by 0) denote data missing through INR, with $\vec{z}_j^{\omega} = z_{1,j}^{\omega}, z_{2,j}^{\omega}, ..., z_{i,j}^{\omega}, ..., z_{n,j}^{\omega}$ identifying the INR on a variable j. In presence of PMD, $z_{i,j}^{\omega}$ is defined only if $z_{i,j}^{\psi} = 0$, leaving $z_{i,j}^{\omega}$ missing whenever $z_{i,j}^{\psi} = 1$.

There can be various causes for INR: Respondents may not understand the question, not know or be sure about the correct response, lack motivation to form an opinion, forget to respond, refuse to answer a sensitive question, or their response may get lost due to an error during data collection or processing (Bech and Kristensen, 2009; Berinsky, 2008; de Leuw et al., 2003; Montagni et al., 2019; Shoemaker et al., 2002). Correspondingly, various missingness mechanisms generating INR are worth considering.

MCAR is a particularly strong assumption which may be realistic for INR only in specific exceptions. For example, data losses at coding or data processing could result in MCAR. Usually, however, social survey research considers the MCAR assumption untenable (de Leuw et al., 2003; Durrant, 2009).

MAR often appears as a more realistic assumption since it allows the INR to depend on respondent characteristics: INR generally occurs more often among respondents that are older (Bech and Kristensen, 2009; Blumenberg et al., 2018; Callens and Loosveldt, 2018; Eliott et al., 2005; Klein et al., 2011; Meitinger and Johnson, 2020; Messer et al., 2012), less educated (Blumenberg et al., 2018; Callens and Loosveldt, 2018; Meitinger and Johnson, 2020; Messer et al., 2012), belong to a (particularly ethnic) minority group (Eliott et al., 2005; Klein et al., 2011; Meitinger and Johnson, 2020) or are not that interested in the survey topic (Callens and Loosveldt, 2018; Kmetty and Stefkovics, 2021). INR rates can also differ considerably between geographic regions (Callens and Loosveldt, 2018; Bech and Kristensen, 2009). Yet, the role of respondent characteristics for INR often varies between different questions and surveys: Some surveys report higher INR among women than men (Bech and Kristensen, 2009; Callens and Loosveldt, 2018; Eliott et al., 2005; Klein et al., 2011; Meitinger and Johnson, 2020; Washington Community Survey, see Messer et al., 2012). Some surveys show a negative association

between income and INR (Klein et al., 2011), while others show no clear association, especially in online surveys (Messer et al., 2012). There may be additional, potentially unknown variables that are associated with the INR on a variable. However, the MAR assumption is reasonable only if one is confident that all variables relevant for the nonresponse mechanism are available in the observed data.

INR can also result from a MNAR mechanism. This may occur when respondents deem their potential answer sensitive or socially undesirable (Copas and Farewell, 1998; de Leuw et al., 2003; Rässler and Riphahn, 2006; Tourangeau and Yan, 2007). For example, respondents with high income tend to refuse reporting their income (see for example Rässler and Riphahn, 2006; Yan et al., 2010).

4.2.4 Imputation

Imputation in general refers to the approach of replacing missing values $X|(Z = 1)$ with non-missing values from an imputation model. This allows for applying standard complete data analysis methods on the completed data.

MI (Rubin, 1987; van Buuren, 2018) is one of the current state of the art procedures for imputation. It aims to both preserve relations in the data and ensure variability. To impute univariate missing data in a variable j using MI, for each missing value a number of m (multiple) imputations are drawn based on an imputation model. This imputation model estimates the distribution of $\vec{x}_j$ conditional on other variables in X_{-j} using a pre-specified imputation method. Drawing m imputations from the conditional distribution yields m imputed datasets and m varying imputed values for each missing value. These multiple datasets are then analyzed separately and the resulting estimates pooled into combined estimates according to Rubin's rules (Rubin, 1987; see also van Buuren, 2018, pp. 145-147).

For multivariate missing data, a common solution is MI by fully conditional specification (FCS; van Buuren et al., 2006). This approach relies on looping through different imputation models that impute missing data in each variable separately. For each variable to be imputed, this involves specifying an imputation method and the relevant predictor variables.

The general procedure of FCS is as follows: We initially replace all missing values by starting values randomly drawn from the marginal distributions of the variables to be imputed. Then we impute the first variable, $\vec{x}_1$, based on the observed data and initial starting values of the predictor variables, replacing the initial starting values in $\vec{x}_1$ by the new imputed ones. We proceed by imputing $\vec{x}_2$ using the observed and imputed values in $\vec{x}_1$ (provided that $\vec{x}_1$ is in the predictor set) and

the observed and initial starting values in the remaining predictor variables, replacing the initial starting values in $\vec{x}_2$ by new imputed ones. This continues until all variables in X are imputed. Subsequently, we repeat this procedure with the previously imputed values instead of the initial starting values: Again, we begin with imputing $\vec{x}_1$, $\vec{x}_2$ up to $\vec{x}_k$ and steadily replace the old imputations by new ones. This looping procedure is repeated for a small (prespecified) number of iterations for convergence, after which the final imputations are drawn. To create m multiple imputations, this entire procedure is repeated m times.

When both PMD and INR appear in a survey, the imputation task might be affected adversely. As described above, SQD surveys tend to generate PMD already at large scale. In practice, this could lead to enormous proportions of missing data in total, since the amount of INR is not under the researchers' deliberate control. This is important because it means the imputation model may need to rely on little observed data. Especially for the imputation of INR this is far from ideal since we would prefer to have as much information on the missing data and its mechanism as possible. Furthermore, more missing data also means a larger impact of imputed values on the estimation, suggesting greater potential for bias from a poor imputation model.

As noted above, an additional challenge may be that PMD and INR may stem from different missingness mechanisms (MCAR and potentially not MCAR). In this context, one might want to account for the different nature of both types of missingness. This would mean to impute a variable j conditional on $\vec{z}_j^{\,\omega}$ or $\vec{z}_j^{\,\psi}$ (for example, by imputing both types of missingness separately). However, in our view this is not meaningful. First, separate imputation models for INR and PMD would likely have to rely on the same observed data $X|(Z = 0)$ that neither experienced planned missingness nor INR, as we only have observations on $x_{i,j}$ when $z_{i,j}^{\omega} = z_{i,j}^{\psi} = 0$. Moreover, being affected by INR ($\vec{z}_j^{\,\omega} = 1$), the remaining available data $\vec{x}_j|(\vec{z}_j = 0)$ may not be subject to a randomness comparable to the PMD anymore without conditioning on the variables determining the INR. Thus, an attempt to impute $\vec{x}_j|(\vec{z}_j^{\,\psi} = 1)$ separate from $\vec{x}_j|(\vec{z}_j^{\,\omega} = 1)$ cannot legitimately be considered MCAR. Finally, even if these challenges were overcome, imputation models conditioning on $\vec{z}_j^{\,\omega}$ or $\vec{z}_j^{\,\psi}$ would likely imply considerably more model parameters to be estimated or (in case of separate models) considerably smaller case numbers. This might be difficult considering we have limited case numbers available but potentially many predictor variables to consider. Therefore, for each variable to be

imputed we build one imputation model imputing all missing values together based on Z without conditioning on Z^{ψ} or Z^{ω}.

Thus, we may face a complex missing data problem with (a) potentially very large proportions of missing data and (b) a potentially complex, heterogeneous missingness mechanism. This complicated missingness mechanism needs to be represented in one single imputation model per variable. This model needs to include all variables predicting the INR despite a potential lack of available cases to support such an extensive model. Thus, the question is how well the imputation can reproduce relevant data structures in spite of these challenges.

4.3 Data and Methods

To examine the impact of INR and PMD on estimates after imputation, we apply a Monte Carlo (MC) simulation study using real social survey data.[2] To allow for a realistic simulation of INR, we first investigate how frequently INR occurs and identify its determinants in the survey dataset that subsequently serves as population data for the simulation study. In each simulation run we draw a random sample from this population dataset and use the information from the preliminary analysis to simulate INR using a procedure similar to Enderle et al. (2013). We also simulate PMD from an SQD with random modules (see Chapter 2). Thus, each simulation run involves stochastically generating both PMD and INR. Through this repeated procedure, we can measure robustly to what extent estimates from our data would be MC biased depending on different PMD and INR scenarios.

4.3.1 Data

The population dataset for this study stems from the German Internet Panel (GIP; Blom et al., 2015; Cornesse et al., 2022), an online panel survey of the German general population. We use items from waves 37 and 38 (Blom et al., 2019a,b) primarily on sociodemographic characteristics, political opinions, organization mem-

[2] All analyses in this paper are carried out in R (R Core Team, 2021) using the following packages (if not cited elsewhere): DescTools (Signorell et al., 2020), doMPI (Weston, 2017), dplyr (Wickham et al., 2021), foreach (Microsoft and Weston, 2020), ggplot2 (Wickham, 2016), glmnet (Friedman et al., 2010), gridExtra (Auguie, 2017), haven (Wickham and Miller, 2019), MASS (Venables and Ripley, 2002), and Rmpi (Yu , 2002). The R code is available as supplementary material to this paper for replication purposes.

bership and the Big Five personality traits. Thereby, we obtain a dataset with 61 variables (see Chapters 2 and 3) that are all categorical, mostly ordinal or binary.[3]

In the MC study, all missing data need to be simulated stochastically. Thus, we need an initially fully observed dataset. To this end, we exclude all unit nonrespondents from the data, reducing the number of cases to 4,061. Furthermore, we complement some further missing values with data from waves 1 and 13 (Blom et al., 2016a,b). Finally, we single-impute the remaining INR using predictive mean matching.

Beyond that, we combine rare events in variables (i.e., categories with < 100 cases) into broader categories. This is necessary because observed case numbers in each simulation run are up to 6.3 times smaller than in the population.

4.3.2 MC Simulation Procedure

In this study, for each parameter specification, the following tasks are repeated over 1,007 simulation runs:

1. draw a simple random sample of 2,000 respondents from the GIP population data
2. simulate PMD, Z^{ψ}
3. simulate missing data by INR, Z^{ω}
4. complete all the missing data using MI
5. obtain estimates with the completed (imputed and observed) data

Using this procedure, we manipulate (a) the proportion of PMD, (b) the proportion of INR, and (c) the missingness mechanism of the INR. The following paragraphs expand on steps (2) through (5) of the simulation procedure.

Simulating PMD

We simulate PMD according to an SQD. For doing so, all items are allocated to modules. 11 sociodemographic items constitute a core module, which is assigned to all respondents. In each simulation run, the remaining 50 items are randomly distributed to five split modules of each 10 items. Each respondent receives a random

[3]This is the same dataset as used in the previous two chapters.

subset of these five split modules. Accordingly, all the PMD are MCAR. We manipulate the proportion of PMD by varying how many split modules are assigned to each respondent: either two, three, four, or all five split modules. This results in either 60%, 40%, 20%, or no PMD in the split modules, while the core module remains completely observed.

Simulating INR

We simulate INR based on the real INR in the GIP. A preliminary analysis shows that overall, 5% of the GIP data are missing due to INR (excluding the sociodemographic items, which are almost completely observed). INR propensities vary heavily by item, ranging from 1% to 19%. Furthermore, to determine how INR propensities vary by survey participant, we estimate elastic net logistic regression models (Zou and Hastie, 2005) of the variables' INR indicators on all other variables in the dataset. This provides us estimated nonresponse propensities specific for each observation in the population data. More detailed information on the preliminary analysis can be found in Appendix A.

These nonresponse propensities are used for simulating INR: We draw values from a uniform distribution $U(0;1)$ and set a value missing if its nonresponse propensity is larger than the value drawn from $U(0;1)$ (see Enderle et al., 2013).

We implement four scenarios with different proportions of INR: one with INR approximately as frequent as in the GIP (overall proportion of INR in the split modules: 5%), and three with INR two times (10%), three times (15%), or four times (20%) as frequent as in the GIP. The sociodemographic core module and further six variables in the split modules remain completely observed, as they show no noteworthy INR. As in the real data, the proportions of INR vary considerably by variable with a minimum of 0% and a maximum of 19% (considering the scenario with overall 5% INR).

Hence, the total proportion of missing data in the simulation study depends on the combination of INR and PMD. To illustrate this, Table 4.1 depicts the combined overall proportion of missing data from both simulation steps for the various scenarios. Accordingly, our simulation scenarios cover overall proportions of missing data ranging from 0% to 68%. This table again highlights why INR and PMD cannot clearly be separated in the imputation: 60% PMD and 20% INR, for instance, do not result in 80% but 68% missing data. Hence, there is a 12% overlap of observations that would be missing both by design and nonresponse.

Table 4.1: Overall proportion of missing data in split modules by simulation scenario.

	Proportion of PMD				
		0%	20%	40%	60%
Proportion of INR	0%	0%	20%	40%	60%
	5%	5%	24%	43%	62%
	10%	10%	28%	46%	64%
	15%	15%	32%	49%	66%
	20%	20%	36%	52%	68%

We also implement different potential nonresponse mechanisms (MCAR, MAR, and MNAR) through adapting the nonresponse propensities.

Under MCAR, INR occurs purely by random chance. Thus, each variable j has nonresponse propensities equal to the proportion of INR on variable j (not varying between respondents). For larger proportions of INR, the propensities are multiplied by 2, 3, or 4. In principle, this procedure can lead to nonresponse propensities larger than 1. However, since all variables in the GIP dataset have proportions of INR smaller than 25%, this is not the case here.

Under MAR, the nonresponse in a variable j depends on data in other variables in X_{-j}. Thus, for a MAR scenario with INR as frequent as in the GIP, we use the nonresponse propensities estimated in the preliminary analysis using logistic regression models . For the scenarios with more INR, we manipulate the intercepts of these models increasing them such that the resulting propensities turn out two, three, or four times larger on average.

Yet, the MAR mechanism in our data might be too modest to differ substantially from an MCAR scenario. This is why we also consider an amplified MAR mechanism. In these scenarios, we multiply the regression coefficients of the logistic models by 2 and subsequently adjust the intercepts such that the proportion of INR on each variable remains the same as in the GIP-like MAR scenarios. Under MNAR, we assume that INR on variable j depends only on variable j itself. For this, we set up the following MNAR model

$$p(z_{i,j}^{\omega} = 1) = \frac{e^{\gamma_0^j + \gamma_1^j x_{i,j}}}{1 + e^{\gamma_0^j + \gamma_1^j x_{i,j}}} \tag{4.1}$$

where γ_0^j is the intercept and γ_1^j is the regression coefficient of $\vec{x}_j$ determining the INR in a variable j. In doing so, γ_0^j and γ_1^j are specified such that the mean and the standard deviation of the nonresponse propensities are approximately the same as

under the (GIP-like) MAR scenario. For the scenarios with more INR, the intercept γ_0^j is adjusted as described in the MAR scenario.

In consequence, we end up with $4 + 4 \times 4 \times 4 = 68$ simulation scenarios:

- four scenarios with varying prevalence of PMD (0%, 20%, 40%, 60% PMD) without INR, plus
- four scenarios with varying prevalence of INR (5%, 10%, 15%, 20%), times
- four missingness mechanisms for INR (MCAR, GIP-like MAR, amplified MAR, MNAR), times
- four scenarios with varying prevalence of PMD (0%, 20%, 40%, 60%).

Imputation

The missing data are imputed using the mice and miceadds packages (van Buuren and Groothuis-Oudshoorn, 2011; Robitzsch and Grund, 2021) with 20 imputations drawn after 10 iterations. In doing so, we use predictive mean matching with dimensionality reduction of the predictor space through a partial least squares regression (Robitzsch et al., 2016). We opt for this method because it can deal with a sample size of 2,000 without dropping some of the many potentially relevant predictor variables from imputation models. Correspondingly, this approach has shown to perform comparatively well with the data at hand compared to alternative techniques, such as logistic regression models and classification and regression trees (see Chapter 3).

Estimation

To examine the imputation's ability to preserve distributions and relations in the data with the various scenarios, in each simulation run and for each scenario we calculate two types of MI estimates:

- Univariate frequencies for all 285 categories of all 44 variables with INR
- Bivariate Spearman correlations between all 88 pairs of variables that have a correlation of 0.2 or stronger in the original population data and feature INR on at least one of the two variables.

For this purpose, these measures are calculated separately in each of the 20 imputed datasets and subsequently pooled according to Rubin's rules.

In order to evaluate the accuracy of a frequency or correlation estimate, we calculate its percentage MC bias. This entails the following operation:

$$\%Bias^{MC}(\hat{\theta}) = 100 \times \frac{1}{S} \sum_{s=1}^{S} (\hat{\theta}_s - \theta) / / \theta \tag{4.2}$$

where s refers to one of $1, 2, \ldots, S$ simulation runs, $\hat{\theta}_s$ is a pooled MI estimate in simulation run s, and θ is the true population benchmark for this estimate.

4.4 Results

4.4.1 Univariate Frequencies

Figure 4.1 displays the percentage MC biases averaged over all simulation runs for each univariate frequency estimate (displayed on the horizontal axis) under the different INR and PMD scenarios. Each of the displayed data points refers to the average bias of one specific category of a variable. To simplify the analysis, boxplots are drawn over the average biases. For each mechanism, Figure 4.1 shows several of these plots referring to the percentage biases obtained with different proportions of INR and PMD. In addition, the exact numbers for the percentage biases discussed below are displayed in an appendix (Table 4.B1).

Note that, mathematically, all percentage biases for univariate frequencies have a lower limit at -100% (because frequencies cannot be negative) but upper limits often exceeding $+100\%$, depending on the size of the frequency ($1/\theta - 1$). Thus, the phenomenon that Figure 4.1 tends to depict more pronounced percentage biases in the positive than in the negative results from their calculation and represents no finding in itself.

The first boxplot in Figure 4.1 depicts percentage MC biases when no missing data at all occurs (and consequently, no data are imputed). Correspondingly, all biases are approximately zero. The following three boxplots show the percentage MC biases for 20%, 40%, and 60% PMD (still without INR). We can observe biases increasing with increasing shares of PMD, even without INR: The central 50% of biases (that is, 25% of biases are smaller and another 25% are larger) range from -0.1% to $+0.3\%$ with 20% PMD, from -0.7% to $+1.6\%$ with 40% PMD, and from -1.4% to $+4.5\%$ with 60% PMD.

The plots beneath show the results for 5%, 10%, 15%, and 20% INR that is MCAR, again separately for 0%, 20%, 40%, and 60% PMD. Each of these INR sce-

narios replicates the finding that percentage MC biases increase with more PMD. Similarly, it also shows that biases increase with the proportion of INR despite the MCAR mechanism. With 60% PMD, for example, the central 50% of biases range from -1.4% to $+4.5\%$ when there is no INR, from -1.4% to $+5.3\%$ with 5% INR, from -1.6% to $+6.2\%$ with 10% INR, from -1.7% to $+7.6\%$ with 15% INR, and from -2.1% to $+8.9\%$ with 20% INR. In comparison to the scenarios without INR, we also observe that percentage biases for a few categories take extreme values. This is because the prevalence of INR varies heavily between variables. For example, in the most extreme scenario (60% PMD and 20% INR), three extreme outliers with percentage biases of each more than 70% stand out. These refer to categories at the tails of the variables CE38256 and CE38260, which have the highest proportions of INR (in the scenario with 20% INR and 60% PMD 68% of cases are unobserved).

The subsequent plots show the results for INR that is MAR and as frequent as in the GIP or according to the amplified mechanism. The general patterns observed before recur in both scenarios: Percentage MC biases increase with larger proportions of both PMD and INR. Yet, INR appears to cause somewhat larger biases under MAR than under MCAR, especially with the amplified MAR mechanism. In the 20% INR scenario with no PMD, for instance, the central 50% of biases range from -1.2% to $+1.1\%$ for the GIP-like MAR mechanism and from -1.7% to $+1.6\%$ for the amplified MAR mechanism, as opposed to -0.3% to $+0.3\%$ for the MCAR mechanism. Interestingly, the presence of PMD (although in general yielding increased biases) seems to attenuate the effect of the INR mechanism to some extent: In the most extreme scenario with 60% PMD and 20% INR, the central 50% of percentage biases range from -2.1% to $+8.9\%$ under MCAR, from -2.4% to $+8.7\%$ under GIP-like MAR, and from -2.7% to $+8.9\%$ under amplified MAR. Hence, under 60% PMD and 20% INR the amplified MAR mechanism increases the range of the central 50% of biases by only 0.6 percentage points compared to MCAR, as opposed to 2.7 percentage points under 0% PMD and 20% INR.

For INR that is MNAR (displayed in the bottom of Figure 4.1), we observe a different pattern. The percentage MC biases generally are much larger than under MCAR or MAR (note that the scale of the horizontal axis for MNAR differs from the rest because otherwise, many biases would fall out of display range). For example, the central 50% of biases with 40% PMD and 20% INR range from -26.0% to $+14.1\%$ under MNAR, as opposed to -2.8% to $+5.2\%$ under amplified MAR, -1.9% to $+5.0\%$ under GIP-like MAR, and -1.3% to $+4.3\%$ under MCAR. In

consequence, we observe some extreme cases with larger proportions of INR (15% and 20%), with some frequencies being biased upwards by more than $\pm 100\%$. This indicates that some categories of variables are not observed at all throughout the simulation due to the MNAR mechanism.

Due to the large effect of the INR under MNAR, the proportion of PMD affects the accuracy of estimates less than under the other mechanisms. With 10% INR, for example, the central 50% of percentage MC biases range from -9.6% to $+6.1\%$ when there is no PMD, from -10.4% to $+6.3\%$ with 20% PMD, from -9.6% to $+6.5\%$ with 40% PMD, and from -8.4% to $+9.4\%$ with 60% PMD.

4.4.2 Bivariate Correlations

Figure 4.2 shows the results for the average percentage MC biases of bivariate Spearman correlations that are larger than 0.2 in the population data. In doing so, it follows the same structure as Figure 4.1. Here, each data point refers to the Monte Carlo bias of the correlation of one variable pair. Unlike Figure 4.1, Figure 4.2 also covers values below -100%, as correlations can be both positive and negative. Again, exact numbers for the percentage biases are also displayed in the appendix (Table 4.B2).

As for the univariate frequencies, we can observe percentage MC biases increase with increasing proportions of PMD, with a clear tendency towards underestimating relationships between variables. This effect is especially severe for the scenario with the highest share of PMD: Considering the scenarios without INR, the central 50% of biases range from -1.7% to $+0.0\%$ with 20% PMD, from -5.0% to $+0.3\%$ with 40% PMD, and from -18.2 to -11.5% with 60% PMD. Thus, the results are slightly different for frequencies and correlations: Given large proportions of PMD, almost all correlations are considerably biased downwards, while at least some frequencies still have percentage biases close to zero (see Figure 4.1).

Again, increasing proportions of INR also yield increasing percentage MC biases, even under MCAR. For each of the INR mechanisms, the largest biases emerge when the proportions of both PMD and INR is high. For example, with 60% PMD and 20% INR that is MCAR, the central 50% of biases range from -48.0% to -35.2%, as opposed to from -18.2 to -11.5% with 60% PMD but no INR. This means that biases are roughly doubled in size despite the total proportion of missing data increases only from 60% to 68% (see Table 4.1).

We also observe a slight tendency towards increasing percentage MC biases when the INR is MAR (GIP-like or amplified, respectively) as compared to MCAR.

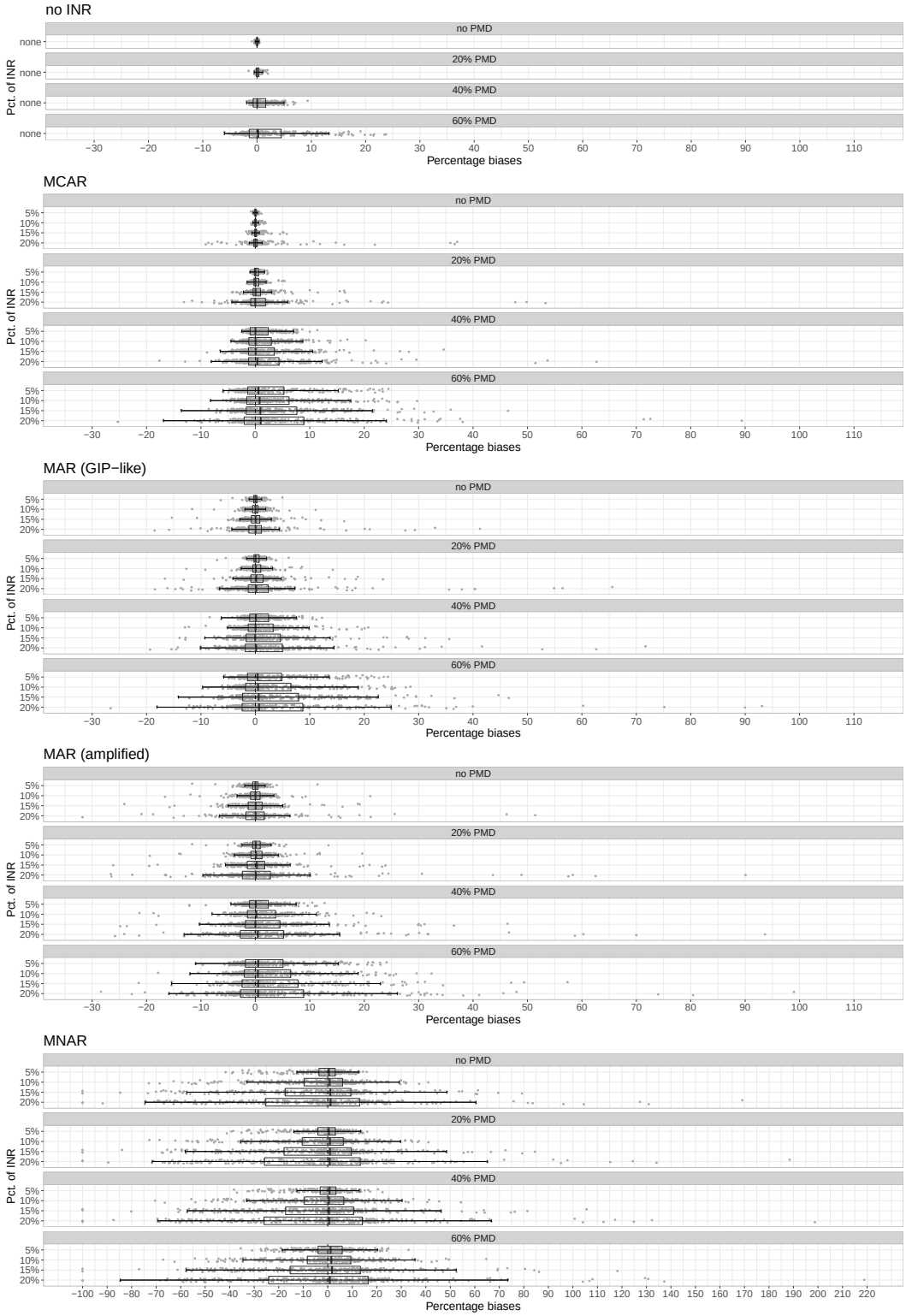

Figure 4.1: Average percentage Monte Carlo biases of univariate frequency estimates for 285 categories of 44 variables, by response mechanism and proportions of item nonresponse and planned missing data.

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases each.

However, the differences are much less pronounced as with the frequency estimates. With 20% INR and 40% PMD, for example, the central 50% of biases range from -21.4% to -8.1% under amplified MAR and from -22.5% to -8.4% under the GIP-like MAR, as opposed to -22.2% to -7.7% under MCAR.

For INR that is MNAR, we again observe some tendency towards larger percentage MC biases compared to both the MCAR and MAR scenarios. With 20% INR and no PMD, for example, the central 50% of biases range from -6.3% to $+1.5\%$ under MNAR, as opposed to -6.1% to $+0.1\%$ under amplified MAR, -4.8% to $+0.1\%$ under GIP-like MAR, and -4.0% to -0.3% under MCAR. However, this effect is less pronounced and less clear than with the frequency estimates. There also tend to be more MC biases in the area around zero than under the MAR scenarios. This suggests that in this simulation study, MNAR affects some correlations considerably while leaving others largely intact. Apart from that, we again observe some extreme biases exceeding -100% with 15% or 20% INR that is MNAR, implying that the direction of these relationships reverses systematically due to the INR. These extreme biases occur primarily in the correlation of the variables BG38001 and BG38002, which have the strongest variability in nonresponse propensities throughout all variables due to their good nonresponse model fit in the preliminary analysis.

Compared to the results on univariate frequencies, the proportion of PMD exhibits a larger effect on the accuracy of correlations under MNAR. With 20% INR that is MNAR, for example, the central 50% of percentage MC biases range from -6.3% to $+1.6\%$ when there is no PMD, from -11.1% to -1.14% with 20% PMD, from -30.6% to -5.6% with 40% PMD, and from -54.8% to -32.7% with 60% PMD.

4.5 Summary

In this paper, we have examined the accuracy of univariate frequency and bivariate Spearman correlation estimates after imputation in data with two sources of missing data: planned missingness from an SQD and INR by survey participants. In doing so, we have manipulated both the proportions of PMD and INR as well as the mechanism causing the INR. Several major findings stand out:

First, the combined presence of INR and PMD in a social survey can affect the estimation adversely. A major reason for this is that both types of missing data combined increase the total proportion of missing data, challenging the imputation:

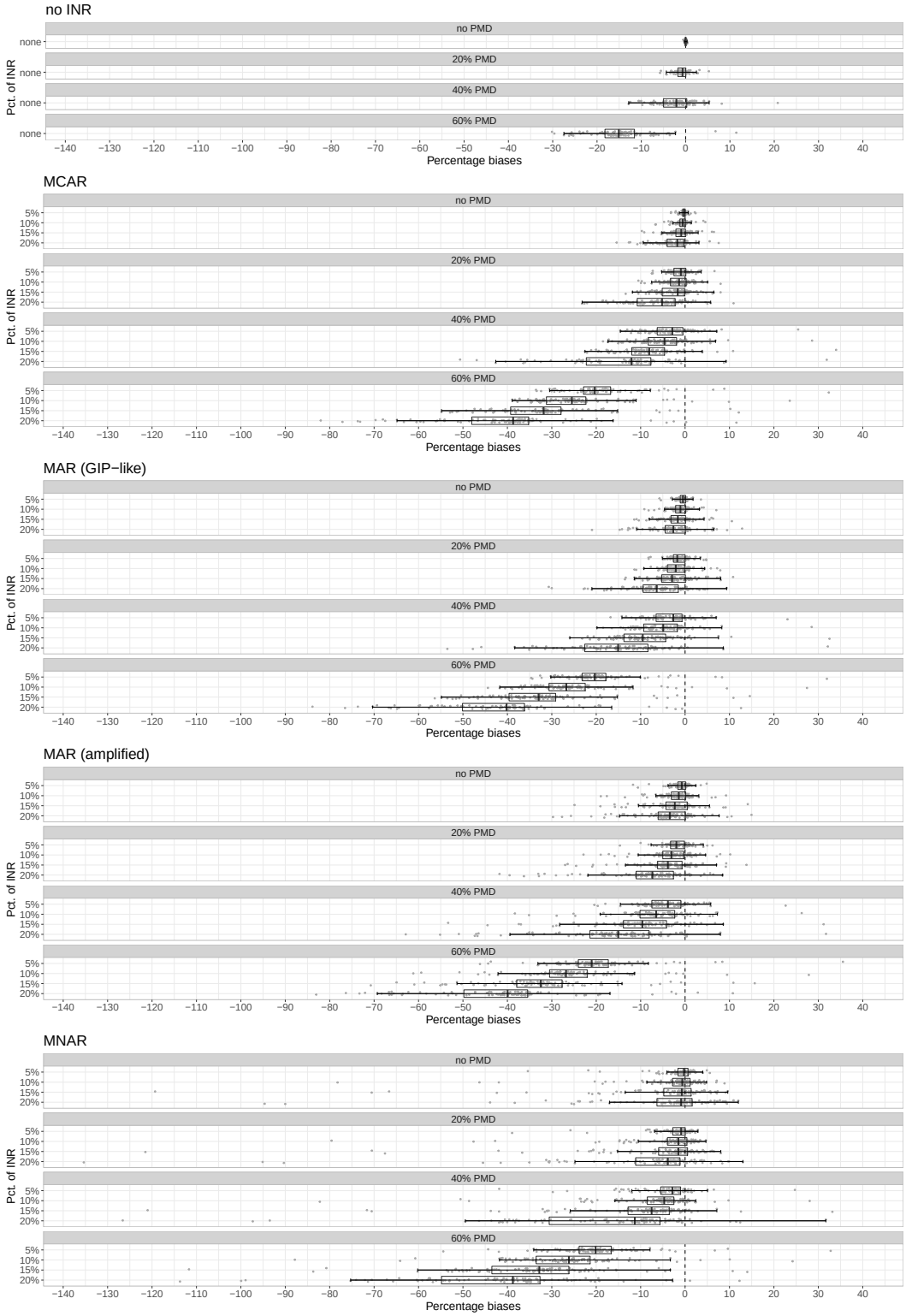

Figure 4.2: Average percentage Monte Carlo biases of bivariate Spearman correlation estimates for 88 variable pairs correlated by 0.2 or more in the population data, by response mechanism and proportions of item nonresponse and planned missing data.

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases each.

In our simulation study, large proportions of missing data led to large Monte Carlo biases even if the INR is MCAR. In particular, rampant increases in Monte Carlo biases emerged when the combined proportion of missing data from both sources exceeded about 40%. Perhaps, this is caused by a lack of pairwise observations available for the imputation model under large amounts of missingness: Whereas 40% PMD in two variables in different split modules would mean 36% of cases being pairwise observed (given 0% nonresponse), 60% PMD would result in only 16% pairwise observed cases. Under all examined nonresponse mechanisms, many frequency estimates (yet not necessarily all of them) turn out considerably overestimated or underestimated when the proportion of missing data is high. Meanwhile, correlation estimates appear especially severely affected by large amounts of missing data, being almost consistently shifted downwards with only few exceptions having Monte Carlo biases close to zero.

Second, under the conditions of our simulation study, MAR caused only slightly larger Monte Carlo biases than MCAR. The effects of INR under MCAR and MAR even tended to converge the more PMD was introduced. Thus, in our simulation study differences between MCAR and MAR appear only as a minor factor affecting the quality of MI estimates, especially compared to the overall proportion of missing data.

Third, under MNAR we observe different effects. In our simulation study, univariate frequency estimates under MNAR were affected much more by the proportion of INR than by the overall proportion of missing data. Thus, the amount of PMD had hardly an effect on univariate frequency estimates. This is presumably because the imputation could not deal adequately with this nonresponse mechanism. For correlations, though, the effect of MNAR over MAR and MCAR was rather small, and the overall amount of missing data also had a considerable impact on the quality of estimates. We could imagine that in the real world this may especially depend on the specific data context, considering that real-world MNAR mechanisms might sometimes affect correlations more directly than in this simulation study. Yet, despite the result that both MNAR nonresponse and large amounts of PMD may cause estimation problems, the combination of both effects seems not to create any further damage beyond (at worst) adding up.

Fourth, in all scenarios the estimates for a few categories or correlations were affected substantially more by INR than most others. These outliers appear because, as our preliminary analysis of real INR in the GIP data showed, INR varies heavily between items both in its prevalence and dependence on other variables in the data.

4.6 Discussion

This study has certain limitations but may also allow some important conclusions for future research. Both aspects deserve broader discussion here.

The most important limitation is that the study's findings rely on a simulation based on specific social survey data. Therefore, their external validity may depend on how similar real data-collection scenarios would be to our simulation setup. Through relying on real social survey data and the INR observed in this dataset, we attempted to create a realistic environment. Nevertheless, INR in the real world could work differently. For example, unlike modeled in this study, INR could follow non-linear mechanisms (see for example Collins et al., 2001). Furthermore, INR was modeled separately for each item based on the other variables in the dataset. In reality, though, INR could also depend on the combination of several variables. This might explain why effects appeared weaker for correlations than for univariate frequencies. All these issues might affect real world applications of imputation, potentially making it harder to model missingness mechanisms accurately than in this study.

Furthermore, the variables in our dataset were discrete. In continuous variables, by contrast, single outliers could have considerable leverage on correlation estimates. Therefore, MNAR mechanisms in continuous variables might potentially affect correlation estimates more severely than found in this study. Moreover, we treated INR as a single uniform missing data source. Yet, in real surveys there are different subtypes of INR (e.g., refusals, data collection errors, etc.) that might behave differently regarding their response mechanism (see, for example, Shoemaker et al., 2002). Apart from all that, response mechanisms could also behave differently in surveys on different substantive topics. Therefore, this study should be replicated with different data in the future.

Furthermore, our study focuses on INR as one of several manifestations of missing data that commonly occur in social surveys. Other important sources of missing data, such as unit nonresponse, were out of scope. However, we encourage future research on how these other missing data sources in surveys interact with the imputation of PMD.

A final limitation is that we examined the accuracy of univariate and bivariate but not multivariate estimates. Yet, for substantive researchers the performance of multivariate models under different planned missingness scenarios may also be highly relevant. Thus, future research should address this issue as well.

Our findings may also guide future research in several other ways. First of all, they allow some direct conclusions for survey design. In particular, survey designers are recommended to carefully evaluate how much PMD is necessary and not introduce more than that, considering that the quality of estimates tends to plummet when the proportion of missing data becomes too large. This is especially the case for items that can be expected to produce considerable amounts of INR. In such items, to allow for an appropriate imputation one may consider reducing the proportion of PMD or allocating them to the core module.

Similarly, it seems particularly important in SQD surveys to keep INR at a low level. For example, this is especially relevant considering the way modules are constructed. For instance, earlier research shows that items of one topic should be allocated to different split questionnaire forms rather than all to the same in order to support the imputation (see Chapter 2 of this dissertation; Imbriano and Raghunathan, 2020; Raghunathan and Grizzle, 1995). It is still an open empirical question how (and if so, when) this would affect INR rates or response quality in general compared to procedures allocating items of one topic to the same questionnaire form. Therefore, future research should investigate this issue, such that INR can be taken into account when designing split questionnaires.

Interactions between SQDs and the participants' response behavior may also play a role in evaluating the costs and benefits of an SQD for a specific survey. By reducing respondent burden in terms of questionnaire length, SQDs are supposed to decrease unit nonresponse, breakoff, and measurement error (Galesic and Bosnjak, 2009; Peytchev and Peytcheva, 2017) at the cost of additional planned missingness (Graham et al., 1996; Raghunathan and Grizzle, 1995; Peytchev and Peytcheva, 2017). This notion highlights key empirical questions for survey researchers considering to implement an SQD in a survey: How much PMD is needed to obviate a given amount of unit nonresponse, breakoff, or measurement error? Is the averted nonresponse considered MNAR, or is it MCAR or MAR? For example, on the one hand, if introducing a limited amount of PMD can prevent a considerable amount of unit nonresponse that is MNAR, the benefits of the SQD may outweigh its costs. On the other hand, if large amounts of PMD can inhibit relatively little nonresponse that can also be expected to be MAR, the opposite may be the case. To allow reasonable claims about the expectable usefulness of an SQD for a specific survey, however, our study would need to be replicated with a broad variety of different survey datasets first. Furthermore, experimental research would be needed to investigate if and how different strategies to design split questionnaires affect response

behavior. First evidence on this domain shows differences in respondents' evaluation of split questionnaires with more versus less frequent switches between topics (Adigüzel and Wedel, 2008). Despite the need for more research, this simulation study may provide a first piece of evidence to help researchers assess to what extent an SQD might make sense for a given survey.

References

Adigüzel, F., & Wedel, M. (2008). Split questionnaire design for massive surveys. *Journal of Marketing Research*, *45*(5), 608-617.

Auguie, B. (2017). *gridExtra: Miscellaneous functions for 'Grid' graphics*. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Bahrami, S., Aßmann, C., Meinfelder, F., & Rässler, S. (2014). A split questionnaire survey design for data with block structure correlation matrix. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.) *Improving survey methods: Lessons from recent research* (pp. 368-380). Routledge.

Bech, M. & Kristensen, M. B. (2009). Differential response rates in postal and Web-based surveys in older respondents. *Survey Research Methods*, *3*(1), 1-6.

Berinsky, A. E. (2008). Survey non-response. In W. Donsbach & M. W. Traugott (Eds.), *The SAGE handbook of public opinion research* (pp. 309-321). Sage.

Blom, A. G., Bossert, D., Funke, F., Gebhard, F., Holthausen, A., & Krieger, U.; SFB 884 "Political Economy of Reforms" Universität Mannheim (2016a). *German Internet Panel, wave 1 - core study (September 2012)*. GESIS Data Archive, ZA5866 Data file Version 2.0.0. https://doi.org/ 10.4232/1.12607.

Blom, A. G., Bossert, D., Gebhard, F., Funke, F., Holthausen, A., & Krieger, U.; SFB 884 "Political Economy of Reforms" Universität Mannheim (2016b). *German Internet Panel, wave 13 - core study (September 2014)*. GESIS Data Archive, ZA5924 Data file Version 2.0.0. https://doi.org/10.4232/1.12619.

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

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

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. *Field Methods*, *27*(4), 391-408.

Blumenberg, C., Zugna, D., Popovic, M., Pizzi, C., Barrios, A. J. D., & Richiardi, L. (2018). Questionnaire breakoff and item nonresponse in web-based questionnaires: Multilevel analysis of person-level and item design factors in a birth cohort. *Journal of Medical Internet Research*, *20*(12), e11046.

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web Survey Methodology*. Sage.

Callens, M. & Loosveldt, G. (2018). 'Don't Know' responses to survey items on trust in police and criminal courts: A word of caution. *Survey Methods: Insights from the Field*. https://surveyinsights.org/?p=9237

Carpenter, J. R., Kenward, M. G., & White, I.R. (2007). Sensitivity analysis after multiple imputation under missing at random: A weighting approach. *Statistical Methods in Medical Research*, *16*(3), 259-275.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330-351.

Copas, A. J. & Farewell, V. T. (1998). Dealing with non-ignorable non-response by using an 'enthusiasm-to-respond' variable. *Journal of the Royal Statistical Society: Series A*, *161*(3), 385-396.

Cornesse, C., Felderer, B., Fikel, M., Krieger, U., & Blom, A. G. (2021). Recruiting a probability-based online panel via postal mail: Experimental evidence. *Social Science Computer Review*, *40*(5), 1259-1284.

de Leeuw, E. (2008). Self-administered questionnaires and standardized interviews. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *Handbook of social research methods* (pp. 313-327). Sage.

de Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, *19*(2), 153-176.

Durrant, G. B. (2009). Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates. *International Journal of Social Research Methodology*, *12*(4), 293-304.

Elliott, M. N., Edwards, C., Angeles, J., Hambarsoomians, K., & Hays, R. D. (2005). Patterns of unit and item nonresponse in the CAHPS® Hospital Survey. *Health Services Research*, *40*(6p2), 2096-2119.

Enderle, T., Münnich, R., & Bruch, C. (2013). On the impact of response patterns on survey estimates from access panels. *Survey Research Methods*, *7*(2), 91-101.

Frick, J. R. & Grabka, M. M. (2005). Item nonresponse on income questions in panel surveys: Incidence, imputation and the impact on inequality and mobility. *Allgemeines Statistisches Archiv*, *89*(1), 49-61.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1-22.

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 115-142). Springer.

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

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological methods*, *11*(4), 323-343.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*(4), 475-92.

Imbriano, P. M. & Raghunathan, T. E. (2020). Three-form split questionnaire design for panel surveys. *Journal of Official Statistics*, *36*(4), 827-854.

Klein, D. J., Elliott, M. N., Haviland, A. M., Saliba, D., Burkhart, Q., Edwards, C., & Zaslavsky, A. M. (2011). Understanding nonresponse to the 2007 Medicare CAHPS Survey. *The Gerontologist*, *51*(6), 843-855.

Kmetty, Z. & Stefkovics, A. (2021). Assessing the effect of questionnaire design on unit and item-nonresponse: Evidence from an online experiment. *International Journal of Social Research Methodology*, *25*(5), 659-672.

Little, R. J. A. (2009). Selection and pattern-mixture models. In G. M. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 409-431). CRC Press.

Little, R. J. A. & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). Wiley.

Luijkx, R., Jónsdóttir, G. A., Gummer, T., Ernst Stähli, M., Fredriksen, M., Reeskens, T., Ketola, K., Brislinger, E., Christmann, P., Gunnarsson, S. Þ., Hjaltason, Á. B., Joye, D., Lomazzi, V., Maineri, A. M., Milbert, P., Ochsner, M., Ólafsdóttir, S., Pollien, A., Sapin, M., ... Wolf, C. (2021). The European Values Study 2017: On the way to the future using mixed-modes. *European Sociological Review*, *37*(2), 330-346.

Meitinger, K. & Johnson, T. P. (2020). Power, culture and item nonresponse in social surveys. In P. S. Brenner (Ed.), *Understanding survey methodology: Sociological theory and applications* (pp. 169-191). Springer.

Messer, B., Edwards, M., & Dillman, D. (2012). Determinants of item nonresponse to web and mail respondents in three address-based mixed-mode surveys of the general public. *Survey Practice*, *5*(2).

Microsoft & Weston, S. (2020). *foreach: Provides foreach looping construct*. R package version 1.5.0. https://CRAN.R-project.org/package=foreach

Montagni, I., Cariou, T., Tzourio, C., & González-Caballero, J-L. (2019). 'I don't know', 'I'm not sure', 'I don't want to answer': A latent class analysis explaining the informative value of nonresponse options in an online survey on youth health. *International Journal of Social Research Methodology*, *22*(6), 651-667.

Munger, G. F. & Loyd, B. H. (1988). The use of multiple matrix sampling for survey research. *The Journal of Experimental Education*, *56*(4), 187-191.

OECD (2014). *PISA 2012 technical report*. OECD.

Peytchev, A. & Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods*, *11*(4), 361-368.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

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

Rässler, S. & Riphahn, R. T. (2006). Survey item nonresponse and its treatment. *Allgemeines Statistisches Archiv*, *90*(1), 217-232.

Robitzsch, A. & Grund, S. (2021). *miceadds: Some additional multiple imputation functions, especially for 'mice'*. R package version 3.10-28. https://CRAN.R-project.org/package=miceadds

Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der Österreichischen Bildungsstandardüberprüfung [Methodological foundation of standard achievement testing]* (pp. 259-293). facultas.

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

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87-94.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

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

Shoemaker, P. J., Eichholz, M., & Skewes, E. A. (2002). Item nonresponse: Distinguishing between don't know and refuse. *International Journal of Public Opinion Research*, *14*(2), 193-201.

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., . . . Zeileis, A. (2020). *DescTools: Tools for descriptive statistics*. R package version 0.99.36.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the national health and nutrition examination survey. *Survey Methodology*, *32*(2), 217-231.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859-883.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC press.

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049-1064.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer.

Weston, S. (2017). *doMPI: foreach parallel adaptor for the Rmpi package*. R package version 0.2.2. https://CRAN.R-project.org/package=doMPI

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *dplyr: A grammar of data manipulation*. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

Wickham, H. & Miller, E. (2019). *haven: Import and export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.1.1. https://CRAN.R-project.org/package=haven

Yan, T., Curtin, R., & Jans, M. (2010). Trends in income nonresponse over two decades. *Journal of Official Statistics*, *26*(1), 145-164.

Yu, H. (2002). Rmpi: Parallel statistical computing in R. *R News*, *2*(2), 10-14.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.

Appendix

Appendix A. Preliminary Analysis: INR in the GIP Data

In order to attempt a realistic simulation of INR, an initial analysis investigates prevalence and determinants of the original INR in the GIP dataset. It reveals that overall, the GIP dataset has not very much INR. However, this varies considerably by item (see Table A.1 for the detailed numbers): Some items (like the sociodemographic characteristics) have no or miniscule INR. Meanwhile, some other items (especially on political opinions) have considerably more INR up to 19%. When ignoring the sociodemographic characteristics, the average prevalence of INR is 5%.

To investigate the determinants of INR of a variable j in our GIP population dataset, we estimate regression models of the INR indicator variable $\vec{z}_j^w$ of variable j on all other variables X_{-j}. This procedure is applied to outcome variables with more than 25 cases of INR (thereby excluding all sociodemographic variables and six of the remaining variables).

To select the most relevant predictors, we first estimate logistic elastic net regressions (Zou and Hastie, 2005) using the glmnet package (Friedman et al., 2010) with cross-validated λ parameters. We set $\alpha = 1.00$, meaning that we use only the lasso penalty (and no ridge penalty). For some outcome variables the lasso regression failed to converge. In this case, we set $\alpha = 0.95$ (or $\alpha = 0.90$ if it fails again). This means we get closer to a ridge regression, but there remains a predominant lasso penalty for variable selection. Figure 4.A1 shows which variables are selected as predictors for each variable with missing values by INR.

Subsequently, we apply conventional logistic regressions of the nonresponse indicator variable $\vec{z}_j^w$ on the previously selected predictor variables. This serves to obtain models with valid effect sizes predicting INR for all the variables in the data. The median Efron's R^2 of these models is 0.08. Some outcome variables have R^2 values considerably under 0.10, potentially in part because many of these variables have little INR. Larger R^2 values (mostly between 0.10 and 0.15) can be found for some political-opinion items, in which INR is more prevalent. The largest R^2 values refer to two items on the perceived influence of lobbying over EU and German domestic politics (BG38001 with $R^2 = 0.33$ and BG38002 with $R^2 = 0.35$). Efron's R^2 values for all outcome variables can also be found in Figure 4.A1.

Using these logistic models, we calculate nonresponse propensities p for each observation i in variable j:

$$p(z^{\omega}_{i,j} = 1) = \frac{e^{L^j_i}}{1 + e^{L^j_i}} \tag{4.3}$$

with L^j_i being the log-odds of INR for the (i, j)-th element of X according to the logistic regression of the j-th outcome variable ($\vec{z}^{\omega}_j$) on predictor variables in $X_{(} - j)$, estimated with the linear index

$$L^j_i = \beta^j_0 + \beta^j_1 x_{i,1} + \beta^j_2 x_{i,2} + ... + \beta^j_{j-1} x_{i,j-1} + \beta^j_{j+1} x_{i,j+1} + ... + \beta^j_k x_{i,k} \tag{4.4}$$

where β^j_0 is the regression intercept and β^j_1 is the regression coefficient of predictor variable $\vec{x}_1$ (and β^j_2 the regression coefficient of predictor variable $\vec{x}_2$, etc.). In these models, the coefficients for variables excluded by the elastic net regression are set to zero.

Figure 4.A2 plots the predicted nonresponse propensities (on the vertical axis) and the observed response behaviour (black = item nonresponse, gray = response) for four exemplar variables. All four graphs show rather small nonresponse propensities for most respondents, but rather large nonresponse propensities for a few. As can be expected, most of the observations with real INR seem to have rather high nonresponse propensities. The nonresponse propensities show more variation in *BG38001* and *BG38002*, which is most likely because of the relatively good model fit.

Table 4.A1: Prevalence of data missing through item nonresponse in the GIP data.

Topic	Item	Number of missing values	Percentage share of missing values
Sociodemographic & sampling cohort	gender_18	0	0%
	year_of_birth_cat_18	1	0%
	educ_school_18	1	0%
	educ_job_18	0	0%
	marital_status_18	21	1%
	number_hh_members_18	15	0%
	occupation_18	19	0%
	german_citizenship_18	1	0%
	internet_usage_18	12	0%
	state_18	17	0%
	sample	0	0%
Organization membership	AA37027	35	1%
	AA37028	35	1%
	AA37029	35	1%
	AA37030	34	1%
	AA37031	36	1%
	AA37032	35	1%
	AA37033	37	1%
	AA37034	35	1%
	AA37035	37	1%
	AA37036	189	5%
Big 5 personality traits	AAxx044	26	1%
	AAxx045	26	1%
	AAxx046	27	1%
	AAxx047	26	1%
	AAxx048	27	1%
	AAxx049	28	1%
	AAxx050	27	1%
	AAxx051	26	1%
	AAxx052	27	1%
	AAxx053	26	1%
Lobbying in EU politics	BG38001	419	10%
	BG38002	341	8%

	BG38006_a	23	1%
	BG38006_b	23	1%
	BG38006_c	23	1%
	BG38006_d	23	1%
	BG38006_e	23	1%
	BG38006_f	23	1%
	BG38007	529	13%
	BG38009	657	16%
Domestic and	CE38153	235	6%
party politics	CE38154	292	7%
	CE38155	349	9%
	CE38156	694	17%
	CE38312	489	12%
	CE38347	218	5%
	CE38348	221	5%
	CE38350	219	5%
	CE38351	245	6%
	CE38352	198	5%
	CE38056	148	4%
	CE38250	223	5%
	CE38252	288	7%
	CE38254	320	8%
	CE38256	764	19%
	CE38258	416	10%
	CE38260	782	19%
	CE38262	570	14%
	CE38280	347	9%
	CE38329	205	5%
Total		10168	4%

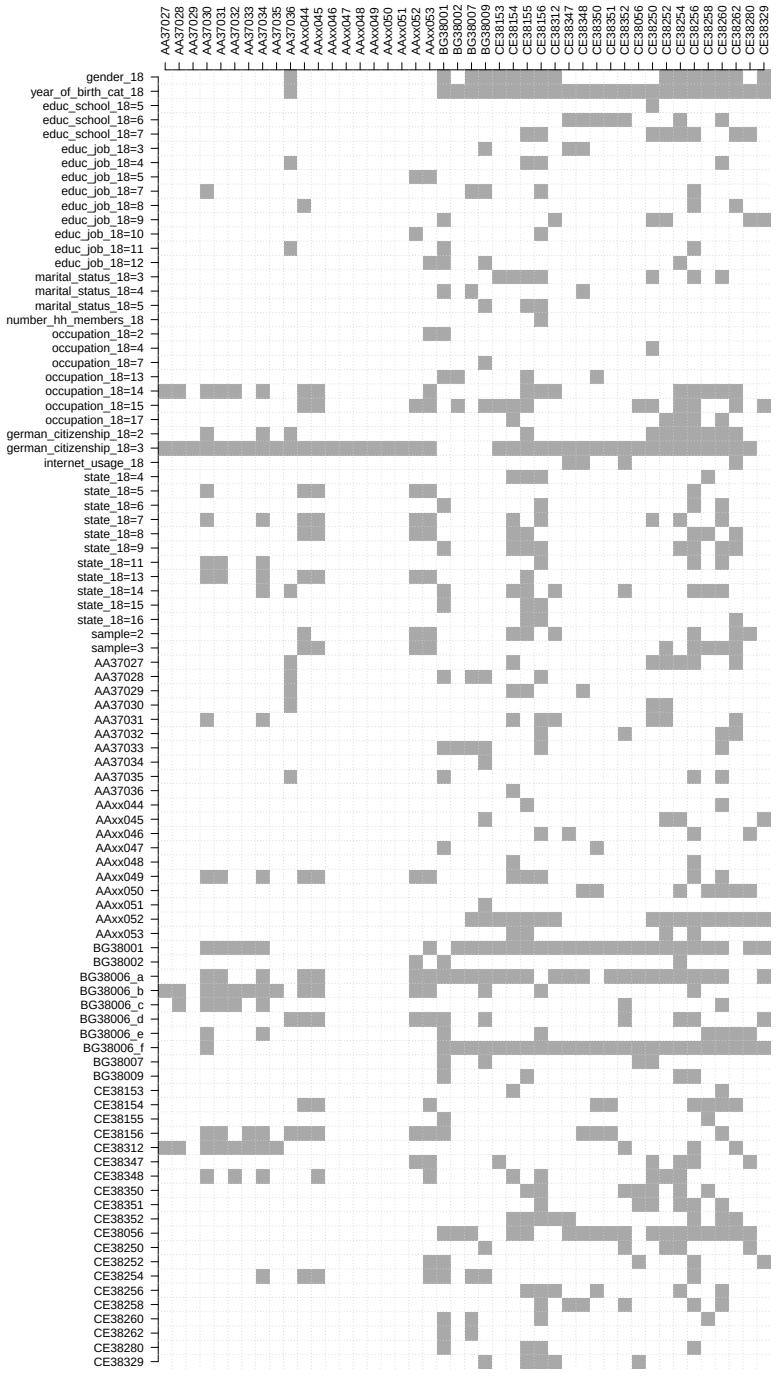

Figure 4.A1: Predictors (on the vertical axis) for item nonresponse on the variables in the GIP data (on the horizontal axis) as selected by elastic net regressions (gray = selected; white = not selected).

Note: The sociodemographic variables and sampling cohort as well as BG38006_a*,* BG38006_b*,* BG38006_c*,* BG38006_d*,* BG38006_e*, and* BG38006_f *each have fewer than 25 values missing by item nonresponse and are therefore excluded from the analysis. The plot displays only predictor variables that were selected for at least one item nonresponse model.*

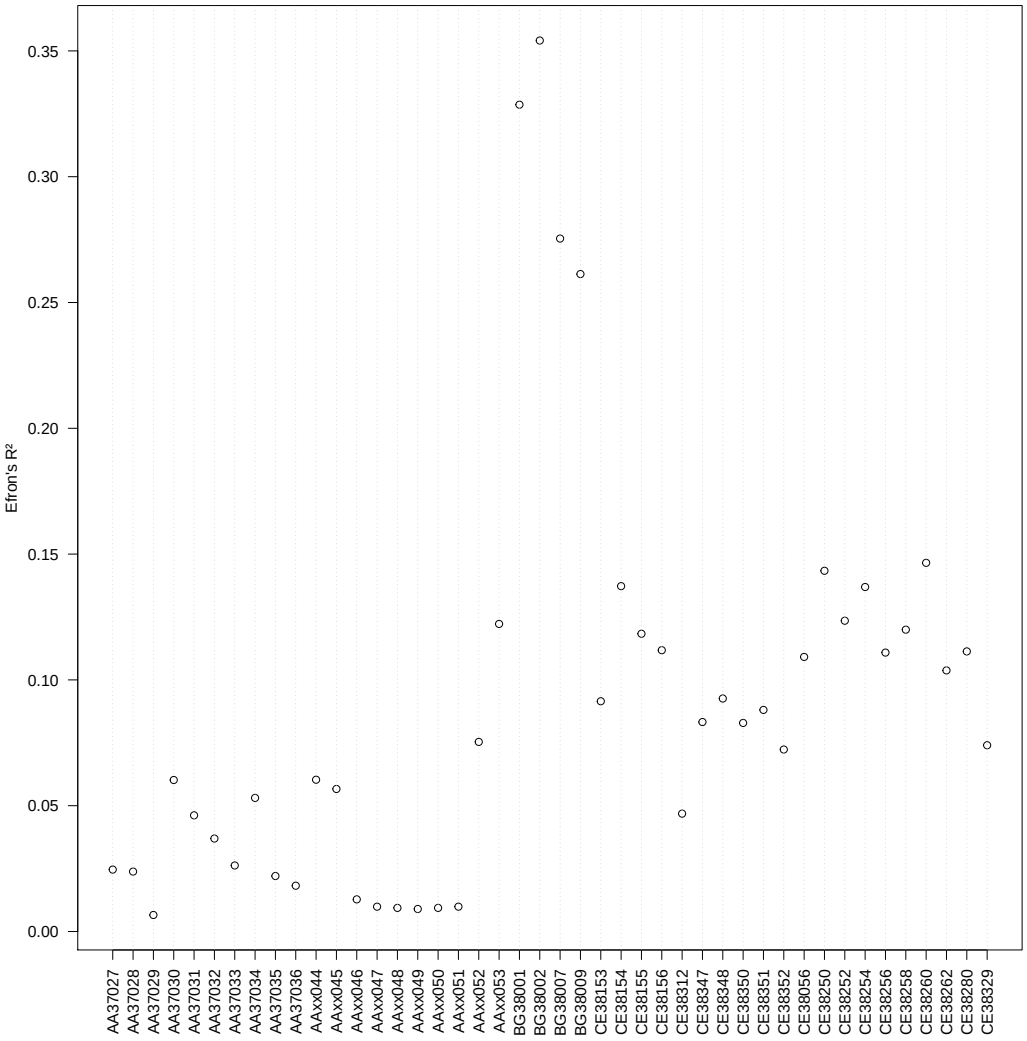

Figure 4.A2: Efron's R^2 values of the item nonresponse models, by outcome variable.

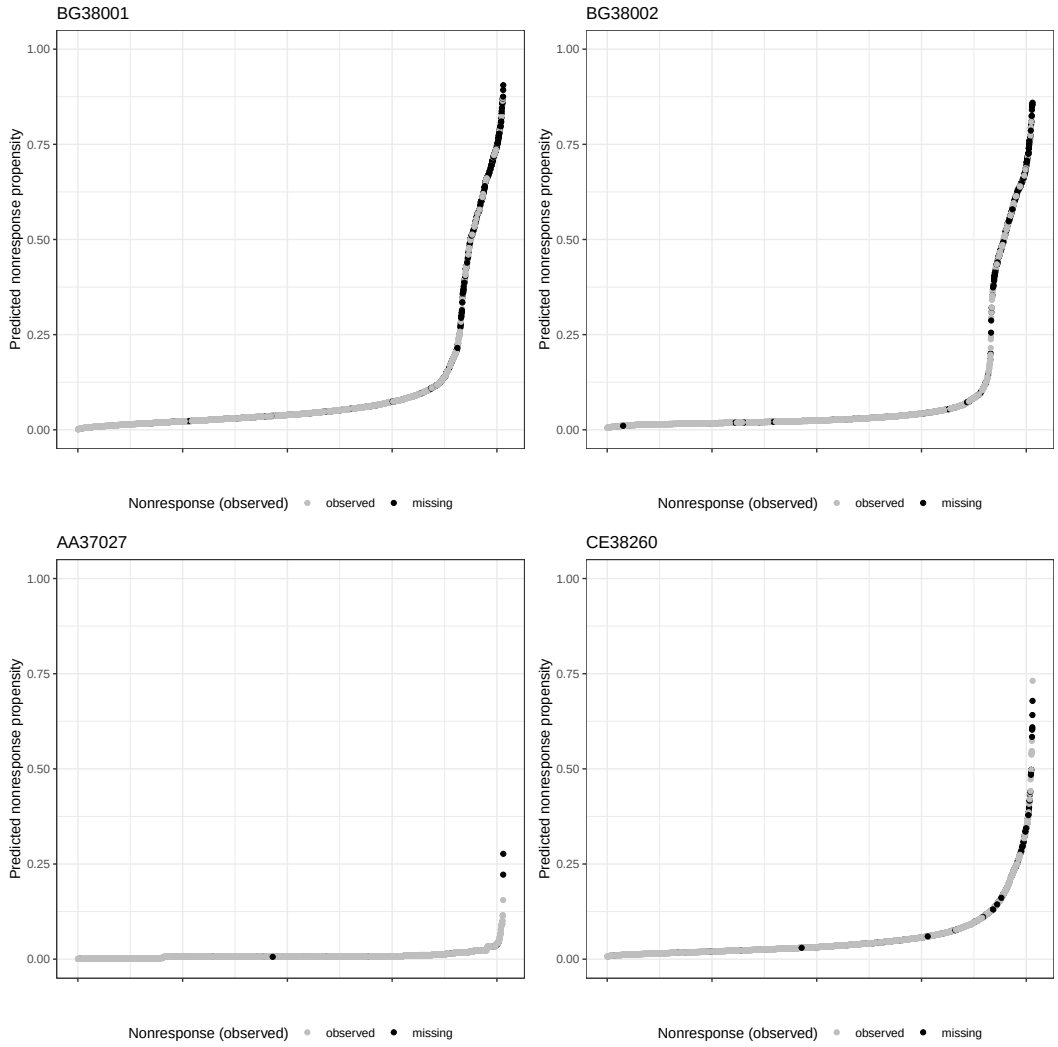

Figure 4.A3: Distribution of predicted nonresponse propensities by actually observed nonresponse in the GIP data for four exemplary variables: BG38001, BG38002, AA370027, and CE38260.

Note: BG38001 *and* BG38002 *have modest proportions of INR but the comparatively best model fit among all variables;* AA37027 *has only a small proportion of INR and a relatively poor model fit of* $R^2 = 0.02$*; and* CE38260 *has the highest proportion of INR,* $R^2 = 0.15$.

Appendix B. Results of the Simulation Study in Tabular Form

Table 4.B1: Quantile distribution of average percentage Monte Carlo biases of univariate frequency estimates for 285 categories of 44 variables, by response mechanism and proportions of item nonresponse and planned missing data.

INR	%PMD	%INR	Quantile						
mechanism			Min.	5%	25%	50%	75%	95%	Max.
-	0	0	-0.9	-0.2	-0.1	0.0	0.1	0.3	0.4
	20	0	-1.5	-0.3	-0.1	0.0	0.3	0.9	2.0
	40	0	-1.9	-1.2	-0.7	0.0	1.6	4.7	9.3
	60	0	-6.0	-2.9	-1.4	0.2	4.5	15.6	23.8
MCAR	0	5	-0.8	-0.3	-0.1	0.0	0.1	0.4	1.1
	0	10	-0.8	-0.3	-0.1	0.0	0.2	0.7	1.9
	0	15	-1.7	-0.6	-0.1	0.0	0.2	1.4	5.8
	0	20	-9.1	-2.3	-0.3	0.0	0.3	5.1	37.1
	20	5	-1.0	-0.5	-0.2	0.0	0.6	1.3	2.3
	20	10	-1.5	-0.8	-0.3	0.0	0.6	1.9	5.3
	20	15	-4.3	-1.6	-0.4	0.0	0.9	3.7	16.5
	20	20	-13.1	-3.8	-0.9	0.0	1.9	8.8	53.3
	40	5	-2.5	-1.7	-1.0	0.0	2.3	6.9	11.4
	40	10	-4.5	-2.4	-1.2	0.0	2.9	8.6	20.3
	40	15	-8.9	-3.5	-1.3	0.0	3.5	11.7	34.6
	40	20	-17.6	-5.1	-1.3	0.4	4.3	14.0	62.7
	60	5	-5.9	-3.2	-1.4	0.6	5.3	16.5	24.5
	60	10	-8.2	-3.8	-1.6	0.8	6.2	17.5	29.7
	60	15	-13.6	-4.3	-1.7	0.9	7.6	19.6	46.4
	60	20	-25.3	-7.0	-2.1	1.0	8.9	25.7	89.4
MAR	0	5	-6.4	-1.1	-0.3	0.0	0.3	1.2	5.0
(GIP-like)	0	10	-11.6	-2.1	-0.5	0.0	0.5	2.2	11.1
	0	15	-15.3	-2.9	-0.7	0.0	0.8	3.4	16.3
	0	20	-18.5	-4.3	-1.2	0.0	1.1	6.1	41.3
	20	5	-7.0	-1.2	-0.3	0.0	0.7	1.9	6.1
	20	10	-12.7	-2.1	-0.5	0.1	1.0	3.0	14.2
	20	15	-16.6	-3.4	-0.8	0.2	1.4	5.1	23.4
	20	20	-18.4	-6.1	-1.3	0.1	2.4	11.3	65.6
	40	5	-8.4	-2.4	-1.1	0.1	2.4	7.0	12.3
	40	10	-12.9	-3.7	-1.3	0.0	3.3	9.3	20.8
	40	15	-13.9	-5.1	-1.7	0.0	4.6	12.9	35.6

	40	20	-19.3	-6.8	-1.9	0.1	5.0	18.3	71.7
	60	5	-5.8	-3.1	-1.5	0.4	4.8	16.7	24.4
	60	10	-9.7	-4.1	-1.8	0.5	6.5	18.5	28.5
	60	15	-14.1	-5.8	-2.4	0.6	7.9	20.9	46.6
	60	20	-26.6	-8.7	-2.4	0.7	8.7	27.3	93.1
MAR	0	5	-11.6	-2.3	-0.5	0.0	0.5	1.8	11.4
(stronger)	0	10	-16.6	-3.8	-0.9	0.0	0.9	3.6	21.1
	0	15	-24.1	-4.9	-1.3	0.0	1.2	5.0	19.2
	0	20	-31.7	-6.3	-1.7	0.0	1.6	8.2	51.4
	20	5	-11.6	-2.1	-0.5	0.1	0.9	2.3	13.4
	20	10	-19.4	-3.8	-0.8	0.1	1.2	4.0	19.6
	20	15	-26.1	-5.1	-1.5	0.3	1.7	6.0	23.9
	20	20	-26.6	-8.6	-2.4	0.1	2.7	14.6	90.1
	40	5	-12.1	-2.7	-1.1	0.1	2.4	7.3	12.7
	40	10	-21.3	-4.5	-1.5	0.2	3.7	9.9	23.2
	40	15	-24.8	-6.3	-1.8	0.0	4.5	13.3	46.3
	40	20	-25.8	-9.0	-2.8	0.4	5.2	20.3	93.6
	60	5	-11.0	-3.9	-1.8	0.5	5.1	16.3	24.1
	60	10	-17.9	-4.9	-2.0	0.5	6.5	18.7	32.4
	60	15	-21.4	-7.0	-2.4	0.5	7.8	20.8	57.4
	60	20	-28.4	-11.3	-2.7	0.5	8.9	30.3	111.9
MNAR	0	5	-41.7	-26.3	-3.6	0.5	3.1	9.1	15.9
	0	10	-73.0	-45.8	-9.6	0.8	6.1	20.1	41.1
	0	15	-100.0	-58.3	-17.4	1.1	9.4	32.3	79.3
	0	20	-100.0	-62.8	-25.4	1.1	13.0	44.4	169.1
	20	5	-41.2	-26.1	-4.0	0.5	3.1	9.1	16.1
	20	10	-72.7	-45.7	-10.4	0.9	6.3	21.1	41.1
	20	15	-100.0	-56.9	-17.8	0.9	9.6	32.3	84.5
	20	20	-100.0	-60.6	-25.8	0.8	13.3	47.8	188.5
	40	5	-38.2	-23.0	-3.1	0.8	3.3	11.3	22.3
	40	10	-70.5	-40.9	-9.6	0.5	6.5	23.4	54.3
	40	15	-100.0	-51.2	-17.2	0.5	10.6	38.4	105.6
	40	20	-100.0	-59.7	-26	0.6	14.1	48.6	198.8
	60	5	-35.5	-19.1	-3.9	1.2	5.9	17.2	33.0
	60	10	-67.1	-38.5	-8.4	1.5	9.4	26.6	64.6
	60	15	-100.0	-50.6	-15.4	1.7	13.3	43.0	118.2
	60	20	-100.0	-57.6	-24.2	0.8	16.5	51.9	218.9

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases each.

Table 4.B2: Quantile distribution of average percentage Monte Carlo biases of Spearman correlation estimates for 88 variable pairs correlated by 0.2 or more in the population data, by response mechanism and proportions of item nonresponse and planned missing data.

INR	%PMD	%INR	Quantile						
mechanism			Min.	5%	25%	50%	75%	95%	Max.
-	0	0	-0.5	-0.2	-0.1	0.0	0.1	0.3	0.4
	20	0	-5.8	-3.4	-1.7	-0.7	0.0	1.8	5.3
	40	0	-12.8	-9.9	-5.0	-2.1	0.3	4.4	20.8
	60	0	-30.2	-26.0	-18.2	-15.1	-11.5	-3.6	40.3
MCAR	0	5	-3.2	-1.6	-0.5	-0.2	0.0	0.7	2.4
	0	10	-6.7	-3.4	-1.2	-0.5	0.1	1.2	4.6
	0	15	-9.7	-5.3	-2.1	-0.9	0.1	2.2	6.5
	0	20	-15.4	-9.8	-4.0	-1.7	-0.3	2.0	7.6
	20	5	-7.5	-5.1	-2.5	-0.9	0.2	2.8	6.6
	20	10	-10.5	-7.2	-3.3	-1.3	0.3	3.2	8.0
	20	15	-13.4	-9.6	-5.2	-1.7	-0.1	3.9	8.0
	20	20	-23.2	-18.2	-10.8	-5.2	-2.2	2.2	10.9
	40	5	-14.6	-12.0	-6.3	-2.9	-0.5	4.2	25.5
	40	10	-18.5	-15.0	-8.3	-4.6	-1.9	5.0	28.6
	40	15	-22.5	-19.5	-12	-8.1	-4.6	3.1	34.0
	40	20	-50.6	-36.8	-22.2	-12.1	-7.7	-0.4	31.9
	60	5	-33.6	-30	-22.9	-20.3	-16.8	-4.1	32.4
	60	10	-38.9	-35.2	-31.2	-25.5	-22.3	-4.3	23.6
	60	15	-54.8	-50.4	-39.3	-31.8	-27.9	-6.0	12.1
	60	20	-82.0	-69.1	-48.0	-38.7	-35.2	-10.7	11.6
MAR	0	5	-5.8	-2.7	-1.1	-0.5	0.1	1.5	3.6
(GIP-like)	0	10	-9.2	-4.6	-2.1	-1.0	0.1	3.1	7.0
	0	15	-11.4	-7.8	-3.2	-1.6	0.1	4.3	10.6
	0	20	-20.9	-12.6	-4.5	-2.6	0.1	6.2	12.8
	20	5	-8.2	-4.7	-2.6	-1.8	0.0	2.4	4.8
	20	10	-11.4	-6.2	-4.0	-2.1	-0.1	4.1	7.9
	20	15	-13.5	-10.7	-5.3	-2.9	0.1	5.0	10.8
	20	20	-30.7	-21.6	-9.5	-6.4	-1.5	3.1	9.4
	40	5	-16.8	-11.1	-6.5	-2.7	-0.7	4.4	23.1
	40	10	-19.8	-13.8	-9.3	-4.9	-1.7	3.9	28.5
	40	15	-25.9	-20.2	-13.8	-9.5	-4.4	1.2	32.5
	40	20	-53.5	-35.8	-22.6	-15.1	-8.4	-0.4	32.1

	60	5	-34.6	-30.2	-23.2	-20.3	-17.8	-3.5	31.9
	60	10	-44.4	-38.4	-30.6	-26.7	-22.5	-4.4	27.5
	60	15	-56.3	-51.3	-39.6	-33.0	-29.1	-6.6	14.6
	60	20	-83.9	-68.4	-50.1	-40.2	-36.1	-11	12.6
MAR	0	5	-11.7	-4.2	-1.6	-0.7	0.1	1.2	4.9
(stronger)	0	10	-19.0	-9.3	-3.1	-1.4	0.1	2.3	9.2
	0	15	-24.9	-13.9	-4.3	-2.3	0.5	4.1	14.1
	0	20	-29.7	-18.1	-6.1	-3.4	0.1	5.6	15.0
	20	5	-14.7	-6.8	-3.3	-2.0	-0.1	3.1	6.2
	20	10	-22.9	-12.9	-5.0	-3.1	-0.3	3.1	10.2
	20	15	-27.0	-19.7	-6.2	-3.9	-0.7	4.0	13.8
	20	20	-41.8	-32.4	-11	-7.4	-2.6	4.1	8.5
	40	5	-20.4	-11.9	-7.5	-3.9	-0.9	4.7	22.6
	40	10	-38.3	-22.7	-10.2	-6.5	-2.3	4.8	26.3
	40	15	-53.3	-29.5	-13.9	-9.6	-4.2	3.3	31.2
	40	20	-62.9	-46.9	-21.4	-15.0	-8.1	0.3	31.7
	60	5	-46.2	-33.0	-24.0	-21.0	-17.3	-3.1	35.5
	60	10	-61.2	-44.7	-30.5	-26.8	-22.0	-3.9	27.9
	60	15	-70.7	-54.9	-37.9	-32.5	-27.7	-6.3	15.7
	60	20	-83.0	-68.9	-49.8	-39.9	-35.4	-10.7	10.8
MNAR	0	5	-35.4	-7.5	-1.6	-0.2	0.7	2.3	5.6
	0	10	-78.2	-18.3	-2.8	-0.6	1.1	4.4	8.9
	0	15 -119.3	-30.8	-4.8	-0.7	1.4	6.9	9.6	
	0	20	-135.6	-35.9	-6.3	-0.9	1.6	9.0	12.0
	20	5	-38.9	-8.3	-2.8	-0.9	0.0	1.7	2.9
	20	10	-79.6	-17.7	-4.0	-1.5	0.4	3.0	4.7
	20	15	-121.5	-32.3	-6.0	-1.5	0.6	5.4	8.0
	20	20	-135.3	-39.2	-11.1	-3.9	-1.1	7.3	13.0
	40	5	-41.8	-15.3	-5.5	-2.9	-1.0	1.9	24.7
	40	10	-82.2	-23.5	-8.5	-4.7	-2.5	1.8	28.1
	40	15	-121.0	-38.4	-12.8	-7.5	-3.5	1.3	33.2
	40	20	-126.6	-45.4	-30.6	-11.3	-5.6	2.9	31.7
	60	5	-54.4	-33.9	-23.9	-20.1	-16.6	-4.1	32.8
	60	10	-87.8	-41.4	-33.5	-26.1	-21.4	-4.4	24.2
	60	15	-111.7	-53.6	-43.5	-32.8	-26.1	-6.8	14.0
	60	20	-113.7	-73.3	-54.8	-38.8	-32.7	-12.0	12.3

Note: Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases each.

Chapter 5

Effects of General Purpose Imputations in Planned Missing Survey Data on the Estimation of a Multiple Regression Model: A Case Study

Abstract

Survey designs using planned missingness, such as the split questionnaire design, are increasingly adopted in social survey research to reduce response quality issues with too long questionnaires. This entails leaving out some items of the questionnaire for each respondent based on a random procedure, trading shorter questionnaires for considerable amounts of planned missing data. In consequence, especially multivariate analyses relying only on available cases are prone to fail due to too small case numbers. Often, methods such as multiple imputation may therefore be needed to analyze such data. Yet, it is an unsettled issue whether the imputation should be carried out by the data-collecting research institute once for all research objectives or by the individual researchers each for their specific research objective. While the former may ease the analysis for data users, the latter allows to tailor the imputation model to the specific analysis model, which might improve the quality of estimates. In particular, an analysis-specific imputation model can be restricted to the analysis sample and to specific variables of interest.

This chapter is single-authored work.

In a simulation study based on data from the European Social Survey, I examine the performance of a multivariate model from the social sciences literature with split questionnaire design data that are imputed using partial least squares predictive mean matching. In doing so, I investigate the effects of a general purpose imputation compared to an analysis-specific imputation. Given this imputation method, the results indicate no beneficial effects of restricting imputation models regarding the number of variables included even if they are not correlated with the analysis variables. However, I also find that analysis-specific imputations may yield more accurate estimates than general purpose imputations when the sample of the imputation model and the analysis sample are not congruent.

5.1 Introduction

Lengthy questionnaires can be a significant challenge in social survey research. They tend to reduce response rates and increase survey breakoff and measurement error (see for example Galesic and Bosnjak, 2009; Peytchev and Peytcheva, 2017). Furthermore, more and more surveys are using the online mode, which is cheaper than traditional face-to-face interviewing (Bianchi et al., 2017; Olson et al., 2021). However, online surveys typically also need to be shorter due to high breakoff rates (Galesic and Bosnjak, 2009; Tourangeau et al., 2013, p. 52). Thus, social surveys face pressures to shorten questionnaires while also collecting data on all topics relevant to its research purpose.

One way to encounter this issue may be planned missing data designs such as the split questionnaire design (SQD; Raghunathan and Grizzle, 1995). The SQD entails allocating all survey items to mutually exclusive modules and randomly assigning only a subset of these modules to each respondent. This allows to collect data on a large set of items while ensuring relatively short questionnaires for each individual respondent.

However, this means that shorter questionnaires come at the price of having large amounts of planned missing data. This may be a problem especially when researchers want to estimate multivariate models based on data from SQDs: Deleting observations with incomplete data and running a complete case analysis would usually result in very small or empty analysis samples.

Therefore, to make data from SQDs analyzable, Raghunathan and Grizzle (1995) propose using multiple imputation (MI; Rubin, 1987; van Buuren, 2018). This procedure allows to replace the missing values with multiple values that would

be statistically plausible based on an imputation model. This entails creating multiple complete datasets with varying imputed values. These datasets can be analyzed separately using standard complete data analysis techniques, after which the single estimates from each dataset are combined into an aggregate estimate using Rubin's rules (Rubin, 1987).

For research practice, one could imagine a scenario in which the data would be imputed by the data-collecting research institute to provide readily analyzable data to data users. Otherwise, each data user working on the data would be required to undertake an imputation by themselves. Thus, a general purpose imputation approach carried out once by the data-collecting research institute may save limited resources and also ensure that researchers who are not able to impute the data themselves are not excluded from using the data. However, imputing such data all at once appropriately for general research purposes may be challenging due to the need for universal applicability of the imputed data: The imputation needs to reproduce all relationships in the data rather than only those that are of interest for a specific research objective. Specifically, a general purpose imputation approach needs to always include all observations and all variables in the imputation model, while an analysis-specific imputation strategy provides the flexibility to exclude irrelevant variables and observations not used in the analysis from the imputation model.

Thus, it needs to be examined how imputing social survey data from an SQD for general research purposes may affect the estimation. This includes investigating whether similarly accurate estimates can be obtained with imputation models (a) based on the gross survey sample or restricted to the analysis sample and (b) based on all variables in the data or restricted to the analysis variables. For research practice, this question will be crucial for evaluating if and under which conditions a general purpose imputation strategy for planned missing social survey data makes sense or if each data user would always need to impute the data on their own.

Moreover, a thorough analysis of this issue needs to rely on a realistic, substantive multivariate model and real (rather than synthetic) social survey data. This is relevant because real multivariate models as found in the substantive literature can be quite complex and real social survey data is not always ideally suited for imputation. By contrast, previous research has dealt mostly with simple univariate or bivariate estimates (Chapters 2, 3, and 4 of this dissertation; Peytchev and Peytcheva, 2017; Raghunathan and Grizzle, 1995; Rässler et al., 2002; Thomas et

al., 2006), with simple ad hoc multivariate models (Bahrami et al., 2014; Thomas et al., 2006), or with synthetic data (e.g., Raghunathan and Grizzle, 1995).[1]

In this paper, I report findings from a Monte Carlo simulation study, in which I estimate a multivariate model borrowed from the social sciences literature (Safi, 2010) using split questionnaire designs simulated in originally complete social survey data. In doing so, I compare the quality of estimates after imputation based on data restricted versus data not restricted to the analysis sample. Furthermore, I manipulate the number of variables in the dataset as well as the strength of correlations among them. The findings suggest that by using suitable imputation procedures, the model estimates and conclusions can be reproduced largely even with large numbers of covariates in the dataset. However, when the analysis is restricted to a subset of the gross sample, analysis-specific imputations may outperform general purpose imputations in accuracy.

5.2 Theory

5.2.1 Split Questionnaire Designs in Social Surveys

The phenomenon that long surveys jeopardize data quality is well-documented in survey research. First, there is a negative correlation between the questionnaire length announced to respondents and initial response rates (Crawford et al., 2001; Walston et al., 2006). Second, long questionnaires result in more breakoff (Galesic, 2006; Galesic and Bosnjak, 2009). Finally, long questionnaires also increase measurement error (Peytchev and Peytcheva, 2017). Altogether, this may often mean that potentially important questions need to be removed from the questionnaire to ensure an adequate data quality.

Planned missing data designs, such as the SQD (Raghunathan and Grizzle, 1995), aim to overcome this issue by systematically leaving out different items for each respondent. The SQD procedure (Raghunathan and Grizzle, 1995), in particular, entails:

1. Allocating all survey items to one of three or more mutually exclusive modules (for different techniques how to construct these modules, see Chapter 2 of this dissertation; Bahrami et al., 2014; Bahrami, 2020; Imbriano, 2018; Rässler et al., 2002; Thomas et al., 2006).

[1] An exception is Imbriano (2018, pp. 49-54), who provides a short description of regression analyses using the Health and Retirement Study.

2. Assigning each respondent to a random subset of two or more of these modules.

3. Delivering only the items from the assigned modules to each participant. This yields planned missing data on all items of the non-assigned modules.

This implies that to attain a significant reduction in questionnaire length, much of the data will be missing. Moreover, the planned missingness typically affects all cases and most variables in the dataset. In consequence, sample sizes especially for multivariate analyses may often be insufficient when relying only on the available cases.

5.2.2 Multiple Imputation

As a solution, Raghunathan and Grizzle (1995) propose using MI (Rubin, 1987; van Buuren, 2018) to complete the partially missing datasets from SQD surveys. MI is a procedure in which missing values are replaced by a predefined number of m multiple imputed values based on an imputation model. For multivariate missing data, one can define distinct imputation models for each variable to be imputed. These imputation models are then repeatedly run one after another in an iterative fashion (this procedure is known as *fully conditional specification*; see Brand, 1999; van Buuren et al., 2006). In consequence, MI generates m complete datasets with varying imputed values.

To analyze the imputed data, conventional complete data analysis techniques can be applied separately to each imputed dataset. Subsequently, the resulting m estimates for each imputed dataset are pooled into an aggregate estimate using Rubin's rules (Rubin, 1987). Thus, once properly imputed, the completed SQD data may in principle be used for a wide range of analyses and analysis methods, even if the analysis objectives are unknown during the imputation (e.g., Bahrami et al., 2014, p. 22).

It is important to be aware of the challenges associated with the imputation of social survey data. First, correlations in social survey data tend to be mostly weak. This means that the predictive power of imputation models may often turn out weak as well. Second, social surveys often include many categorical variables. This limits the set of imputation methods from which to choose, as many methods are developed primarily for continuous data. Van Buuren (2018, p. 91) also notes that categorical variables are more difficult to impute than continuous variables, as conventional imputation methods need enormous case numbers to accurately impute

data with many categories. This is especially challenging due to the large proportions of missing data in SQD surveys, as this limits the sample sizes available for the imputation model even further. Correspondingly, a simulation study shows considerable biases particularly when proportions of missing data exceed 40% (see Chapter 4).

These difficulties make constructing accurate imputation models particularly essential for the quality of imputations in an SQD survey and the consequential substantive estimates. This is particularly important because in SQD surveys the imputations can have a big impact on the estimation due to the large proportion of data to be imputed. For that purpose, one needs to specify the relevant predictor variables and select an appropriate imputation method that estimates the imputed variable using the predictor variables.

Predictor variables in imputation models can serve various purposes: to reduce uncertainty in the imputations, to reduce potential bias from a non-random missingness mechanism, and to preserve their relations to the imputed variable in the imputed data for subsequent analyses. To cover as much information as possible on all these domains, the original recommendation is to include all variables available in the dataset as predictors (e.g., Rubin, 1996). However, including too many predictor variables may in practice cause problems with estimating the imputation model given that sample sizes are limited (van Buuren, 2018, pp.167-170, 259-271; White et al., 2011), potentially leading to bias in the substantive estimates (Hardt et al., 2012). Therefore, it is often suggested to exclude irrelevant predictor variables from imputation models (for example, van Buuren et al., 1999).

Regarding imputation methods, researchers have to choose from a wide variety of approaches. In particular, the imputation method must suit the properties of the imputed variable, such as its level of measurement (for a more general discussion of different imputation methods, see for example van Buuren, 2018; Murray, 2018). Previous research has shown that some established methods such as ordinal logistic regression may fail to deliver appropriate results with SQD data (see Chapter 3). Similarly, Bahrami et al. (2014) also report downward biases in regression coefficients with standard imputation methods (predictive mean matching for continuous variables and ordinal and polytomous logistic regression for categorical variables).

Meanwhile, methods that limit the amount of parameters to be estimated in the imputation model performed more favorably (see Chapter 3). This applies particularly to partial least squares predictive mean matching (PLS-PMM; see Robitzsch et al., 2016), which could reduce Monte Carlo biases decisively. This multi-step

procedure first involves estimating a partial least squares (PLS) regression (see for example de Jong, 1993) of the imputed variable on all predictor variables to extract a predefined number of PLS components optimized to predict the outcome variable. Then, these PLS components are used as predictors in a predictive mean matching (PMM; see Little, 1988; Rubin, 1986) imputation model, in which predictive values for the imputed variable are estimated using a regression model. Subsequently, cases with missing values on the imputed variable are matched to a set of nearest neighbors with observed values based on the previously estimated predictive values. Finally, one of the observed values of the nearest neighbors is drawn by random as an imputation for the missing value. Thereby, all variables in the data can be included in the imputation models despite limited sample sizes.

Being originally developed for continuous data, standard PMM methods can be used for ordinal and dichotomous outcome variables as well (Koller-Meinfelder, 2009, pp. 48-68; van Buuren, 2018, p. 166). However, this does not apply to nominal data with more than two categories, for which no adaptation of PLS-PMM has been developed yet.

For nominal data with more than two categories, classification trees (Breiman et al., 1984; Burgette and Reiter, 2010; Doove et al., 2014) seem to be the best option currently available (see Chapter 3). This simple non-parametric technique repeatedly splits the data binarily into subregions along the values of predictor variables (e.g., individuals over vs. under 60 years old) that are locally optimal for predicting the outcome variable. This leads to smaller and smaller subregions the more splits are applied. When a minimum subregion size is reached, missing values are imputed by randomly drawing an observed value from the same final subregion. Since this procedure does not necessarily account for all relations in the data adequately due to limited sample sizes, however, this method can still result in moderate biases (see Chapter 3).

5.2.3 General Purpose vs. Analysis-Specific Imputation of Planned Missing Data

For imputing planned missing social survey data, two possible scenarios may be considered: The data could either be imputed by the institute administering the survey or by the individual researchers analyzing it.

In the first scenario, multiple imputation would most likely be applied only once with a broad scope (van Buuren, 2018, p. 46) for the imputed data to be delivered to researchers for general research purposes. This may appear promising from a

research-pragmatic perspective: An appropriate imputation of missing data can be computationally intensive, cumbersome for the individual researcher, and may require specialized training in the first place. Thus, a general purpose imputation strategy may save resources compared to imputation applied over and over by individual researchers with a narrow scope (van Buuren, 2018, p. 46) for their specific analyses: It reduces financial costs as well as energy consumption and lets researchers focus more on their substantive research.

Meanwhile, relying on the individual researchers to impute the data may exclude researchers who are not able to implement imputation by themselves. Moreover, due to a lack of resources or training, they might adopt inappropriate imputation strategies. The data-collecting research institute, by contrast, could acquire or might already have computational resources and experts for imputation within their organization.

However, properly imputing data with a broad scope is more difficult than with a narrow scope: Imputation models "perform best when the analysis objectives are known" (Peytchev and Peytcheva, 2017, p. 367). With a general purpose imputation strategy, however, analysis objectives are usually unknown. In this regard, the main issue is congeniality (Meng, 1994). An imputation model is considered congenial only if it is at least as general as the analysis model, hence covering all relations between variables within the analysis model with the same or less strict modeling assumptions. If the imputation model is not congenial, estimates of interest may be inaccurate (Bartlett et al., 2015; Meng, 1994). A general purpose imputation model, however, needs to ensure congeniality not only for one given analysis model, but for all possible analysis models—and must therefore cover all relations between variables in the data adequately.

Thus, in a general purpose imputation all variables need to be adequately included as predictors in the imputation model. While this may be problematic for many imputation methods, through dimensionality reduction PLS-PMM could in principle include a huge number of parameters without breaking down. However, it has not been tested yet whether estimates stay equally accurate with increasing numbers of predictors that might actually not improve the imputation model. Furthermore, due to the underlying regression model, PLS-PMM might have difficulties preserving non-linear relations particularly when they are not continuous.

Restrictions of the analysis sample (e.g., excluding certain age groups) may be another factor potentially jeopardizing the congeniality under a general purpose imputation strategy. Excluding cases from the analysis may behave implicitly as

if the analysis conditioned on an additional variable (i.e., being vs. not being in the analysis sample): The model's coefficients are estimated only for observations within the analysis sample. This is analogous to including interaction terms between all analysis model predictors and the additional variable. Therefore, in a strict sense a fully congenial general purpose imputation model would need to account for each possible analysis subset of the survey sample. As this is clearly not possible in practice, more realistic solutions would need to be considered. For instance, the PLS-PMM implementation by Robitzsch and Grund (2021) provides researchers with an option to include interaction terms of two variables into the PLS model if its correlation to the imputed variable exceeds a pre-specified value. Thereby, the imputation model might be able to account for differential effects in different subgroups of the sample. At the same time, including only interaction terms with a certain correlation may help to prevent overfitting. Meanwhile, not including interaction terms with an effect equal to zero may not yield bias in estimates after imputation (Bartlett et al., 2015).

5.3 Data and Methods

To test the accuracy of model estimates properly, I perform a Monte Carlo simulation with the real survey data that was originally used for estimating the model in question.[2] This simulation study repeats the following major steps in each simulation run:

1. Draw a random sample of cases from the full survey dataset, which is used as a population dataset for this study.

2. Simulate SQD survey data using the sample previously drawn from the population.

3. Apply MI to the simulated SQD data.

4. Estimate the multivariate model based (a) on the imputed SQD data and (b) on the complete sample data (for comparison).

[2]The analyses for this paper are primarily carried out in R (R Core Team, 2021) using the following packages (if not cited elsewhere in this paper): DescTools (Signorell et al., 2020), doMPI (Weston, 2017), dplyr (Wickham et al., 2021), foreach (Microsoft and Weston, 2020), ggplot2 (Wickham, 2016), haven (Wickham and Miller, 2021), MASS (Venables and Ripley, 2002), Rmpi (Yu , 2002). Some data preparation has been done in Stata 14 (sta2016).

Before expanding on the steps of the simulation study listed above, the following paragraphs describe the examined multivariate model and the data preparation.

5.3.1 Analysis Model

This paper draws on an analysis model from the research article "Immigrants' Life Satisfaction in Europe: Between Assimilation and Discrimination" by Safi (2010), published in the *European Sociological Review*. Using data from the first three rounds of the European Social Survey (ESS), Safi studies the life satisfaction of first and second generation immigrants.

The ESS is a probability-based social survey that collects data on a broad range of topics such as political opinion, social trust, subjective well-being, national identity, and religion in a biennial cross-section of the population over 15 years old in more than 21 European countries. For her analysis, Safi (2010) uses the ESS rounds 1 through 3 data (European Social Survey Round 1 Data, 2002; European Social Survey Round 2 Data, 2004; European Social Survey Round 3 Data, 2006; for the questionnaire documentation see European Social Survey, 2018a,2018b,2018c), which were collected between 2002 and 2007.

The paper by Safi (2010) serves the purpose of the study intended here exceptionally well for several reasons. First, it empirically investigates a research question highly relevant for the social sciences using data from a well-established social survey. Second, it has an exceptionally large analysis sample of almost 60,000 cases that can be used as a population dataset for a Monte Carlo simulation. This is especially important to preserve sufficient statistical power for the analysis model in the simulation, which relies on random samples drawn from the population dataset in each individual simulation run. Third, the complexity of this analysis model is exemplary for survey research in the social sciences: Analysis variables can be combinations of different survey variables (e.g., interaction effects) as well as categorized or transformed versions of the original survey variables.

5.3.2 Definition of Variables

The definition of the analysis variables mostly follows Safi (2010). The ESS item on respondents' general life satisfaction serves as the central outcome variable of a linear regression model. Life satisfaction is measured on an eleven-point scale and, as in Safi (2010), interpreted as a continuous variable.

The main predictor variable for life satisfaction is the respondents' immigrant status. In the ESS, all respondents were presented three items asking if they themselves, their mother or their father were born abroad. Thereby, four classes of respondents are differentiated: natives (i.e., neither respondent nor parents born abroad), first generation immigrants (i.e., respondent born abroad), second generation immigrants (i.e., both parents born abroad, but not the respondents themselves), and 2.5 generation immigrants (i.e., only one parent born abroad).

The wide range of control variables used to account for a potential confounding of the main effects covers some socio-demographic items. Gender is included as a dummy variable with *male* being the reference category. The originally continuous age variable is categorized into a factor with four age groups: respondents under 28 years old (reference category), 28 to 40 years old, 40 to 55 years old, and 55 to 65 years old. Years of education are included untransformed as a continuous variable. Furthermore, a family life factor variable with four categories is constructed as a combination of three original survey variables (having a partner, having children currently living in the household, and having had children in the household in the past): respondents that have neither partner nor children (reference category), respondents that have children but no partner, respondents that have a partner but no children, and respondents that have both partner and children.

Beyond that, two socio-economic control variables are included as well. The annual income, originally measured in deciles, is simplified into four categories: below 18,000 Euros (the reference category), 18,000 to 30,000 Euros, 30,000 to 60,000 Euros, and more than 60,000 Euros. Safi (2010) also includes a leftover category for respondents with missing income data. As this procedure is highly impractical in an imputation context, in this study I single-impute these missing values using PMM. The effects of this procedure on the model coefficients (except for the leftover category that is now omitted) are very limited (see appendix Table 5.A1). A further control variable on the respondents' employment status is deducted from three survey items: whether respondents were unemployed the past seven days (yes or no), if they have been unemployed in the past three months (yes or no), and an item on the respondents' occupation using the ISCO-88 classification (Elias, 1997). These are combined into a four-category variable: Those who were unemployed in the last seven days, those who are currently employed but were previously unemployed during the past three months, those who are continuously employed in the previous three months in professional or managerial positions, and a reference category with all other continuously employed individuals.

Another control variable is the respondents' subjective health, a variable with five categories (very good, good, fair, bad, and very bad) that is included as a factor variable. The analysis also includes dummies for each country (with France serving as reference category) and ESS round 2 and 3 (with round 1 being the reference category) as further control variables.

Cases that are incomplete with respect to the analysis variables are removed from the dataset. As already described, an exception is made here for the annual income variable, in which missing observations are single-imputed. Another exception is the ISCO-88 variable, in which missing values are interpreted as an additional category to account for the fact that not everyone has an occupation. I also refrain from imputing the whole four-digit ISCO-88 variable in the simulation, as this would mean there are thousands of categories with many empty cells to be imputed. Therefore, I confine the imputation to the one-digit ISCO-88 major occupational groups.

5.3.3 Preparation of Population Data for the Simulation

To test whether it makes a difference to impute the whole data or only the analysis data, two population datasets are created: The first one is the analysis population to which the analysis model attempts to infer. In line with Safi (2010), this population is restricted to data from Austria, Belgium, France, Denmark, Germany, the Netherlands, Norway, the United Kingdom, Sweden, Switzerland, Portugal, Spain, and Ireland. Furthermore, this population excludes all respondents under 18 or over 65. This results in an analysis population of n = 58,399. The second population dataset is the out-of-analysis population. This dataset includes all the remaining cases, i.e., the respondents from all other countries and those younger than 18 or older than 65 (n = 65,751).

To examine the effect of additional items in the survey beyond those related to the analysis, I simulate additional normally distributed variables attached to the analysis population dataset. In doing so, I consider three different scenarios, in which these additional variables are not correlated or strongly correlated to the analysis variables. Consequently, each of the additional variables is designed to correlate to one of the original analysis variables by $r = 0.00$ or $r = 0.60$. Through this procedure, I include scenarios with 16, 32, and 48 additional variables. An additional scenario with moderate correlations ($r = 0.30$) can be found in appendix B (see Figures 5.B1 and 5.B1), showing results generally ranging in between the other two scenarios.

5.3.4 Population Model

Table 5.1 displays the regression coefficients of interest in this study with their standard errors as observed in the analysis population data. Coefficients for the country and wave dummies are also included in the model, but not examined in the simulation study (and thus not included in Table 5.1 either). For the full model, see Table 5.A1 in Appendix A.

This population model provides the benchmarks for examining the imputed SQD data. In this model, migrant status is associated with lower life satisfaction. This effect is the most pronounced for first generation migrants (-0.233 scale points), somewhat reduced for second generation migrants (-0.172), and almost vanishes for generation 2.5 migrants (-0.048). Most control variables also clearly have non-zero effects on life satisfaction (exceptions are the dummies for the ESS rounds and the combination of continuous employment in professional or managerial occupations, which have effects close to zero). There are also some notable non-linear effects. Most prominently, age has a U-shaped effect on life satisfaction. Furthermore, having children has a slightly positive effect for couples but a negative effect for singles.

5.3.5 Samples From the Population

In each simulation run, I randomly draw an analysis sample of 5,000 from the analysis population dataset. This sample data is used to simulate and impute SQD data and to estimate the regression model. Results of an additional simulation with smaller sample sizes (2,500) can be found in Appendix C (Figures 5.C1-5.C6), yielding the same overall conclusions as the main simulation study.

To investigate effects of an imputation model sample incongruent to the analysis sample, I also implement a scenario in which I append an additional sample of 5,633 from the out-of-analysis population to the analysis population sample. In doing so, I maintain the size ratio of the analysis and out-of-analysis datasets in the sample data. Based on these samples, I simulate SQDs and impute the corresponding planned missing data as if they were one dataset. The regression model, however, is estimated using only the (imputed) analysis sample data.

Table 5.1: Coefficients of interest as observed in the analysis population data with standard errors, obtained from a linear regression of general life satisfaction.

Variable	Coefficient		Standard error
Migrant status (ref: native)			
Gen. 1 migrant	-0.233	*	0.027
Gen. 2 migrant	-0.172	*	0.055
Gen. 2.5 migrant	-0.048	n.s.	0.034
Gender (ref: male)			
Female	0.147	*	0.015
Age (ref: <28)			
28-40	-0.229	*	0.025
40-55	-0.287	*	0.026
>55	0.008	n.s.	0.028
Years of education	0.007	*	0.002
Family status (ref: single w/o children)			
Single parent	-0.270	*	0.029
Childless couple	0.293	*	0.029
Couple with children	0.358	*	0.023
Income (ref: <18k)			
18k-30k	0.311	*	0.022
30k-60k	0.493	*	0.023
>60k	0.563	*	0.030
Employment (ref: employed)			
Currently unemployed	-1.056	*	0.033
Unemployed in past 3 months	-0.334	*	0.020
Employed: professional/managerial	0.019	n.s.	0.020
Health (ref: very good)			
good	-0.477	*	0.018
fair	-1.086	*	0.023
bad	-1.953	*	0.041
very bad	-3.055	*	0.091

Note: ** = significant on a 5% level; n.s. = not significant.*
This model also controls for countries and ESS rounds. The coefficients for these variables are listed in the appendix, Table 5.A1.

5.3.6 Simulation and Imputation of SQD Data

I simulate SQDs based on the sample data by randomly allocating each original survey item to one of eight evenly large modules. The variables for the respondents' country of residence and ESS round are exempted from this procedure, as they would be known in advance for all respondents. Each respondent is randomly assigned to five out of the eight modules. By dropping all data from the non-assigned modules, this results in 37.5% missing data on each questionnaire item.

The simulated SQD data are imputed using the *mice* package (van Buuren and Groothuis-Oudshoorn, 2011) with 20 imputations drawn for each missing value after 10 iterations. For most variables, PLS-PMM as implemented in the *miceadds* package (Robitzsch and Grund, 2021) with 20 PLS components included in the imputation model is used as imputation method. Interaction terms of predictor variables are included into the PLS model if their absolute correlations to the imputed variable exceed 0.1. Classification trees are used as imputation method for the ISCO-88 variable, which is the only non-binary nominal variable with missing data in this study. As a non-parametric method, classification trees do not require an explicit specification of interaction terms.

5.3.7 Measures

After the imputation, the analysis model is estimated in each of the 20 imputed datasets. The model coefficients and standard errors are combined using Rubin's rules.

The first central measure represents the percentage biases of the 21 combined regression coefficients of interest:

$$\%Bias^{\text{MC}}\left(\hat{\beta}_i^{\text{imputed}}\right) = 100 \times \frac{1}{S}\sum_{s=1}^{S}\frac{\left(\hat{\beta}_{i,s}^{\text{imputed}} - \beta_i\right)}{\beta_i}, \tag{5.1}$$

where β_i is the i-th regression coefficient of the analysis model in the population, $\hat{\beta}_{i,s}^{\text{imputed}}$ is the i-th regression coefficient estimate in a Monte Carlo sample s after MI, and S is the total number of simulation runs.

Another measure tests to what extent standard errors increase with imputed SQD data compared to standard errors based on the complete sample, in which no SQD has been applied and all data are observed. To this end, I take the average percentage difference of standard error estimates compared to standard errors obtained from

complete samples:

$$\%\Delta\left(\hat{\sigma}_i^{\text{imputed}}\right) = 100 \times \frac{1}{S} \sum_{s=1}^{S} \frac{\left(\hat{\sigma}_{i,s}^{\text{imputed}} - \hat{\sigma}_{i,s}^{\text{complete}}\right)}{\hat{\sigma}_{i,s}^{\text{complete}}}\,, \tag{5.2}$$

where $\hat{\sigma}_{i,s}^{\text{imputed}}$ is the standard error estimate for the i-th coefficient in a simulation run s after imputation and $\hat{\sigma}_{i,s}^{\text{complete}}$ is the standard error estimate for the i-th coefficient in a simulation run s based on the complete sample data in this simulation run.

5.4 Results

5.4.1 Effects of Incongruent Imputation Models

Figure 5.1 shows the results of the simulation study for the accuracy of the regression coefficient estimates depending on whether the imputation model relied on the gross sample or the net analysis sample, which excludes cases not used in the analysis.

Overall, some coefficients are reproduced well, while others have substantial biases. When data are imputed using the net analysis sample, small biases (less than $\pm 5\%$) are obtained especially for the income dummies and good, fair and bad health. Furthermore, biases up to $\pm 10\%$ are observed for first and second generation immigrants, gender, couples with children, current and previous unemployment. Large biases occur especially in the two middle age dummies (-42% and -45%), single parents and childless couples (-30% and 23%), and for employment in professional and managerial occupations (-67%). Note, however, that the latter coefficient is very small in the population (0.019), so its bias is not big in absolute terms.

The majority of estimates appear more accurate with imputations relying on the net sample rather than the gross sample. This applies specifically to the coefficients for first generation and second generation immigrants, all age categories, years of education, all income categories, couples with children, current unemployment, unemployment within the past three months, and fair health. This is especially apparent in the years of education coefficient, which already has a moderate bias in the net sample scenario but is reduced by 98% in the gross sample scenario. Furthermore, while the age effect is not well captured in either scenario, the gross sample scenario yields glaringly inaccurate estimates: The effects of the two middle age

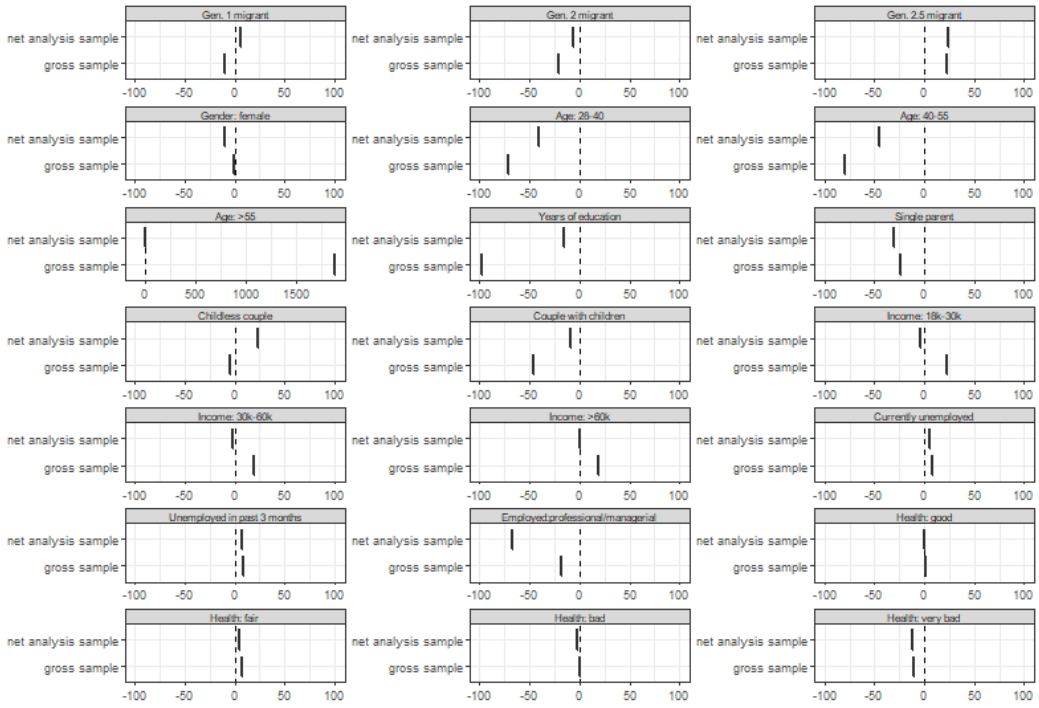

Figure 5.1: Percentage Monte Carlo biases of regression coefficient estimates using (a) the net analysis sample or (b) the gross sample. Solid vertical lines indicate the population benchmarks.

Note: Based on a Monte Carlo simulation with 1,023 runs on 5,000 cases (net analysis sample) or 5,000 + 5,633 = 10,633 cases (gross sample) (each 37.5% missing data).

groups are underestimated both by about 71% and 80%, and the initially very small effect for the higher age group (originally 0.008) is overestimated by 1884%. This distorts the originally U-shaped effect substantially, now showing almost no effect for the younger and middle age groups and a steep increase in life satisfaction for the oldest age group. Meanwhile, five coefficients are estimated somewhat more accurately based on the gross sample: gender, single parents, childless couples, employment in managerial or professional occupations, and bad health. Furthermore, three coefficients have a similar bias in both scenarios: 2.5 generation immigrants, good health, and very bad health.

Figure 5.2 displays the average percentage increase in standard errors due to the SQD compared to model estimates obtained with the complete sample data, again depending on whether the imputation relied on gross or net samples. Expectedly, the SQD yields a substantial increase in standard errors due to the planned missing data. These average increases range from 32% to 81% across the different coefficients and scenarios.

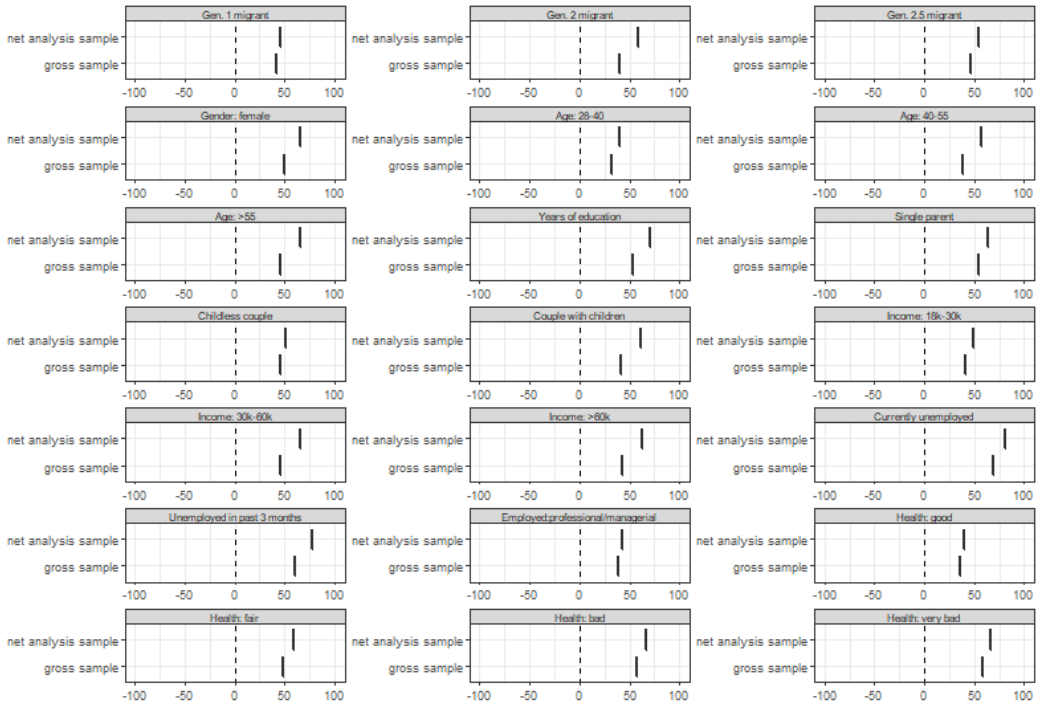

Figure 5.2: Percentage increase in standard errors with imputed SQD data compared to complete sample data, based on imputation models using (a) the net analysis sample or (b) the gross sample.

Note: Based on a Monte Carlo simulation with 1,023 runs on 5,000 cases (net analysis sample) or 5,000 + 5,633 = 10,633 cases (gross sample) (each 37.5% missing data).

Increases in standard errors are consistently more pronounced for the net sample scenario than for the gross sample scenario. However, this increase varies by coefficient, ranging from a 5 percentage points (first generation immigrants) to more substantial increases of up to 19 percentage points (income 30-60k).

5.4.2 Effects of Additional Covariates

Figure 5.3 displays the percentage biases of coefficient estimates obtained with SQDs comprising only the analysis variables (i.e., the net analysis sample scenario from the previous section) versus those with SQDs including 16, 32, or 48 additional variables that are not correlated with the analysis variables. Figure 5.4 does the same, except here each of the additional variables is designed to correlate by 0.60 to one of the analysis variables.

Figure 5.3 shows that coefficients are not substantially affected by additional uncorrelated variables in the imputation model. Only minor differences can be observed: the biases of gender, the three income dummies, current unemployment, and

good health slightly increase with increasing numbers of variables in the dataset. Apart from gender and current unemployment, however, these increases are very small. The other coefficients have biases that are either mostly constant across the scenarios (the two middle age groups, single parents, couples with children), show no clear pattern towards an increase or decrease with increasing numbers of variables (the immigration dummies, age over 55, education, childless couples), or even a slight tendency towards less bias when the number of variables increases (previous unemployment, bad health, and very bad health).

With the correlated additional variables (Figure 5.4), many biases tend to decrease with increasing numbers of variables in the dataset (generation 2.5 immigrants, gender, the middle age categories, single parents, childless couples, fair health, and very bad health). The other biases remain constant (couples with children) or show no clear patterns (first generation and second generation immigrants, age over 55, education, the income dummies, current unemployment, employment in professional and managerial occupations, good health, and bad health). No coefficient shows a bias increasing with higher numbers of variables.

Figure 5.5 displays the percentage increase in standard errors with SQDs comprising only the analysis variables (i.e., the net analysis sample scenario from the previous section) versus those with SQDs including 16, 32, or 48 additional variables that are not correlated with each other. It shows no major effects of the number of variables on standard errors. Surprisingly, however, increases in standard errors appear slightly less pronounced the more variables are added.

Figure 5.6 does the same for the additional variables correlated to the analysis variables by 0.60. As expected, standard errors increase less the more variables are added to the data. For example, while standard error estimates for the first generation migrants coefficient increase at the median by 48% without additional variables, they increase by only 23% at the median with 48 additional variables. The only exception is the standard error for the second generation immigrants coefficient, which shows no such pattern with increasing numbers of variables.

5.5 Summary

Using a simulation study with real survey data, I tested the potential effects of a general purpose imputation of SQD data on the estimation of a multivariate regression model compared to an analysis-specific imputation. Such a general purpose imputation would be bound to include all cases and all variables of the dataset into

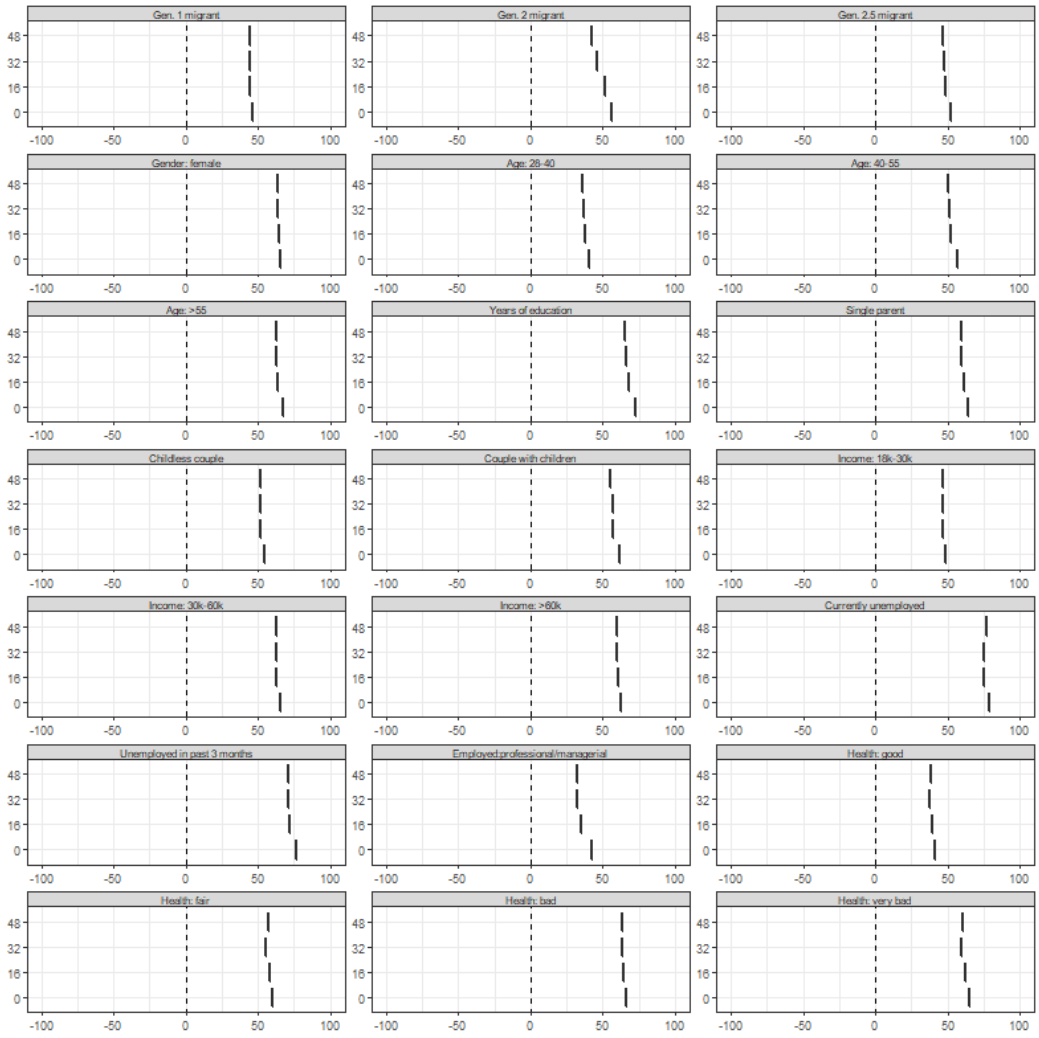

Figure 5.3: Percentage Monte Carlo biases of regression coefficient estimates based on the net analysis sample, with either 0, 16, 32, or 48 additional uncorrelated variables in the dataset to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 5,000 cases (37.5% missing data) each.

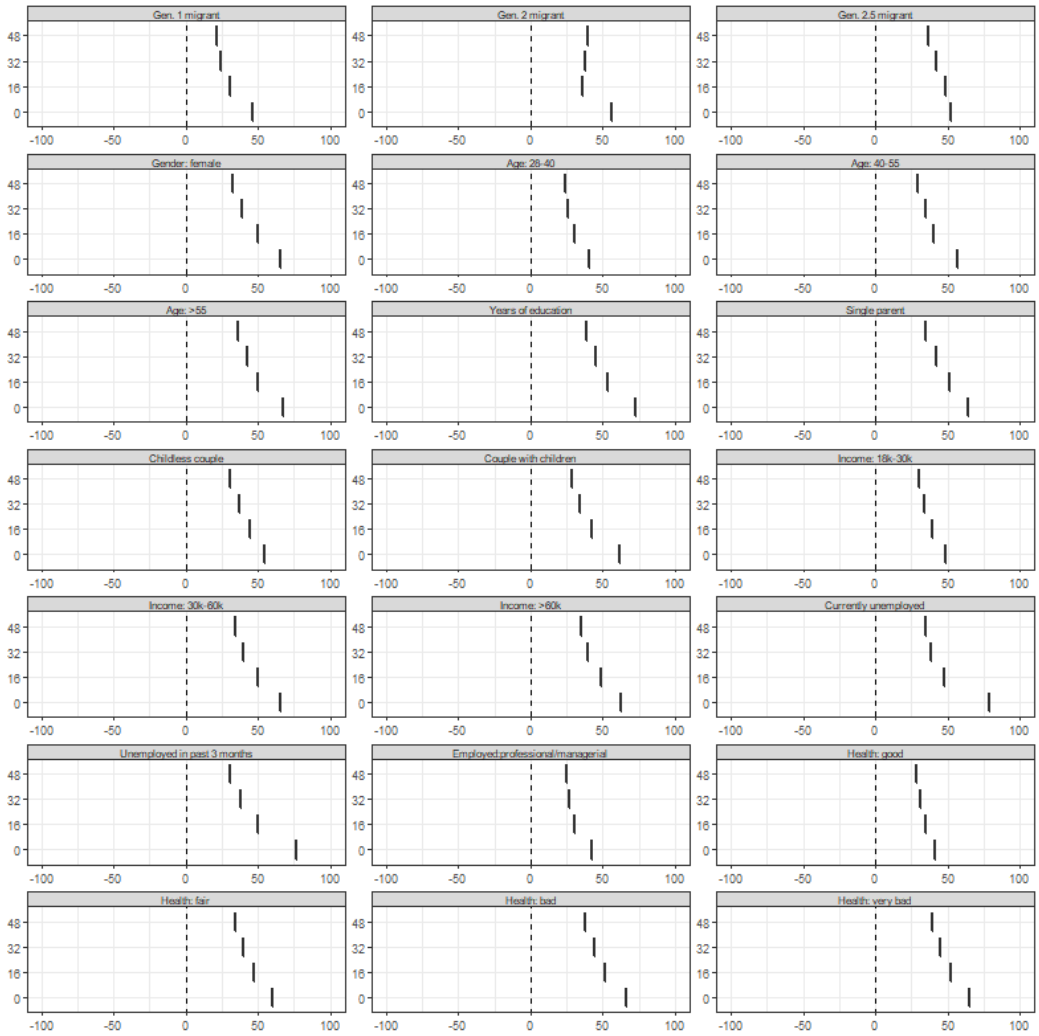

Figure 5.4: Percentage Monte Carlo biases of regression coefficient estimates based on the net analysis sample, with either 0, 16, 32, or 48 additional variables in the dataset highly correlated to the analysis variables ($r = 0.60$) to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 5,000 cases (37.5% missing data) each.

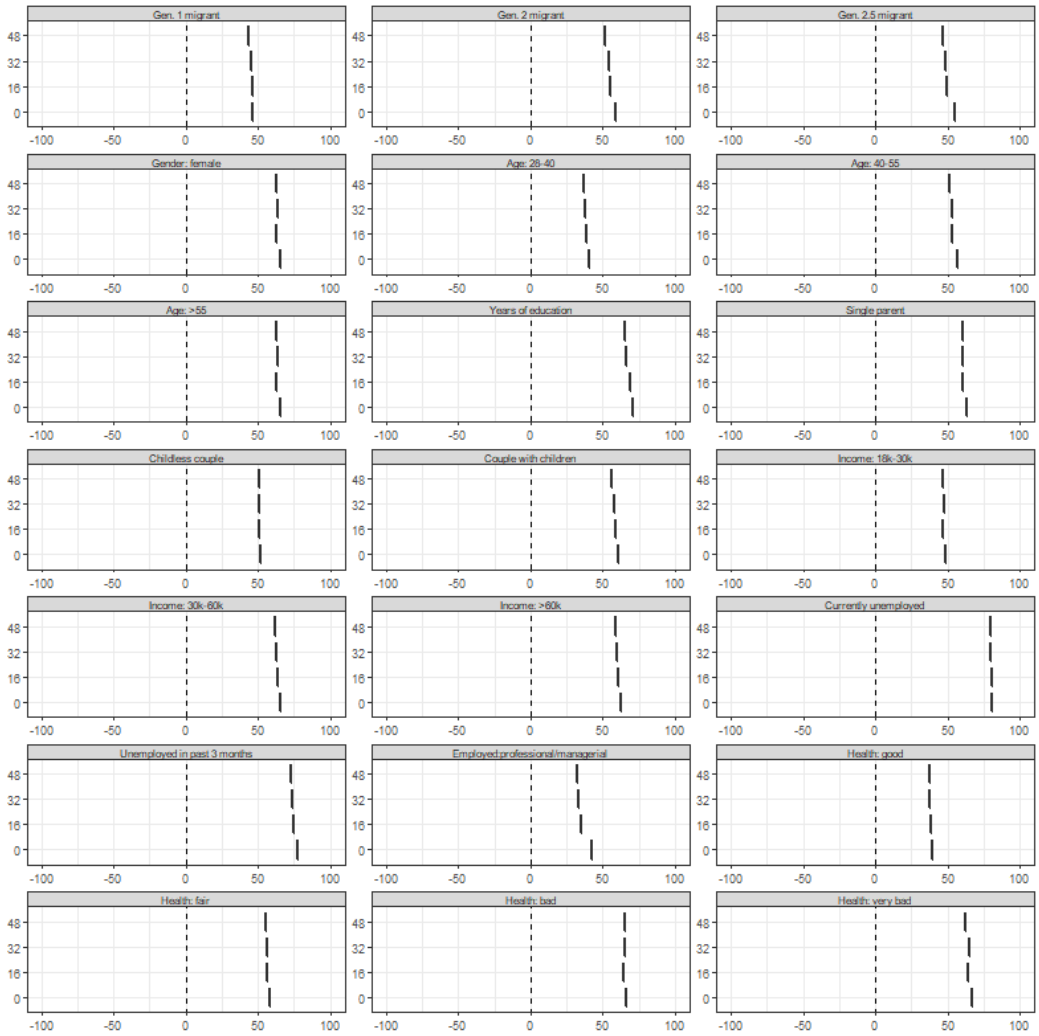

Figure 5.5: Percentage increase in standard errors with imputed SQD data compared to complete sample data after multiple imputation based on the net analysis sample, with either 0, 16, 32, or 48 additional uncorrelated variables in the dataset to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 5,000 cases (37.5% missing data) each.

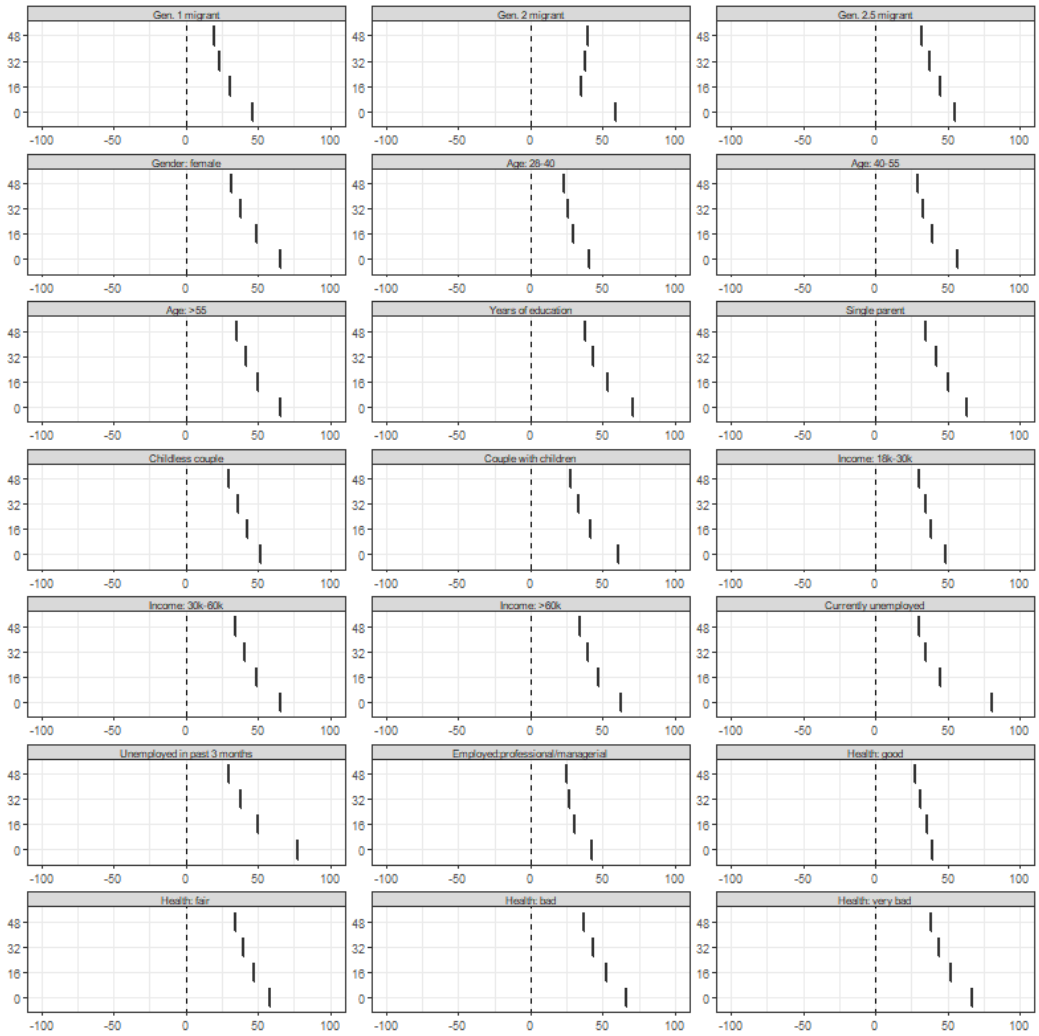

Figure 5.6: Percentage increase in standard errors with imputed SQD data compared to complete sample data after multiple imputation based on the net analysis sample, with either 0, 16, 32, or 48 additional variables in the dataset highly correlated to the analysis variables ($r = 0.60$) in the dataset to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 5,000 cases (37.5% missing data) each.

the imputation models. Meanwhile, with an analysis-specific imputation one would have some freedom to exclude certain cases or variables from the imputation models insofar they are not part of the analysis model.

I found that given the PLS-PMM imputation procedure, the number of variables in the data seems to pose no major problem for the estimation of the model after imputing the SQD data. Regression coefficient and standard error estimates mostly remain similar with increasing numbers of additional variables even if these variables are uncorrelated to the analysis variables. Furthermore, as suggested by the imputation literature (see for example Collins et al., 2001; Raghunathan and Grizzle, 1995; Rubin, 1987, 1996), if the additional variables are correlated with the analysis variables, they can help decrease the size of standard errors and thereby reduce the uncertainty due to the planned missing data. Note, however, that this advantage is not exclusively inherent to general purpose imputation: In research practice, analysis-specific imputation strategies may also make use of preselected highly correlated variables as predictors in the imputation model (e.g., van Buuren et al., 1999).

Incongruent samples for the imputation model and analysis model turned out considerably more challenging for a general purpose imputation. Estimates for many of the regression coefficients were less accurate when imputations relied on the gross sample rather than the net analysis sample, as the imputation is not strictly congenial regarding the analysis model. Apparently, the strategy to include two-way interactions of variables exceeding a correlation threshold did not manage to recover all relevant differences between the gross sample and the analysis sample. At the same time, imputations based on the gross sample also yielded smaller standard errors. Both effects combined could make general purpose imputations yield particularly erroneous inference: Not only are coefficients inaccurate, but the comparatively small standard error estimates could mean that confidence intervals may often not cover the true value.

Furthermore, estimating the non-continuous effect of age and (to a lesser extent) the family status interaction of having a partner and having children proved difficult in all imputation scenarios. This sheds light on a persistent issue with imputation by predictive mean matching: It can only account for non-linearity as long as the relation is still continuous, excluding for instance U and inverted-U shapes. Therefore, when aiming to estimate such effects, an analysis-specific imputation might be a preferable solution.

In sum, this study suggests that for specific analyses a general purpose imputation strategy as examined here might work only under certain conditions: First, the imputations may only be valid for analyses of the whole survey sample. For analyses of specific subgroups within the sample, the data would need to be imputed anew for this specific analysis sample. Second, due to the semiparametric imputation procedure, the extent to which non-continuous or interaction effects could be modeled with this data is also very limited. Again, analyses of such effects would require imputing the data anew, taking the functional form explicitly into account in the imputation model.

5.6 Discussion

The findings from this study may have consequences on the imputation of planned missing survey data. First, when one intends to supply readily analyzable survey data with general purpose imputations for planned missingness, data users should be carefully informed about how the imputed data may be analyzed. From the context of this study, data users should be cautioned to, whenever possible, desist from restricting the analysis sample and from estimating non-continuous effects. Second, data users might be advised to carry out an analysis-specific imputation on their own if they are confident in doing so, even when the data are delivered with general purpose imputations.

In this regard, this study may as well have significance for other areas of research than planned missing data designs. Oftentimes, large-scale social surveys provide imputations in their published data for variables that suffer severe nonresponse issues (see for example U.S. Census Bureau, 2022; Frick and Grabka, 2005). Like in the present study, these imputations are intended for general research purposes rather than a specific model of interest. For generating such imputations, thus, there might be great value in testing their robustness against analysis sample restrictions.

This study has some limitations. First, its findings rely on a simulation. Thus, although the analysis variables are real survey data, the data were originally not collected using a planned missing data design. Moreover, the other variables were simulated based on a normal distribution to ensure controllable conditions in the simulation study. Therefore, the imputation of real planned missing data might to some extent behave differently. Second, this study is also confined to a single exemplary multivariate model. The findings obtained for this model are not necessarily

representative of all other models that may be estimated based on imputed planned missing survey data. This refers to aspects such as a model's analysis method, its variable relations of interest, and potential restrictions of the analysis sample. Therefore, future research may want to extend the findings from this paper through testing the performance of other types of analysis models with imputed planned missing data.

Future research should also focus on further developing imputation procedures to offset the problems that currently still exist with a general purpose imputation of planned missing survey data. A first question may be how the need to model relations with and between lots of predictor variables can be reconciled with the need to maintain non-continuous relationships. The former can be solved by the PLS-PMM algorithm, while the latter would usually require non-parametric methods such as decision tree learning techniques. A solution on this domain may help both to deal with non-continuous effects and potential analysis sample restrictions. Moreover, the good performance of PLS-PMM in this study regarding the number of variables in the data emphasizes the need for an equivalent solution for imputing nominal variables. As nominal variables are widespread in social surveys, advances on this domain may be critical for the contribution of imputation for analyzing survey data with planned missingness.

References

Bahrami, S. (2020). *Missing by design patterns for optimizing survey response by efficient and consistent data collection* [Doctoral dissertation]. University of Bamberg.

Bahrami, S., Aßmann, C., Meinfelder, F., & Rässler, S. (2014). A split questionnaire survey design for data with block structure correlation matrix. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.) *Improving survey methods: Lessons from recent research* (pp. 368-380). Routledge.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462-487.

Bianchi, A., Biffignandi, S., & Lynn, P. (2017). Web-face-to-face mixed-mode design in a longitudinal survey: Effects on participation rates, sample composition, and costs. *Journal of Official Statistics*, *33*(2), 385-408.

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Burgette, L. F. & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070-1076.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330-351.

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review*, *19*(2), 146-162.

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251-263.

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

Elias, P. (1997). Occupational classification (ISCO-88): Concepts, methods, reliability, validity and cross-national comparability. *OECD Labour Market and Social Policy Occasional Papers*, 20. https://doi.org/10.1787/304441717388

European Social Survey (2018a). *ESS-1 2002 documentation report. Edition 6.6*. European Social Survey Data Archive, Sikt - Norwegian Agency for Shared Services in Education and Research, Norway for ESS ERIC. https://doi.org/10.21338/NSD-ESS1-2002

European Social Survey (2018b). *ESS-2 2004 documentation report. Edition 3.7*. European Social Survey Data Archive, Sikt - Norwegian Agency for Shared Services in Education and Research, Norway for ESS ERIC. https://doi.org/10.21338/NSD-ESS2-2004

European Social Survey (2018c). *ESS-3 2006 documentation report. Edition 3.7*. European Social Survey Data Archive, Sikt - Norwegian Agency for Shared Services in Education and Research, Norway for ESS ERIC. https://doi.org/10.21338/NSD-ESS3-2006

European Social Survey Round 1 Data (2002). *Data file edition 6.6*. Sikt - Norwegian Agency for Shared Services in Education and Research, Norway – Data archive and distributor of ESS data for ESS ERIC. https://doi.org/10.21338/NSD-ESS1-2002

European Social Survey Round 2 Data (2004). *Data file edition 3.6*. Sikt - Norwegian Agency for Shared Services in Education and Research, Norway – Data archive and distributor of ESS data for ESS ERIC. https://doi.org/10.21338/NSD-ESS2-2004

European Social Survey Round 3 Data (2006). *Data file edition 3.7*. Sikt - Norwegian Agency for Shared Services in Education and Research, Norway – Data archive and distributor of ESS data for ESS ERIC. https://doi.org/10.21338/NSD-ESS3-2006

Frick, J. R. & Grabka, M. M. (2005). Item-non-response on income questions in panel surveys: Incidence, imputation and the impact on the income distribution. *Allgemeines Statistisches Archiv*, *89*(1), 49-61.

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, *22*(2), 313-328.

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360.

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology*, *12*(184).

Imbriano, P. (2018). *Methods for improving efficiency of planned missing data designs* [Doctoral dissertation]. University of Michigan.

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

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

Meng, X.-L. (1994). Multiple imputation with uncongenial sources of input. *Statistical Science*, *9*(4), 538-558.

Microsoft & Weston, S. (2020). *foreach: Provides foreach looping construct*. R package version 1.5.0. https://CRAN.R-project.org/package=foreach

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, *33*(2), 142-159.

Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N. A., McCarthy, J. S., O'Brien, E., Opsomer, J. P., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z. T., Turakhia, C., & Wagner, J. (2021). Transitions from telephone surveys to self-administered and mixed-mode surveys: AAPOR task force report. *Journal of Survey Statistics and Methodology*, *9*(3), 381-411.

Peytchev, A. & Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods*, *11*(4), 361-368.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

Rässler, S., Koller, F. & 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*. https://www.statistik.rw.fau.de/files/2016/03/d0042b.pdf

Robitzsch, A. & Grund, S. (2021). *miceadds: Some additional multiple imputation functions, especially for 'mice'*. R package version 3.11-6. https://CRAN.R-project.org/package=miceadds

Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der Österreichischen Bildungsstandardüberprüfung [Methodological foundation of standard achievement testing]* (pp. 259-293). facultas.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87-94.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputations after 18 plus years. *Journal of the American Statistical Association*, *91*(434), 473-489.

Safi, M. (2010). Immigrants' life satisfaction in Europe: Between assimilation and discrimination. *European Sociological Review*, *26*(2), 159-176.

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., ... Zeileis, A. (2020). *DescTools: Tools for descriptive statistics*. R package version 0.99.36.

StataCorp (2015). *Stata statistical software: Release 14*. StataCorp LP.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the national health and nutrition examination survey. *Survey Methodology*, *32*(2), 217-231.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford University Press.

U.S. Census Bureau (2022). *2021 Survey of Income and Program Participation users' guide*. https://www2.census.gov/programs-surveys/sipp/tech-documentation/methodology/2021_SIPP_Users_Guide_AUG22.pdf

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC press.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*(6), 681-694.

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049-1064.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer.

Walston, J. T., Lissitz, R. W., & Rudner, L. M. (2006). The influence of web-based questionnaire presentation variations on survey cooperation and perceptions of survey quality. *Journal of Official Statistics 22*(2), 271-291.

Weston, S. (2017). *doMPI: foreach parallel adaptor for the Rmpi package*. R package version 0.2.2. https://CRAN.R-project.org/package=doMPI

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *dplyr: A grammar of data manipulation*. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

Wickham, H. & Miller, E. (2021). *haven: Import and export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.4.1. https://CRAN.R-project.org/package=haven

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377-399.

Yu, H. (2002). Rmpi: Parallel statistical computing in R. *R News*, *2*(2), 10-14.

Appendix

Appendix A: Regression Model in the Population, With Missing Income Imputed or Included as an Additional Dummy

Table 5.A1: Population model: with missing income as separate income vs. single-imputed income.

Variable	Additional category	Single-imputed
Intercept	6.609	6.665
Migrant status (ref: native)		
Gen. 1 migrant	-0.233	-0.233
Gen. 2 migrant	-0.188	-0.172
Gen. 2.5 migrant	-0.046	-0.048
Country (ref: France)		
Austria	0.965	0.961
Belgium	0.913	0.893
Switzerland	1.248	1.118
Germany	0.430	0.414
Denmark	1.732	1.700
Spain	0.978	0.992
Britain	0.509	0.480
Ireland	0.828	0.740
Netherlands	1.047	1.028
Norway	1.017	0.981
Portugal	-0.409	-0.348
Sweden	1.212	1.185
ESS round (ref: round 1)		
Round 2	-0.022	-0.018
Round 3	-0.012	-0.018
Gender (ref: male)		
Female	0.148	0.147
Age (ref: <28)		
28-40	-0.213	-0.229
40-55	-0.257	-0.287
>55	0.029	0.008
Years of education	0.009	0.007

Variable	Additional category	Single-imputed
Family status (ref: single w/o children)		
Single parent	-0.263	-0.270
Childless couple	0.323	0.293
Couple with children	0.387	0.358
Income (ref: <18k)		
18k-30k	0.271	0.311
30k-60k	0.439	0.493
>60k	0.500	0.563
missing	0.294	
Employment (ref: employed)		
Currently unemployed	-1.072	-1.056
Unemployed in past 3 months	-0.323	-0.334
Employed:professional/managerial	0.040	0.019
Health (ref: very good)		
good	-0.477	-0.477
fair	-1.088	-1.086
bad	-1.954	-1.953
very bad	-3.034	-3.055

Appendix B: Additional Covariates With Moderate Correlations to the Analysis Variables ($r = 0.30$)

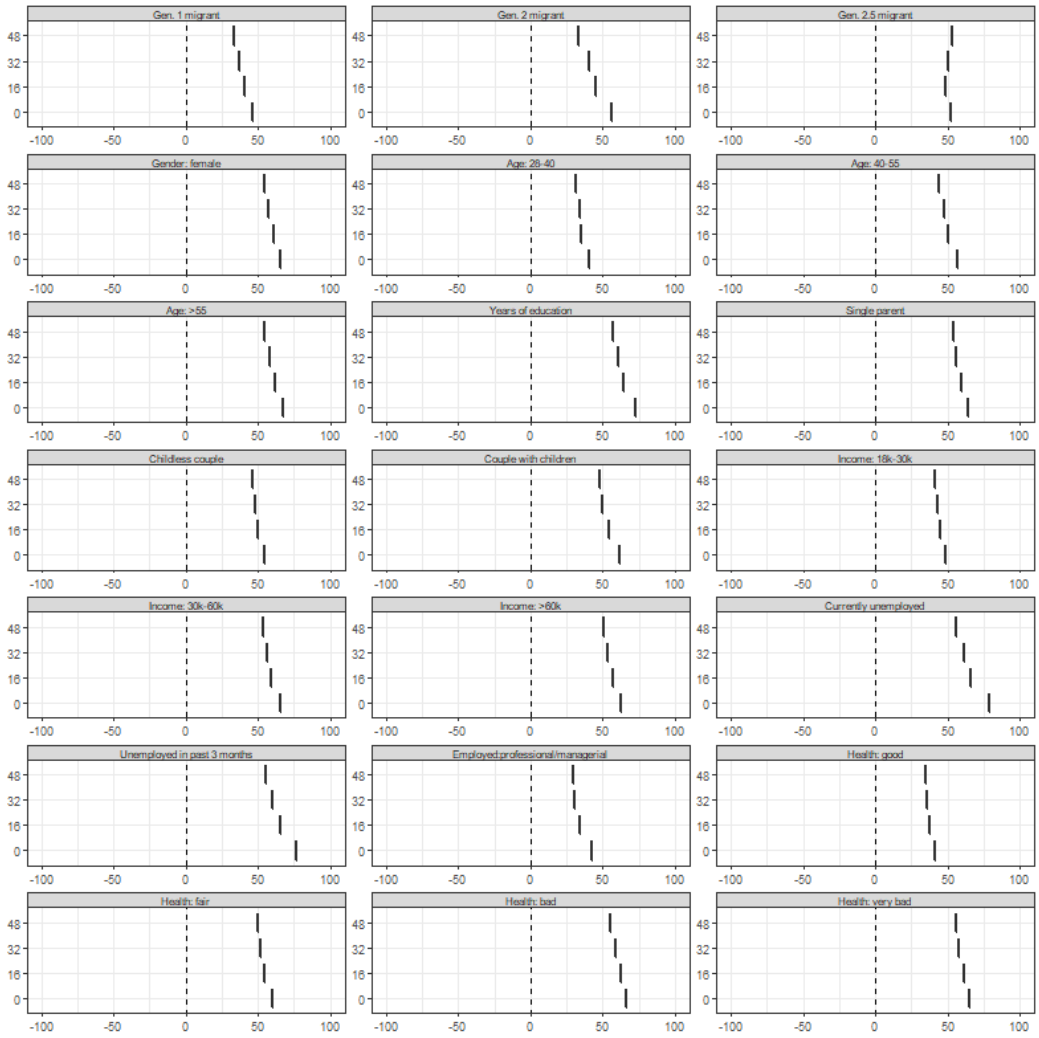

Figure 5.B1: Percentage Monte Carlo biases of regression coefficient estimates based on the net analysis sample, with either 0, 16, 32, or 48 additional variables in the dataset highly correlated to the analysis variables ($r = 0.30$) to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 5,000 cases (37.5% missing data) each.

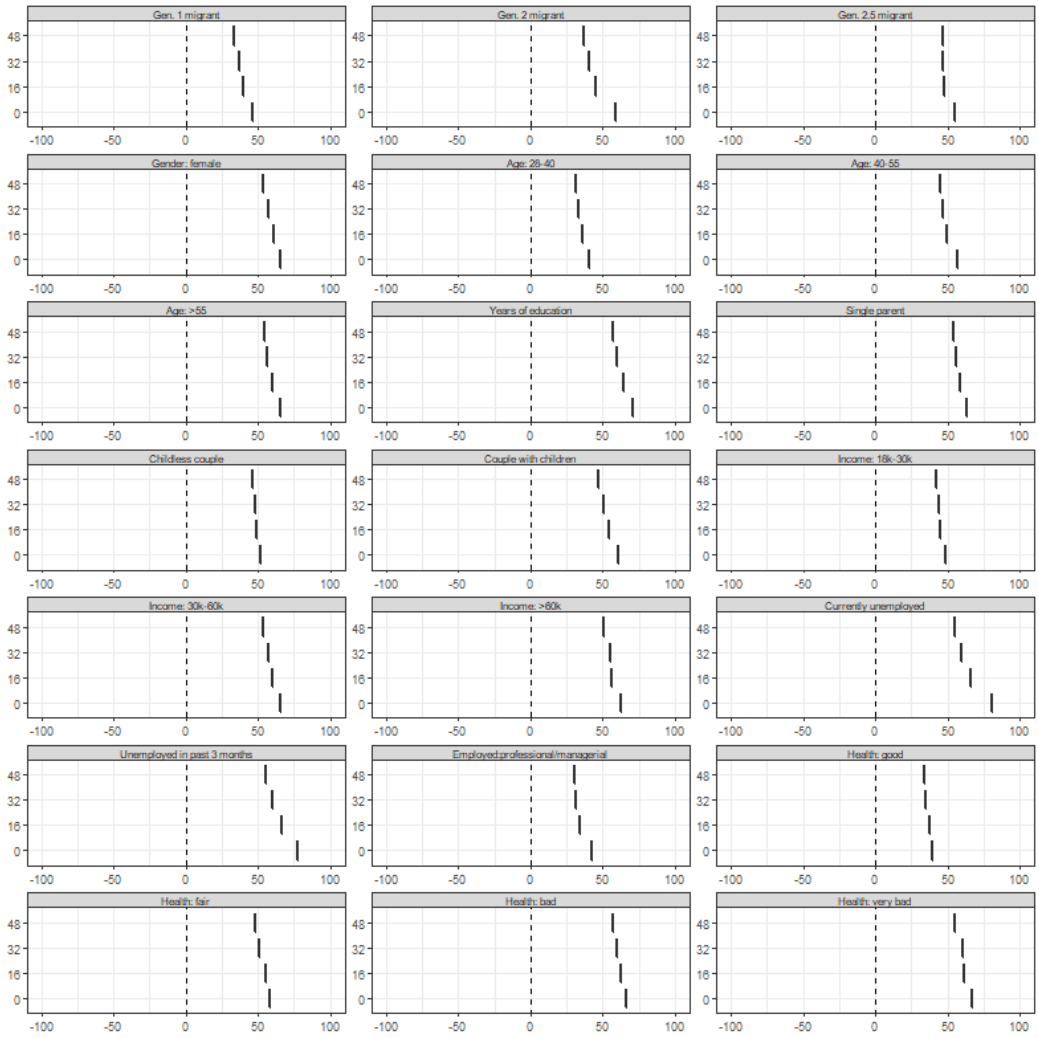

Figure 5.B2: Percentage increase in standard errors with imputed SQD data compared to complete sample data after multiple imputation based on the net analysis sample, with either 0, 16, 32, or 48 additional variables in the dataset highly correlated to the analysis variables ($r = 0.30$) in the dataset to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 5,000 cases (37.5% missing data) each.

Appendix C: Simulation With Smaller Samples (n = 2,500)

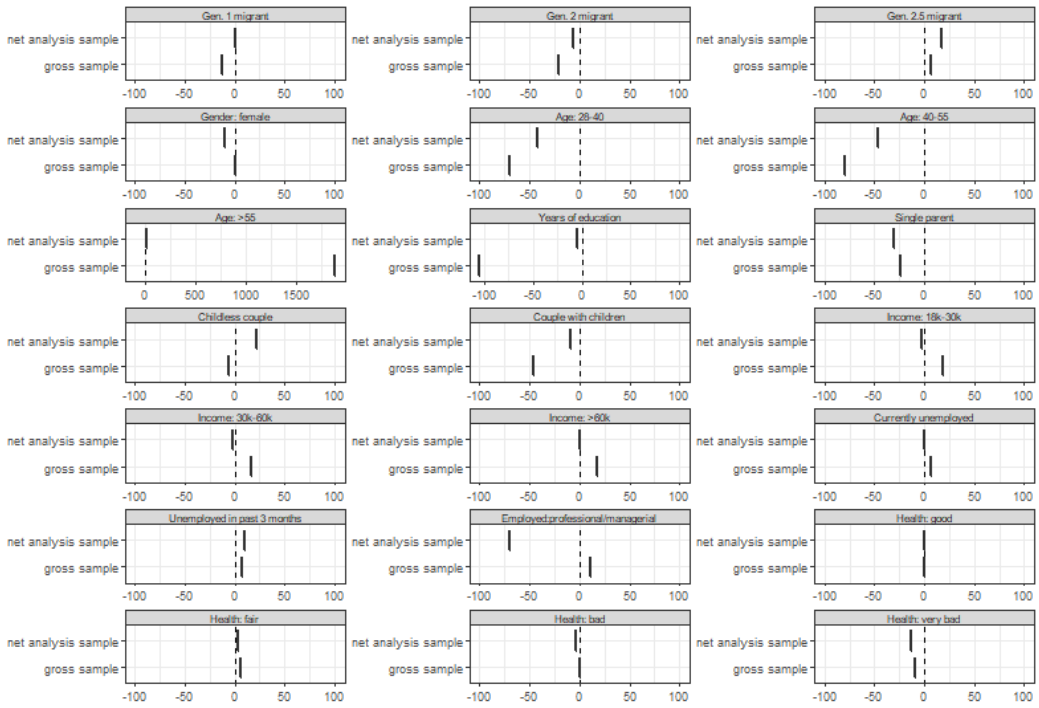

Figure 5.C1: Smaller samples: percentage Monte Carlo biases of regression coefficient estimates using (a) the net analysis sample or (b) the gross sample. Solid vertical lines indicate the population benchmarks.

Note: Based on a Monte Carlo simulation with 1,023 runs on 2,500 cases (net analysis sample) or 2,500 + 2,817 = 5,317 cases (gross sample) (each 37.5% missing data).

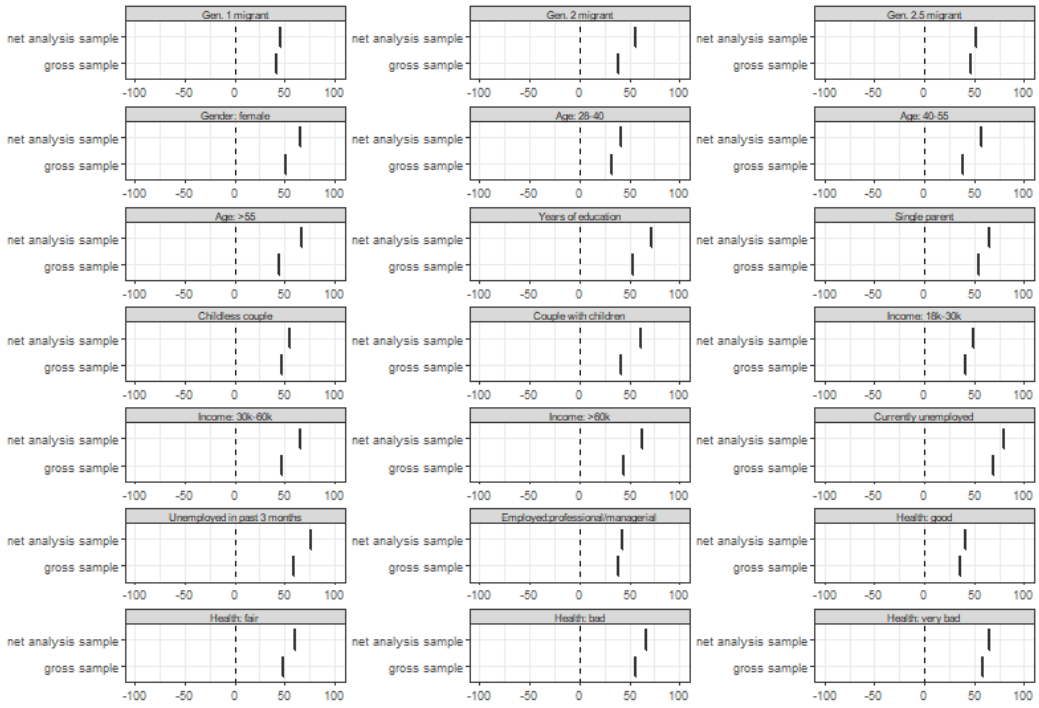

Figure 5.C2: Smaller samples: percentage increase in standard errors with imputed SQD data compared to complete sample data, based on imputation models using (a) the net analysis sample or (b) the gross sample.

Note: Based on a Monte Carlo simulation with 1,023 runs on 2,500 cases (net analysis sample) or 2,500 + 2,817 = 5,317 cases (gross sample) (each 37.5% missing data).

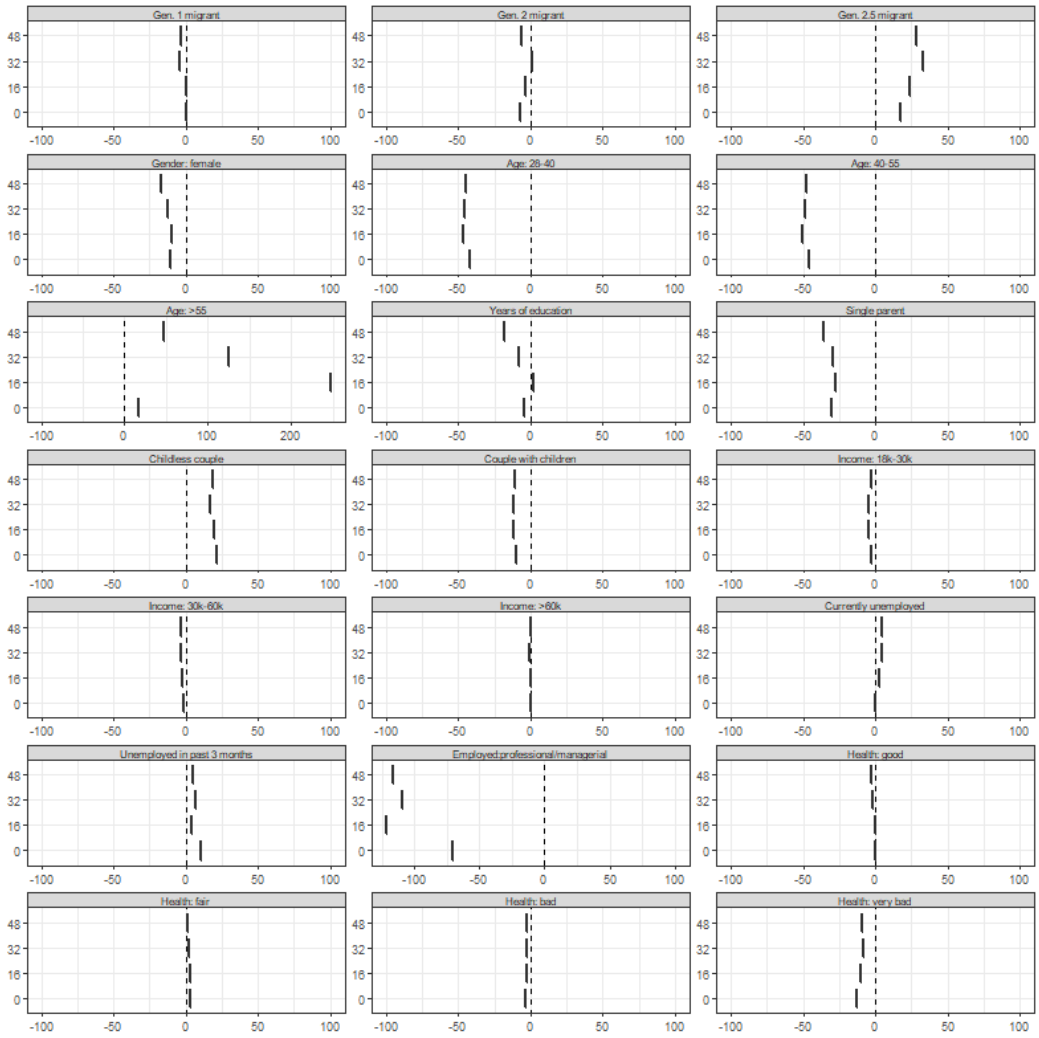

Figure 5.C3: Smaller samples: percentage Monte Carlo biases of regression coefficient estimates based on the net analysis sample, with either 0, 16, 32, or 48 additional uncorrelated variables in the dataset to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 2,500 cases (37.5% missing data) each.

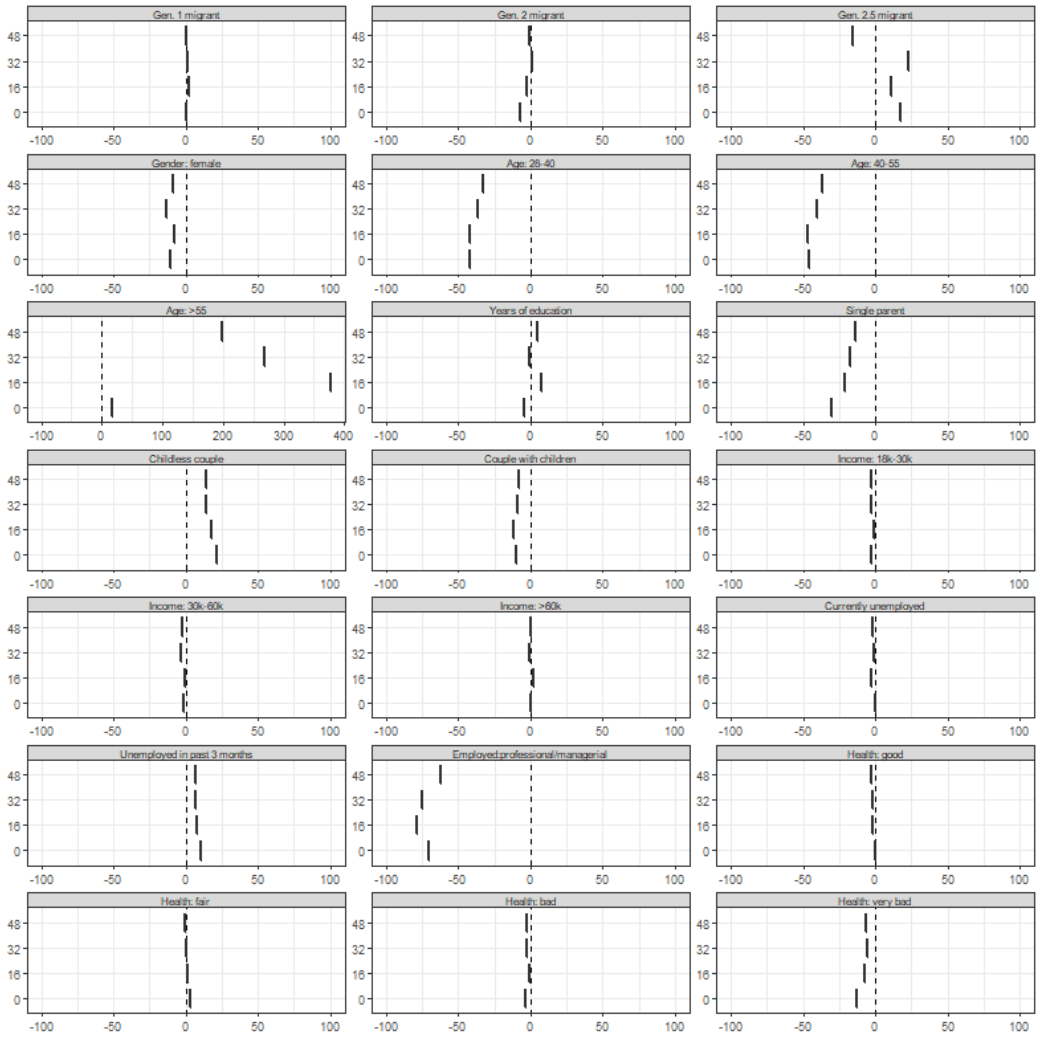

Figure 5.C4: Smaller samples: percentage Monte Carlo biases of regression coefficient estimates based on the net analysis sample, with either 0, 16, 32, or 48 additional variables in the dataset highly correlated to the analysis variables ($r = 0.60$) to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 2,500 cases (37.5% missing data) each.

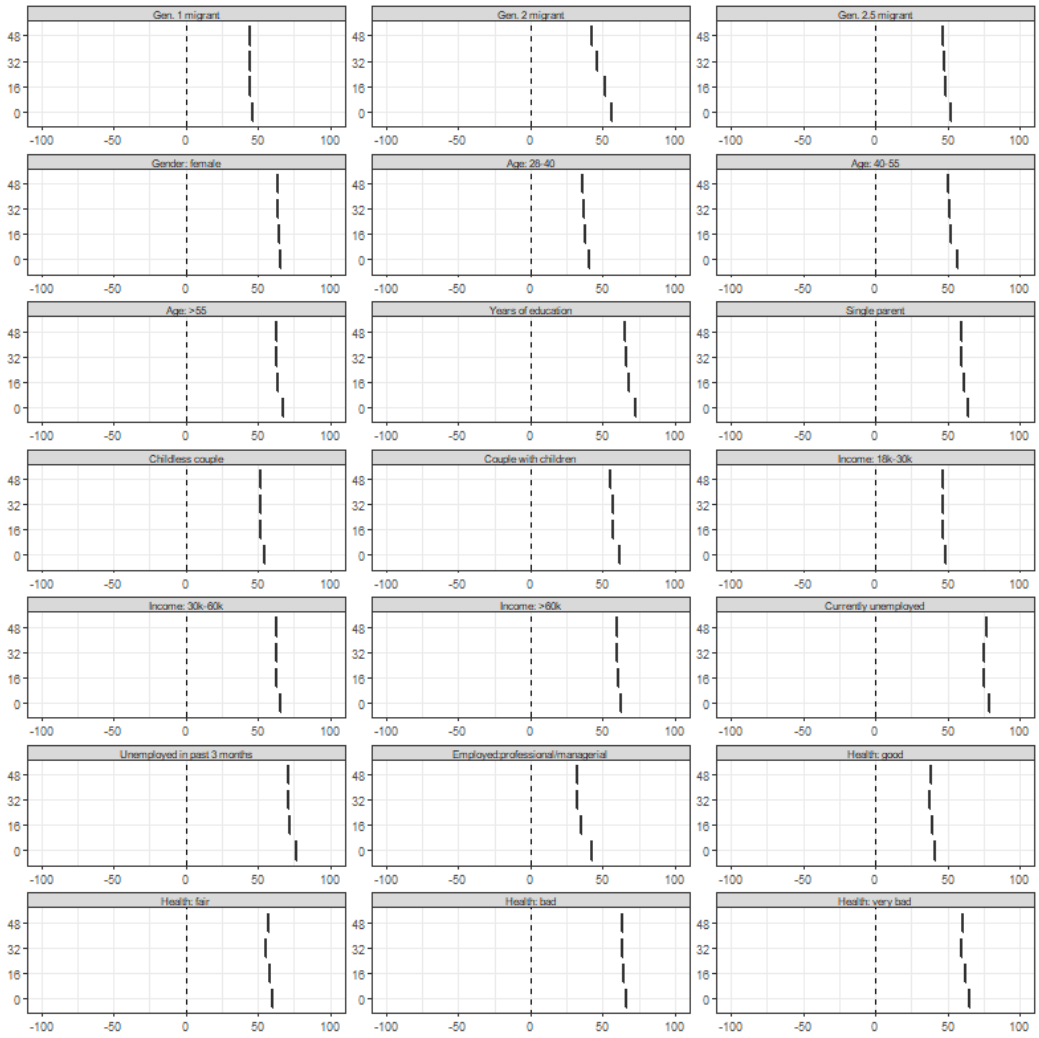

Figure 5.C5: Smaller samples: percentage increase in standard errors with imputed SQD data compared to complete sample data after multiple imputation based on the net analysis sample, with either 0, 16, 32, or 48 additional uncorrelated variables in the dataset to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 2,500 cases (37.5% missing data) each.

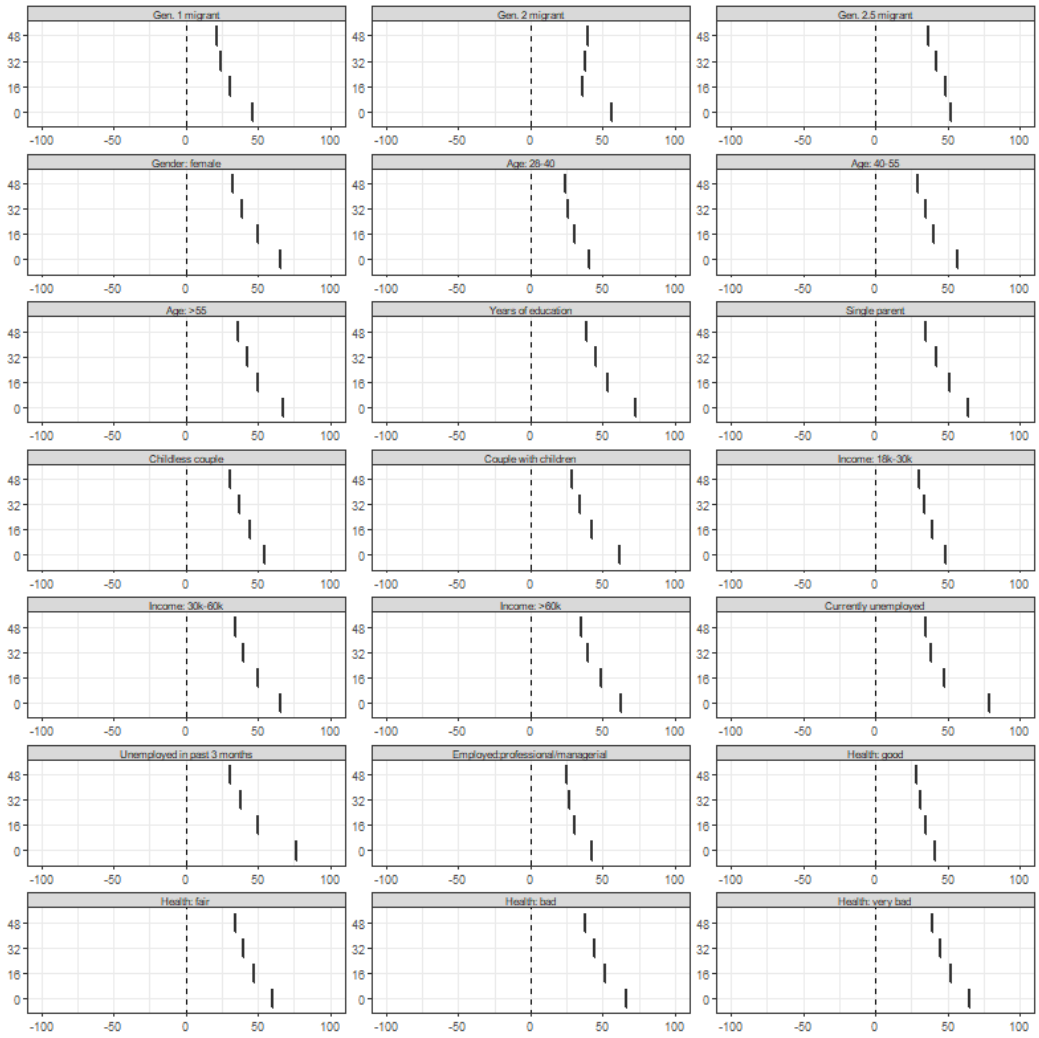

Figure 5.C6: Smaller samples: percentage increase in standard errors with imputed SQD data compared to complete sample data after multiple imputation based on the net analysis sample, with either 0, 16, 32, or 48 additional variables in the dataset highly correlated to the analysis variables ($r = 0.30$) in the dataset to be considered in imputation models.

Note: Based on a Monte Carlo simulation with 1,023 runs on 2,500 cases (37.5% missing data) each.

Chapter 6

Conclusion

In an era of declining response rates and staggering survey costs, survey research is characterized by constant attempts to find sustainable ways for collecting high-quality data at reasonable costs. Ever since lengthy questionnaires have been identified as one of the causes of low response rates (Heberlein and Baumgartner, 1978; Dillman et al., 1993), limiting questionnaire length is commonly considered an important piece of this puzzle. With the more recent rise of self-administered online surveys that are often considered especially responsive to lengthy questionnaires (see for example Revilla and Höhne, 2020), the aspect of questionnaire length has become even more relevant. If long questionnaires are cut down too much, however, consequences for survey research could be serious as well. In this case, important research projects may be rejected or canceled, in the long run obstructing the empirical and theoretical advancement in the social sciences and related fields.

Approaches utilizing planned missingness, such as the split questionnaire design (Raghunathan and Grizzle, 1995), could play a role in resolving this conflict. This is highlighted by large-scale implementations of such designs in the past years, most prominently the PISA context questionnaire 2012 (OECD, 2014, chap. 3) and the European Values Study 2017 (Luijkx et al., 2021). However, the data resulting from these designs are often hard to analyze without proper missing data analysis techniques. So far, there has been little research on how to design the split questionnaires and apply these missing data techniques with real social survey data such that the estimates obtained with this data are satisfactorily accurate.

Building on a series of Monte Carlo simulation studies, this dissertation addresses the effects of various methodological decisions associated with the implementation of split questionnaire designs in social surveys on the accuracy of estimates. In doing so, I covered all the process from designing module structures for the split questionnaires, to factoring in item nonresponse, and to determining viable imputation strategies for the resulting missing data. The results highlight challenges associated with split questionnaire designs as well as conditions and strategies under which satisfactorily accurate estimates can be expected. In the conclusion of this dissertation, I summarize the findings and discuss in more detail which situations may facilitate an accurate estimation and map out where more research is needed for the future.

6.1 Summaries of the Four Papers

6.1.1 Paper I

The first paper focused on how modules in a split questionnaire design may be constructed from the perspective of imputation quality. Three potential strategies were compared: random modules, single topic modules, and diverse topics modules. The expectation was that diverse topics modules would perform the most favorably, followed by random modules and single topic modules.

Three major conclusions can be derived: First, as expected, single topic modules overall led to the poorest estimates, both regarding univariate frequencies and bivariate correlations. Hence, with regard to the quality of imputations, modules should be constructed in such a way that they cover not only one but different survey topics. Second, there were almost no differences in performance between diverse topics and random modules. We also found evidence that even if the within-topic correlations were considerably stronger, the advantage of diverse topics modules over random modules might still be very limited. This finding is especially striking because a large portion of the previous research on split questionnaire designs specifically revolves around optimizing the allocation of items to modules beyond random chance (Adigüzel and Wedel, 2008; Bahrami et al., 2014; Bahrami, 2020; Chipperfield and Steel, 2009, 2011; Imbriano, 2018; Rässler et al., 2002; Thomas et al., 2006). Furthermore, this is also important because optimization strategies often require data collected in advance. Collecting this data specifically for this purpose would likely be a costly endeavor that is probably not worth it. Third, contrary to

the overall pattern, one set of estimands is estimated the most accurately with single topic modules: correlations of two items of the same topic. This also is a novel finding not anticipated by previous research. In consequence, questionnaire developers might consider single topic modules if they are willing to assume that each subsequent analysis primarily covers only a single survey topic. In this case, however, one might even be able to forgo the imputation of the planned missing data in many instances, as all analysis variables would be located in the same module and hence be either pairwise observed or completely unobserved for a given respondent.

6.1.2 Paper II

In the second paper, a wide range of different imputation procedures were examined regarding their ability to reproduce correlations in data from a split questionnaire design in a general purpose imputation scenario, including both different imputation methods and different predictor set specifications. This paper goes beyond previously existing research on different imputation procedures (e.g., Akande et al., 2017; Burgette and Reiter, 2010; Doove et al., 2014; Seaman et al., 2012; Slade and Naylor, 2020; Wu and Leung, 2017) in various respects: First, it evaluates not only a few but a wide variety of different state of the art imputation procedures, providing a comprehensive and direct comparison of their performance. Second, in contrast to the imputation scenario of most studies on imputation, this paper deals with data specifically from a split questionnaire design and aims explicitly at a general purpose imputation rather than confining the imputation to a small set of variables. Finally, unlike most prior research, the Monte Carlo simulations rely on real social survey data rather than simulated data. In doing so, the challenges of real survey data are accounted for while maintaining the robustness of a Monte Carlo simulation study.

The results show that in such a scenario several established imputation methods, especially generalized linear models for categorical variables and classification and regression trees, can lead to poor correlation estimates. This finding may be critical information for researchers imputing data from split questionnaire designs in the future: Generalized linear models are the default imputation method for categorical data in many software implementations of multiple imputation, such as *mice* (van Buuren and Groothuis-Oudshoorn, 2011). Furthermore, were it not for the findings of this paper, tree-based imputation techniques may be a tempting technique for a general purpose imputation: They claim to automatically model the data based on the relevant predictor variables with all its linear and non-linear relations, and

prior research not specifically tailored to the present imputation scenario generally suggests a good performance of these techniques (Akande et al., 2017; Burgette and Reiter, 2010; Doove et al., 2014).

Surprisingly, Bayesian linear regressions and joint modeling via the multivariate normal distribution yielded comparably accurate estimates despite the data being ordered categorical (at least when not transforming the imputations to discrete categorical values). Furthermore, simplifying imputation models through restricting predictor sets proved promising. This refers both to removing predictors with near-zero correlations to the imputed variable and to partial least squares regression reducing the dimensionality of the predictor space. These techniques worked especially well with predictive mean matching. In contrast to Bayesian linear regression and joint modeling, these techniques also have the advantage that they preserve the discrete scale of the categorical data.

The major conclusion from this paper is that to allow for a general purpose imputation strategy for missing survey data from split questionnaire designs to work out, imputation models need to be simplified to some extent. This can be done by choosing an "undemanding" imputation method such as Bayesian linear regression or the multivariate normal model, which can describe the relation between two variables using only one coefficient. However, this is associated with strong normality assumptions that are often unrealistic especially for categorical data, and they generate continuous imputations that are not compatible with the discrete scales of categorical variables. This issue may be averted by using predictive mean matching combined with predictor sets that are restricted to variables with clearly non-zero correlations to the imputed variable. However, this relies on the strong assumption that the removed predictors indeed have no relation at all to the imputed variable in the real world (Bartlett et al., 2015). Partial least squares regression, in contrast, relaxes this assumption, as it does not remove predictor variables from the imputation model per se. Thus, for imputing survey data from a split questionnaire design for general research purposes, partial least squares predictive mean matching might be the most promising procedure so far. This also suggests that future research may focus on further developing this procedure, such as generalizing it to unordered categorical data or implementing more data-driven ways of determining the number of partial least squares components included into the imputation model.

6.1.3 Paper III

The third paper enhanced the previous simulations of planned missing data by taking into account the additional item nonresponse by survey participants. In doing so, this study examines the accuracy of estimates under a wide range of scenarios, implying the manipulation of the item nonresponse mechanism and of the proportion of both planned missingness and item nonresponse.

The most important conclusion from this paper is that the combined proportions of missing data jointly produced by item nonresponse and planned missingness may lead to partly strong biases in univariate and bivariate estimates. This can be the case even if the item nonresponse (like the planned missingness) is missing completely at random. Especially combined proportions of missing data exceeding 40% proved challenging. This finding is noteworthy, also because the literature on split questionnaire designs often reports reductions in questionnaire length that are significantly higher, such as 50% (Bahrami et al., 2014), up to 60% (Raghunathan and Grizzle, 1995), or even up to 74% (Adigüzel and Wedel, 2008). Yet, readers should be aware that this number may be specific to the data used for this study and should be evaluated with different data in the future. Moreover, as the level of item nonresponse varies heavily between different items, some items (those with the highest amount of item nonresponse) are more affected by large amounts of planned missingness than others.

A less concerning finding from this paper is that the heterogeneity of the missingness mechanism as introduced by planned missingness and item nonresponse occurring jointly seems not to be a particular problem. On the one hand, having item nonresponse that is missing at random (instead of missing completely at random) seems to increase biases only slightly, and this increase does not get larger with increasing amounts of planned missing data. Item nonresponse that is missing not at random, on the other hand, turns out to be a challenge irrespective of the amount of planned missing data.

In consequence, the expected degree of item nonresponse on each item must be taken into account in the process of designing split questionnaires. If an item can be expected to show a considerable amount of item nonresponse, questionnaire developers may be well advised to reduce the amount of planned missingness on this item or even allocate it to a core module.

6.1.4 Paper IV

The fourth and final paper of this dissertation investigated the effects of a general purpose imputation strategy on estimates compared to an analysis-specific imputation strategy given the procedures identified in the previous papers. Two central differences of both strategies were highlighted: In contrast to analysis-specific imputation, a general purpose imputation strategy offers neither the flexibility to restrict the imputation to the analysis sample nor to the analysis variables. To compare the effect of general purpose imputation with analysis-specific imputation on analysis models, this paper also went beyond evaluating univariate and bivariate estimates towards a case study using a multiple regression model from substantive social scientific research. The results suggest no detrimental effect of larger numbers of predictor variables in a general purpose imputation model, but partly considerable biases when imputation model and analysis model are based on a different sample. Furthermore, a non-continuous effect of age as well as a family status interaction effect were not adequately reproduced in either of the scenarios.

This research provides evidence in which situations general purpose imputations might work for social research based on a split questionnaire design and in which situations an analysis-specific imputation strategy may be superior. In sum, general purpose imputations can be expected to work well for analyses relying on the full survey sample that attempt to model continuous effects only. Analysis-specific imputation therefore is preferable either when a subgroup of the whole survey sample is to be analyzed, or if the estimated effects are expected to be strongly non-linear (e.g., interaction effects, squared terms, dummification of continuous variables). Moreover, this implies that if general purpose imputations are supplied by the data provider, the documentation should emphasize clearly how the imputed data can be used and in which cases data users would have to impute the data by themselves tailored specifically to their analysis.

Yet, these conclusions relate to the status quo imputation procedures used in this paper and should be considered subject to constant change. Specifically, the weaknesses of general purpose imputations revealed in this paper may trigger further development in imputation procedures to overcome these obstacles. For instance, if partial least squares predictive mean matching could be refined such that it allows for detecting non-linear effects more accurately while maintaining its efficiency in capturing linear effects, one might potentially be able to alleviate all the constraints of general purpose imputations discussed above to some extent. At the same time, it is important to remember that this analysis is a case study of only one multivariate

model. In consequence, future research should also evaluate other types of analysis models regarding the accuracy of general purpose imputations. For example, it should be elaborated whether the conclusions from this study also hold in a logistic regression model.

6.2 Contribution to the Literature

The present dissertation makes a significant contribution to our knowledge about split questionnaire designs by addressing several gaps in previous research.

First, most previous research on split questionnaire designs and planned missingness had focused on the development of methods (e.g., Raghunathan and Grizzle, 1995; Rässler et al., 2002; Thomas et al., 2006; Adigüzel and Wedel, 2008; Bahrami et al., 2014), but we had limited evidence on how these methods perform in the real world with actual social survey data. In contrast, the present dissertation provides a thorough examination of split questionnaire design techniques. To this end, using Monte Carlo simulation studies and real social survey data allows evaluating such techniques robustly while also accounting for the challenges of real data. In doing so, this work contributes to closing the gap between methods development and application in empirical research. Therefore, this dissertation significantly enhances the current understanding of how well multiple imputation may perform in a context with social survey data from a split questionnaire design in practice, especially for general purpose imputation strategies.

Second, despite intense research activity regarding the development of methods and testing them using simulation techniques, previous research had offered little comparison of how imputation performs with different methods and scenarios. For instance, the few previously existing small-scale attempts to compare different item-allocating strategies under real-data conditions had not involved evaluating estimates regarding their accuracy after imputation (see Adigüzel and Wedel, 2008; Rässler et al., 2002). Furthermore, while some previous research on imputation (such as Akande et al., 2017; Slade and Naylor, 2020; Wu and Leung, 2017) had compared the performance of different imputation procedures in general, these studies commonly do not account for the challenges and scenarios specific to social survey data from a split questionnaire design. This dissertation contributes to closing this gap by thoroughly evaluating different modularization strategies, imputation procedures, and scenarios regarding the amounts of planned missing data and item nonresponse.

This reveals under which conditions and strategies one may expect the imputation to yield acceptable results: Modules should not consist of single topics, and the overall proportion of missing data should be kept at moderate levels in spite of the planned missingness. If these conditions are met, a general purpose imputation strategy may succeed with predictive mean matching if the imputation model is simplified to some extent (e.g., through a partial-least-squares regression), and for subsequent analyses that model continuous effects based on the whole survey sample. For more complex analyses that restrict the analysis sample or model strongly non-continuous effects, an analysis-specific imputation may be needed.

Finally, this dissertation also contributes to the more general body of research on multiple imputation. Most importantly, it provides a comprehensive performance comparison of various imputation procedures using real survey data. This includes one of the first empirical evaluations of partial least squares predictive mean matching (Robitzsch et al., 2016; Robitzsch and Grund, 2021). Moreover, this dissertation also demonstrates potential limitations of imputation models that have a "broad scope" (van Buuren, 2018), such as general purpose imputation models, in a more general context beyond planned missingness research. For example, when item nonresponse is accounted for by providing data users with already imputed data (e.g., U.S. Census Bureau, 2022; Frick and Grabka, 2005), it might be worthwhile to think about how estimates may be affected if data users restrict their analysis samples. Therefore, this dissertation emphasizes the importance of transparent communication on this issue to ensure the value and integrity of social survey research with imputed data.

6.3 Implications for Future Research

The findings obtained in this work contribute implications for future survey research in a twofold manner: First, they may affect the way when and how one may choose to implement planned missingness in a survey; and second, they may stimulate further methodological research on the design and imputation of planned missingness in surveys.

6.3.1 Future Implementations of Planned Missingness

First, the imputation of planned missing data in surveys has proved to be a challenging endeavor. Even with the most suitable imputation procedures, in practice

the imputation can still be some source of bias for survey estimates. Therefore, researchers who consider implementing a split questionnaire design (or similar approach) should carefully assess whether the expected benefits in response rates and response quality are significant enough to make up for this additional challenge.

Second, when implementing a split questionnaire design with the aim of imputing the data subsequently, the module construction method should factor in the subsequent imputation. Here, it is important to resist the intuitive notion of constructing modules along the boundaries of survey topics. Rather, modules should cover a diverse range of the different survey topics to allow for an appropriate imputation. For instance, this could be achieved by randomly allocating items to modules. Meanwhile, the evidence of this work suggests no clear advantages of any more sophisticated methods for optimizing modules with respect to the imputation. Apart from this, in this dissertation, imputation procedures suitable for imputing nominal variables have not performed favorably. Hence, researchers may consider avoiding planned missing data on such items, for example by allocating them to a core module.

Also in the planning stage of the data collection, researchers should make sure that the split questionnaire design does not cause too many missing values per variable. This primarily entails limiting the number of modules missing per questionnaire form. However, expectations about item nonresponse by the participants should also be taken into account in this process, ensuring sufficient case numbers even with unplanned missing data.

Finally, when the data are to be provided with general purpose imputations to ease the analysis for data users, researchers involved in the imputation may take account of some key findings from this work. Most importantly, a general purpose imputation strategy would need to reduce complexity in imputation models to some extent. At least for imputing ordered categorical data, partial least squares predictive mean matching has shown promising behavior for this purpose in this dissertation. Restricted predictor sets may be an alternative if one is willing to assume that the near-zero correlations in the data correspond to true null relationships in the population. Researchers should also take care to communicate clearly how the imputed data can be analyzed. As discussed above, partial least squares predictive mean matching ,for instance, works primarily well for estimating continuous effects based on the whole survey sample. For estimating non-continuous effects or analyzing a subgroup of the survey sample, data users would need to impute the

data anew using an analysis-specific imputation strategy that includes the non-linear terms explicitly in the imputation model.

6.3.2 Future Methodological Research

The findings and limitations of this dissertation may also help to direct the focus of future methodological research on new important aspects.

First, all studies in this work are entirely simulation-based. This is a useful strategy to isolate effects of different strategies under controllable conditions and also ensures that clear and valid benchmarks exist for evaluating their performance. However, simulations can never fully imitate the real world. Thus, it would be intriguing to see how the different strategies perform in real-world applications of split questionnaire designs. This implies the need for experimental studies, for example testing different modularization strategies against each other. Such a field study may also include a proper investigation of the effects of different split questionnaire designs on relevant response-related measures, such as response rates, breakoff, and measurement error, and evaluate a split questionnaire design's benefits on these domains compared to the information loss from planned missingness.

Second, the findings obtained in this work might depend significantly on the parameter settings of the simulation study. However, it is simply not feasible to manipulate all possible dimensions at the same time. Perhaps most importantly, this includes the data used for the analyses. Thus, the methodology of split questionnaire designs may benefit hugely from replications with different datasets or diverging specifications regarding module construction and imputation procedures. Future research may also cover new dimensions such as different definitions of core modules or varying numbers of modules assigned to each participant.

Third, future research might also extend our knowledge about the performance of split questionnaire designs with measures other than bias and variability of univariate frequency estimates, bivariate Spearman correlation estimates, and multiple regression estimates as used in this work. For example, future research may study the accuracy of inference statistics with data from split questionnaire designs, such as t tests or confidence intervals, in more detail.

Fourth, this dissertation demonstrates the need for further development in imputation procedures for social survey data. This refers not only to imputation scenarios with planned missing data, but more generally to survey data imputation scenarios in which lots of predictor variables need to be accounted for. For instance, there is a pressing need for more suitable imputation procedures for nominal data that can

adequately preserve relations to other variables in such a demanding imputation scenario. Future research may therefore explore whether partial least squares predictive mean matching or a related method can be generalized to categorical data. Furthermore, even with ordinal or continuous data, detecting strongly non-linear relations while also preserving all the linear relations in the data remains a challenge for general purpose imputation strategies still to be resolved.

Fifth, the focus of this work is on the relatively simple case of non-hierarchical, cross-sectional data. Especially longitudinal data may behave differently especially regarding the imputation of planned missing data, since here the imputation needs to account for relations between variables not only at one but multiple points of time. Pioneering research on this domain has already been done by Imbriano and Raghunathan (2020) particularly regarding the construction of modules with longitudinal data. However, optimal strategies for the imputation of longitudinal or multilevel planned missing data still need to be identified.

Finally, this dissertation has dealt with planned missing data from a multiple imputation perspective. However, there are different methods for coping with missing data as well. In particular, future research might want to investigate to what degree the findings from the present work can be generalized to other methods such as single imputation or full information maximum likelihood techniques.

References

Adigüzel, F. & Wedel, M. (2008). Split questionnaire design for massive surveys. *Journal of Marketing Research*, *45*(5), 608-617.

Akande, O., Li, F. & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, *71*(2), 162-170.

Bahrami, S. (2020). *Missing by Design Patterns for Optimizing Survey Response by Efficient and Consistent Data Collection* [Doctoral dissertation]. University of Bamberg.

Bahrami, S., Aßmann, C., Meinfelder, F., & Rässler, S. (2014). A split questionnaire survey design for data with block structure correlation matrix. In Engel, U., Jann, B. Lynn, P. , Scherpenzeel, A. & Sturgis, P. (Eds.), *Improving survey methods: Lessons from recent research* (pp. 368-380). Routledge.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462-487.

Burgette, L. F. & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070-1076.

Chipperfield, J. O. & Steel, D. G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, *25*(2), 227-244.

Chipperfield, J. O. & Steel, D. G. (2011). Efficiency of split questionnaire surveys. *Journal of Statistical Planning and Inference*, *141*(5), 1925-1932.

Dillman, D. A., Sinclair, M. D., & Clark, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*, *57*(3), 289-304.

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

Frick, J. R. & Grabka, M. M. (2005). Item-non-response on income questions in panel surveys: Incidence, imputation and the impact on the income distribution. *Allgemeines Statistisches Archiv*, *89*(1), 49-61.

Heberlein, T. A. & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review*, *43*(4), 447-462.

Imbriano, P. (2018). *Methods for improving efficiency of planned missing data designs* [Doctoral dissertation]. University of Michigan, Ann Arbor.

Imbriano, P. M. & Raghunathan, T. E. (2020). Three-form split questionnaire design for panel surveys. *Journal of Official Statistics*, *36*(4), 827-854.

Lee, K. J. & Carlin, J. B. (2010). Multiple imputation in the presence of non-normal data. *Statistics in Medicine*, *36*(4), 624-632.

Luijkx, R., Jónsdóttir, G. A., Gummer, T., Ernst Stähli, M., Fredriksen, M., Reeskens, T., Ketola, K., Brislinger, E., Christmann, P., Gunnarsson, S. Þ., Hjaltason, Á. B., Joye, D., Lomazzi, V., Maineri, A. M., Milbert, P., Ochsner, M.,

Ólafsdóttir, S., Pollien, A., Sapin, M., ... Wolf, C. (2021). The European Values Study 2017: On the way to the future using mixed-modes. *European Sociological Review*, *37*(2), 330-346.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(75).

OECD (2014). *PISA 2012 Technical Report*. OECD.

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

Rässler, S., Koller, F. & 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*. https://www.statistik.rw.fau.de/files/2016/03/d0042b.pdf

Revilla, M. and Höhne, J. K. (2020). How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research*, *62*(5), 538-545.

Robitzsch, A. & Grund, S. (2021). *miceadds: Some additional multiple imputation functions, especially for 'mice'*. R package version 3.11-6. https://CRAN.R-project.org/package=miceadds

Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der Österreichischen Bildungsstandardüberprüfung [Methodological foundation of standard achievement testing]* (pp. 259-293). facultas.

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

Slade, E. & Naylor, M. G. (2020). A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Statistics in Medicine*, *39*(8), 1156-1166.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the national health and nutrition examination survey. *Survey Methodology*, *32*(2), 217-231.

U.S. Census Bureau (2022). *2021 Survey of Income and Program Participation users' guide*. https://www2.census.gov/programs-surveys/sipp/tech-documentation/methodology/2021_SIPP_Users_Guide_AUG22.pdf

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC press.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

Wu, H. & Leung, S.O. (2017). Can Likert scales be treated as interval scales?—A simulation study. *Journal of Social Service Research*, *43*(4), 527-532.