*Journal of Survey Statistics and Methodology (2022) **10**, 1236–1262*

SPLIT QUESTIONNAIRE DESIGNS FOR ONLINE SURVEYS: THE IMPACT OF MODULE CONSTRUCTION ON IMPUTATION QUALITY

JULIAN B. AXENFELD*
ANNELIES G. BLOM
CHRISTIAN BRUCH
CHRISTOF WOLF

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

JULIAN B. AXENFELD is a Research Associate at the Mannheim Centre for European Social Research (MZES), University of Mannheim, 68131 Mannheim, Germany. ANNELIES G. BLOM is a Professor of Data Science at the Department of Political Science, School of Social Sciences, the Principal Investigator of the German Internet Panel (GIP) at the Collaborative Research Center (SFB 884) "Political Economy of Reforms"; and a Project Director at the Mannheim Centre for European Social Research (MZES) at the University of Mannheim, 68131 Mannheim, Germany. She also is an Associate Professor of Political Science at the Digital Social Science Core Facility (DIGSSCORE), Department of Administration and Organization Theory, University of Bergen, Norway. CHRISTIAN BRUCH is a Postdoctoral Researcher at the GESIS Leibniz Institute for the Social Sciences, B6, 4-5, 68159 Mannheim, and External Fellow and Project Director at the Mannheim Centre for European Social Research (MZES), University of Mannheim, 68131 Mannheim, Germany. CHRISTOF WOLF is the President of GESIS Leibniz Institute for the Social Sciences, B6, 4-5, 68159 Mannheim, and a Professor for Sociology at the Department of Sociology, School of Social Sciences, and a Project Director at the Mannheim Centre for European Social Research (MZES), University of Mannheim, 68131 Mannheim, Germany.

The authors like to thank Azim Selvi and Lisa Wellinghoff for their assistance in preparing the manuscript. This study design and analysis was not preregistered.
This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) [project numbers: BL 1148/1–1, BR 5869/1–1, WO 739/20–1]. The Monte Carlo simulations were run on the High Performance Computing facilities of the state of Baden-Württemberg (bwHPC). This paper uses data from the German Internet Panel (GIP) funded by the DFG through the Collaborative Research Center (SFB) 884 "Political Economy of Reforms" (SFB 884) [Project-ID: 139943784].
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
*Address correspondence to Julian B. Axenfeld, Mannheim Centre for European Social Research (MZES), University of Mannheim, 68131 Mannheim, Germany; E-mail: julian.axenfeld@uni-mannheim.de

https://doi.org/10.1093/jssam/smab055 Advance access publication 11 April 2022

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

KEYWORDS: Long surveys; Missing data; Monte Carlo simulation; Multiple imputation; Split Questionnaire Design.

Statement of Significance

Face-to-face surveys—until recently considered the gold standard for high-quality surveys—are increasingly replaced by online surveys requiring shorter questionnaires. Split questionnaire designs can be used to reduce individual questionnaire length while collecting data on questions from a longer questionnaire. The resulting dataset contains a large share of intentionally missing data. One way to analyze these data is to replace the missing values by imputing them. This study compares different strategies for constructing split questionnaires with respect to the quality of the imputed datasets. More precisely, we study how different ways to distribute questions to respondents affect the quality of imputed data regarding frequencies of and correlations between variables. We find that distributing questions belonging to the same topic across different questionnaires improves imputation quality.

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, Davern, Freese, and Morgan 2019) or the British Social Attitudes survey (Curtice, Clery, Perry, Phillips, and Rahim 2019) stimulate scientific discourse with high-quality data. However, face-to-face surveys are increasingly under pressure due to decreasing response rates (De Leeuw, Hox, and Luiten 2018) and increasing costs (e.g., Calinescu, Bhulai, and Schouten 2013; Roberts, Vandenplas, and Ernst Stähli 2014).

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, Biffignandi, and Lynn 2017; Olson et al., 2021). Several large-scale probability-based online surveys have been established across the world (e.g., the KnowledgePanel in the United States (Ipsos 2021), the LISS Panel in the Netherlands (Knoef and de Vos 2009), and the German Internet Panel (GIP; Blom, Gathmann, and Krieger 2015)).

Consequently, survey projects face pressures to switch to the less expensive online mode (e.g., Jäckle, Lynn, and Burton 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; Peytchev 2009; Tourangeau, Conrad, and Couper 2013, p. 52; Mavletova and Couper 2015; Revilla 2017), 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 considerably longer at approximately 1 hour (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 multitopic questionnaires and nonexact 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 (nonsynthetic) 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. ADMINISTRATION OF SQDs

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, Koller, and Mäenpää 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 Multiple Imputation

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, Hughes, Tilling, and Heron 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.

2.3 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, that is, 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, Oldendick, and Tuchfarber 1983 or Kreuter, McCulloch, Presser, and Tourangeau 2011 for question-filter effects on data quality), prevent differential order effects (e.g., McFarland 1981; Silber, Höhne, and Schlosser 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 1 illustrates these three strategies with a small example questionnaire.

2.3.1 Random modules. The upper part of figure 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

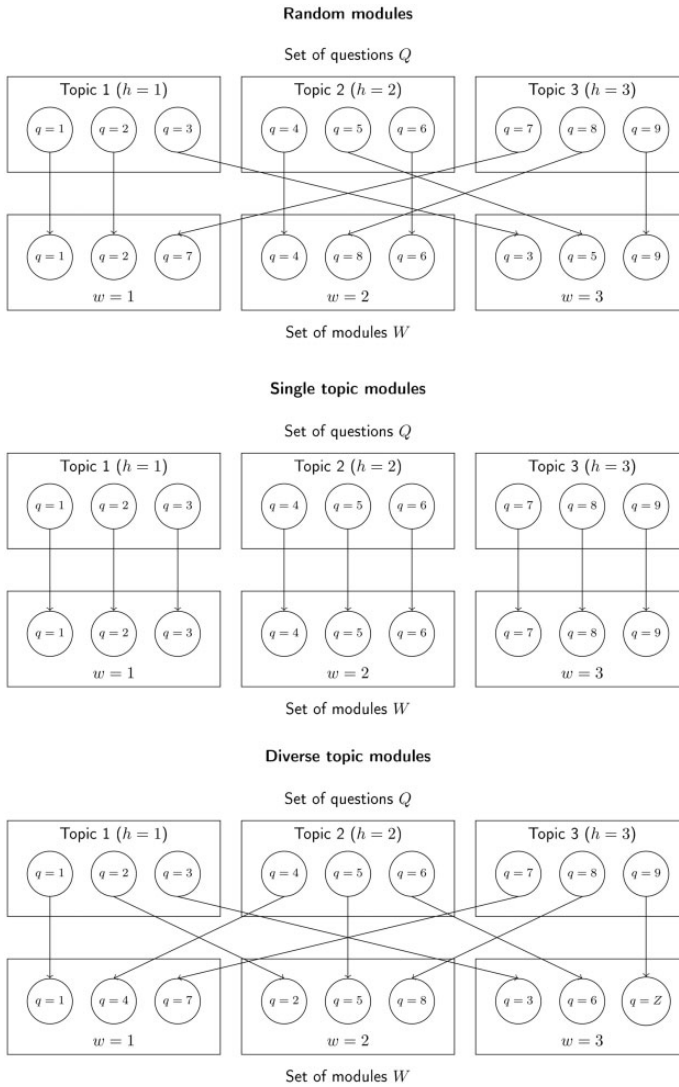

Figure 1. Illustration of modularization strategies.

$< M \leq Z$ (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, Raghunathan, Schenker, Katzoff, and Johnson 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 $M(Z/M - \lfloor Z/M \rfloor)$ modules. For these two subsets, we define different module sizes:

$$B_w = \begin{cases} \lceil Z/M \rceil & if \;\; w \in V \\ \lfloor Z/M \rfloor & if \;\; w \notin V \end{cases}. \tag{1}$$

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.

2.3.2 Single topic modules. STM's procedure is illustrated in the middle part of figure 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. 2021) and the 2012 PISA study (OECD 2014, Chapter 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.

2.3.3 Diverse topics modules. 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, Barr, and

Steel 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, that is, 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, Aßmann, Meinfelder, and Rässler (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 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 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:

$$B_{w,h} = \begin{cases} \lceil A_h/M \rceil & \text{if } w \in U_h \\ \lfloor A_h/M \rfloor & \text{if } w \notin U_h \end{cases}. \tag{2}$$

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 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.

2.4 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 (Raghunathan and Grizzle 1995; Thomas et al. 2006; Bahrami et al. 2014) and case studies that explore different modularization strategies (Rässler et al. 2002; Adigüzel and Wedel 2008; Imbriano and Raghunathan 2020).

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 regression 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, or Raghunathan and Grizzle 1995, study 1).

3. DATA AND METHODS

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, Felderer, Fikel, Krieger, and Blom 2021). The GIP is particularly suited, because it has a reasonably large number of cases (5,411) and a multitopic 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, 2019b). Table 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 A.1 and A.2 in the online supplementary materials.

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, Bossert, Funke et al., 2016a; Blom, Bossert, Gebhard et al., 2016b). 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 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 A.1 in the online supplementary materials).

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.

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 A.2 in the online

1. Other R packages used for this paper are as follows: DescTools (Signorell et al. 2020), doMPI (Weston 2017), dplyr (Wickham, François, Henry, and Müller 2021), faux (DeBruine 2020), foreach (Microsoft and Weston 2020), ggcorrplot (Kassambara 2019), MASS (Venables and Ripley 2002), Matrix (Bates and Maechler 2019), Rmpi (Yu 2002), and tidyverse (Wickham et al. 2019).

Table 1. Variables Used in Monte Carlo Simulation

Topic	No. of 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 and 5

supplementary materials for an illustration). Absolute values of correlations range from 0.000 to 0.702 with 81.6 percent smaller than 0.1. 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 percent of within-topic correlations and 89.8 percent of between-topic correlations are below 0.1.

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.

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.

3.3.1 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

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).

to their stochastic procedure, STMs are predefined (see table 1) and thus do not vary.

3.3.2 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 percent of the original size.

3.3.3 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 the 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 B.1 in the online supplementary materials for details). Considering the possibility that the poor performance of the default methods could be due to a violation of the proportional-odds assumption for ordinal logistic regression, we also tested polytomous logistic regression as an alternative method (also displayed in figure B.1 in the online supplementary materials). However, the shifts in estimates with this imputation method seem even somewhat larger than with the mice default. These findings comply with prior research revealing difficulties with imputation using categorical regression methods (White, Royston, and Wood 2011; Wu, Jia, and Enders 2015; Van Buuren 2018, p. 91) and recommending PMM at least as a fallback option (Koller-Meinfelder 2009, pp. 48–68; Van Buuren 2018, p. 166).

Small-scale tests also showed that restricting imputation models to predictor variables with Spearman correlations stronger than $|0.1|$ in the nonimputed 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 B.2 and B.3 in the online supplementary materials, 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.

3.3.4 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 ($\widehat{\pi}_{c,s}^{complete}$) and imputed data ($\widehat{\pi}_{c,s}^{imputed}$).

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 $\widehat{\rho}_{i,j,s}^{complete}$ for the complete sample data and $\widehat{\rho}_{i,j,s}^{imputed}$ 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.

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:

$$\widehat{\Delta}\left(\widehat{\pi}_{c,s}\right) = \widehat{\pi}_{c,s}^{imputed} - \widehat{\pi}_{c,s}^{complete}, \tag{3}$$

$$\widehat{\Delta}\left(\widehat{\rho}_{i,j,s}\right) = \widehat{\rho}_{i,j,s}^{imputed} - \widehat{\rho}_{i,j,s}^{complete}. \tag{4}$$

3. Dividing $\hat{\Delta}$ 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.

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

3.4.1 Bias. If a given estimate is Monte Carlo unbiased, we expect the average of its deviations $\bar{\widehat{\Delta}}$ 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 $\widehat{\pi}$ for a category c is obtained through the average over its deviations in all $S = 1,007$ simulation runs:

$$\bar{\widehat{\Delta}}(\widehat{\pi}_c) = \frac{1}{S} \sum_{s=1}^{S} \widehat{\Delta}(\widehat{\pi}_{c,s}). \tag{5}$$

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

$$\bar{\widehat{\Delta}}(\widehat{\rho}_{i,j}) = \frac{1}{S} \sum_{s=1}^{S} \widehat{\Delta}(\widehat{\rho}_{i,j,s}). \tag{6}$$

3.4.2 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 one 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; Mashreghi, Léger, and Haziza 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 (3) and (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 C.1 and C.2 in the online supplementary materials) 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) $\widehat{\sigma}\left\{\widehat{\Delta}(\widehat{\pi}_c)\right\}$:

$$\widehat{\sigma}\left\{\widehat{\Delta}(\widehat{\pi}_c)\right\} = \sqrt{\frac{1}{S-1} \sum_{s-1}^{S} \{\widehat{\Delta}(\widehat{\pi}_{c,s}) - \bar{\widehat{\Delta}}(\widehat{\pi}_c)\}^2}. \tag{7}$$

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:

$$\widehat{\sigma}\left\{\widehat{\Delta}(\widehat{\rho}_{i,j})\right\} = \sqrt{\frac{1}{S-1}\sum_{s-1}^{S}\{\widehat{\Delta}(\widehat{\rho}_{i,j,s}) - \widehat{\bar{\Delta}}(\widehat{\rho}_{i,j})\}^2}. \quad (8)$$

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.

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.

4. RESULTS

4.1 Univariate Frequencies

Figure 2 displays the distribution of average Monte Carlo biases of univariate frequencies for the imputed data for RM (first boxplot), STM (second boxplot), and

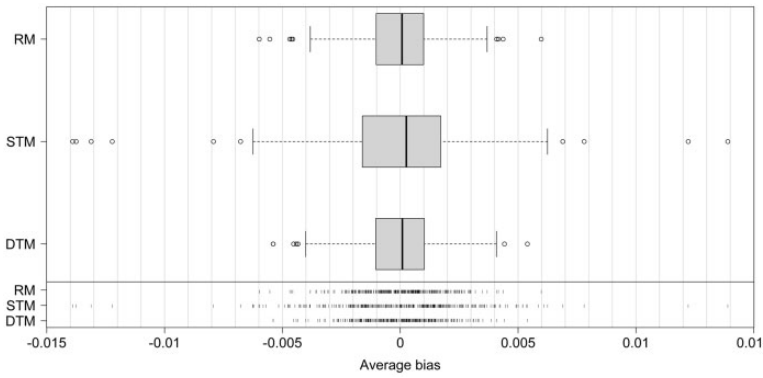

Figure 2. Average biases for 297 univariate frequencies according to (5), by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM). Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40 percent missing data) each.

DTM (third boxplot). The rug plots in the second section of figure 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 percent 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 percent of biases range from −0.004 to +0.003 with outliers of up to ±0.014.

Figure 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 preferable, 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.

4.2 Bivariate Correlations

Figure 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 percent of average biases range from −0.006 to +0.006, 90 percent from −0.017 to +0.017, and the most extreme bias is

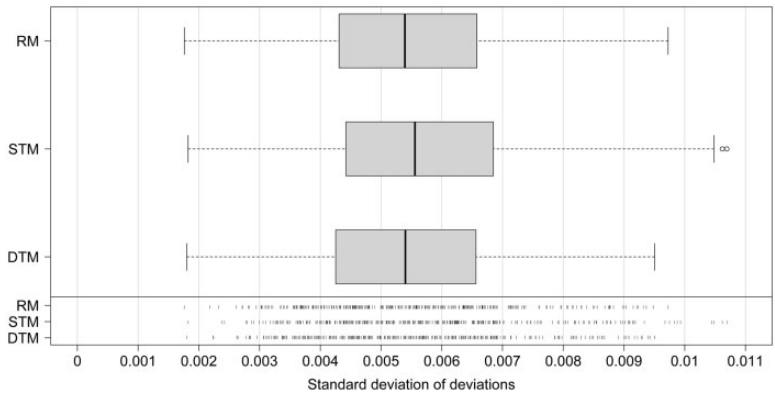

Figure 3. Standard deviations of deviations (SDDs) of 297 univariate frequencies according to (7), by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM). Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40 percent missing data) each.

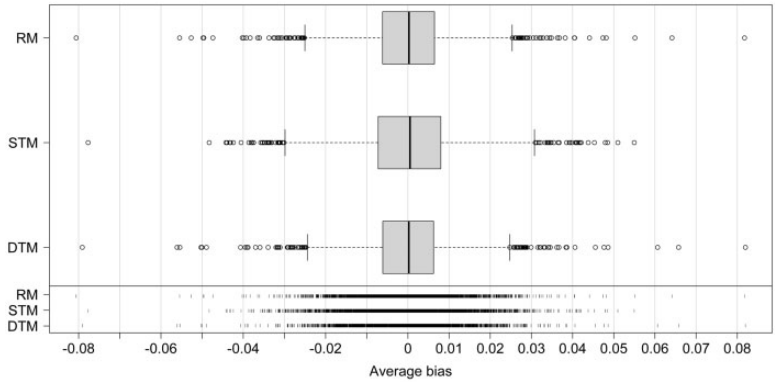

Figure 4. Average biases of 3,675 bivariate correlations according to (6), by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM). Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40 percent missing data) each.

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 percent, 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 percent. Furthermore, the rug plots also show some average biases in the area closely around zero. STM has a different pattern: 50

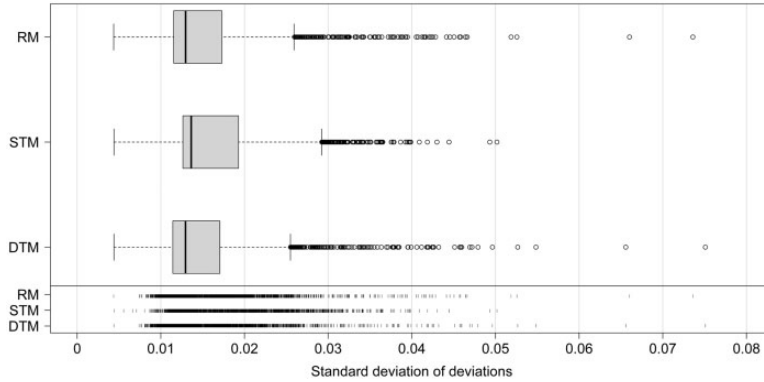

Figure 5. Standard deviations of deviations (SDDs) of 3,675 bivariate correlations according to (8), by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM). Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40 percent missing data) each.

percent range from −0.007 to +0.008 and 90 percent from −0.020 to +0.020. Furthermore, STM produces fewer extreme outliers larger than ±0.05 (three correlations) than RM (six correlations) and DTM (eight correlations).

Figure 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.

4.2.1 Analysis by topic. To further investigate effects of the modularization on biases in bivariate correlations, figure 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 percent of average biases with RM and DTM are between −0.008 and +0.009. Biases with STM are larger with 50 percent 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 percent of average biases with RM and DTM are between −0.015 and +0.005 and 50 percent of biases with STM between −0.009 and +0.007. STM leads to fewer extreme biases of larger than ±0.05 (two with STM, 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 percent of within-topic correlations have biases smaller than zero, 60.0 percent with STM and 68.0 percent with DTM.

Figure 7 shows the sizes of SDDs for different-topic and within-topic correlations. For different-topic correlations, small SDDs are again less common with

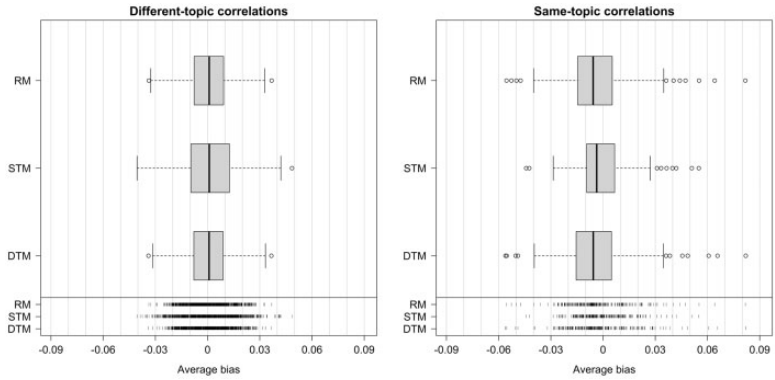

Figure 6. Average biases of 3,675 bivariate correlations according to (6), separating correlations of variables of different versus same topics, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM). Based on a Monte Carlo simulation with 1,007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

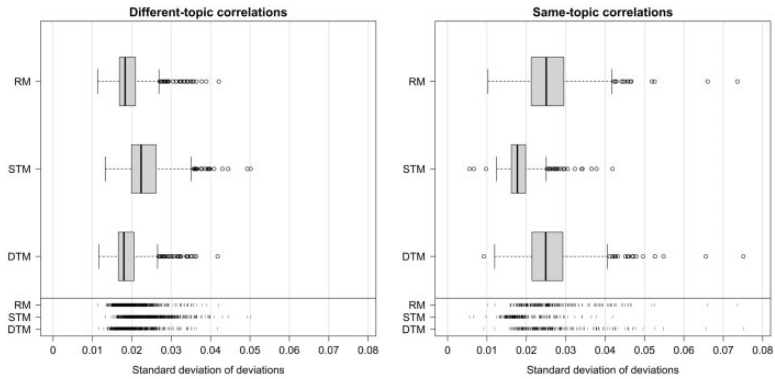

Figure 7. Standard deviations of deviations (SDDs) of 3,675 bivariate correlations according to (8), separating correlations of variables in different versus same topics, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM). Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40 percent missing data) each.

STM than with RM or DTM: With RM and DTM, the majority of SDDs are smaller than 0.02, while with STM, the majority of SDDs are larger than 0.02. For same-topic correlations, however, STM tends to produce smaller SDDs.

4.2.2 Subset by representation in the imputation models. Figure 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

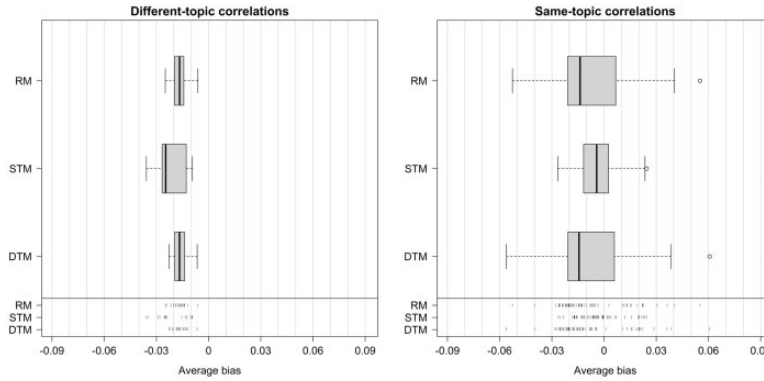

Figure 8. Average biases of 72 bivariate correlations according to (6) for correlations represented in every imputation model throughout the simulation, separately for correlations of variables of different versus same topics, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM). Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40 percent missing data) each.

(72 correlations) of all correlations. These correlations are generally stronger, as imputation models only included correlations stronger than 0.1. 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 73.2 percent (RM and DTM) and 71.4 percent (STM) of same-topic correlations are underestimated.

Fifty percent 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 percent 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 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.

4.3 Alternative Correlation Structures

In contrast to our expectations, DTM and RM generally performed similarly. 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

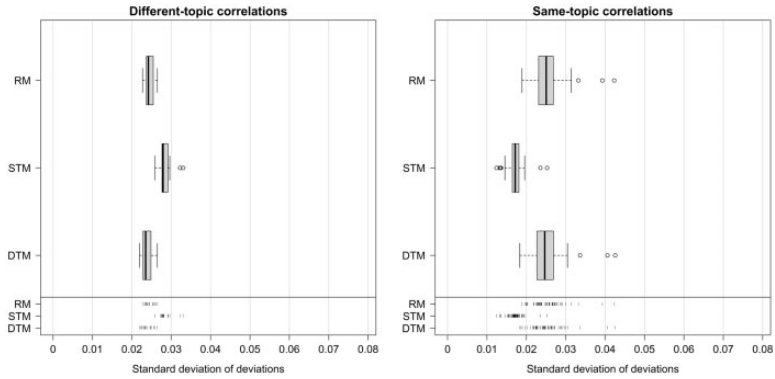

Figure 9. Standard deviations of deviations (SDDs) of 72 bivariate correlations according to (8), for correlations represented in every imputation model throughout the simulation, separately for correlations of variables in different versus same topics, by modularization technique: Random modules (RM), single topic modules (STM), and diverse topics modules (DTM). Based on a Monte Carlo simulation with 1,007 runs on 2,000 cases (40 percent 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 in the online supplementary materials for a description of the data-generating process). Scenario 1 (control condition) largely adopts the original correlation structure but with maximum correlations of $|0.2|$. Scenario 2 is the same except for one same-topic correlation per imputed variable increased towards ± 0.9.

Results (see Figures D.1 through D.4 in the online supplementary materials) indeed show somewhat smaller 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.

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.

Correlations tend to be attenuated. Most correlations that are positive in the population data have biases smaller than zero (RM: 81.0 percent; STM: 81.0 percent; DTM: 81.5 percent). However, most correlations that are *negative* in the population data have biases larger than zero (RM: 84.2 percent; STM: 86.1 percent; DTM: 83.9 percent). (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.

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), although 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 to be 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. In particular, Relationships between variables are not fully retained. This finding is compatible with Bahrami et al. (2014), who report small downwards slants in regression estimates. 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.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

Adigüzel, F., and M. Wedel (2008), "Split Questionnaire Design for Massive Surveys," *Journal of Marketing Research*, 45, 608–617.

Bahrami, S., C. Aßmann, F. Meinfelder, and S. Rässler (2014), "A Split Questionnaire Survey Design for Data with Block Structure Correlation Matrix," in *Improving Survey Methods: Lessons from Recent Research*, ed. U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis, pp. 368–380, New York: Routledge.

Bates, D., and M. Maechler (2019), "*Matrix*: *Sparse and Dense Matrix Classes and Methods*," *R Package Version*, 1, 2–18.

Bianchi, A., S. Biffignandi, and P. Lynn (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, 385–408.

Bishop, G. F., R. W. Oldendick, and A. J. Tuchfarber (1983), "Effects of Filter Questions in Public Opinion Surveys," *Public Opinion Quarterly*, 47, 528–546.

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

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

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

———. (2019b), SFB 884 "Political Economy of Reforms," Universität Mannheim, *German Internet Panel, Wave 38 (November 2018)*, Cologne: GESIS Data Archive, ZA6958 Data file Version 1.0.0, DOI: 10.4232/1.13391.

Blom, A. G., C. Gathmann, and U. Krieger (2015), "Setting up an Online Panel Representative of the General Population: The German Internet Panel," *Field Methods*, 27, 391–408.

Blom, A. G., J. M. E. Herzing, C. Cornesse, J. W. Sakshaug, U. Krieger, and D. Bossert (2017), "Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence from the German Internet Panel," *Social Science Computer Review*, 35, 498–520.

Bruch, C. (2016), *Varianzschätzung Unter Imputation Und Bei Komplexen Stichprobendesigns*, Trier: University of Trier.

Calinescu, M., S. Bhulai, and B. Schouten (2013), "Optimal Resource Allocation in Survey Designs," *European Journal of Operational Research*, 226, 115–121.

Chipperfield, J. O., and D. G. Steel (2009), "Design and Estimation for Split Questionnaire Surveys," *Journal of Official Statistics*, 25, 227–244.

———. (2011), "Efficiency of Split Questionnaire Surveys," *Journal of Statistical Planning and Inference*, 141, 1925–1932.

Chipperfield, J. O., M. L. Barr, and D. G. Steel (2018), "Split Questionnaire Designs: Collecting Only the Data That You Need through MCAR and MAR Designs," *Journal of Applied Statistics*, 45, 1465–1475.

Cornesse, C., B. Felderer, M. Fikel, U. Krieger, and A. G. Blom (2021), "Recruiting a Probability-Based Online Panel via Postal Mail: Experimental Evidence," *Social Science Computer Review*, DOI: 10.1177/08944393211006059.

Curtice, J., E. Clery, J. Perry, M. Phillips, and N. Rahim (eds.) (2019), *British Social Attitudes: The 36th Report*, London: National Centre for Social Research.

De Leeuw, E., J. Hox, and A. Luiten (2018), "International Nonresponse Trends across Countries and Years: An Analysis of 36 Years of Labour Force Survey Data," Survey Insights: Methods from the Field [online], available at https://surveyinsights.org/?p=10452.

DeBruine, L. (2020), *faux: Simulation for Factorial Designs [R package version 0.0.1.5]*, Vienna: R Foundation for Statistical Computing.

Galesic, M. (2006), "Dropouts on the Web: Effects of Interest and Burden Experienced during an Online Survey," *Journal of Official Statistics*, 22, 313–328.

Galesic, M., and M. Bosnjak (2009), "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey," *Public Opinion Quarterly*, 73, 349–360.

Gonzalez, J. M., and J. L. Eltinge (2007), "Multiple Matrix Sampling: A Review," in *JSM Proceedings, Survey Research Methods Section*, Alexandria, VA: American Statistical Association, 3069–3075.

Honaker, J., and G. King (2010), "What to Do about Missing Values in Time-Series Cross-Section Data," *American Journal of Political Science*, 54, 561–581.

Imbriano, P. (2018), "Methods for Improving Efficiency of Planned Missing Data Designs," Ph.D. dissertation, University of Michigan, Ann Arbor.

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

International Telecommunication Union (2019), *World Telecommunication/ICT Indicators Database* (23rd ed.), Geneva: ITUPublications,

Ipsos (2021), *KnowledgePanel* [online], available at https://www.ipsos.com/en-us/solutions/public-affairs/knowledgepanel.

Jäckle, A., P. Lynn, and J. Burton (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, 57–70.

Kassambara, A. (2019), *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2' [R package version 0.1.3]*, Vienna: R Foundation for Statistical Computing.

Knoef, M., and K. de Vos (2009), *The Representativeness of LISS, an Online Probability Panel*, Tilburg: CentERdata [online], available at 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*, Bamberg: University of Bamberg.

Kreuter, F., S. McCulloch, S. Presser, and R. Tourangeau (2011), "The Effects of Asking Filter Questions in Interleafed versus Grouped Format," *Sociological Methods & Research*, 40, 88–104.

Luijkx, R., G. A. Jónsdóttir, T. Gummer, M. Ernst Stähli, M. Fredriksen, T. Reeskens, K. Ketola, et al. (2021), "The European Values Study 2017: On the Way to the Future Using Mixed-Modes," *European Sociological Review*, 37, 330–346.

Madley-Dowd, P., R. Hughes, K. Tilling, and J. Heron (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., C. Léger, and D. Haziza (2014), "Bootstrap Methods for Imputed Data from Regression, Ratio and Hot-Deck Imputation," *Canadian Journal of Statistics*, 42, 142–167.

Mavletova, A., and M. P. Couper (2015), "A Meta-Analysis of Breakoff Rates in Mobile Web Surveys," in *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies*, ed. D. Toninelli, R. Pinter, and P. de Pedraza, pp. 81–98, London: Ubiquity Press.

McFarland, S. G. (1981), "Effects of Question Order on Survey Responses," *Public Opinion Quarterly*, 45, 208–215.

Microsoft and Weston, S. (2020), *foreach: Provides Foreach Looping Construct* [R package version 1.5.0], Vienna: R Foundation for Statistical Computing.

Munger, G. F., and B. H. Loyd (1988), "The Use of Multiple Matrix Sampling for Survey Research," *The Journal of Experimental Education*, 56, 187–191.

Münnich, R., and S. Rässler (2005), "PRIMA: A New Multiple Imputation Procedure for Binary Variables," *Journal of Official Statistics*, 21, 325–341.

OECD (2014), *PISA 2012 Technical Report*, Paris: OECD.

Olson, K., J. D. Smyth, R. Horwitz, S. Keeter, V. Lesser, S. Marken, N. A. Mathiowetz, et al. (2021), "Transitions from Telephone Surveys to Self-Administered and Mixed-Mode Surveys: AAPOR Task Force Report," *Journal of Survey Statistics and Methodology*, 9, 381–411.

Peytchev, A. (2009), "Survey Breakoff," *Public Opinion Quarterly*, 73, 74–97.

Peytchev, A., and E. Peytcheva (2017), "Reduction of Measurement Error Due to Survey Length: Evaluation of the Split Questionnaire Design Approach," *Survey Research Methods*, 11, 361–368.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing [online], available at https://www.R-project.org/.

Raghunathan, T. E., and J. E. Grizzle (1995), "A Split Questionnaire Survey Design," *Journal of the American Statistical Association*, 90, 54–63.

Rässler, S., F. Koller, and C. Mäenpää (2002), "A Split Questionnaire Survey Design Applied to German Media and Consumer Surveys," in *Friedrich-Alexander University Erlangen-Nuremberg, Chair of Statistics and Econometrics Discussion Papers* [online], available at https://www.statistik.rw.fau.de/files/2016/03/d0042b.pdf.

Revilla, M. (2017), "Analyzing Survey Characteristics, Participation, and Evaluation across 186 Surveys in an Online Opt-In Panel in Spain," *Methods, Data, Analyses*, 11, 135–162.

Revilla, M., and J. K. Höhne (2020), "How Long Do Respondents Think Online Surveys Should Be? New Evidence from Two Online Panels in Germany," *International Journal of Market Research*, 62, 538–545.

Roberts, C., C. Vandenplas, and M. Ernst Stähli (2014), "Evaluating the Impact of Response Enhancement Methods on the Risk of Nonresponse Bias and Survey Costs," *Survey Research Methods*, 8, 67–80.

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

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

Signorell, A., K. Aho, A. Alfons, N. Anderegg, T. Aragon, C. Arachchige, A. Arppe, et al. (2020), *DescTools: Tools for Descriptive Statistics* [R package version 0.99.36], Vienna: R Foundation for Statistical Computing.

Silber, H., J. K. Höhne, and S. Schlosser (2016), "Question Order Experiments in the German-European Context," *Survey Methods: Insights from the Field* [online], available at https://surveyinsights.org/?p=7645.

Smith, T. W., M. Davern, J. Freese, and S. L. Morgan (2019), *General Social Surveys, 1972-2018, Cumulative Codebook*, Chicago: NORC.

Thomas, N., T. E. Raghunathan, N. Schenker, M. J. Katzoff, and C. L. Johnson (2006), "An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey," *Survey Methodology*, 32, 217–231.

Tourangeau, R., F. G. Conrad, and M. P. Couper (2013), *The Science of Web Surveys*, Oxford: Oxford University Press.

Van Buuren, S. (2018), *Flexible Imputation of Missing Data*, Boca Raton, FL: CRC press.

Van Buuren, S., and K. Groothuis-Oudshoorn (2011), "Mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45, 1–67.

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

Weston, S. (2017), *doMPI: Foreach Parallel Adaptor for the Rmpi Package [R package version 0.2.2]*, Vienna: R Foundation for Statistical Computing.

White, I. R., P. Royston, and A. M. Wood (2011), "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice," *Statistics in Medicine*, 30, 377–399.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, et al. (2019), "Welcome to the Tidyverse," *Journal of Open Source Software*, 4(43), 1686.

Wickham, H., R. François, L. Henry, and K. Müller (2021), *dplyr: A Grammar of Data Manipulation [R package version 1.0.6]*, Vienna: R Foundation for Statistical Computing.

Wu, W., F. Jia, and C. Enders (2015), "A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables," *Multivariate Behavioral Research*, 50, 484–503.

Yu, H. (2002), "Rmpi: Parallel Statistical Computing in R," *R News*, 2(2), 10–14.