INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
https://doi.org/10.1080/13645579.2025.2561653

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Composition of core modules and item allocation in split questionnaire designs: impact on estimates from imputed data

Julian B. Axenfeld [a], Christian Bruch [b] and Christof Wolf [b,c]

[a]Research Infrastructure 'Socio-Economic Panel (SOEP)', DIW Berlin, Berlin, Germany; [b]Survey Design and Methodology, GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany; [c]School of Social Sciences, Department of Sociology, University of Mannheim, Mannheim, Germany

ABSTRACT
An increasing number of social science surveys use split questionnaire designs to reduce questionnaire length, presenting only a subset of several questionnaire modules to each respondent while leaving out others. This approach results in large amounts of planned missing data that necessitates imputation. Research shows that imputation is most effective when each module covers various topics. Yet, single-topic modules may sometimes be preferable from a questionnaire-design perspective. A potential alternative from survey practice is using single-topic modules with an extended core module presented to all respondents that includes key items from all topics. This study investigates whether this strategy yields outcomes comparable to mixed-topic modules. Using Monte-Carlo simulations based on the German Internet Panel, we simulate split questionnaire designs, impute the missing data, and calculate estimates based on these data. Findings suggest that while an extended core module improves single-topic module outcomes, it is inferior to randomly allocated mixed-topic modules.

ARTICLE HISTORY
Received 26 June 2024
Accepted 10 September 2025

KEYWORDS
Imputation; module construction; planned missing data; social surveys; split questionnaire design

1. Introduction

An increasing number of surveys are testing or implementing survey designs with planned missing data, especially split questionnaire designs (SQDs; Raghunathan & Grizzle, 1995), to reduce questionnaire length. SQDs entail allocating all items of a survey to several so-called split modules and presenting each respondent only a randomly selected subset of the modules. Thus, respondents do not receive the complete questionnaire. The downside of this procedure is that the modules not presented to respondents result in planned missing data. To make the data analyzable, Raghunathan and Grizzle (1995) propose to impute these planned missing data. Here, the imputation describes the procedure of assigning values to the (planned and unplanned) missing values of the variables of the dataset with a possible use of auxiliary variables to receive a

CONTACT Julian B. Axenfeld jaxenfeld@diw.de Research Infrastructure 'Socio-Economic Panel (SOEP)', DIW Berlin, Anton-Wilhelm-Amo-Straße 58, Berlin 10117, Germany

complete data set (see Kim & Rao, 2009; see also Bruch, 2023; Chauvet et al., 2011; Haziza, 2009).

An essential aspect of designing split questionnaires is how to construct the modules. This can be especially important for the imputation: To ensure an acceptable quality of the imputed data, highly correlated variables should be allocated to *different* split modules (Raghunathan & Grizzle, 1995). Then, imputation models can utilize the predictive information of the observed highly correlated data to impute a variable if it has not been presented to a respondent. Empirical research (Axenfeld et al., 2022a; Imbriano & Raghunathan, 2020; Rässler et al., 2002) generally supports this notion. In our research (Axenfeld et al., 2022a), we could show that modules consisting only of one single survey topic lead to a reduced quality of the imputed data in comparison to a random allocation or distributing questions of one topic over different split modules. This is due to the typically higher correlation between variables of the same topic (Axenfeld et al., 2022a).

From a practical perspective, however, allocating variables of the same topic to different modules might have negative effects on the quality of answers. As Smyth (2016, p. 229) notes, 'grouping topically similar questions gains efficiencies because respondents can use retrieved information to answer all questions on a topic before moving to a different topic'. Similarly, Krosnick and Presser (2010, p. 292) recommend to design questionnaires in such a way that 'items [...] flow coherently, which usually requires that items on related topics be grouped together.' In other words, survey respondents may perceive a split questionnaire as inconsistent if items belonging to the same topic appear in several places throughout the questionnaire. Thus, from a questionnaire design perspective single-topic modules may be preferred.

To reconcile the questionnaire design perspective with the requirements for a successful imputation, one might consider including a so-called core module (Raghunathan & Grizzle, 1995; sometimes also called X block, see; Graham et al., 1996). The core module constitutes a special module that is presented to all respondents. One potential strategy could be to include key items that are highly correlated to the items in the split modules, aiming to obtain an effect similar to allocating highly correlated items to different split modules. This might be especially relevant in contexts in which survey designers are afraid of dismantling their carefully constructed meaningful questionnaire structure by distributing survey topics across different split modules. For example, the European Values Study uses such an extended core module strategy combined with single-topic modules (Luijkx et al., 2021). The PISA 2012 students' context questionnaire's core module (the so-called 'common part') includes the most central variables and variables needed for estimating interaction effects, while split modules are also created largely along topic allocations (OECD, 2013, Chapter 6). The European Working Conditions Survey 2021 (Ipsos NV, 2022) has a large core module covering a wide range of topics, including some key job quality questions. This is followed by an SQD part with more in-depth questions on job quality, which is again modularized along topic allocations. So far, however, the evidence on the utility of approaches like this – supplementing single-topic modules with a more or less extensive core module – is very limited: Does an extended core module actually improve the quality of estimates after imputation in an SQD with single-topic modules? And if so, does this extended core module strategy yield a performance comparable to a random allocation of items to modules, revoking the need to distribute survey topics across split modules and allowing to use single-topic modules instead?

However, using such an extended core module would also yield the disadvantage of increasing questionnaire length for the individual respondents. Thus, the idea of including an extended core should be carefully considered and evaluated to clarify whether the potential positive effects of an extended core module would be worth the longer questionnaires. This is especially relevant because the imputation perspective may also compete with other legitimate reasons to include questions into a core module. For example, correct questionnaire routing might necessitate certain filter questions to be available for all respondents. Similarly, variables that are needed to stratify or restrict eventual analysis samples to relevant strata of the population may also be required to be observed for all respondents.

In this study, we will examine if an SQD with single-topic modules and an extended core module is able to perform (or outperform) similarly well as a random allocation of variables to modules. This extended core module will include, in addition to sociodemographic items, key items that are highly correlated to the variables in the split modules. We will compare this to a situation in which the core module only includes sociodemographic variables. In addition, we will also examine if an extended core module yields better outcomes than a simple sociodemographic core module under split modules generated through random allocation (hereinafter called random modules). Largely building on the dataset and the procedures implemented and described in Axenfeld, Blom, et al. (2022a), we use a Monte Carlo simulation study with real social survey data, in which we simulate SQDs by deleting data from the complete dataset. Through such a simulation with real survey data, we are able to test the effects of different SQD strategies on estimates while also maintaining a realistic setting.

2. Module design strategies

In this study, we examine the cross-classification of two strategies to construct split modules and two strategies to design a core module (i.e. four distinct scenarios). Table 1 gives an overview on these strategies along with an illustration of their assumed advantages and disadvantages for various aspects of a survey. In this paper, we focus on testing the hypotheses related to the imputation performance.

Concerning split module construction, we evaluate single-topic modules and random modules. Under single-topic modules, variables of the same topic are allocated to the same module. As in Axenfeld et al. (2022a), we assume that correlations between variables of the same topic are higher than variables of different topics, which may adversely affect the imputation (see also Figure 1).

Under the random modules strategy, by contrast, items are randomly allocated to split modules irrespective of survey topic. The random modules, in this context, represent a heuristic form of assigning highly correlated items to different modules. This is done because Axenfeld et al. (2022a) found that regarding the imputation, this simple strategy performs very similar to more elaborated strategies, such as diverse topics modules through stratified sampling.

Regarding the core module design, we include two different strategies depending on two groups of variables that might be relevant: a simple core module with only

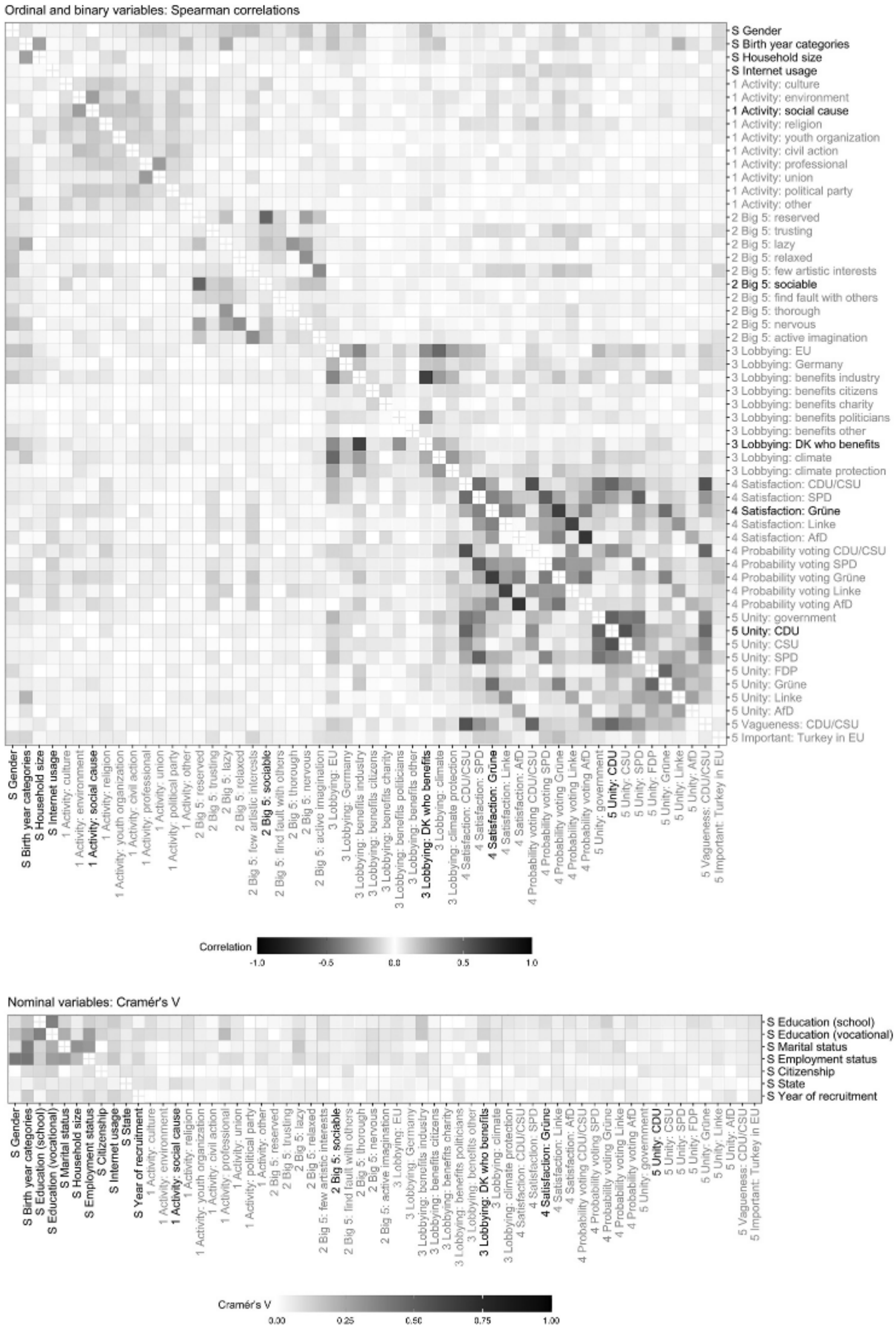


Figure 1. Associations between variables in the GIP data: Spearman correlations (for ordinal and binary variables, upper panel) and Cramér's V (for nominal variables, lower panel). Gray variable labels = variables in the split modules; black variable labels = variables in the (extended) core module. Variables sorted by survey topic: S = sociodemographic variables, 1 = topic 1 (activities), 2 = topic 2 (big five personality), 3 = topic 3 (lobbying), 4 = topic 4 (politics I), 5 = topic 5 (politics II).

Table 1. Assumptions on the effects of module design strategies on imputation, questionnaire structure, and questionnaire length.

Core module	Simple		Extended	
Split modules	Single-topic	Random	Single-topic	Random
Effect on …				
… imputation performance	−	+	+	++
… questionnaire structure	+	−	+	−
… questionnaire length reduction	+	+	−	−

Favorable effects are marked by "+". Adverse effects are marked by "−". Stronger effects are marked by more "+" or "−"symbols.

sociodemographic questions, and an extended core module enriched with additional key items. The goal of the extended core module is to have predictive information on split module variables available in completely observed key variables that can be used in the imputation of the planned missing data.

Sociodemographic variables play a central role in social science research. They are often used to restrict the sample to relevant strata of the population or as breakdown variables in substantive analyses, and beyond that, often included as control variables in multivariate analyses. Thus, one may assume that survey designers would deem sociodemographic items important enough to allocate them to a core module per se.

However, sociodemographic variables need not necessarily be strongly correlated to all the other substantive survey variables. In this case, their value for the imputation of items in the split modules may be limited. Therefore, a common idea is to go beyond the sociodemographic items by selecting specific *key items* from the other survey topics for the core module (e.g. Graham et al., 1996; Moore et al., 2020). However, evidence on the utility of such an approach in social survey research is scarce. To our knowledge, only one study (Moore et al., 2020) explores the effects of extended core modules as opposed to a simple, sociodemographic core module empirically in the context of sports psychology. The authors find that an extended core module tends to outperform the simple core module regarding the estimation of a confirmatory factor analysis throughout different variations of sample size and split module specifications.

Thus, for a strong test of an extended core module strategy, we need to identify key items from the specific survey topics suitable for an extended core module. For this purpose, we define key items as the variables with the highest average correlations to the other variables within each survey topic (except for the sociodemographic block, which is in the core module anyways). To this end, we calculate each variable's average correlation with the other variables from its topic, based on absolute values of Fisher's Z transformation of the correlations. From each topic block, we take the variable with the highest average correlation and add it to the extended core module.

3. Data

As population data for the Monte Carlo simulation study, we use 4,061 observations on 61 items from the German Internet Panel (GIP)[1] waves 37 and 38 (Blom et al., 2019a, 2019b). Some missing data were replaced by data from waves 1 and 13 (Blom et al., 2016a, 2016b), and a small amount of remaining missing data were imputed by stochastic single imputation. Rare events in categorical variables (less than 100 observations in the population dataset) were combined into broader categories to prevent problems imputing the reduced data in the simulation.

The data includes blocks on sociodemographic questions (11 items), personality psychology (10 items), membership in organizations (10 items), lobbyism in EU politics (10 items), and two blocks on German domestic politics (each 10 items) (see Figure 1). The sociodemographic items will serve as a simple core module. The extended core module will consist of all sociodemographic items enriched by key items from the other substantive blocks. All 50 items from the other five blocks will be allocated either to the split modules or to the extended core module. These 50 items are all categorical with an ordinal level of measurement. For a more detailed description of this analysis dataset see Axenfeld et al. (2022a).

Figure 1 gives an overview of the correlations between all variables in our GIP dataset, sorted by survey topic. The upper panel shows the Spearman correlations between all the ordinal variables in the data, and the lower panel reports Cramér's V for all bivariate relationships in which at least one of the variables is only nominal. Both subplots indicate that high correlations occur primarily within survey topics, while correlations between different survey topics are overwhelmingly very low. Yet, even within topics, some variables are correlated weakly or not at all.

Figure 1 also highlights variables that have been selected as part of the simple and extended core module using the procedure described above. The key items selected for extending the core module (with labels in black color) on average share stronger correlations with the variables of their survey topic than other variables. Yet, even these variables are not all consistently strongly correlated to all other variables of their topic. Hence, it is questionable whether they can effectively cover the whole correlation structure of their topic.

4. Simulation procedure

Each of the 1,023 simulation runs[2] of the Monte Carlo simulation consists of the following steps:

(1) Draw a sample of 2,000 from the population dataset. This is a common sample size realized in nationwide probability-based surveys.[3]
(2) Define five split modules (either by survey topics or by random allocation).
(3) Assign all respondents to three of five split modules randomly and (always) to the core module.
(4) Delete all data of the non-assigned modules.
(5) Use multiple imputation (Rubin, 1987) to complete the missing data resulting from step 4.
(6) Estimate and pool univariate frequencies and means based on the imputed data and compare the estimates to benchmarks from the population data.

Using this procedure, we examine four different scenarios (1,023 simulation runs each):

(1) Simple core module + single-topic split modules
(2) Extended core module + single-topic split modules

(3) Simple core module + random split modules
(4) Extended core module + random split modules

In the simple core module condition, the core module consists only of the 11 sociodemographic items. In the extended core module condition, the core module covers the sociodemographic items and five key items previously identified as described above. Split modules are constructed as described in Axenfeld et al. (2022a): The single topic split modules correspond to the five topic blocks with 10 items (simple core module scenario) or 9 items (extended core module scenario) each. Meanwhile, the random split modules are constructed by assigning the split items randomly to five split modules with 10 items (simple core scenario) or 9 items (extended core scenario) each. This yields an overall share of missing data of 33% in the simple core module scenarios and 30% in the extended core module scenarios.

The missing data are imputed with mice (van Buuren & Groothuis-Oudshoorn, 2011) in R (R Core Team, 2021) using predictive mean matching (Little, 1988; Rubin, 1986) with 20 imputations drawn after 15 iterations. As in Axenfeld et al. (2022a), the imputation models' predictor sets only include variables that are correlated by more than 0.1 to the imputed variable because earlier experimentation with the data has shown that otherwise estimation quality might be reduced (see also Axenfeld et al., 2022b).

We assess the quality of the imputed data by evaluating the quality of estimates calculated from the imputed data. In each simulation run s, we estimate 268 relative univariate frequencies for all categories i of the 45 variables that are imputed throughout all the four scenarios:

$$\hat{\pi}_{i,s} = \frac{1}{M}\sum_{m=1}^{M}\left(f_{i,m,s}/n\right) \tag{1}$$

Here, $f_{i,m,s}$ represents the absolute frequency of a category i for the m-th of M imputed datasets in a simulation run s of which is divided by the sample size $n = 2,000$. Variables that are used for the extended core module are excluded to ensure comparability. We then compute the deviations of these estimates from calculated benchmarks $\pi_i = F_i/N$ based on the complete population dataset (F_i demarking the absolute frequency of a category i in the population, and N the population size).

In applied social research, ordinal variables like those in the GIP data are sometimes treated as continuous rather than categorical to ease analysis. To take this into account, we also estimate means for all 45 variables:

$$\widehat{\bar{x}}_{k,s} = \frac{1}{M}\sum\nolimits_{m=1}^{M}\left(\frac{1}{n}\sum\nolimits_{j=1}^{n} x_{j,k,m,s}\right), \tag{2}$$

where $x_{j,k,m,s}$ is the value of the j-th person on the k-th variable for the m-th imputed dataset in a simulation run s, and $\widehat{\bar{x}}_{k,s}$ is the mean of variable k as estimated based on simulation run s.

We use two measures for evaluating the quality of estimates: their absolute average deviations from the population benchmarks and their variances, as approximated through the 1,023 runs of the Monte Carlo simulation.

For conciseness, the following paragraphs describe the calculation of the absolute average Monte Carlo deviations and variances with respect to univariate frequencies; the calculation of the absolute average Monte Carlo deviations and variances for means is analogous.

As a first measure, we obtain the average Monte Carlo deviation for each of the 268 estimands by taking the mean of deviations $\hat{\pi}_{i,s} - \pi_i$ of an estimated frequency from the population benchmark over all simulation runs S for each of the estimands:

$$\widehat{\text{dev}}_{MC,i} = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\pi}_{i,s} - \pi_i\right) \tag{3}$$

This means we end up with 268 average Monte Carlo deviations for each scenario to evaluate the effects of the different strategies.

Since we are only interested in the magnitude of the average Monte Carlo deviations (and not their direction), we analyze their absolute values:

$$\widehat{\text{dev}}_{MC,i}{}^{*} = \left|\widehat{dev}_{MC,i}\right| \tag{4}$$

This measure is also commonly referred to as an indicator of bias. However, identifying bias in a strict sense requires full convergence across simulation runs. Despite the relatively large number of simulation runs, we cannot guarantee that all average deviations – 268 for frequencies and 45 for means in each scenario – have fully converged. Furthermore, it is not feasible to increase the number of simulation runs indefinitely to ensure complete convergence is achieved. Consequently, deviations from the benchmark may reflect a combination of bias and some remaining variance across the simulation runs. Variances of estimates across simulation runs can be theoretically expected to differ between the SQD strategies, as they induce varying levels of imputation uncertainty (Raghunathan & Grizzle, 1995). We therefore recommend interpreting differences in average Monte Carlo deviations as indicators of general estimation quality, rather than as definitive evidence of bias.

Although we summarize absolute average Monte Carlo deviations at an aggregate level, each individual $\widehat{\text{dev}}_{MC,i}{}^{*}$ (assuming sufficient precision in its approximation across the simulation runs) can be understood as representing the systematic error associated with a specific category i. Consequently, a large $\widehat{\text{dev}}_{MC,i}{}^{*}$ for even a single category may signal a critical weakness in a given SQD strategy. To aid in the assessment of potential systematic deviations, Appendix Figures A1–A8 present the absolute average Monte Carlo deviations for all four strategies along with their corresponding 95% confidence intervals, based on standard errors calculated as described by Morris et al. (2019).

As a second measure, we calculate Monte Carlo variances $\hat{\sigma}^2_{MC,i}$ for all 268 estimands across the 1,023 simulation runs:

$$\hat{\sigma}^2_{MC,i} = \frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\pi}_{i,s} - \overline{\pi_i}\right)^2, \tag{5}$$

where $\overline{\pi_i}$ represents the average of all estimates for a frequency i. Like the absolute average Monte Carlo deviations, the variances are analyzed collectively, but still, each of the 268 points in the graph represents the variance of one variable category.

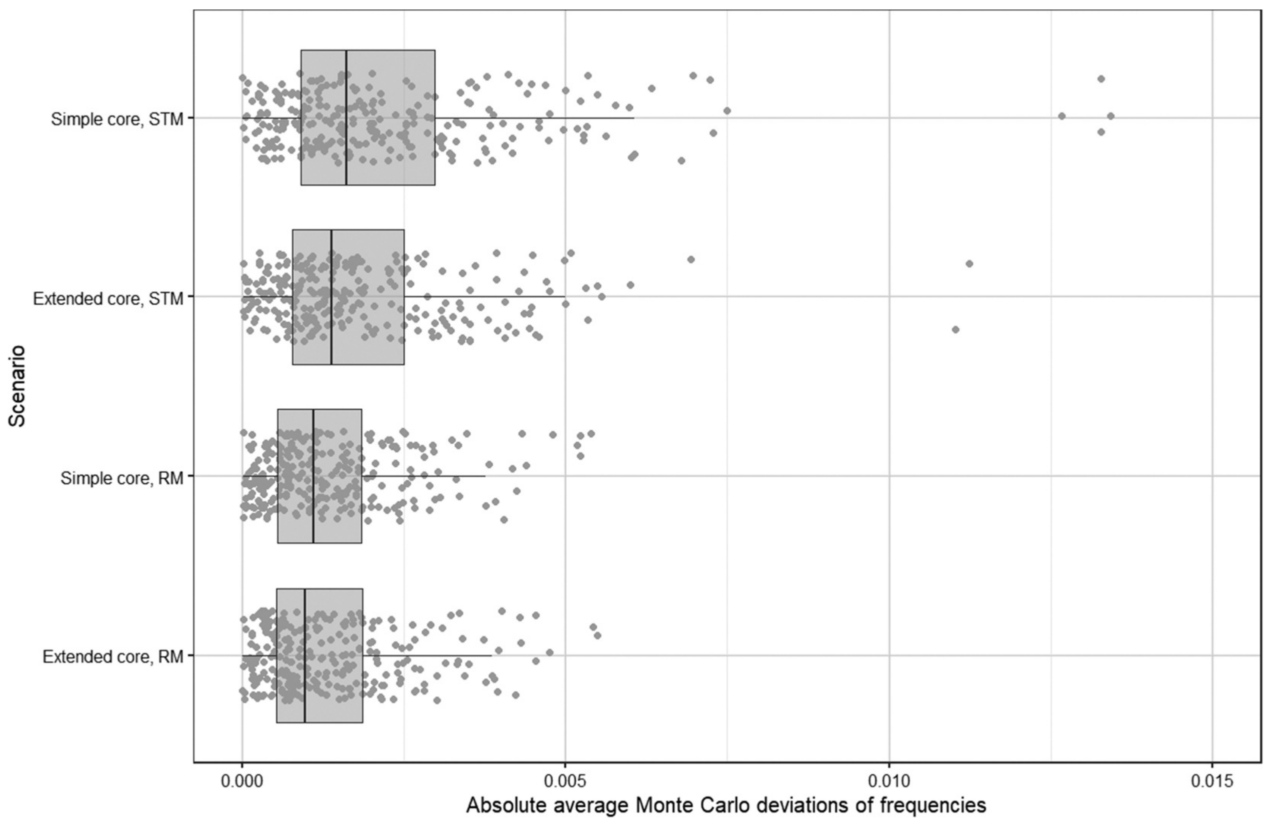
5. Results

5.1. Absolute average Monte Carlo deviations

Figure 2 displays the absolute values of the average Monte Carlo deviations for all univariate frequency estimates within the four different scenarios. Every data point shown represents the average deviation of one specific category of a variable. To simplify the analysis, boxplots are superimposed over these average deviations.

As shown by Figure 2, for single-topic modules an extended core module leads to estimations on average closer to the benchmark. For example, while 32.5% of absolute average Monte Carlo deviations are larger than 0.0025 (that is, a quarter of a percentage point) with a simple core module, it is only 24.6% with an extended core module. This was expected since the key items in the extended core module were selected based on their comparatively high correlation with the remaining items of the respective topic and thus supporting the imputation process.

However, even when using an extended core module, absolute average Monte Carlo deviations with single-topic modules are, overall, still higher than with random modules. This applies to both random modules with and without an extended core module: Both random modules strategies outperform single-topic modules with an extended core module substantially. For example, only 13.4% and 14.2% of absolute average Monte Carlo deviations are larger than 0.0025 in these scenarios, as compared to 24.6% in the single-topic modules scenario with an extended core module.



Interestingly, the extended core module shows barely any effect when using random modules. We found that an extended core module only improves estimate quality with single-topic modules, but hardly so when using random modules.

Figure 3 displays the absolute values of average Monte Carlo deviations for 45 mean estimates. As for frequency estimates, absolute average Monte Carlo deviations turn out the largest with STM and a simple core module, with values up to 0.07, while RM have the lowest absolute average Monte Carlo deviations regardless of the core module strategy. STM with an extended core module again performs better than without the extended core module, but still inferior to RM.

STM = single-topic modules; RM = random modules. Based on a Monte Carlo simulation with 1,023 runs per scenario.

These results might be explained by the covariance structure of data. A covariance structure that favors the application of an extended core module would include key variables that are particularly highly correlated to all the other variables of its topic, such that their predictive information on the other variables can be exploited in the imputation. In reality, however, high correlations within a topic may not occur on one specific key item in particular: High correlations within a survey topic may be distributed over different items, implying that there is no single key item that would be strongly correlated to all other items of its topic.

For a random modules strategy, a covariance structure is advantageous if higher correlations are distributed across different items within a survey topic. Random modules tend to distribute all the items of a topic across different modules, which may

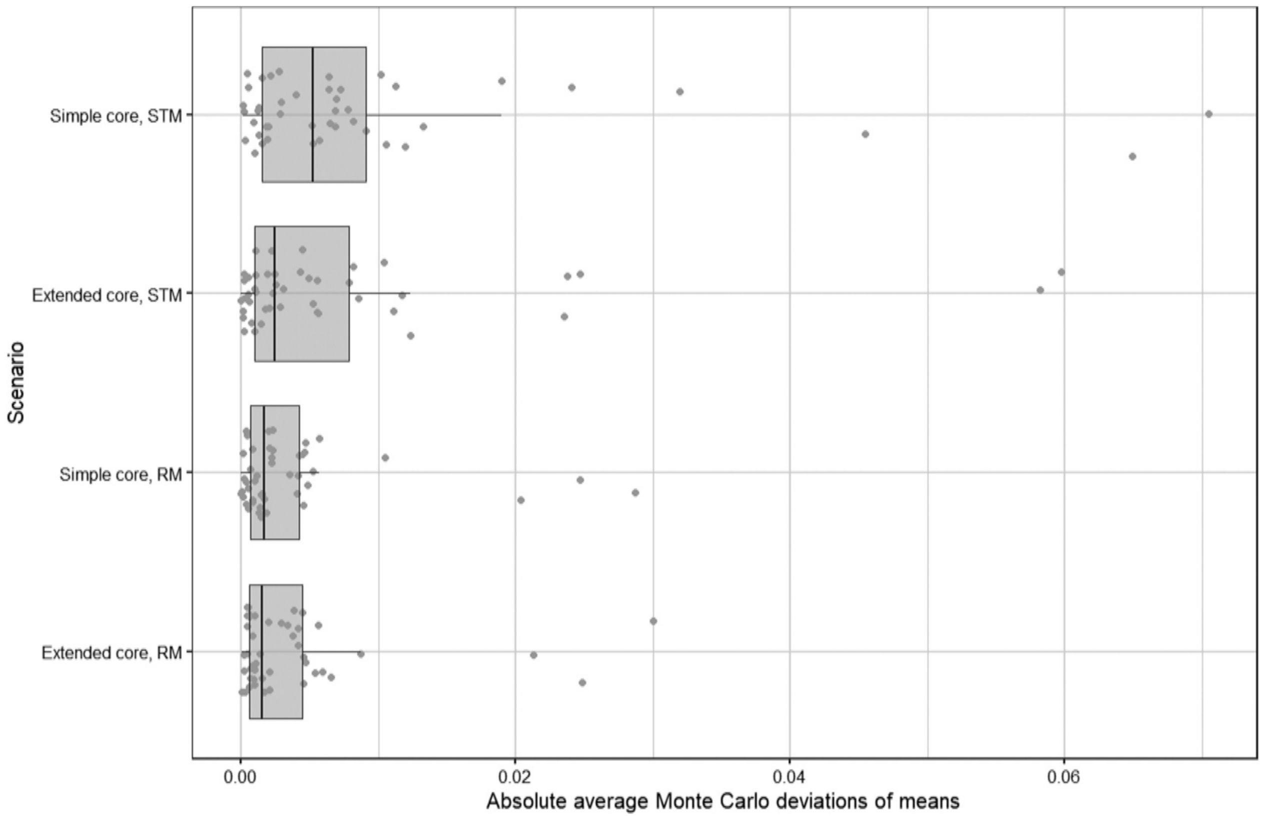


Figure 3. Absolute values of average Monte Carlo deviations of 45 mean estimates, by core module construction strategy and split module construction strategy.

represent this type of covariance structure better than allocating specific key items to a core module. In consequence, here, random modules lead to better results than single-topic modules even with an extended core module.

This might also explain why the extended core module has almost no effect in the random modules condition: The random allocation already entails that highly correlated items tend to be in different modules and are thus capable of supporting each other's imputation. In this situation, it seems not decisive that, in addition, some key items are fully observed.

5.2. Monte Carlo variances

Figures 4 and 5 display the Monte Carlo variances for all univariate frequency estimates (Figure 4) and all mean estimates (Figure 5) across the four scenarios. The presentation format otherwise corresponds to that in the previous figures.

STM = single-topic modules; RM = random modules. Based on a Monte Carlo simulation with 1,023 runs per scenario.

STM = single-topic modules; RM = random modules. Based on a Monte Carlo simulation with 1,023 runs per scenario.

Differences between the scenarios are more difficult to spot in the graph with Monte Carlo variances as compared to the absolute average Monte Carlo deviations. Yet, at least for means, the general pattern still follows the same trends: STM lead to increased

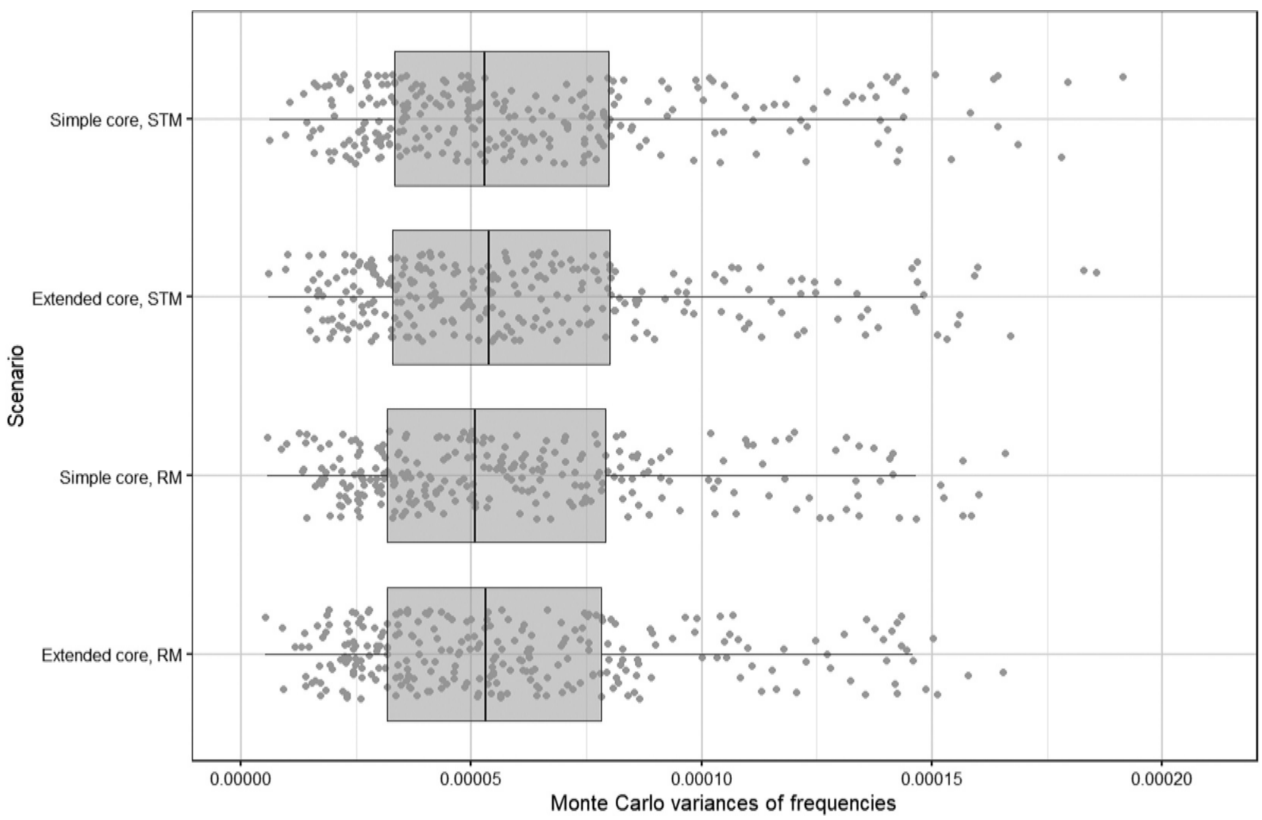


Figure 4. Monte Carlo variances of 268 univariate frequency estimates, by core module construction strategy and split module construction strategy.

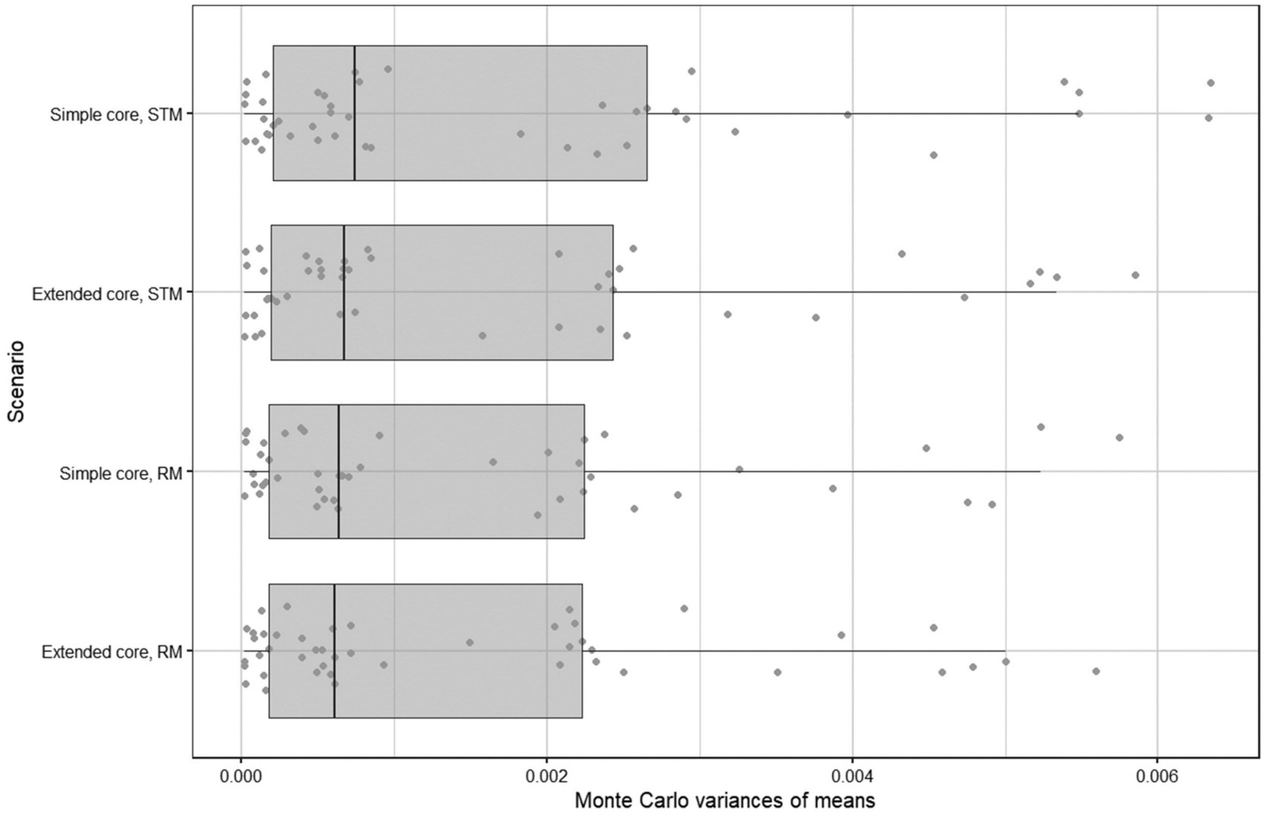


Figure 5. Monte Carlo variances of 45 mean estimates, by core module construction strategy and split module construction strategy.

variance compared to RM, while the use of an extended core module can reduce this difference to some extent. Meanwhile, the variances for frequency estimates are similar across scenarios. However, relatively high variances seem more common with STM (both with simple and extended core module) compared especially to RM with simple core module, at least with respect to estimates of means.

6. Discussion

This paper examined two strategies to design a core module for an SQD both with single-topic modules and random modules using a Monte Carlo simulation study based on the GIP data. The study yields three main findings:

First, an extended core module strategy leads to estimations on average closer to the benchmark and, at least when estimating means, variances for single-topic modules compared to the simple core module strategy with only sociodemographic items. This finding replicates previous evidence from a different data context by Moore et al. (2020). Single-topic modules using only sociodemographic items in the core module, by contrast, fail to capture the covariance structure of the data sufficiently and yield the least favorable outcomes. Thus, in cases in which avoiding single-topic modules is no viable option, extending a core module through key items from substantive survey topics should be considered. Furthermore, this finding raises the question whether sociodemographic items always need to be part of a core module or could be included in split modules as well. When only arguing from an imputation perspective, their rather low correlations with the other variables in this dataset substantiate this notion. However, testing strategies without the sociodemographic items in the core module explicitly was out of scope for this study. Furthermore, even if the sociodemographic variables have no effect on the imputation, there may be other reasons to keep these items fully observed. First, sociodemographic variables are typically used to assess nonresponse bias and calculate weights. Second, they often serve as grouping variables in social analysis, for example, to compare young vs. old, poor vs. rich, or less educated vs. highly educated respondents; in the same vein, these variables are typically used as 'control variables' in multivariate analyses. Third, if an analysis should only pertain to subpopulations, these usually are defined based on sociodemographic characteristics as, for example, only citizens with voting rights or only people in employment. Of course, these functions of sociodemographic variables could also be fulfilled if they were imputed, that is, part of a rotating module or randomly assigned to respondents. However, the question is if sociodemographic variables can be imputed satisfyingly in an SQD due to their often nominal scale of measurement with many rather infrequent categories and their rather limited correlations with the other variables.

Second, we observed that even with an extended core module, the single-topic module strategy could not keep up with random modules in terms of the quality of estimates. This applies to both random modules with an extended core module or even a simple core module. This might result from the covariance structure of the dataset: Even key items are not consistently highly correlated to all other variables of their topic. Furthermore, within survey topics many correlations are rather small but still large enough to have an influence on the imputation. Distributing these slightly higher correlations of variables within a topic across different modules can improve the imputation and, in consequence, the estimation. Thus, the covariance structure of this dataset favors the application of random modules compared to single-topic modules with an extended core module.

Third, the extended core module has almost no effect when using random modules. Thus, random modules may have the additional advantage that the core module might be kept short, which reduces questionnaire length.

Altogether, we find that from an imputation perspective, random modules outperform single-topic modules independent of core module specification. This is in line with earlier research, which generally stresses the importance to allocate highly correlated items to different modules (Axenfeld et al., 2022a; Imbriano & Raghunathan, 2020; Raghunathan & Grizzle, 1995) and corroborates this argument to hold even with an extended core module. This means that survey designers should construct split modules such that each module covers the different topics of the survey rather than only a single topic. If this is done, there may be no need for (and no avail in) an extensive core module. Meanwhile, if single-topic modules are the only option available, an extended core module may help. Yet, when considering an extended core module, survey designers should still carefully assess the implications the resulting prolonged questionnaire may have, as for instance on survey costs or respondent fatigue and thereby on response quality.

What do our results imply for the construction of questionnaires for SQDs? From a survey practitioner's point of view, the relatively small absolute differences of the absolute average Monte Carlo deviations for frequencies and means may seem to indicate that the choice of module construction strategy has little substantive impact on the quality of (univariate) estimates after imputation. It should be pointed out, however, that single-topic modules resulted in considerably more outliers among absolute average Monte Carlo deviations, indicating that following this strategy implies a greater risk of undermining estimation quality based on the imputed data. Combining the single-topic modules strategy with an extended core module reduces this risk but not as much as choosing a random distribution of items across modules. Thus, if priority is given to maximizing the quality of estimates as resulting from the imputation of incomplete data, the clear recommendation is to distribute survey items of a topic across modules rather than allocating all of them to the same module. This does not mean that questionnaire coherence must be neglected. Rather, as the module structure of the SQD does not dictate the order in which items are presented to respondents, we believe that coherent questionnaires may be possible even with a random-modules strategy. For practice, this means that items do not have to be presented sorted by module but can be delivered to respondents in their original thematic order. As compared to a complete questionnaire, this would mean that while some items are left out in each questionnaire form, the overarching structure of the questionnaire remains untouched. Hence, respondents should – by and large – experience the shorter questionnaire just as coherent as the longer, complete questionnaire. Such a strategy would be compatible with the demand for suitable imputations as well as practitioners' concern with presenting questionnaires that make sense to respondents.

This study has some limitations. First, it deals exclusively with the statistical question of when the imputation works best using pre-existing data. Hence, it does not include implementing an SQD in an actual data collection. This means that examining how the different module construction strategies affect response behavior or respondents' experience of the questionnaires is out of scope. This especially includes the question whether single-topic modules in fact have advantages regarding perceived questionnaire consistency and, if so, whether this perception has any effect on data quality. For investigating issues like that, field experiments are needed in the future.

Second, the measures we used to compare estimation quality under the different scenarios – absolute average Monte Carlo deviations and Monte Carlo variances – have some limitations and particularly require a sufficient number of simulation runs. Even though we had more than a thousand simulation runs, some observed Monte Carlo deviations might appear different from zero only due to a lack of convergence. To some extent, absolute average Monte Carlo deviations may therefore be the result of both bias and/or variance of estimates after imputation. Yet, the confidence intervals of the deviations displayed in Appendix Figures A1–A8 show that depending on SQD scenario and estimand type, between 60% and 88% of absolute average Monte Carlo deviations are significantly different from zero. This suggests that to some degree, observed deviations might reflect systematic deviations for many of the estimands, that is, a bias. This may not be an isolated finding of the specific imputation strategy in this simulation, as previous research finds similar or higher average deviations with a wide range of multiple-imputation methods and models (Axenfeld et al., 2022b). Determining the exact statistical cause of this potential bias is out of scope for this study. On the one hand, statistical theory suggests that systematic bias should not arise with multiple imputation in a missing-completely-at-random scenario (as it is examined in this study). On the other hand, this expectation presupposes a correctly specified imputation model. In practice, imputation models that are correctly specified in an absolute sense may be rarely achievable, particularly with real-world data, which rarely conform exactly to the distributional assumptions underlying any standard statistical technique. As a result, bias might emerge from the practical difficulty of specifying a correct model, even when all analysis variables are included. This challenge might also contribute to differences in estimation quality across the various module construction strategies within an SQD due to their different amounts of available information. The apparent discrepancy between theory and practice points to a need for further research into the mechanisms by which real-data issues may manifest as systematic bias in multiple imputation settings. At the same time, it remains plausible that observed performance differences between the SQD strategies may, at least in part, stem directly from differences in variance: Given that the amount of information available in the data with each of the SQD strategies varies, corresponding differences in imputation uncertainty are to be expected (Raghunathan & Grizzle, 1995), potentially resulting in differing levels of estimate variability across simulation runs. Ultimately, it is not possible to disentangle with certainty whether the observed differences in performance are primarily driven by bias or by variance. Nonetheless, the primary objective of this study was to provide a general comparison of the four SQD strategies with regard to their overall estimation quality, rather than focusing exclusively on bias or variance. We presume that, for practitioners in survey research, overall estimation quality is likely to represent the more relevant criterion when selecting an appropriate SQD strategy for their specific application.

Third, strategies for core module construction differing from those tested here could yield diverging outcomes. This involves two main aspects. On the one hand, through mathematical optimization procedures, one might improve the selection of key items beyond what is done in this paper. Yet, in survey practice the information on the covariance structure of the data needed for such procedures may rarely be available at the stage of questionnaire design (see also Axenfeld et al., 2022a). On the other hand, allocating even more key items to a core module might help single-topic modules to catch up with random modules in terms of the imputation's performance. However, large core modules also mean longer questionnaires, which may at some point contradict the overall purpose of a split questionnaire design.

Fourth, our simulation study only examined planned missing data from split questionnaire designs, which are usually missing completely at random. For severe item nonresponse problems with more complex missingness mechanisms, some key items might also determine the selection into item nonresponse and thus might be required to be fully observed to correct for nonresponse bias. In this situation, including key items in a core module might improve the imputation even with random modules or equivalent strategies.

Fifth, this study tests the performance of multiple imputation only with respect to univariate estimates. Bivariate and multivariate estimates were out of scope. Previous research, however, has shown similar patterns in effects of different module construction strategies on univariate and bivariate estimates, with single-topic modules performing overall inferior to random modules (Axenfeld et al., 2022a).

Finally, this study is based on a single social survey dataset. The question is to which extent the findings can be transferred to surveys in social science in general. For instance, our dataset is rather small in comparison to other surveys with several hundred variables in which SQDs are applied. However, we suspect that the covariance structure is comparable to many other datasets in social survey practice. In this regard, one open question is whether there may be surveys with specific covariance structures that favor an extended core module strategy. This would require data with a small number of key variables that collectively cover the covariance structure to the other variables in the split modules to a sufficient degree. However, we assume that this situation may be unlikely in survey practice. Rather, as in our dataset, there will probably be many moderate or small but non-zero correlations within a survey topic which are often higher than correlations between variables of different topics. Nonetheless, to expand the external validity of the findings obtained in this study, more research is needed with data that covers different topics and measurement instruments, particularly with surveys in other disciplines of the social sciences.

Notes

1. For a more detailed information on the GIP data and its data collection procedure, see Blom et al. (2015, 2017) and Cornesse et al. (2021).
2. According to Morris et al. (2019), common numbers of simulation runs in Monte Carlo simulations are 500 and 1,000. In our study, 1,023 simulation runs were favored over a round number (e.g. 1,000) because we had access to 1,024 processor cores (one core per simulation run, except for one consumed by setting up the simulation).
3. For example, the European Social Survey 2020 has a mean net sample size of 1,925 (median: 1577) persons per country (ESS Round 11, 2024), and the International Social Survey Programme 2021 has on average 1,485 (median: 1,361) persons in each country sample (ISSP Research Group, 2024). In our simulation study, the sample size of 2,000 is equivalent to a net sample size, as unit nonrespondents were excluded from the data beforehand.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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 authors also acknowledge support by the state of Baden-Württemberg through bwHPC. Assistance with formatting by Julia Schäfer is sincerely appreciated.

Notes on contributors

Julian B. Axenfeld is a survey methodologist and head of the German Social Cohesion Panel at DIW Berlin (German Institute for Economic Research). His research interests include split questionnaire designs, imputation, and survey nonresponse. He has published, for instance, in the *Journal of Survey Statistics and Methodology, Survey Research Methods*, and *Statistics Surveys*.

Christian Bruch is a postdoctoral researcher at GESIS – Leibniz Institute for the Social Sciences, Mannheim. His current research interests are variance estimation, imputation, weighting, and Bayesian multilevel modeling. His research is published, for example, in the journals *Journal of*

Survey Statistics and Methodology, *Communications in Statistics - Simulation and Computation*, *Statistics Surveys*, *Computational Statistics* and *Journal of Statistical Computation and Simulation*.

Christof Wolf is President of GESIS – Leibniz Institute for the Social Sciences, Mannheim, and Professor for Sociology at the University of Mannheim. His research focuses on data collection and analysis, comparative analysis of social stratification, religion in modern societies, and society and health.

ORCID

Julian B. Axenfeld http://orcid.org/0000-0003-3728-3828
Christian Bruch http://orcid.org/0000-0003-0926-6609
Christof Wolf http://orcid.org/0000-0002-9364-9524

References

Axenfeld, J. B., Blom, A. G., Bruch, C., & Wolf, C. (2022a). Split questionnaire designs for online surveys: The impact of module construction on imputation quality. *Journal of Survey Statistics and Methodology*, *10*(5), 1236–1262. https://doi.org/10.1093/jssam/smab055

Axenfeld, J. B., Bruch, C., & Wolf, C. (2022b). General-purpose imputation of planned missing data in social surveys: Different strategies and their effect on correlations. *Statistics Surveys*, *16*, 182–209. https://doi.org/10.1214/22-SS137

Blom, A. G., Bossert, D., Funke, F., Gebhard, F., Holthausen, A., & Krieger, U. (2016a). Sfb 884 "political economy of reforms" Universität Mannheim. *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. (2016b). SFB 884 "Political Economy of Reforms" Universität Mannheim. *German Internet Panel, Wave 13 - GESIS Data Archive, 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. (2019a). SFB 884 "Political Economy of Reforms", Universität Mannheim. *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. (2019b). SFB 884 "Political Economy of Reforms", Universität Mannheim. *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. https://doi.org/10.1177/1525822X15574494

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. https://doi.org/10.1177/0894439316651584

Bruch, C. (2023). *Imputation of missing values in survey data*. GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey guidelines). https://doi.org/10.15465/gesis-sg_en_044

Chauvet, G., Deville, J.-C., & Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, *98*(2), 459–471. https://doi.org/10.1093/biomet/asr011

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. https://doi.org/10.1177/08944393211006059

ESS Round 11 (2024). *European Social Survey European Research Infrastructure (ESS ERIC). ESS11 - integrated file, edition 2.0* [Data set]. Sikt - Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess11e02_0

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. https://doi.org/10.1207/s15327906mbr3102_3

Haziza, D. (2009). Chapter 10 - Imputation and inference in the presence of missing data. In C. R. Rao (Ed.), *Handbook of statistics. Sample surveys: Design, methods and applications: Vol. 29*, Part A (Vol. 29, pp. 215–246). Elsevier. https://doi.org/10.1016/S0169-7161(08)00010-2

Imbriano, P. M., & Raghunathan, T. E. (2020). Three-form split questionnaire design for panel surveys. *Journal of Official Statistics*, *36*(4), 827–854. https://doi.org/10.2478/jos-2020-0040

Ipsos NV. (2022). *European working conditions telephone survey, 2021: Technical report*. *European Foundation for the Improvement of Living and Working Conditions*. Eurofound reference number WPEF22050. https://www.eurofound.europa.eu/system/files/2022-12/wpef22050.pdf

ISSP Research Group. (2024). *ZA800 international social survey programme: Health and health care II - ISSP 2021*. ZA8000 Data file Version 2.0.0, GESIS. https://doi.org/10.4232/5.ZA8000.2.0.0

Kim, J.-K., & Rao, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, *96*(4), 917–932. https://doi.org/10.1093/biomet/asp041

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 263–314). Emerald Group Publishing Limited.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, *6*(3), 287–296. https://doi.org/10.1080/07350015.1988.10509663

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

Moore, E. W. G., Lang, K. M., & Grandfield, E. M. (2020). Maximizing data quality and shortening survey time: Three-form planned missing data survey design. *Psychology of Sport & Exercise*, *51*, 101701. https://doi.org/10.1016/j.psychsport.2020.101701

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. https://doi.org/10.1787/9789264190511-en

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90*(429), 54–63. https://doi.org/10.1080/01621459.1995.10476488

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

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

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, *4*(1), 87–94. https://doi.org/10.1080/07350015.1986.10509497

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

Smyth, J. D. (2016). Designing questions and questionnaires. In C. Wolf, D. Joye, T. W. Smith, & Y.-C. Fu (Eds.), *The SAGE Handbook of survey methodology* (pp. 218–235). SAGE Publications.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Appendix. Confidence intervals of average absolute Monte Carlo deviations

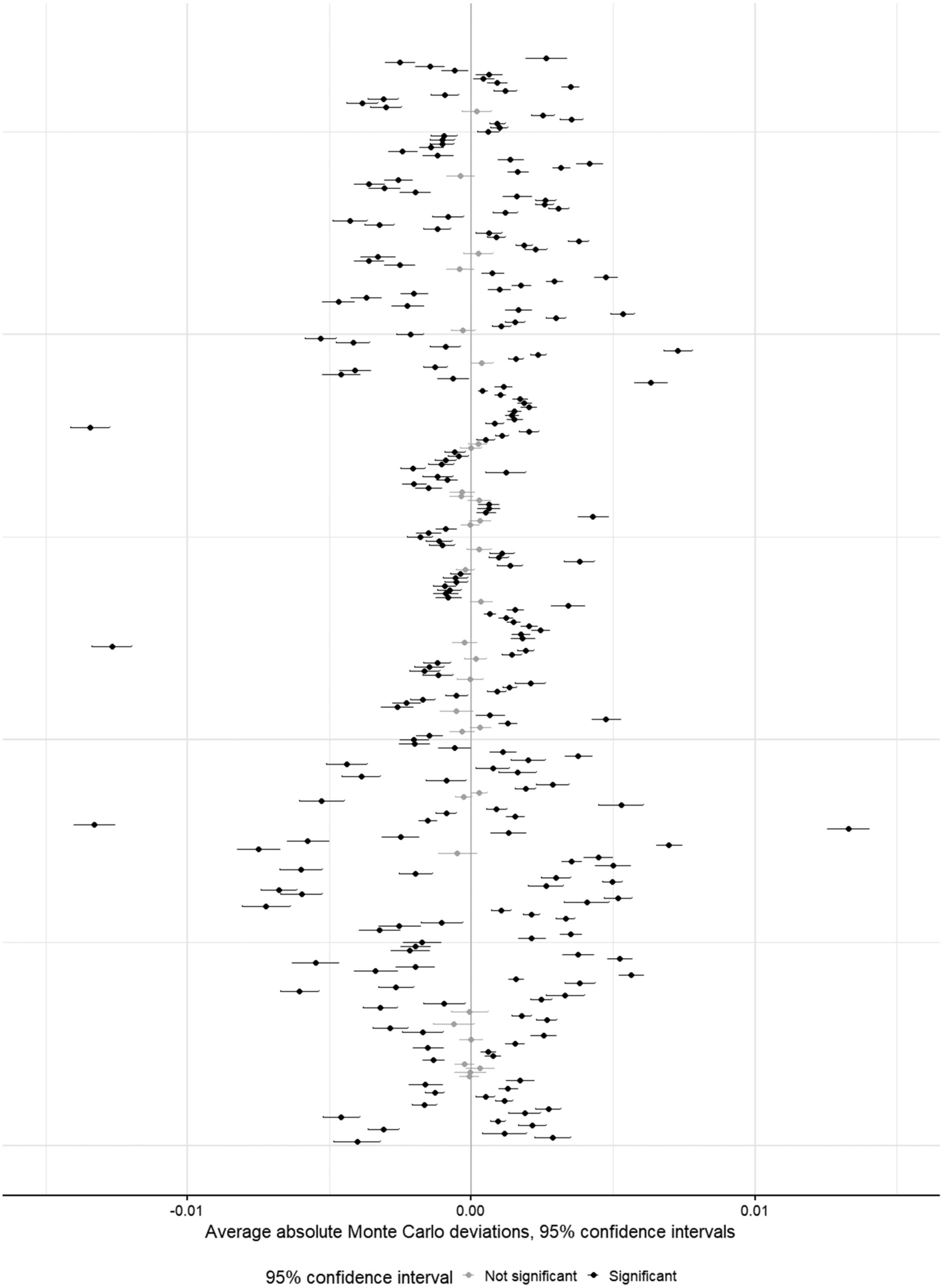


Figure A1. Simple core module and single-topic modules: absolute values of average Monte Carlo deviations of 268 univariate frequency estimates with 95% confidence intervals. Based on a Monte Carlo simulation with 1,023 simulation runs.

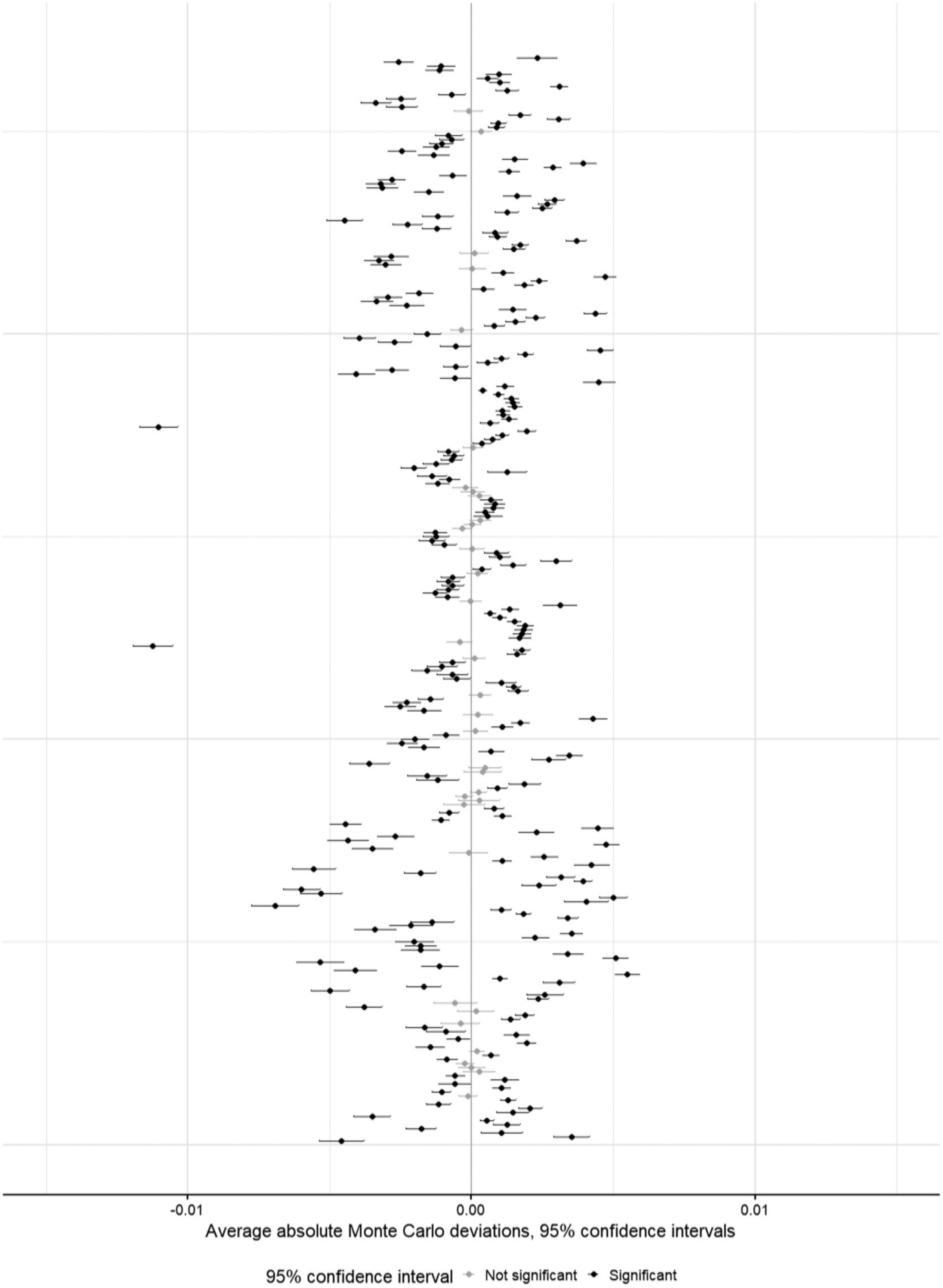


Figure A2. Extended core module and single-topic modules: absolute values of average Monte Carlo deviations of 268 univariate frequency estimates with 95% confidence intervals. Based on a Monte Carlo simulation with 1,023 simulation runs.

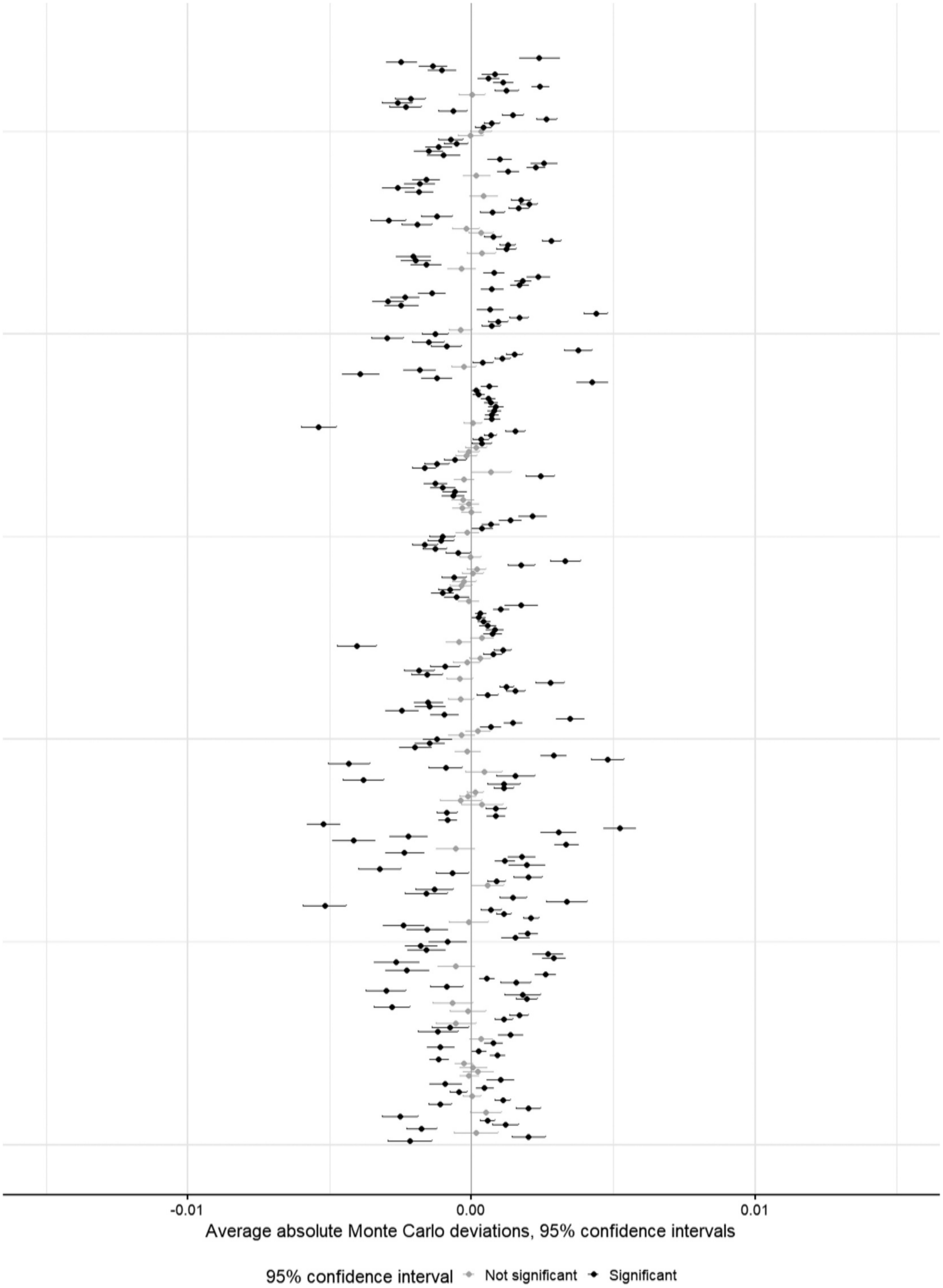


Figure A3. Simple core module and random modules: absolute values of average Monte Carlo deviations of 268 univariate frequency estimates with 95% confidence intervals. Based on a Monte Carlo simulation with 1,023 simulation runs.

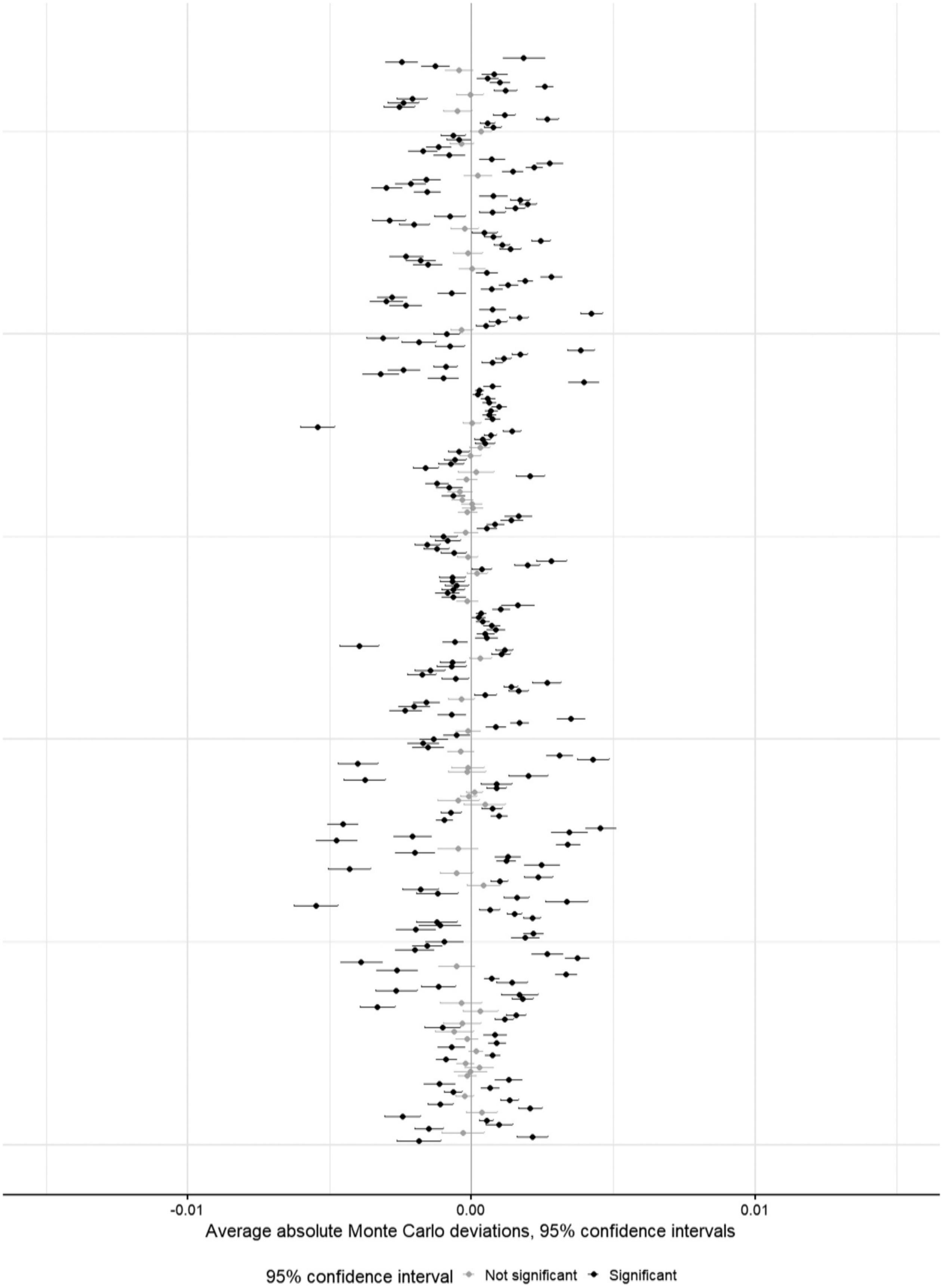


Figure A4. Extended core module and random modules: absolute values of average Monte Carlo deviations of 268 univariate frequency estimates with 95% confidence intervals. Based on a Monte Carlo simulation with 1,023 simulation runs.

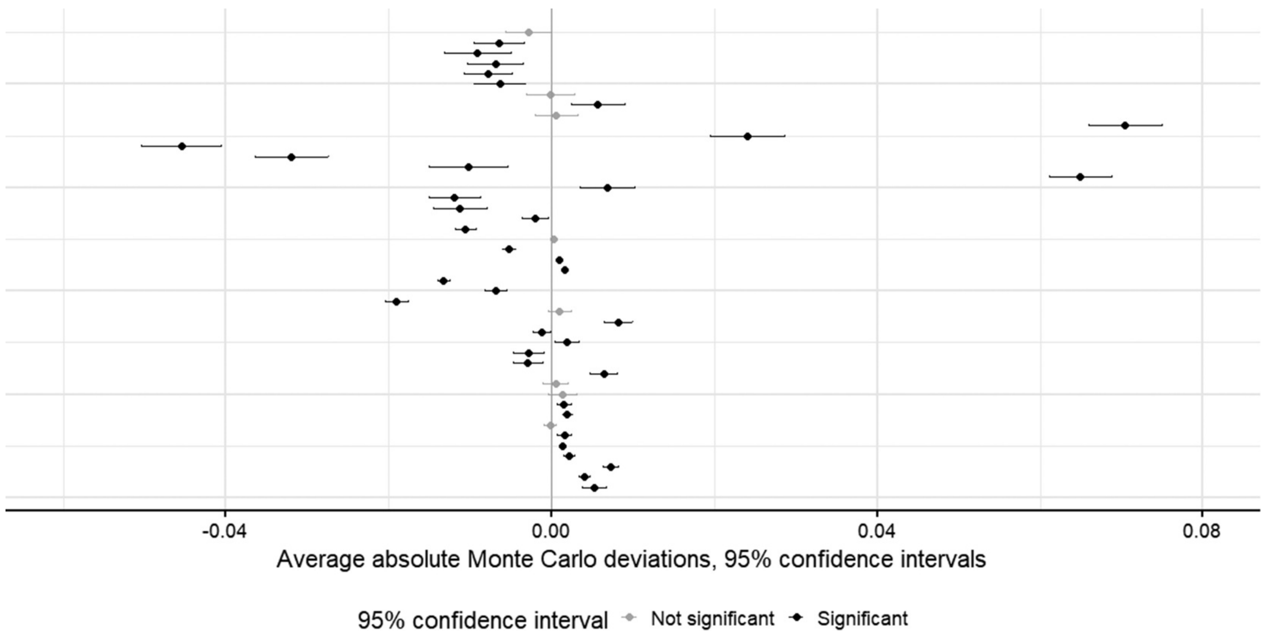


Figure A5. Simple core module and single-topic modules: absolute values of average Monte Carlo deviations of estimates with 95% confidence intervals. Based on a Monte Carlo simulation with 1,023 simulation runs.

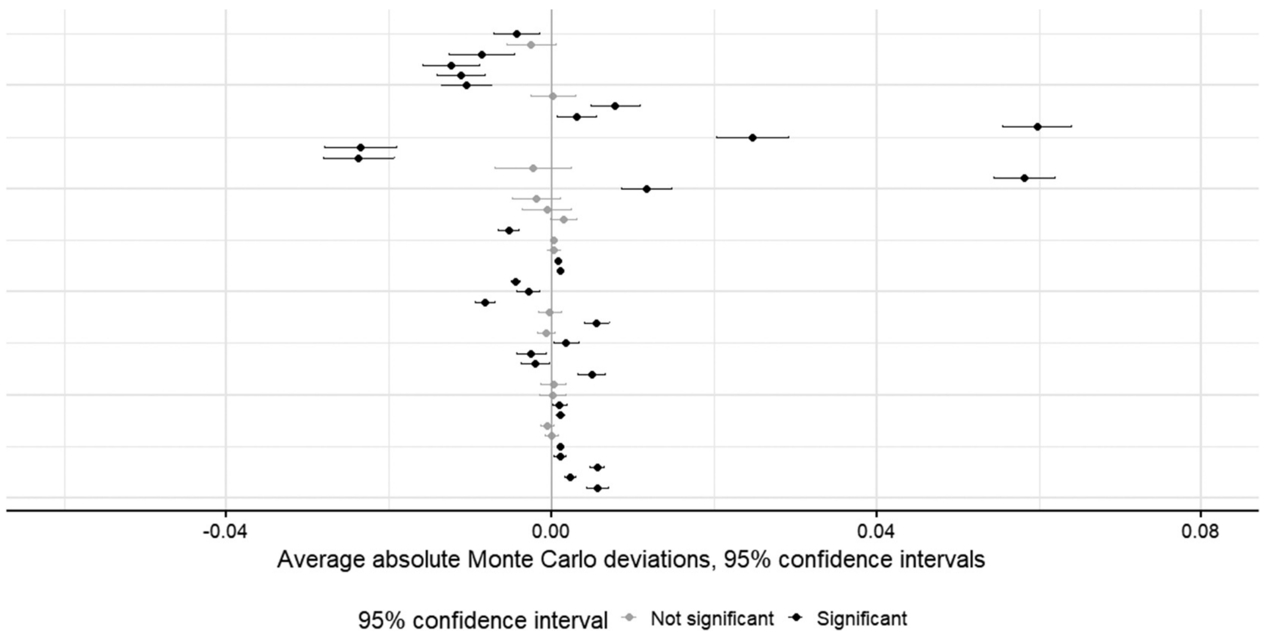


Figure A6. Extended core module and single-topic modules: absolute values of average Monte Carlo deviations of 45 mean estimates with 95% confidence intervals. Based on a Monte Carlo simulation with 1,023 simulation runs.

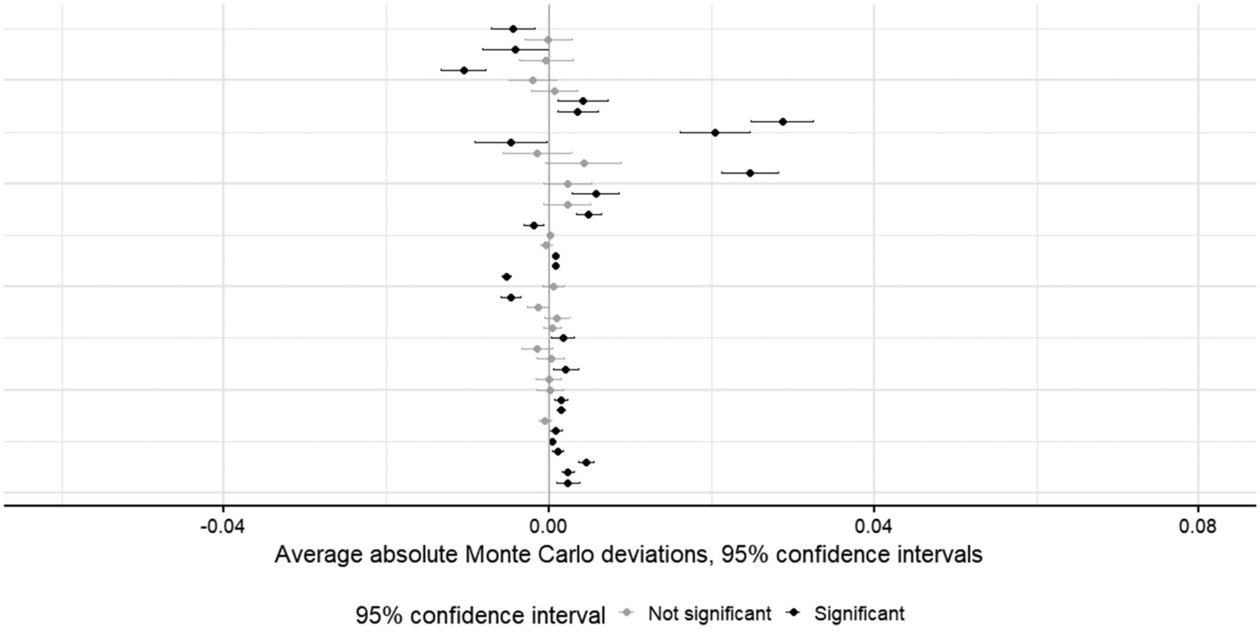


Figure A7. Simple core module and random modules: absolute values of average Monte Carlo deviations of 45 mean estimates with 95% confidence intervals. Based on a Monte Carlo simulation with 1,023 simulation runs.

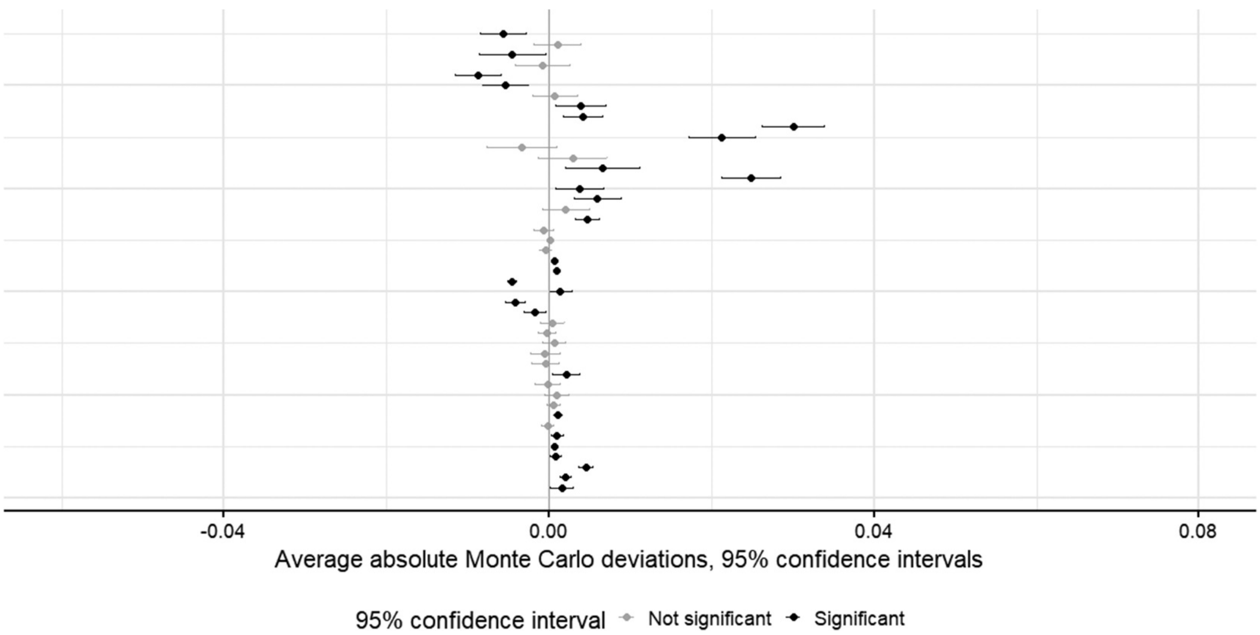


Figure A8. Extended core module and random modules: absolute values of average Monte Carlo deviations of 45 mean estimates with 95% confidence intervals. Based on a Monte Carlo simulation with 1,023 simulation runs.