National Center for Science and
Engineering Statistics

Title: Split Questionnaire Design Literature Review

Date: December 2020
Final Report

Contractor Awardee: RTI International
Contract Number: 49100420C0021

Disclaimer: Broad Agency Announcement (BAA) awards provide research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF). This report documents research funded through an NCSES BAA and is being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed in this report do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov.

RTI INTERNATIONAL
3040 E. Cornwallis Road • PO Box 12194 • Research Triangle Park, NC 27709-2194 • USA
Telephone +1.919.541.6000 • Fax +1.919.541.5985 • www.rti.org

Split Questionnaire Design Literature Review

December 31, 2020

Darryl Cooney, Darryl Creel, Andy Peytchev, Emilia Peytcheva

The idea for split questionnaire design is not new—in the early 70s, Good (1969, 1970) called for the development of split-questionnaire methods to maximize data collection efficiency. A decade later, Herzog and Bachman (1981) advised long questionnaires to be split in parts that are administered in different order. Yet, little attention has been focused on split questionnaire designs until Raghunathan and Grizzle (1995) demonstrated the potential advantages of split questionnaire designs in two simulation studies. The method is based on multiple matrix sampling design (Shoemaker, 1973; Munger and Lloyd, 1988) where a random sample of items is presented to each sampled individual. In contrast, split questionnaire design imposes restrictions on which items are administered to which individuals and shares some similarities with educational test designs (Holland and Wightman, 1982; Holland and Thayer, 1985). Split Questionnaire Design (SQD) has the promise to reduce respondent burden, increase response rates, and possibly improve quality due to reduced respondent fatigue.

More recently, another motivation for SQD has been argued as more important. Studies have shown mixed evidence on the expected association between survey length and nonresponse (e.g., Bogen, 1996; Heerwegh and Loosveldt, 2006), with some studies showing significant associations, especially in self-administered surveys (e.g., Dillman, Sinclair, and Clark 1993, Galesic and Bosnjak 2009, Heberlein and Baumgartner 1978). Depending on the study design, sample members may be equally likely to respond to longer or shorter surveys but may provide different responses by survey length due to fatigue. Peytchev and Peytcheva (2017) demonstrated that a key strength of SQD is reduction of measurement error related to survey length. This has implications not only for the motivation whether to use SQD, but also for how to create and order the modules within questionnaire splits.

This review is organized in three sections: (1) a brief overview of SQD, (2) a review of methods to split questionnaires and assign to respondents, and (3) a review of the most relevant imputation methods, issues, and findings related to these methods.

1. Basic overview of SQD

When SQD was first explored, the general approach was to split the full questionnaire into modules, which were then combined into smaller forms (questionnaires). The “full questionnaire”, “module” and “form” terminology will be used for this rest of this review. Graham et al. (2006) discussed efficiency gains due to using a three-form SQD where missingness was assumed to be Missing Completely At Random (MCAR). They examined a four-module design (X, A, B, C) where module X was asked of all respondents and the three pair-wise combinations between A, B and C were used to create the three forms as shown in Table 1 below (table 2 from the original manuscript). The module asked of all respondents, X, consisted of survey items that required higher precision or items deemed important to inform imputation of omitted modules. The three-form approach allowed for the collection of more items than a single form survey of the same length and retained adequate or better power to examine associations between survey items. The authors noted that the order in which modules were assigned could change and that the total number of modules could be increased. Both approaches would lead to an increase in the total number of SQD instrument forms beyond the 3-form design, if all combinations and permutations are realized.

Table 1. The 3-Form Design, With X Set

	Module			
Form	X	A	B	C
1	1	1	1	0
2	1	1	0	1
3	1	0	1	1

Note. 1 = questions asked; 0 = questions not asked.

Many studies, including Raghunathan and Grizzle (1995) and Graham et al. (2006), showed promising results for SQD compared to a full survey design although there was often a loss in precision relative to the full questionnaire. Additionally, Ahmed et al. (2015) described the use of SQD on a survey in Pakistan where they compared the results directly to a full questionnaire. Means and standard deviations were calculated for all items; estimates from the SQD survey were not found to be significantly different from those based on the full questionnaire. All of these studies consider the full questionnaire as the gold standard; this is a limitation to which we later return.

The allocation of sample members to each form does not have to be equal. Chipperfield and Steel (2009) noted that a strong advantage of SQD is the flexibility to vary the sample size for different survey items, as some items may not need the same level of precision as others. The authors introduced a best linear unbiased estimator (BLUE) design-based approach for optimal allocation of survey items in a SQD based on minimizing variance subject to a fixed-cost or minimizing cost subject to a fixed variance.

The loss of precision in estimates due to SQD was explored by Merkouris (2015). The study explored modules consisting of overlapping subsets of questions which led to more efficient estimation because the correlations between all items could be leveraged. The process was based on BLUE but implemented a calibration weighting procedure on the sampling weights.

2. Creating Questionnaire Splits for SQD

2.1. Approaches to Creating the Questionnaire Splits

Statistical perspective

Gonzalez and Eltinge (2007) summarized the methods for creating the split questionnaire forms in the Consumer Expenditure Survey. They discussed the concepts of determining the number of modules and split forms, summarizing research into assignment of questions to different modules based on survey item logic, item themes and correlations between questions. One approach for splitting the full questionnaire was to first assign questions into modules based on logical constraints or thematic groups, and then randomly subsample those modules into forms. A second approach was the use of statistical criteria such as correlations among questions. Items that were highly correlated could be spread across modules so that not all modules needed to be included in each form. The strong correlations added more information to imputation models for item missingness to account for modules that were not part of the form. Another statistical approach was automatic allocation of questions using an algorithm, such as those discussed later in this section. Questions that are high priority, or those that require more precision, could be incorporated into a core set of questions that were included in all forms. The review noted that the number of modules to include in a form should be based on the targeted length of the interview and cognitive demand on the respondent.

Thomas et al. (2006) discussed a method to allocate questions to modules based on their correlations with other survey questions in the National Health and Nutrition Examination Survey (NHANES) data. The method involved an automated allocation of questions to modules using a predictive value based on comparing variance between the imputed estimator and non-imputed estimator. The authors found the approach to yield low bias compared to the full survey, but larger loss in precision than expected attributed to choosing a variety of NHANES health characteristics without giving enough consideration to the fact that they were not strongly correlated with other survey items, or represented rare events.

Ioannidis et al. (2016) started with the basic SQD approach where modules of exclusive questions were split across a set number of forms. They examined how to allocate modules to reduce survey cost and respondent burden based on overall precision requirements and building on Chipperfield and Steel (2009), where some modules were included in more forms than others to meet power requirements. The use of a random search optimization model was explored to minimize cost, while meeting precision and survey design requirements. The authors were able to reduce computational burden in their optimization model by reducing the set of considered forms. As described in the next section, one of the downsides of this approach is that the random search optimization model does not make use of correlations between items to inform the allocation of questions to modules.

Cognitive perspective

None of these approaches are consistent with what we know about cognitive processing in the context of survey design and administration. The popular models for human memory structure suggest that it would be optimal to organize a series of questions by domain (e.g., topic and chronology) rather than go back and forth across domains (Schank and Abelson's theory of scripts (1977); Shank's theory on memory organization packets (1982); Tulving's model for memory structure (1983); Kolodner's CYRUS model (1985), and Conway's multilayered model (1996)). The memory organization would also suggest that once a person is thinking on a topic, it would be easier and more accurate to recall information on this topic as opposed to switching to another, unrelated topic, which is what the sampling of questions from each domain would lead to.

Disrupting the structure of a questionnaire can result in unexpected context effects, especially in attitudinal questions. Context effects in surveys can be associated with the mode of data collection, question order, or response option order. Context effects due to question order are of focal interest within the context of SQD. Previous questions can provide context for subsequent questions in a number of ways—by establishing a norm of reciprocity (e.g., Hyman and Sheatsley's (1950) experiment with Communist reporters in the U.S. and U.S. reporters in a Communist country like Russia); by altering the frame in which subsequent questions are interpreted (e.g., knowledge questions followed by an attitude question and vice versa); by changing the salience of different alternatives when open-ended questions are asked (e.g., the "most important problem facing the country" is influenced by what questions were asked first); by creating part-whole contrast effects (e.g., Kalton, Collins and Brook's (1978) general-specific and specific-general questions), and by inducing response set bias in series of questions that use the same response format.

There are no procedures for eliminating question order effects. Randomization is often used for series of questions that use the same response format to minimize the impact of question order effects within the series—this approach does not eliminate the effect itself but ensures the response set bias is spread evenly. However, in SQD with a limited number of forms, it means that questions on the same topic can be asked in a different context for all respondents in one form verses another. This has multiple potential implications, including making estimates dependent on the sample allocation to each form. In SDQ, split assignments should follow the best practices of questionnaire design—modules should be grouped by topic and presented in a logical order.

2.2. Assignment of Questionnaire Splits to Sample Members

In response to Ioannidis et al. (2016), Chipperfield (2016) noted that the use of a Horvitz-Thompson estimator meant correlations among questions between different modules were ignored. He argued that the accuracy of the estimates, whether using a model-based or sampling model-assisted approach, could have been improved if estimation accounted for such correlations and if the SQD was designed to split items that were highly correlated across modules. Furthermore, limiting module assignment to a set number per form reduced the ability to analyze higher order interactions between questions that are spread out across multiple modules. Finally, Chipperfield (2016) argued that assignment to modules based on known respondent characteristics could lead to a more efficient design (Missing at Random instead of MCAR) due to gains from assigning modules to individuals based on who would contribute the most information to the imputation model. For example, when considering a health outcome such as diabetes Chipperfield noted that an individual with diabetes would contribute as much information to a maximum likelihood model as 400 people without diabetes.

In addition to the survey questions themselves, Gonzalez (2012) examined the use of prior information about respondents when determining the splits. His research was based on panel surveys used to produce cross-sectional population estimates, where information on sampled individuals was available from prior rounds. All participants were given a full questionnaire as their first interview and then five different approaches were tested to split the questionnaire using prior information known for panel members. He found that using prior information led to higher levels of precision compared to SQD designs that were not informed by such data and allowed for accurate targeting of questions, saving time and burden. This approach expanded on prior studies (e.g., Thomas et al., 2006) which only focused on the survey questions to split the questionnaire. Additional considerations and implications for longitudinal surveys (meant to be analyzed as a dataset with measures at multiple points in time) are addressed in the next section.

The use of respondent information to split forms for SQD was also explored for the New South Wales Health Survey by Chipperfield et al. (2018). They allocated SQD modules to individuals based on their characteristics. This meant that missing survey items due to module assignment were MAR, instead of the default MCAR, which improved efficiency. One important finding was that individual characteristics would allow to determine if a respondent would contribute meaningfully to the precision of the regression coefficient in a logistic regression model. This approach could guide the splits of the questionnaire, as individuals who are less likely to contribute to the precision of an estimate would not be asked the item associated with it. The approach focused on binary variables and requires prior information on the survey from a similar proxy survey or a prior round. Information loss due to SQD was simulated. The authors noted that all covariates should be collected for respondents with rare outcome characteristics, but for those without such rare characteristic, the covariates should be collected with a lower probability. In their evaluation, the impact of SQD was measured using cost models for time spent interviewing and estimated variance. It is important to note that this research focused on analytic modeling and not the estimation of means or totals. Consideration should be given to the impact on bias for means and totals due to prioritizing covariate responses from sample members with more rare outcomes. Bias could likely be reduced through proper implementation of imputation methods.

2.3. Extensions to longitudinal survey designs

There are a few studies that examine SQD in longitudinal surveys. Jorgensen et al. (2014) split a questionnaire using a standard 3-form approach and examined the impact on a longitudinal panel study. They tested different approaches to assign questions to forms and found that rotating the forms assigned to individuals across data collection waves performed the best with respect to relative efficiency and bias.

Imbriano (2018) and Imbriano and Raghunathan (2020) expanded upon the work in Jorgensen et al. (2014) using a 3-form approach for a longitudinal panel survey. Six different methods were tested to rotate the forms across waves, *de facto* varying three factors: assigning different or the same modules across waves after the initial assignment, combing both approaches, and the number of forms (this condition simply used random assignment). Evaluation used simulated data for the Health and Retirement Study (HRS) data. The performance of the six methods in the simulation was highly dependent on the correlation structure between the survey questions. As part of this research, Imbriano and Raghunathan (2020) also examined the use of Kullback-Leibler divergence to optimize survey question allocation to the forms. Using Bayesian approaches, they measured differences between the posterior distributions of questions including imputed missing data and the posterior distributions without using missing data. The conclusion was that the optimal approach depended on the correlational structure of the data and the estimands of interest. In all likely scenarios (presence of correlations among items in different modules and some autocorrelation), the authors acknowledged that assigning different modules across waves and using a limited number of forms outperformed all other approaches to assignment of the modules in the simulation. Not surprisingly, it was also the approach that performed best in terms of point estimates, variances, and estimates of change.

Ioannidis et al. (2016) investigated the use of SQD to integrate multiple separate and independent surveys. They expanded the approach to allocate modules based on precision constraints to account for the periodicity and overlap of different instruments. Their approach used a balancing matrix to align all instruments that overlapped on the same time schedules to combine their individual samples. A main strength of this approach is in its flexibility to incorporate new modules, or eliminate redundant ones, adjust the precision parameters for certain modules and cross modules with each other.

3. Imputation for Split Questionnaire Designs in Complex Samples

By definition, a split questionnaire survey design has missing data values by design, i.e., planned missingness. There are four general, and not mutually exclusive, categories of methods to account for the missing data: (1) complete or available case analysis, (2) weighting procedures, (3) imputation-based methods, and (4) model-based methods (Little and Rubin 2002, pp. 19-20). This section focuses on the imputation-based methods as a flexible method that does not ignore the missing data, leverages associations between variables, and yields complete rectangular datasets to which usual analytic approaches could be applied.

Ford (1983) provided a definition of imputation-based methods as a procedure that imputes a value for each missing value. How the value is calculated is not important to the definition. For example, a hot-deck imputation procedure is one that uses an observed value to replace the missing value. Other methods, e.g., regression models, may use a fully parametric model and draw imputed values for a posterior predictive distribution.

Historically, some of the main imputation methods included: use of dummy indicators, using the mean or mode as the imputed value, using expected values from regression models, and using the last observation carried forward in longitudinal designs. All of these methods often involve unrealistic assumptions, risk of bias in resulting estimates, and often distortion of distributions. Better imputation methods can preserve the distributional properties of the observed data and reduce the risk of bias in different estimands. Two very different approaches are hot deck and regression-based methods; both can employ sophisticated models, but one imputes with observed values and the other does not.

In this section, we review developments in hot deck and regression-based methods in the context of multiple imputation. We also include comparisons between the hot deck and multiple imputation methods. We note efforts to incorporate the complex survey design in the imputation process. Finally, we briefly discuss the incorporation of paradata into the imputation process, as a way to augment the imputation models.

3.1. Imputation Methodology

Multiple Imputation

Multiple imputation (MI) was first proposed by Rubin (1978) to calculate variances that capture the uncertainty due to imputation. To calculate the MI variance of an estimate of some parameter θ, e.g., a mean, one creates M independent imputations to create M complete data sets. Using standard statistical procedures, one estimates θ from each of the M complete data sets. The point estimate for θ is

$$\overline{\theta_M} = \sum_{m=1}^{M} \widehat{\theta_m},$$

where $\widehat{\theta_m}$ are the estimates of θ from each of the M complete data sets. Calculate the variance associated with $\widehat{\theta_m}$, and call it $\widehat{W_m}$. The average of the M variances is

$$\overline{W_M} = \sum_{m=1}^{M} \widehat{W_m}.$$

Calculate the between imputation variance as

$B_M = \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\theta_m} - \overline{\theta_M}\right)^2$.

The total variance of $\overline{\theta_M}$ is

$$V(\overline{\theta_M}\,) = \overline{W_M} + \frac{M+1}{M} B_M$$

SQD was first proposed by Raghunathan and Grizzle (1995), it included an MI method to produce imputed values for analysis. In that article they assumed simple random sampling. The authors mention complex surveys and state, "In complex surveys, the assignment of components to individuals sampled may depend on the sample design. For example, in stratified sample design, we may want to tailor the assignment mechanism so that the desired stratum specific population quantities (e.g., correlations coefficients) can be estimated. *The imputation method may also require modification in complex surveys* [Italics added]" (Raghunathan and Grizzle 1995, p. 61). Although, they do not explain what modifications may be required.

The inclusion of survey design information as independent variables in MI goes back to at least Rubin (1987). He states, "… if there exists any auxiliary information available on the units and we wish to incorporate it in our survey design or data analysis, then without loss of generality this information can be used to create covariates or stratification variables" (Rubin 1987, pp. 27-28).

Fay (1992) pointed out how MI variance estimates can be biased when the imputation and analysis models differ. Meng (1994) provided a framework for this bias and termed it "uncongeniality." Others have shown how this problem arises—particularly due to the complex survey design—and have suggested potential solutions (e.g., Robins and Wang, 2000; Kim et al., 2006). Xie and Meng (2017) provide a comprehensive description of the issue and solutions. For many, if not most, the potential bias in variance estimates due to uncongeniality is secondary, at best. The primary concern should be the unbiasedness of key estimates, and any bias in variance estimates can only be considered in this context.

Hot Deck Multiple Imputation (HDMI)

Ford (1983) presents three “general statistical characteristics which are major arguments for the use of hot-deck procedures: reduction of the nonresponse bias, production of clean data set [a data set that appears complete and consistent], and preservation of the distribution of the population as represented by a sample. The first characteristic is included in the third, but it is important enough to deserve separate attention” (Ford 1983, p. 187). The third characteristic encompasses preserving joint and marginal distributions. That is, internal consistency by examining relationships among variables for a given sampling unit and external consistency by examining the relationships among the sampling units for a given variable, i.e., the distribution properties of each variable (Ford 1983, p. 190).

In hot-deck imputation, classes—mutually exclusive groups—of sample records are created so that the sample members are homogenous within each class and heterogenous across classes. Missing values for a record are imputed using observed values from other records in the same class.

Analogous to regression-based methods that depend on the model specification, the burden here is on the selection of classification variables (Ford 1983, p. 186). However, a key difference is that disjoint groups need to be selected, and the number of cells needs to be controlled.

A comprehensive contemporary review article for hot deck imputation is *A Review of Hot Deck Imputation for Survey Non-response* by Andridge and Little (2010). It discusses methods for creating donor pools such as adjustment cell methods and metrics for matching donors to recipients; multivariate missing data describing approaches to different patterns of missingness; and variance estimation such as explicit variance formulae, resampling methods for single imputation, and multiple imputation for hot deck using a “proper” multiple imputation method. It also discusses the role of sampling weights, properties of hot deck estimates, and detailed examples.

As stated above, Andridge and Little (2010, p. 54) mention multiple imputation for hot deck in the context of variance estimation. They cite Rubin (1978) that when the same donor pool is used for a respondent’s missing value for all MI data sets, the method is not a proper MI procedure. To properly implement HDMI, Bayesian Bootstrap (Rubin 1981) or Approximate Bayesian Bootstrap (Rubin and Schenker 1986) methods could be considered when creating the multiple imputations.

Cranmer and Gill (2013) present a method they call *multiple hot deck imputation*. “This tool is designed specifically to work well in situations where (traditional) parametric multiple imputation falls short” (Cranmer and Gill 2003, p. 426). Their major criticism is that parametric multiple imputation assumes a continuous outcome, produces continuous imputed values, and transforms the imputed continuous values to categorical data. This criticism may not be as valid now that multiple imputation procedures incorporate a wider variety models, which include models for categorical data. A caution with this method is that their procedure uses affinity scores to create imputation classes and hot deck to select donors for missing values across the multiple data sets. Consequently, their procedure may suffer from the caution provided in the preceding paragraph about using the same donor for a respondent across all data sets.

Sequential Regression Multiple Imputation (SRMI)[1]

For each of the M imputation data sets created, the sequential regression multiple imputation approach (van Buuren, *et al*. 1999, Raghunathan *et al*. 2001, van Buuren and Groothuis-Oudshoorn 2011, Raghunathan 2016, Raghunathan *et al*. 2016, and van Buuren 2018) implements an approach that eventually uses all variables to imputed the missing values for each variable. To begin the process, the complete response variables on the data set are used as possible independent variables in the imputation model for the "first" variable with missing data. There are different variable visit sequences defining the first variable, e.g., smallest amount of missingness to largest amount of missingness or as they appear on the data file. The process proceeds sequentially through the other variables with missing data using all the complete response variables and any previously imputed variables as possible independent variables in the imputation model. Once all the variables on the data have been imputed, the first cycle of the process is complete.

For any additional cycles, the imputed values for the first variable imputed are replaced by new imputations using all the other variables on data set, both complete response and imputed variables, as possible independent variables in the imputation model. The process proceeds sequentially through the other variables as it did in the first cycle using all the other variables on the data set, including the re-imputed variables, as possible independent variables in the imputation model. Once all the variables with initial missing data have been re-imputed, the second cycle is complete. The process proceeds through any additional specified cycles to complete the model-based imputation process for the first imputed data set. Raghunathan (2016, p. 69) states, "Empirical analysis show that 5 to 10 iterations are sufficient to condition the imputed values on any variable on all other variables." Van Buuren (2018, p. 120) concurs with Raghunathan but later states that the MICE algorithm should be monitored for convergence (van Buuren 2018, pp. 187-188) by using a combination of tools to assess convergence. This process is repeated for each multiply-imputed data set.

[1] *SRMI is also known as fully conditional specification or multivariate imputation by chained equations.*

Complex Survey Design

Accounting for the complex survey design and differential weighting in imputation is often not addressed when imputing data from a complex survey design. Most imputation methods are applied to data from simple random samples or complex survey samples treated as simple random samples. There has been some focus on how to incorporate the complex survey design and differential weighting in imputation methods. The general approach to incorporating the complex survey design in MI is to include the design variables as covariates in the imputation models. For example, in the context of split questionnaire design from the National Health and Nutrition Examination Survey, Thomas *et al*. (2006) provided a more detailed explanation of how they accounted for the complex survey design for their simulation using MI. In their simulation they included strata and clusters nested within strata as main effects in their imputation model. In addition, they include the logarithm of the sample weight as a main effect in the model along with core data items, which are especially important or are predictive of many of the spit items, and split items, the items only administered to a set of respondents.

For a more detailed look at the use of sampling weights, Andridge and Little (2009) investigated the preferred way to incorporate the sampling weights into hot deck imputation. Their conclusion was that the correct approach is to use the sampling weight as a stratifying variable alongside additional adjustment variables when forming adjustment cells.

Other Alternatives

A more recent article by Brunton-Smith and Tarling (2017) mentions a third imputation approach, doubly robust inverse probability weight (DRIPW). “Of the three methods listed, [MI, FIML, and DRIPW], multiple imputation (MI) is emerging as the most popular” (Brunton-Smith and Tarling 2017, p. 710). In addition, they focus on a multilevel imputation procedure in Stat-JR using the 2LevelImpute templet developed by the Centre for Multilevel Modeling (Charlton et al. 2013, Goldstein and Parker 2014).

Thomas *et al*. (2006) provide an alternative to multiple imputation for SQD. They state, “An alternative to the multiple imputation estimation is two-phase weighting based on core item estimators and their differences between blocks. Any advantage in efficiency from multiple-imputation estimation would be due to the additional information from the split items” (Thomas *et al*. 2006, p. 227). Thomas *et al*. (2006) also conclude that, for regression problems, “the multiple-imputation estimators are generally more efficient than the no-imputation estimators for regression problems. Nevertheless, the losses in efficiency of the multiple-imputation estimators relative to the complete-data estimators from some coefficients as well as gains in efficiency for other coefficients are worth further investigation" (Thomas *et al*. 2006, p. 229).

3.2. Comparisons

In the context of democratization and modernization theory, Cranmer and Gill (2013) used a Monte Carlo simulation to demonstrate their "*multiple hot deck imputation*," which was specifically designed to work well in situation where (traditional) parametric multiple imputation falls short, i.e., using a continuous distributional assumption for a distribution with a few discrete values. Under the conditions of the Monte Carlo simulation, multiple hot deck imputation was less biased than casewise deletion and parametric multiple imputation using the Amelia package implemented in R. Although, as noted previously, their procedure may suffer from using the same donor for a respondent across all data sets. In addition, more recent advances in MICE have include a wide variety of methods that can handle discrete data. Consequently, this may not be as much of a problem as when this article was published.

Using the Current Population Survey Annual Social and Economic Supplement data, Hokayem, Raghunathan, and Rothbaum (2015) compared estimates, e.g., median income and poverty estimates, between hot deck and two SRMI procedures. One of the SRMI procedures uses only the survey data; the other procedure uses the survey data and additional auxiliary information. Their general conclusion is, even with the added variance introduced by accounting for the imputation uncertainty, both SRMI models have more precise estimates of poverty than hot deck.

A 2017 study (Center for Behavioral Health Statistics and Quality, 2017) that compared predictive mean neighborhood, weighted sequential hot deck, sequential regression multivariate imputation, and a modified predictive mean neighborhood multiple imputation for the *National Survey on Drug Use and Health* (NSDUH) concluded that the results did not differ much by method, for estimates of substance use. However, the report noted that this could have been due to the low item nonresponse that was being imputed—an average of less than 5 percent. From a practical perspective, the simpler methods with fewer consistency checks yielded more inconsistent data that makes it more difficult for analysis.

A comparative study of modern imputation techniques by Zaninotto and Saker (2017), investigated FIML and three MI imputation techniques, multivariate normal imputation, multiple imputation by chained equations (MICE), and, for repeated measures, the two-fold fully conditional specification. Their conclusion was that, under the conditions of the simulation, MICE is the preferred method. For a continuous outcome, they showed in longitudinal studies with non-monotone missing, FIML and MI techniques all perform well. Comparing MI and FIML techniques, MI might be better at incorporating interaction terms. It was also shown that MICE and two-fold FCS produced estimates that were more accurate and precise than those obtained from FIML and multivariate normal imputation techniques, especially when dealing with non-continuous variables and interaction terms. The results of this study showed that MICE in general showed slightly better precision and accuracy than two-fold FCS.

Bertsimas, Pawlowski, and Zhuo (2018) propose an imputation method based on a general optimization framework with a predictive model-based cost function that explicitly handles both continuous and categorical variables and can be used to generate multiple imputations. They review an extensive list of different imputation methods. Of the approximately 20 imputation methods they review, five methods use single imputation, (i.e., mean imputation, K-nearest neighbors (KNN), iterative KNN, Bayesian principal component analysis, and predictive mean matching), and one imputation method uses multiple imputation, multivariate imputation by chained equations. The apply their method and the other imputation methods to 84 data sets taken from the UCI Machine Learning Repository and compare the imputation methods. They show that their method yields statistically significant gains in imputation quality over the alternative imputation methods, which leads to improved performance for classification and regression models based on the imputed data.

3.3. Paradata

Paradata (Couper, 1998) was first defined as data that are created in the process of collecting survey data, such as keystroke data. It has since been expanded to include a variety of sources, including deliberately designed interviewer observations. Paradata have received little attention in imputation. Peytchev (2012) used different types of paradata to impute for nonresponse and measurement error in the National Survey of Family Growth data. He found that interviewer-related paradata such as education and religiosity were related to nonresponse, while respondents' self-reports about comfort with providing sensitive information were related to measurement error. These variables were then used in multiple imputation using SRMI to augment the models.

In the context of a longitudinal study of prison inmates, Brunton-Smith and Tarling (2017) describe their use of paradata to understand unit nonresponse and create variables for use in the imputation process. They explored the reasons for data being missing, details of the timings for each interview attempt, and the prison in which the interview was to take place. In addition to all the survey items collected during the initial interview, administrative data related to reoffending, and prison characteristics, the paradata “were used as the basis for two multilevel logistic regression models: the factors associated with non-contact; and the factors associated with non-compliance conditional on successful contact. As unit response rates varied between prisons, a multilevel structure (Goldstein 2011) was specified, with prisoners grouped within prisons, reflecting additional structural constraints on prisoners’ involvement in the follow up interview” (Brunton-Smith and Tarling 2017, p. 713). “The models confirmed the importance of structural processes outside of the control of individual prisoners in driving unit missingness. For example, in the contact model, sufficient lead time to secure contact with prisoners who had shorter sentences, those serving theft, drug offences were more difficult to contact. For the cooperation model, prisoners who had a previous prison sentence or who had previously convicted of burglary had higher odds of contact (Brunton-Smith and Tarling 2017, p. 714).

4. Summary

A literature review is helpful in identifying not only relevant findings, but also omissions that merit further research. There are some notable limitations in the literature related to SQD, related to the almost exclusive focus on statistical aspects and limited experimentation:

- Almost all studies are based on simulations, so the methods used rely on model assumptions rather than design.
- Simulation studies do not allow us to study the nonresponse and measurement properties of SQD.
    - Nonresponse due to survey length cannot be simulated.
    - Changes in responses due to length cannot be simulated.
    - Context effects that are due to different questionnaire structure cannot be examined.

To elaborate, when evaluating alternative splits for a questionnaire, simulation studies use generated data or observed full survey data and impose different scenarios. These simulations require assumptions that are unlikely to hold true, based on the findings in the research literature on questionnaire design. Question order effects would suggest that questions cannot be randomly sampled into a form, yet simulations assume no variation in responses related to how the splits are created.

The second set of related limitations is the inability to demonstrate the potential benefits of using a SQD beyond reduction in variance. The common motivation is the reduction in respondent burden, but empirical experiments (note that in some disciplines simulations are referred to as experiments or simulated experiments) are needed to examine potential reduction in nonresponse bias and measurement error.

There are some direct implications for our future research:

- We will start with constraints imposed by questionnaire design and only then consider additional factors such as inter-item correlations.
- We still find greatest promise in the SRMI method, and find the HDMI method to be the most promising major alternative. This review did not find any new major alternative.

Despite the introduction of SQD a quarter of a century ago, there are substantial gaps in our understanding of when SQD is feasible and desirable. Additional research comparing alternative imputation procedures and possible covariates (e.g., paradata) is also needed.

References

Ahmed, Alia, Lodhi, Suleman A., and Ahmad, Munir. (2015). Using Split-Questionnaire Design: An Empirical Analysis. *Pakistan Journal of Statistics* vol. 31, no. 2, pp. 211-218.

Andridge, Rebecca R. and Roderick J. A. Little. (2009). The Use of Sample Weights in Hot Deck Imputation. *Journal of Official Statistics*, Vol. 25, No. 1, pp. 21-36.

Andridge, Rebecca R. and Roderick J. A. Little. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, Vol. 78, No. 1, pp. 40-64.

Bertsimas, Dimitris, Colin Pawlowski, and Ying Daisy Zhuo. (2018). From Predictive Methods to Missing Data Imputation: An Optimization Approach. *Journal of Machine Learning Research*, Vol. 18, No. 196.

Bogen, K. (1996). *The effect of questionnaire length on response rates -- A review of the literature.* Paper presented at the Survey Research Methods Section of the American Statistical Association.

Brunton-Smith, Ian and Roger Tarling. (2017). Harnessing Paradata and Multilevel Multiple Imputation When Analysing Survey Data: A Case Study. *International Journal of Social Research Methodology*, Vol. 20, No. 6, pp. 709-720.

Center for Behavioral Health Statistics and Quality. (2017). *Evaluation of Imputation Methods for the National Survey on Drug Use and Health*. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Charlton, C. M. J., D. T. Michaelides, R. M. A. Parker, B. Cameron, C. Szmaragd, H. Yang, …, and W. j. Brown. (2013). *Stat-JR Version 1.0*. Centre for Multilevel Modeling, University of Bristol & Electronics and Computer Science, University of Southhampton. URL http:/www.bristol.ac.uk/cmm/software/statjr/

Chipperfield, James O. and Steel, David G. (2009). Design and Estimation for Split Questionnaire Surveys. *Journal of Official Statistics* vol. 25, no. 2, pp. 227-244.

Chipperfield, James O., Barr, Margo L., and Steel, David G. (2018). Split Questionnaire Designs: collecting only the data that you need through MCAR and MAR designs. Journal of Applied Statistics vol. 45, no. 8, pp 1465-1475.

Chipperfield, James O. (2016). Discussion. *Journal of Official Statistics* vol. 32, no. 2, pp. 287-289.

Conway, M. A. (1996). Autobiographical Knowledge and Autobiographical Memories. In D. C. Rubin (Ed.), Remembering Our Past, pp. 67-93. Cambridge: Cambridge University Press

Couper, M. P. (1998). *Measuring Survey Quality in a CASIC Environment.* Paper presented at the Invited paper presented at the Joint Statistical Meetings of the American Statistical Association, August, Dallas, TX.

Cox, Brenda G. (1980). The Weighted Sequential Hot Deck Imputation Procedure. *Proceedings of the American Statistical Association Section on Survey Research Methods*, pp. 721–726.

Cranmer, Skyler J. and Jeff Gill (2013). We Have to be Discrete about This: A Non-Parametric Imputation Technique for Missing Categorical Data. *British Journal of Political Science*, Vol. 43, No. 2, pp. 425-449.

Department of Education, National Center for Education Statistics (1997)*. National Postsecondary Student Aid Study, 1995-96 (NPSAS:96), Methodology Report*, NCES 98-073, by John A. Riccobono, Roy W. Whitmore, Timothy J. Gabel, Mark A. Traccarella, and Daniel J. Pratt, Research Triangle Institute; Lutz K. Berkner, MPR Associates, Inc. Andrew G. Malizio, project officer. Washington, DC. URL https://nces.ed.gov/pubs98/98073.pdf

Dillman, Don A., Michael D. Sinclair, and Jon R. Clark. 1993. "Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates of occupant-addressed census mail surveys." *Public Opinion Quarterly* 57 (3):289-304. doi: 10.1086/269376.

Fay, R. E. (1992). When are Inferences from Multiple Imputation Valid? Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 227-232

Ford, Barry L. (1983). An Overview of Hot-Deck Procedures. *Incomplete Data in Sample Surveys: Theory and Bibliographies*, Vol. 2, pp. 185-207. New York, NY: Academic Press, Inc.

Galesic, Mirta, and Michael Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73 (2):349-360. doi: 10.1093/poq/nfp031.

Goldstein, H. (2011). *Multilevel Statistical Models*. 4th ed. London: Arnold.

Goldstein, H. and R. Parker (2014). *Missing Data via Multiple Imputation in Stat-JR*. Stat-JR templet (2LevelImpute). URL http:/www.bristol.ac.uk/cmm/software/statjr/

Gonzalez, Jeffrey M. and Eltinge, John L. (2007). Multiple Matrix Sampling: A Review. *Section on Survey Research Methods*, December.

Gonzalez, Jeffrey M. (2012) The use of Responsive Split Questionnaires in a Panel Survey. Dissertation submitted to University of Maryland College Park for Doctor of Philosophy.

Good, Irving J. (1969), “Split Questionnaires I”, *The American Statistician*, 23(4), 53-54.

Good, Irving J. (1970), “Split Questionnaires II”, *The American Statistician*, 24(2), 36-37.

Heberlein, Thomas A., and Robert Baumgartner. 1978. "Factors Affecting Response Rates to Mailed Questionnaires: A Quantitative Analysis of the Published Literature." *American Sociological Review* 43 (4):447-462.

Heerwegh, D., & Loosveldt, G. (2006). An experimental study on the effects of personalization, survey length statements, progress indicators, and survey sponsor logos in Web Surveys. *Journal of Official Statistics, 22*, 191-210.

Herzog, Regula A. and Jerald G. Bachman (1981), “Effects of Questionnaire Length on Response Quality,” *Public Opinion Quarterly*, 45 (4), 489-504.

Hokayem, Charles, Trivellore Raghunathan, and Jonathan Rothbaum (2015). *Sequential Regression Multivariate Imputation in the Current Population Survey Annual Social Economic Supplement*. Proceedings of the Joint Statistical Meetings, Survey Research Methods Section.

Holland, P. W., and Thayer, D. T. (1985), "Section Pre-Equating in the Presence of Practice Effects," *Journal of Educational Statistics,* 10, 109-120.

Holland, P. W., and Wightman, L. E. (1982), "Section Pre-Equating: A Preliminary Investigation," in *Test Equating,* eds. P. W. Holland and D. B. Rubin, New York: Academic Press, pp. 271-297.

Hyman, H. and Sheatsley, P. (1950). The Current Status of American Public Opinion. In J.C. Payne (ed.) The Teaching of Contemporary Affairs. Twenty-first Yearbook of the National Council of Social Studies, pp.11-34.

Imbriano, Paul M. Methods for Improving Efficiency of Planned Missing Data Designs. Dissertation submitted to The University of Michigan for Doctor of Philosophy (2018).

Imbriano, P. M., & Raghunathan, T. E. (2020). Three-Form Split Questionnaire Design for Panel Surveys. *Journal of Official Statistics, 36*(4), 827-854.

Ioannidis, Evangelos; Merkouris, Takis; Zhang, Li-Chun; Karlberg, Martin; Petrakos, Michalis; Reis, Fernando; and Stavropoulos, Photis. (2016). On a Modular Approach to the Design of Integrated Social Surveys. *Journal of Official Statistics* vol. 32, no. 2, pp. 259-286.

Jorgensen T, Rhemtulla M, Schoemann A, McPherson B, Wu W, Little T. (2014). Optimal assignment methods in three-form planned missing data designs for longitudinal panel studies. *International Journal of Behavioral Development*., 38(5):397-410.

Kalton, G., Collins, M. and Brook, L. (1978). Experiment in Wording Opinion Questions. *Journal of the Royal Statistical Society Series C*, 27, pp.149-161.

Kim, J. K., Michael Brick, J., Fuller, W. A., and Kalton, G. (2006). On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(3), 509-521.

Kolodner, J. (1985). Memory for Experience. In G. H. Bower (Ed.), The Psychology of Learning and Motivation, Vol. 9, pp. 1-57. Orlando, FL: Academic Press

Little, Roderick J. A. and Donald B. Rubin (2002). *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Meng, X. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9 , 538-558.

Merkouris, Takis. An efficient estimation method for matrix survey sampling. (2015). *Survey Methodology,* 41 , no. 1, pp. 237 —262.

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

Peytchev, A. (2012). Multiple Imputation for Unit Nonresponse and Measurement Error. *Public Opinion Quarterly, 76*(2), 214-237. doi:10.1093/poq/nfr065

Peytchev, A., & Peytcheva, E. (2017). Reduction of Measurement Error due to Survey Length: Evaluation of the Split Questionnaire Design Approach. *Survey Research Methods, 11*(4), 361-368.

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

Raghunathan, Trivellore E. and James E. Grizzle. (1995). A Split Questionnaire Survey Design. *Journal of the American Statistical Association*, Vol. 90, No. 429, pp. 54-63.

Raghunathan, Trivellore (2016). *Missing Data Analysis in Practice*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Raghunathan, T., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*. Vol. 27. Pp. 85-95.

Raghunathan, T., Solenberger, P., Berglund, P., and Van Hoewyk, J. (2016). *IVEware: Imputation and Variance Estimation Software (Version 0.3)* Accessed June 14, 2020 from https://www.src.isr.umich.edu/wp-content/uploads/iveware-manual-Version-0.3.pdf

Robins, J., and Wang, N. (2000). Inference for Imputation Estimators. *Biometrika, 87*(1), 113-124.

Rubin, Donald B. (1978). Multiple Imputations is Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*. Pp. 20-34.

Rubin, Donald B. (1981). The Bayesian Bootstrap. *Annuals of Statistics*, Vol. 9, pp. 130-134.

Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse is Surveys*. Hoboken, NJ: John Wiley & Sons.

Rubin, Donald B., and Nathaniel Schenker. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Non-Response. *Journal of the American Statistical Association*, Vol. 81, pp. 366-374.

Schank, R. C. and Abelson, R. P. (1977). Scripts, Plans, Goals, and Understanding. Hillsdale, NJ: Erlbaum.

Schank, R. C. (1982). Dynamic Memory. Cambridge: Cambridge University Press

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

Thomas, Neal, Trivellore E. Raghunathan, Nathaniel Schenker, Myron J. Katzoff, and Clifford L. Johnson. (2006). An Evaluation of Matrix Sampling Methods Using Data form the National Health and Nutrition Examination Survey. *Survey Methodology*, Vol. 32, No. 2, pp. 217-231.

Tulving, E. (1983). Elements of Episodic Memory. Oxford: Oxford University Press van Buuren, Stef (2018). *Flexible Imputation of Missing Data*. 2nd ed. Boca Raton, FL: CRC Press, Taylor & Francis Group.

van Buuren, S., H. C. Boshuizen, and D. L. Knook (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*. Vol. 18. Pp. 681-694

van Buuren, Stef and Karin Groothuis-Oudshoorn (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

Xie, X., and Meng, X. (2017). Rejoinder please visit the wild arboretum of multi-phase inference. *Statistica Sinica*, 27 , 1584-1594.