National Center for Science and Engineering Statistics

Title: Split Questionnaire Design Final Report

Date: February 2023
Final Report

Contractor Awardee: RTI International
Contract Number: 49100420C0021

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

Split Questionnaire Design Final Report

Andy Peytchev, Study PI and RTI Fellow, RTI

February 28, 2023

This memo summarizes the key findings from the Split Questionnaire Design Broad Agency Agreement (SQD BAA) and includes recommendations. Detailed results will be disseminated through a peer-reviewed publication, the manuscript for which is in progress. The team included: Andy Peytchev, Emilia Peytcheva, David Wilson, Darryl Creel, Darryl Cooney, and Jeremy Porter. Collaborators at NCSES were: Jennifer Sinibaldi, Matt Williams, and John Finamore.

Survey length has been found to affect both nonresponse and measurement error. Although findings from very different survey designs show mixed results, longer surveys have been found to lead to greater nonresponse. The findings on survey length and measurement error are more limited yet more definitive, showing greater measurement error as respondents have to answer more questions, including our prior work in this area. Yet, a key strength of surveys is to provide variable-rich data that allow investigations of the relationships between variables.

Split Questionnaire Design (SQD) is an approach to reduce the number of questions asked from each respondent, while yielding the complete dataset with all survey variables for analysis. The premise is to divide the questionnaire into distinct modules—splits—and assign each respondent to be asked a subset of these modules. The missing data from the modules that are not asked are imputed. Multiple imputation is used to reflect the uncertainty in the imputed values—and, conversely, to reflect the certainty gained from informative statistical models.

Implementation on a survey would involve substantial risks without sufficient evaluation of (1) how to create the questionnaire splits, (2) what imputation method is most suitable for a particular data structure, and (3) how to incorporate complex survey design information. These are the research questions that were addressed in this BAA using NCSES' 2019 National Survey of College Graduates (NSCG) data. If the results are deemed sufficiently encouraging, and superior methods are identified, the results would inform an empirical test in an upcoming data collection.

Creation of the Questionnaire Splits. From a cognitive perspective, an optimal survey design would ask questions that are related, together, as it facilitates retrieval processes and follows conversational norms. Surveys such as the 2019 NSCG often include complex skip logic that further forms groups of questions on a topic that are asked together. Form a statistical perspective, similar questions would be assigned to different modules, to better inform imputation models for questions that are not asked. A purely statistical approach to creating the modules could yield splits that are excessively difficult for a respondent to follow

and process, and even infeasible with the extensive skip logic in NSCG. Therefore, we developed a ***logical split*** based on questionnaire design considerations largely following the topical modules in NSCG, and a ***statistically-informed split*** that modified the assignment of subsets of questions to modules based on associations between variables. For this purpose, we developed heatmaps based on the full correlation matrix and used rules to identify when there are insufficient predictors of a particular variable. The resultant split followed reasonable logic from a questionnaire design perspective, yet was further informed by statistical criteria.

Imputation Methods. There are different imputation methods that could be used for multiple imputation in SQD. These methods can be grouped into two types. One type are statistical models that build multivariate distributions from which to draw values, such as ***regression-based imputation***. The other type of methods use models only to identify observed donor values for the missing data, such as ***Hot Deck imputation***.

For regression-based imputation we began using IVEware, a SAS-callable macro library software that can perform multiple imputation for continuous, categorical, and count data and the ability to fully specify bounds and conditional restrictions as those encountered in the NSCG's complex skip logic. Despite our past experience with this software, this was the first instance where it could not be used, generating critical errors that could not be ameliorated. Ultimately, we believe that the data structure, including skip patterns and associations, underly the failure of the application of this software solution to NSCG. While extremely unfortunate, identification of such critical problems was part of the evaluation on this BAA. After many attempts to identify the cause and to work around the issues, and reaching out to experts in the field, the team had to eventually turn to another statistical package. As IVEware was already being used from within a SAS environment, and based on expert opinion, we opted for using SAS PROC MI—the internal SAS regression-based and fully conditional specification (FCS) procedures for multiple imputation. There were numerous deficiencies and issues with implementation of SAS PROC MI, such as an inability to set all necessary restrictions and lack of model diagnostics. In addition to specific issues, this approach was extremely time intensive, initially taking over one week to complete the 32-imputed datasets for the logical or statistically-informed split, with running parallel programs on multiple server nodes. The team worked to identify and implement solutions, such as restructuring the variables, variable transformations, reducing rather than increasing the number of iterations, among others, and we plan to have a paper to detail these limitations and efforts to aid other researchers. However, there were still many variables for which the imputation models did not converge. The bias in the survey estimates was substantial for some of the variables, and that was not limited to variables where the models had not converged.

For Hot Deck imputation, we used bootstrap sampling, recursive partitioning, Weighted Sequential Hot Deck (WSHD), and cycling to implement the multiple imputation process. The process used bootstrap sampling to create the imputation replicate data sets. Independently, for each data set, the process created

imputation classes, imputed for missing values, and cycled through the imputed values. The terminal nodes from a recursive partitioning algorithm were used to create the imputation classes. Within an imputation class, based on the weight and order of the observations in the imputation class, WSHD chose a donor's value for a recipient with a missing value. Once the initial imputation for all variables with missing values had been completed, the process cycled through the data set several times. The cycling consisted of reconstituting the missing values for a variable, using all the other variables on the data set as possible predictors, imputing, and moving to the next variable in the data set with missing values. Although computationally intensive, the WSHD proved unproblematic relative to the model-based approach.

Results

Imputation Method. For the 2019 NSCG data we found that the Hot Deck approach substantially outperformed the fully model-based approach. ***Table 1*** shows the summary statistics across 594 categorical variable estimates (including both outcomes for dichotomous variables) for both imputation approaches and each split creation approach. The average absolute bias for the fully model-based imputation is about 3 (1.12/0.34 and 1.01/0.33) times larger than under the Hot Deck-based method. The standard error ratio, defined as the ratio of the standard error under multiple imputation to the standard error of the observed data is 63% (2.89/1.67) to 73% (2.72/1.67) higher in the fully model-based method. The Fraction of Missing Information (FMI)—defined as the between imputation variance to the total variance (between and within variance)—may seem surprising to be lower in the model-based method, given the preceding summary statistics. Upon further investigation it became apparent that it is a function of how this method failed for some variables. It consistently imputed very biased values for some respondents, across the multiply-imputed datasets, disproportionately increasing the within-imputation variance relative to the between imputation variance.

Table 1. Summary statistics across 594 categorical variable estimates for each imputation approach and each split creation approach

Imputation Method	**Model-Based MI**		**Hot Deck-Based MI**	
Split Creation Method	**Logical**	**Statistically-Informed**	**Logical**	**Statistically-Informed**
Absolute Bias	1.12	1.01	0.34	0.33
Absolute Relative Bias	0.15	0.14	0.03	0.03
Standard Error Ratio	2.89	2.72	1.67	1.67
Fraction of Missing Information (FMI)	0.45	0.44	0.63	0.63

Split Creation Approach. Another surprising finding was with regard to the approach to split creation. A ubiquitous objective and recommendation is to create the splits to maximize association between

variables assigned to different modules. To do so, requires time, effort, and increased risk of measurement error from asking respondents questions in a sequence that switches topics back and forth.

Yet, the summary statistics under the Hot Deck-based method are virtually identical for the logical and statistically-informed splits. The summary statistics are similar under the model-based imputation method, and given the model convergence issues, it is difficult to ascertain whether the small differences are due to the split creation approach or due to the increased variability.

Figure 1 shows that the logical and statistical splits are similar not just on average, but also at the estimate-level. The scatterplot with the regression-based imputation shows much more variability and a slight preference for the statistical split, while the WSHD imputation shows both lower absolute bias and lack of estimates on a particular side of the diagonal.

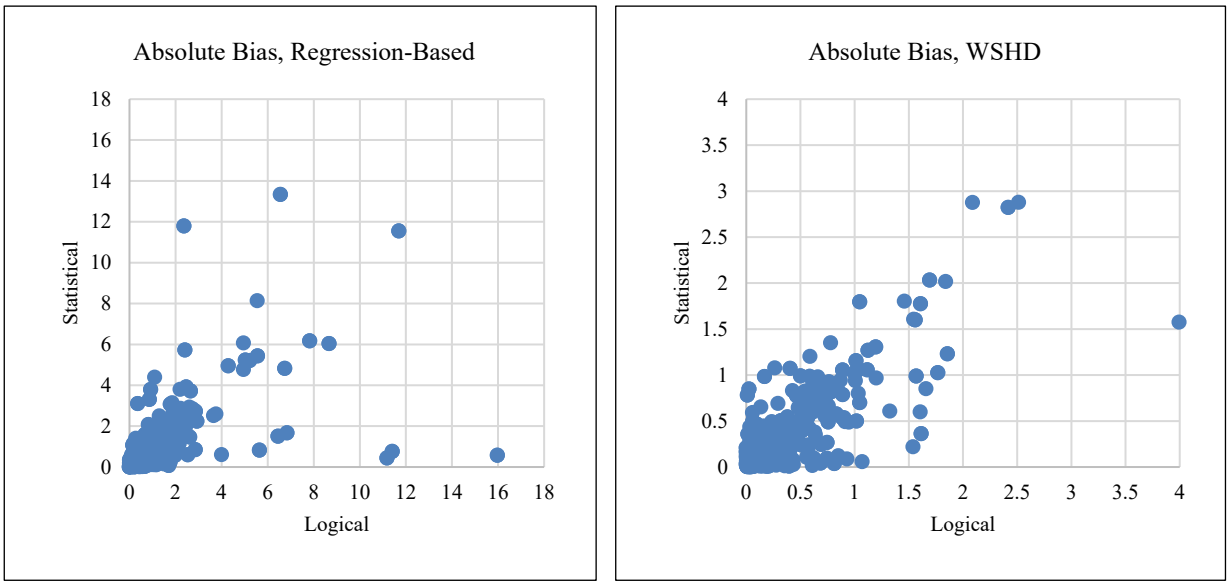

Figure 1. Absolute bias in the statistical and logical splits, for the regression-based and WSHD multiple imputation

Incorporation of the Complex Survey Design. For NSCG the complex survey design proved to be much less of an issue than on other surveys with complex survey design. The sample is not geographically clustered. The design strata have been found to have little impact on variance estimates and are based on demographic, education, and occupation variables that were included in the models. One of the stregths of the selected Hot Deck imputation method is that it uses weights within the imputation classes. In the model-based imputation, inclusion of the weights in the models in different ways made no impact.

Recommendations

There are several recommendations that can be made based not only on the outcomes, but also on the challenges encountered on the study:

1. Creating the questionnaire splits using topical modules is desirable. There were no observable positive impacts from the extensive exercise to identify survey questions to reassign to another split in order to increase cross-module associations. Moreover, such reassignment carries substantial risks with regard to how the respondent interprets, processes, recalls, and responds to questions that deviate from a topical order, that can result in increased measurement error bias and/or variance. This is the first such finding, and should be welcome to survey practitioners working on split questionnaire design.

2. WSHD performed far better than the fully model-based approaches. In addition to yielding low bias and acceptably low variance estimates, it requires somewhat less set up as the skip logic is identified by the nodes from the recursive partitioning model building step. Specifying all skip logic for NSCG is feasible—and we did that for both IVEware and SAS PROC MI—but adds to the time and effort to set up the imputation. That difference in effort does not compare to the amount of effort that the model-based approaches required, for the NSCG data. IVEware could not handle this dataset and imputation problem. SAS PROC MI lacked sufficient control over its implementation, and its diagnostic statistics were severely inadequate. Had IVEware been able to work with these data, it may have been a time-efficient solution as it is design for survey data with skip logic.

3. Incorporation of the design variables did not prove to be problematic with the NSCG data and these imputation methods. We did not find the results to be sensitive to how the design information was incorporated.

These simulations on existing data demonstrate how much losses can be minimized from not asking all questions from all respondents. A shorter survey may yield higher response rates, less nonresponse bias, and less measurement error. These potential benefits can only be evaluated through an experiment. This study has provided feasibility and guidance on methods—using logical/topical splits and WSHD—that can be used for the design of the experiment and for imputation of the resulting data.