Missing by Design Patterns for Optimizing Survey Response by Efficient and Consistent Data Collection

Dissertation

Presented to the Faculty for Social Sciences, Economics, and Business Administration at the University of Bamberg in Partial Fulfillment of the Requirements for the Degree of

DOCTOR RERUM POLITICARUM

by
Sara Bahrami, M.Sc. Phys.
born May 29, 1979 in Shiraz, Iran

Date of submission
7th June 2020

Principal advisor:	Prof. Dr. Christian Aßmann University of Bamberg, Germany
Reviewers:	Prof. Dr. Timo Schmid Free University of Berlin, Germany Prof. Dr. Henriette Engelhardt-Wölfler University of Bamberg, Germany
Date of submission	7th June 2020
Date of defence	11th December 2020

URN: urn:nbn:de:bvb:473-irb-494874
DOI: https://doi.org/10.20378/irb-49487

Acknowledgments

Firstly, I would like to offer special thanks to my former supervisor, Prof. Susanne Rässler, who provided me an opportunity to join her team and, although no longer with us, continues to inspire by her example and dedication to the students she served over the course of her career.

I would like to thank Prof. Christian Aßmann for taking over the supervision of my PhD which was an enormous relief. I am thankful for his guidance and valuable ideas for accomplishing a part of research in this thesis.

I would also like to acknowledge Dr. Florian Meinfelder as the reader of this thesis, and I am gratefully indebted for his very valuable comments. I am thankful for his guidance and help in all time of research and writing this thesis.

I would like to thank my fellow doctoral student, Dr. Ariane Würbach for her friendship and encouragement.

Finally, I must express my very profound gratitude to my husband Amir for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without him.

Eindhoven, May 2020

Sara Bahrami

Abstract

Respondent burden due to long questionnaires in surveys can negatively affect the response rate as well as the quality of responses. A solution to this problem is to use split questionnaire design (SQD). In an SQD, the items of the long questionnaire are divided into subsets and only a fraction of item-subsets are assigned to random subsamples of individuals. This will lead to several shorter questionnaires which are administered to random subsample of individuals. The completed sub-questionnaires are then combined and the missing values due to design are imputed by means of multiple imputation method. Identification problems can be avoided in advance by ensuring that the combination of variables in the analysis model of interest are jointly observed on at least a subsample of individuals. Furthermore, including an appropriate combination of items in each sub-questionnaire is the most important concern in designing the SQD to reduce the information loss, i.e. highly correlated items that explain each other well should not be jointly missing. For this reason, training data must be available from previous surveys or a pilot study to exploit the association between the variables.

In this thesis two SQDs are proposed. In the first study a potential design for NEPS data is introduced. The data consist of items which can be divided and allocated into blocks according to their context, with the objective that the within block correlations are higher relative to the between block correlations. According to the design, the target sample is divided to subsamples. In addition to the items of a whole block which is assigned to each subsample, a fraction of items of the remaining blocks are randomly drawn and assigned to each subsample. Where items that belong to blocks with relatively higher correlations are drawn with lower probability. The design is evaluated by means of several ex-post investigations. The design is imposed on complete data and several models are estimated for both complete data and data deleted by design. The design is also compared with a random multiple matrix sampling design which assigns random subset of items to each sample individual.

In the second study, a genetic algorithm is used to search among a vast number of SQDs to find the optimal design. The algorithm evaluates the designs by the fraction of missing information (FMI) induced by the design. The optimal design is the one with the smallest FMI. The optimal design is evaluated by means of several simulation studies and is compared with a random MMS design.

Basis for this thesis

Earlier publication

This thesis is in part (see Chapter 4) based on work published in earlier paper by,

- Bahrami, S., Aßmann, C., Meinfelder, F., & Rässler, S. (2014). A split questionnaire survey design for data with block structure correlation matrix. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving survey methods: lessons from recent research.* Routledge. Retrieved from https://www.taylorfrancis.com/books/improving-survey-methods-uwe-engel-ben-jann-peter-lynn-annette-scherpenzeel-patrick-sturgis/e/10.4324/9781315756288.

- as well as supplements of the data manuals accompanying the Scientific Use Files of the Starting Cohorts 3 (Grade 5 students) of the NEPS.

Data used

This Thesis uses data from the National Educational Panel Study (NEPS):

- Starting Cohort First-Year Students,, doi:10.5157/NEPS:SC5:3.0.0.

- Starting Cohort 3 – 5th Grade, doi:10.5157/NEPS:SC3:6.0.1.

From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Statistical software used

All analysis provided are based on The R Project for Statistical Computing (R Core Team, 2014). See Appendix E for further information.

Contents

List of Figures **VIII**

List of Tables **IX**

1 Introduction **1**
1.1 Problems with response burden in surveys 1
1.2 Multiple matrix sampling 2
1.3 Split questionnaire design and multiple imputation 2
1.4 Structure of the thesis 4

2 Statistical Analysis with Missing Data **6**
2.1 Introduction 6
2.2 Mechanisms and assumptions for missing data 7
2.2.1 Concepts of MCAR, MAR and MNAR 7
2.2.2 Assumptions for ignorability 8
2.3 Naive solutions to the missing data problem 9
2.3.1 Complete case analysis 9
2.3.2 Available case analysis 10
2.3.3 Mean imputation 10
2.3.4 Regression imputation 11
2.4 Introduction of randomness 11
2.4.1 Hot deck imputation 11
2.4.2 Stochastic regression imputation 11
2.5 Explicit modeling 12
2.5.1 Model-based univariate imputation 12
2.5.2 Model-based multivariate imputation 12
2.6 Implicit modeling imputation 14
2.7 Likelihood-based approaches 15
2.8 Variance estimation with replication methods: resampling . . 16
2.9 Data augmentation and Gibbs sampling 17
2.10 Multiple imputation 18

2.10.1 Multiple imputation framework 19
2.10.2 Rubin's combining rules 19
2.10.3 Benefits of multiple imputation 22

3 Missing Data Patterns 23
3.1 Introduction 23
3.2 General 'archetypes' 23
3.2.1 Univariate and multivariate patterns 24
3.2.2 Item nonresponse and unit nonresponse 24
3.2.3 Monotone and non-monotone patterns 25
3.2.4 General patterns 26
3.2.5 Connected patterns 26
3.3 Missing by design 26
3.3.1 Statistical matching and identification problem 26
3.3.2 Historic roots of MMS 28

4 A First Potential Application of SQD 36
4.1 Introduction 36
4.2 Evaluation of the split design using simulation 38
4.2.1 Data-generating process 38
4.2.2 Application of the design 41
4.2.3 Multiple imputation to recover missing values induced by split design 42
4.2.4 Simulation results 43
4.3 Empirical applications 47
4.3.1 Empirical application using ALLBUS data 47
4.3.2 Empirical application using NEPS data 48
4.4 Discussion 52

5 Relaxing the Ordered Design in SQD 53
5.1 Introduction 53
5.2 Steps towards an optimal SQD 54
5.2.1 Design-generating algorithm 55
5.2.2 Steps of a genetic algorithm 56
5.2.3 Application of GA to find an optimal split questionnaire design 60
5.2.4 Multiple imputations to recover missing values induced by split design 61
5.3 Simulation study 62
5.3.1 First simulation study 62
5.3.2 Second simulation study 64

5.4 Simulation results 65
5.4.1 First simulation study 66
5.4.2 Second simulation study 68
5.5 Empirical application 71
5.6 Discussion 81

6 Summary-Outlook **82**

Bibliography **84**

A Tables **90**

List of Figures

3.1 Examples of nonresponse patterns 24
3.2 Data fusion 27
3.4 Balanced incomplete block design (BIB) 32
3.5 A (4 choose 2) SQD with six versions of split questionnaires . 35

4.1 Implementation of the design on the simulated data set 42

5.1 Distribution of items on individuals in the MMS design 67
5.2 CIs of the estimates of the optimized linear model 69
5.3 CIs of the estimates of an arbitrary linear model 69
5.4 CIs of the estimates of the optimized linear model 71
5.5 CIs of the estimates of an arbitrary linear model 71
5.6 Distribution of items on individuals in the MMS design 75
5.7 CI widths of the linear model estimates under two designs . . 79
5.8 CI widths of the OL model estimates under two designs 80

List of Tables

4.1 Comparison of split and complete data estimates of empirical distribution parameters and their variances as well as the coverage of their 95% confidence intervals for a data set on 1000 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table. 45

4.2 Comparison of the complete and split (caused by block and MMS) data estimates of regression coefficients and their variances as well as the coverage of their 95% confidence intervals for three different models for a data set on 1,000 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table. . 46

4.3 Comparison of split and complete data estimates of regression coefficients and their variances as well as the coverage of their 95% confidence intervals for two different models for *Allbus* Data on 1,000 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table. 48

4.4 Comparison of split and complete data estimates of regression coefficients and their variances as well as the coverage of their 95% confidence intervals for an ordered logistic model for NEPS main survey Data on 2,500 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table. 50

4.5 Comparison of split and complete data estimates of regression coefficients and their variances as well as the coverage of their 95% confidence intervals for an ordered logistic model for NEPS main survey Data on 2,500 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table. 51

5.1 The optimal design suggested by GA for the data set with 4 Items. . . . 64

5.2 The optimal design suggested by GA for the data set with 8 Items. . . . 66

5.3 Comparison of complete data and data reduced by GA-optimal SQD and a random MMS design based on the estimates of two linear models, their relative bias and coverage for 2,000 samples of size 210. 68

5.4 Comparison of complete data and data reduced by GA-optimal SQD and a random MMS design based on the estimates of two linear models, their bias and coverage for 1,000 samples of size 400. 70
5.5 The optimal design suggested by GA for the NEPS data. 74
5.6 Relative bias and coverage for linear model estimates. 77
5.7 Relative bias and coverage for ordered logit model estimates. . 78

A.1 Best and average fitness values for each GA iteration. 91
A.2 Best and average fitness values for each GA iteration. 92
A.3 Best and average fitness values for each GA iteration. 93
A.4 Variables selected for constructing the empirical data. 94
A.4 Variables selected for constructing the empirical data. 95
A.5 Correlation matrix for for variables v1 to v20 of data used for empirical application of GA. 96
A.6 Correlation matrix for variables v17 to v32 of data used for empirical application of GA. 97
A.7 A Submatrix of Correlation matrix for variables v30 to v48 of data used for empirical application of GA. 98
A.8 Variables assigned to each block in empirical data application of GA. 99
A.9 Models used in fitness function of GA to find the optimal SQD. 100
A.9 Models used in fitness function of GA to find the optimal SQD. 101
A.10 polytomous regression Model used in fitness function of GA to find the optimal SQD. 101
A.11 CI widths for linear model estimates. 102
A.12 CI widths for ordered logit model estimates. 103

Chapter 1

Introduction

1.1 Problems with response burden in surveys

In multipurpose surveys, there is often a need for increasing the number of questions in order to assure the coverage of all issues of interest. However, increasing the length of a questionnaire can increase respondent burden. A study by Sharp & Frankel (1983) measures the respondent burden using some behavioral indicators and the responses to a self-administered reaction form. The study indicates that the questionnaire length is the only significant variable which leads to differences in respondent burden. The respondent burden can possibly have a negative influence on the response rate as well as the quality of the responses which subsequently leads to a potential raise of nonignorable nonresponse.

A study by Herzog & Bachman (1981) shows the negative influence of the lengthy questionnaires on the quality of responses. He shows that the motivation to answer all the survey questions decreases if the survey length exceeds a certain amount of time. The responses to most or all of the items in the later parts of a long questionnaire is more likely to be identical compared to the shorter questionnaires. Low motivated respondents are more likely to look for easier ways to finish the responding process in a shorter time and carry out a less burdensome response strategy. An experimental study by Dillman et al. (1993) of alternatives to the current U.S. decennial census, a national study of 17,000 households, addresses the improvement of responses by shortening the questionnaire.

In a study by Adams & Gale (1982) questionnaire return rates were examined under different methodological conditions for three different form of questionnaires. The study indicates that the length of the questionnaire has a negative influence on the return rates. Another study by Blumberg et al.

(1974) compares five forms of different length and concludes that a questionnaire that is not "overly long" leads to a higher response rate. Roszkowski & Bean (1990) also gives an empirical evidence that surveys with long questionnaires tend to have high nonresponse rates. In longitudinal surveys the length of questionnaire can cause dropouts and unit nonresponse.

This problem of lengthy questionnaires is addressed by constructing questionnaires with designed missingness which allow assigning only a subset of questions to a random subsample of individuals to possibly increase the response rate and quality of responses by decreasing the item and unit nonresponse and to increase the efficiency by decreasing the likelihood of misreporting survey questions due to respondent burden and fatigue. The design is also cost-effective.

1.2 Multiple matrix sampling

An idea is to create shorter versions of questionnaires which contain only a fraction of questions in the long questionnaire and assign these shorter questionnaires to random subsamples of individuals. The idea was revealed by the psychometrician Lord and the corresponding statistical procedures were derived by him for the first time in the 1950's.

The details about the initially known as item-examinee sampling was illustrated by Shoemaker (1973). The idea was to select random subsets of items and assign each subset to a random sample of examinees. Statistics obtained from each subset of items were combined subsequently to provide information about the whole item-examinee matrix. Item-examinee sampling enjoys shorter questionnaire length as a trade-off for smaller sample size which indeed results in loss of information. The split questionnaire design introduced by Raghunathan & Grizzle (1995) compensates for a part of information loss by using multiple imputation method for estimating the analysis model parameters.

The item-examinee sampling was later called multiple matrix sampling (MMS) and was applied to data collection in different fields of study, see Gonzalez & Eltinge (2007) for a review on MMS.

1.3 Split questionnaire design and multiple imputation

A remarkable extension to MMS by Raghunathan & Grizzle (1995) is not only to decrease the interview time for the Cancer Risk Behavior Survey in

order to reduce the respondent burden and increase the data quality but also to multiply impute the missing information, in order to minimize the loss of information. They called their design split questionnaire design (SQD) and suggested dividing the long questionnaire to components of items and assign only a fraction of the split components to a random subsample of individuals. The assignment of split components to random subsamples allows the assumption of MAR or even MCAR for the nonresponses due to design. To avoid the identification problem there should always be a subsample which receives all the combination of items that the analysis model of interest contains.

In a study by Adigüzel & Wedel (2008) a modified fedorov algorithm is used to find an optimal SQD among all possible designs. The algorithm seeks for a design with a minimum Kullback-Leibler distance (KLD) relative to a complete training data set.

Not many empirical applications of SQD's exist in practice, therefore we were interested in gathering some experience in applying SQD's on different types of empirical data. Furthermore, apart from the ordered design and the MMS design there are under circumstances more flexible and efficient solutions which are not tested so far. The first empirical data we were confronted with had a block-wise correlation matrix, i.e. every couple of items in the long questionnaire can be considered as a block and the correlation of items (the correlations are assumed to be available e.g. from a training data set) within the blocks are high, whereas the correlation of items in different blocks are not as high. Our first approach was to construct a SQD that captures the mentioned data structure while reducing the number of items according to the design.

The data collected by split questionnaires are combined and the missing values for the questions which were not asked due to the design were imputed by multiple imputation. Statistical analysis models are estimated on the imputed data sets and pooled according to Rubin's combining rules to provide multiple imputation estimates. The design can be considered an "ordered design" because everything about this design e.g. the number of subsamples, the subsample size, the number of items assigned to each subsample is predefined by the designer.

In the second approach for designing a MMS design genetic algorithm (GA) is used to find an optimal design. GA is a heuristic method that belongs to the category of evolutionary algorithms. Its main application is in finding high quality solutions to optimization problems. GA is inspired by natural evolution and is based on operators like selection, elitism, crossover and mutation. GA starts with a population of randomly selected SQDs and evolves this population toward a better one in an iterative process in the

next generation by an optimization criterion, which in our approach is the fraction of missing information (FMI) of some model parameters specified by the designer. After a sufficient number of iterations an optimal design with a minimum FMI for the model parameters is found. The ordered design of the previous approach is somewhat relaxed in this approach. Setting parameters like the number of subsamples or the number of items that should be allocated to each split questionnaire is not necessary but possible. For example setting the number of subsamples or allocating variables in blocks in case there are a large number of items can speed up the algorithm excessively. Overall, the second design gains in flexibility in comparison to the first design, where no constraints on either the size of SQD's or subsample of individuals who receive a set of questions are required.

According to the fact that the majority of a population (especially in developed countries) have internet access, the option of online data collection is growing very fast. Online data collection facilitates the usage of SQDs with regard to the number of questionnaire versions that are assigned to random subsamples of individuals.

1.4 Structure of the thesis

Collecting data by means of SQDs yields incomplete data. The nature of missing values and the techniques developed to handle missing values are therefore of great interest. Chapter 2 gives an insight about the methods used so far for statistical analysis of data with missing values. These methods range from simply removing individuals with incomplete answers and neglecting the related information loss to sophisticated methods developed in recent decades which take the most advantage of the data available on observed values to reduce information loss as much as possible. In Section 2.2 the mechanisms of the creation of missing values are described and accordingly possible solutions to missing values are discussed in Section 2.3 to 2.7. Section 2.10 introduces multiple imputation developed by Rubin (1987a) as the dominant solution for handling missing values induced by design.

Chapter 3 gives an overview of missing data patterns. A brief review of possible known missing data patterns as a result of incomplete answers of survey individuals to specific questions or the individuals who temporarily or permanently drop out of the survey and suitable methods for handling individual missing data patterns are illustrated in Section 3.2. Several planned missing data patterns are discussed in Section 3.3. Most of Section 3.3 is devoted to the MMS design patterns induced by the designer of the questionnaire to reduce the questionnaire length. Historic roots of MMS designs together with

some applications of these designs in recent years in different field of studies are described throughout Subsection 3.3.2. Statistical matching is another type of planned missing data pattern which is the result of combining data from different sources. This design together with the identification problem which is inherent in such designs are discussed in Subsection 3.2.

Chapter 4 illustrates the first approach for designing an ordered SQD for a data set with a block-wise correlation matrix structure. The data set was simulated based on data from a pilot study in starting cohort 5 (students sample) which was conducted prior to the main survey. The design was evaluated by means of several simulation studies Section 4.2. Furthermore, the design was evaluated by means of two empirical studies in Section 4.3.

Chapter 5 illustrates the second approach for designing a SQD which does not follow a particular structure. This approach uses genetic algorithm (GA) to seek for an optimal design. The details of GA and its application for the purpose of this study are demonstrated in Section 5.2. To evaluate this approach two simulation studies are established in Section 5.3 followed by simulation results in Section 5.4. An empirical application is represented in Section 5.5.

Chapter 6 briefly summarizes the suggested approaches throughout this thesis and represents the conclusion.

Chapter 2

Statistical Analysis with Missing Data

2.1 Introduction

Collecting data is most of the time associated with missing values on some subjects that all the information could not be measured for them. For a long time removing the subjects with incomplete measures has been a widespread approach. However, this can lead to invalid inferences, especially if the proportion of missing values is not small (see Little & Rubin, 2002). In the recent decades several methods for handling missing data have been developed to efficiently use the information from the subjects which are partially observed instead of discarding them. Depending on the complexity of the data structure, corresponding the number and type of incomplete variables in the data as well as the proportion of missing values, some methods are more suitable than others. This chapter gives a brief overview of the most prominent methods developed so far.

One of the major concerns of this thesis is how to cope with the planned missingness forced by SQDs. Multiple imputation, proposed by (Rubin, 1978), is considered as the most general method to deal with missing values in recent years. This method account for the uncertainty due to missingness by generating several versions of completed data and combines the inferences of all the versions to create unique valid overall inferences.

2.2 Mechanisms and assumptions for missing data

2.2.1 Concepts of MCAR, MAR and MNAR

Before reviewing the methods for handling missing data, let's take a look at the mechanisms which generate missing data. To explain that, let's consider an $\mathtt{n} \times \mathtt{p}$ data matrix Y, where $\mathtt{n}$ is the number of individuals in the data and $\mathtt{p}$ is the number of variables in the data. Data matrix Y consists of observed values Y_{obs} and missing values Y_{mis}, $Y = (Y_{obs}, Y_{mis})$. Each data point in the matrix, y_{ij}, can belong to either of these two groups. Furthermore, an $\mathtt{n} \times \mathtt{p}$ response indicator matrix R is defined. The elements of the indicator matrix are denoted by r_{ij}. If y_{ij} is observed, $r_{ij} = 1$ and if y_{ij} is missing, $r_{ij} = 0$.

According to Rubin each data point y_{ij} has a probability of being missing. If the probability of missingness does not depend on the observed or missing data, the missing data mechanism is called missing completely at random (MCAR). The missing data model can be denoted by $Pr(R = 0 \mid Y_{obs}, Y_{mis}, \psi)$ according to (Little & Rubin, 2002; Rubin, 1987a), where ψ contains the parameters of the missing data model. In the case of MCAR, The probability of being missing depends only on the parameter ψ,

$$Pr(R = 0 \mid Y_{obs}, Y_{mis}, \psi) = Pr(R = 0 \mid \psi)$$

or in other words each data point in the data set has the same probability of being missing. For example consider a math achievement test given to 8th grade students of a school. Suppose that in September all the students have taken the exam. In January 50% of the students who have taken the test in September are sampled randomly to take another math test. If the missingness in the test in January is completely independent of any variable including the math scores in September, the missingness mechanism is called MCAR (However, this is a very rare occasion in the real world).

Another more realistic mechanism that creates missing data is missing at random (MAR). In MAR, the probability of being missing depends on the missingness parameters as well as the observed data. In this case, the probability of being missing can be written as,

$$Pr(R = 0 \mid Y_{obs}, Y_{mis}, \psi) = Pr(R = 0 \mid Y_{obs}, \psi)$$

or in other words for a group of data under the same conditions, the probability of being missing is the same for all the data points in that group. In the above example suppose that the objective is to learn if the students who have performed not so well in September have made any improvements. Therefore for the test in January only a random sample of students whose

scores in September are less than average is taken. If the missingness in the second sample does not depend on any other variable, it is said to be MAR.

Another mechanism that creates missingness, is missing not at random (MNAR). MNAR happens when the probability of missingness, beside the missingness parameters and the observed data, also depends on the missing data. In this case the missing data model can not be reduced,

$$Pr(R = 0 \mid Y_{obs}, Y_{mis}, \psi).$$

For math achievement test example, imagine that all the students who have taken the test in September are asked to take the test in January. But this time those students who expect to do badly, are less likely to take part in the test.

2.2.2 Assumptions for ignorability

In making inferences about the complete data parameters, θ, it is of advantage if we can ignore the missingness parameter and missingness mechanism. According to (Little & Rubin, 2002; Rubin, 1987a) under the two assumptions of MAR and distinctness the missing data mechanism is said to be ignorable. The MAR assumption indicates that the missing values in the data should be missing at random. Distinctness assumption is about distinctness of complete data parameters θ and the missing data parameters ψ. The Distinctness of θ and ψ means that the joint prior distribution of the two parameters can be factored into the marginal prior distributions of each of the parameters, [1]

$$f(\theta, \psi) = f(\theta)f(\psi).$$

As stated by (Schafer, 1997) the distinctness assumption holds intuitively, because knowing θ does not give much information about ψ.

The point of interest is making inferences about the complete data parameter θ. For likelihood-based inferences the likelihood of the parameters given Y_{obs} and R is given by,

$$P(R, Y_{obs} \mid \theta, \psi) = \int P(R \mid Y, \psi)P(Y \mid \theta)dY_{mis}$$

[1] Frequentists would say that the joint parameter space is the Cartesian cross product of the parameter space for θ and ψ.

under MAR assumption,

$$\begin{aligned} P(R, Y_{obs} \mid \theta, \psi) =& P(R \mid Y_{obs}, \psi) \int P(Y \mid \theta) dY_{mis} \\ =& P(R \mid Y_{obs}, \psi) P(Y_{obs} \mid \theta) \end{aligned}$$

which is factored into a function of θ and a function of ψ. By adding the distinctness assumption, the missingness mechanism can be ignored and the likelihood of observed data depends only on θ.

For Bayesian inferences, the posterior probability of complete data parameters is of interest. Under MAR and distinctness assumptions by (Schafer, 1997) this is given by,

$$P(\theta \mid Y_{obs}) \propto L(\theta \mid Y_{obs}) \pi(\theta)$$

where $L(\theta \mid Y_{obs}) = f(Y_{obs} \mid \theta)$ is the likelihood and $\pi(\theta)$ is the prior probability of θ.

2.3 Naive solutions to the missing data problem

In order to perform the statistical analysis on data containing missing values, the missing values of the data are removed, filled with other reasonable values or based on the underlying model for observed values, parameters of the whole incomplete data are estimated. There are some simple solution as well as some advanced methods to cope with the missing data problem. The simple methods are only reasonable for special cases with small amounts of missing values. Most of the simple solutions work only in the case of MCAR. Some of these simple solutions are listed below.

2.3.1 Complete case analysis

Complete case analysis also known as listwise deletion is the most comfortable way of dealing with missing data. In this approach all the cases with at least one missing value for the variables of the data are removed. Since the analyses are performed on the same sample, this method allows complete data standard analysis. Moreover, the univariate statistics can be compared between different variables. An important disadvantage of this method is the loss of information due to removed incomplete cases which leads to loss of precision, and additionally to bias when the missing data mechanism is not MCAR (Little & Rubin, 2002). However, there are several exceptions where complete case analysis performs slightly better than other imputation methods (see van Buuren, 2012).

2.3.2 Available case analysis

Available case analysis attempts to reduce the information loss of complete case analysis. Complete case analysis is a wasteful approach, especially for multivariate analysis, since all the cases, even if they are observed on the variable of interest are removed if they are missing on any other variable. This leads especially for datasets with many variables to a massive loss of information. Available case analysis, in contrast, uses all the complete cases that are available for the variable of interest for univariate analysis. The disadvantage of available case analysis is that the analyses involving more than one variable work with different samples. Therefore, if the missing data mechanism is not MCAR, performing such analyses becomes problematic (see Enders, 2010; Schafer & Graham, 2002).

Under MCAR assumption, univariate analysis like estimation of mean and variances can be performed using the available case analysis approach. A special case of available case analysis, called pairwise available case analysis, can be used for estimating the covariances or correlations. In this method the available cases on the two variables of interest are used for analyses. According to (Kim & Curry, 1997) only if data are MCAR and the correlations are modest, using available case analysis method has advantages over the complete case analysis method.

Complete case analysis and available case analysis are both methods that eliminate the cases with missing values. Another way of handling missing data is to fill in the missing values using the information from the observed values of that variable or the other variables in the data, that are predictive of the variable of interest. Imputations are, e.g. means or draws from predictive distributions of missing values given the observed data. Predictive distributions are constructed by explicit or implicit models. Explicit models generate predictive distributions based on an explicit statistical model, such as a multivariate normal model, whereas implicit models are algorithms with an underlying model which should be adopted for each missing data problem to prevent unbiased estimates. In the following, several imputation methods based on explicit and implicit modeling are discussed.

2.3.3 Mean imputation

Mean imputation is a single imputation method. In this method the missing values of each variable are replaced with the mean of observed values of that variable. Mean imputation leads to unbiased estimates for mean, only if data are MCAR. It leads to biased estimates for any other parameter and underestimates the variance and consequently disturbs the relation between

the variables. This imputation method is not recommended unless for a fast fix in the case of small amounts of missing values.

2.3.4 Regression imputation

A regression model is fitted to the observed data to use the information from other observed variables in order to obtain predictions for the missing values. However, the imputed values taken from the regression line do not account for the uncertainty about the actual unknown values. Regression imputation can lead to unbiased point estimates for mean and regression coefficient under MCAR and MAR if the predictor variables are complete and if all the sources of missingness are included in the model.

2.4 Introduction of randomness

2.4.1 Hot deck imputation

Hot deck is an imputation technique where each missing value of a variable for a nonrespondent is replaced by the observed value of a similar respondent with respect to some common characteristics. In another version of hot deck, called random hot deck the missing value of a nonrespondent is replaced by a random draw from a pool of observed values of some similar respondents. Hot deck preserves the univariate distribution of data but not the bivariate or multivariate distribution of data and underestimates the standard errors. For more details about hot deck refer to (Andridge & Little, 2010).

2.4.2 Stochastic regression imputation

To account for the uncertainty about the real unknown values to the imputed values of regression imputation, a certain amount of noise is added to the regression line. Consider a normal linear regression model for a univariate $Y_i, i = 1, \dots, n$ containing n_1 observed and $n_0 = n - n_1$ missing values and a set of $q, q < n_1$ predictors X_i which are observed on all n individuals. In stochastic regression imputation method, n_0 number of imputations Y_{i*} are drawn from the predictive distribution $Y_{i*} = X_i\hat{\beta} + z_i\hat{\sigma}$, where z_i are n_0 independent draws from a standard normal distribution. The stochastic regression imputation can lead to unbiased estimates for mean, regression coefficients and the correlation coefficients under MAR. The method however does not account for the uncertainty about the imputation model parameters.

2.5 Explicit modeling

2.5.1 Model-based univariate imputation

To incorporate the uncertainty about the regression parameters in the regression imputation model, one can draw the parameter estimates from the posterior distribution of the parameters given the observed data. (Rubin, 1987a, Example 5.1) explains this imputation method based on a normal linear regression model for a univariate $Y_i, i = 1, \dots, n$ containing n_1 observed and $n_0 = n - n_1$ missing values and a set of $q, q < n_1$ predictors X_i which are observed on all n individuals. First, a linear regression model $Y_i \sim \mathcal{N}(X_i\beta, \sigma^2)$ is fitted to data and an improper prior is assumed for the model parameters, $\theta = (\beta, \log\sigma)$. The imputation procedure is described as follows,

- using a non-informative prior for (β, σ^2), draw σ_* from its posterior distribution given the observed data, $\sigma_*^2 = \widehat{\sigma_1}^2(n_1 - q)/g$, where g is a random variable from the $\chi^2_{n_1-q}$ distribution and $\widehat{\sigma_1}^2 = \sum_{obs}(Y_i - X_i\widehat{\beta_1})^2/(n_1 - q)$ is the estimated residual variance of the observed data.
- draw β_* from its posterior distribution given σ_* and the observed data $\beta_* = \widehat{\beta_1} + \sigma_*[V]^{1/2}Z$, where $\widehat{\beta_1} = V\Big[\sum_{obs} X_i^t Y_i\Big]$ is the estimated regression coefficient from the observed data. The q-variate Z is a set of q independent standard normal variables and $[V]^{1/2}$ is the Cholesky decomposition of $V = \Big[\sum_{obs} X_i^t X_i\Big]^{-1}$.
- draw n_0 number of imputations Y_{i*} from the predictive distribution $Y_{i*} = X_i\beta_* + z_i\sigma_*$, where z_i are n_0 independent draws from a standard normal distribution.

2.5.2 Model-based multivariate imputation

The conditional specification method explained in Subsection 2.5.1 for imputing the univariate Y_i can be extended for more realistic cases of multivariate Y_i, where more than one variable in the data contain missing values. For a monotone missing data pattern as explained by (Rubin, 1987a, p.171-172), see also Subsection 3.2.3, the variables are imputed one by one. Suppose variables Y_i follow a monotone pattern, that means for $i = 1, \dots, p$, Y_1 is the variable with the least number of missing values and Y_p is the variable with

the most number of missing values and variables X are the predictor variables which are observed on all the individuals. The imputation procedure is as follows: Y_1 is imputed first using X ignoring Y_i with missing values by the imputation method explained in e.g. subsection 2.5.1 . Y_2 is the second variable which is imputed using X and the imputed variable Y_1. In each step the Y_i is imputed using X and the so far imputed Y_is. This procedure is continued until all Y_is are imputed.

Joint Modeling

For a general missing data pattern, see Subsection 3.2.4, assuming ignorability, imputations can be drawn from an explicit multivariate model (joint model) fitted to the data. In (Schafer, 1997) various statistical models are developed for different types of incomplete multivariate data. For continuous data a multivariate normal model $Y \sim \mathcal{N}(\mu, \Sigma), \quad \theta = (\mu, \Sigma)$ is considered for the data. In a Bayesian framework With a non-informative prior for θ as $\pi(\theta) \propto |\Sigma|^{-}p + 1/2$, the posterior distributions for μ and Σ are defined as $\mu \mid \Sigma, Y \sim \mathcal{N}(\bar{y}, n^{-1}\Sigma)$ and $\Sigma|Y \sim \mathcal{W}^{-1}(n-1, nS^{-1})$,where $\mathcal{W}^{-1}()$ is the inverted-Wishart distribution. For more details, see (Little & Rubin, 2002, Exp. 6.18., Exp. 6.21) and (Schafer, 1997, Sec 5.2.2.).

Imputation model parameters ϕ are functions of θ and can be extracted from θ by sweep operator (see Schafer, 1997, Subsection 5.2.4). For incomplete categorical data the saturated multinomial and loglinear models and for mixed data a general location model are used. These models however, lead to complexity and are not suitable for large numbers of variables in the data.

In most cases the θ-parameters are unknown. For general patterns of missing data the observed data posterior $P(\theta|Y_{obs})$ is not easy to handle. A well-known solution as explained by (Schafer, 1997, Section 5.4) is to augment Y_{obs} by imputed values of Y_{mis} and draw parameters from the resulting complete data posterior $P(\theta|Y_{obs}, Y_{mis})$. This can be done in an iterative process. Assuming an initial θ_0, in the `i`th step the imputations are drawn from $P(Y_{mis}|Y_{obs}, \theta^{i-1})$. Subsequently, parameters are drawn from $P(\theta|Y_{obs}, Y_{mis})$. After enough number of iterations this will lead to the stationary distributions, $P(\theta|Y_{obs})$ and $P(Y_{mis}|Y_{obs})$. The rows with similar missing pattern are grouped together and the described procedure is performed separately for each group. Data augmentation is explained in more detail in Section 2.9.

Fully conditional specification

Finding an explicit joint model for data is often a cumbersome task. Another imputation method for multivariate data with general missing pattern is the variable by variable imputation. This imputation technique is a generalization of the conditional specification. For each incomplete variable imputations are drawn from the conditional distribution of that variable given some predictor variables. This method was first introduced by (Kennickell, 1991) for continuous data under the name stochastic relaxation. Generalizations of this method are developed by (e.g. Brand, 1999) under the name variable by variable imputation (Raghunathan et al., 2001) under the name sequential regression imputation and is implemented in `IVEware` (Raghunathan et al., 2002) and (van Buuren & Oudshoorn, 2000) under the name chained equations and is implemented in the `R` package `mice` (van Buuren & Groothuis-Oudshoorn, 2011). Fully conditional specification(FCS) is the term used by (van Buuren et al., 2006).

For a set of $Y = (Y_1, Y_2, ..., Y_p)$ variables with a p-variate distribution $P(Y|\theta)$, FCS draws imputations for each incomplete variable separately by fitting a separate model for each incomplete variable Y_j and draws imputations from the conditional distribution $P(Y_j|Y_{-j}, R, \phi_j)$ of Y_j given the rest observed or imputed variables Y_{-j}, the parameters of the univariate distribution ϕ_j and the response indicator matrix R. Gibbs sampling (see section 2.9) is used to iteratively draw model parameters from the complete data posterior $\phi_j^{i+1} \sim P(\phi_j|Y_j^{obs}, Y_{-j}^{i}, R)$ and imputations from the respective posterior predictive distributions $Y_j^{i+1} \sim P(Y_j^{mis}|Y_j^{obs}, Y_{-j}^{i}, R, \phi_j)$. In each iteration `i` all the incomplete variables are imputed once. After a sufficient number of iterations ($5 - 10$ iterations according to (van Buuren, 2012; Raghunathan, 2015) draws from the conditional distributions $P(Y_j|Y_{-j}, R, \phi_j)$ are equivalent to draws from the multivariate distribution $P(Y, R|\theta)$.

2.6 Implicit modeling imputation

In addition to the imputation methods based on some explicit statistical models, there are several implicit modeling imputation methods such as hot deck imputation, see Subsection 2.4.1, substitution and cold deck imputation. In implicit modeling methods the missing values are typically drawn from the observed values of the present sample or another available sample with similar characteristics for the units with missing values. A widely used method is predictive mean matching (PMM). PMM is a nearest-neighbor approach that automatically yields plausible values in the imputation process. Assume

there is a variable X with all values observed and a variable Y which contains missing values. Fitting variable Y on variable X results in predicted values for Y. for each individual with missing values in Y, choose k candidate donors with observed Y values whose predicted values are close to each other. choose one of the candidate donors randomly and impute the missing value with observed value of the selected donor. For $k > 1$ PMM turns into a sophisticated hot deck method (Rubin, 1986; Little, 1988; Little & Rubin, 2002).

2.7 Likelihood-based approaches

Maximum Likelihood (ML) approaches for dealing with missing data problems (referred to as full information maximum likelihood and direct maximum likelihood in the literature) although limited in scope with only limited practical applications were already used in the 1950's (see Anderson, 1957). The real breakthroughs came in the 1970's (Beale & Little, 1975; Dempster et al., 1977).

In ML approach the missing values in the data are not imputed. For the observed values in the incomplete data a model is assumed. Based on this model the likelihoods and posterior distributions are calculated. The parameters for the whole data including the missing values are estimated using different approaches like ML, expectation maximization (EM), sweep operator, Newton-Raphson algorithm and so on. Using likelihood-based methods has advantages over traditional approaches, which remove the variables with missing values. The standard errors of the estimates can also be estimated in many cases. likelihood-based methods take advantage of all the available information from the model underlying the whole data. For MAR data, where traditional approaches fail to work, likelihood-based methods lead to accurate inferences. Even for the case of MCAR these methods lead to better results than the traditional approaches. However likelihood-based methods are not the perfect solution to the missing data problem and can yield biased estimates for the data parameters under MNAR mechanism.

An example by (Enders, 2010, chapter 4) is explained in the following to illustrate the FIML method. Suppose data under investigation come from a multivariate normal distribution $Y \sim \mathcal{N}(\mu, \Sigma), \theta = (\mu, \Sigma)$ with μ and Σ, the mean vector and covariance matrix respectively. The log-likelihood for multivariate normal data for case i is defined as,

$$logL_i = -\frac{k}{2} log(2\pi) - \frac{1}{2} log|\Sigma| - \frac{1}{2}(Y_i - \mu)^T \Sigma^{-1}(Y_i - \mu).$$

If data contain missing values, the function changes to the following,

$$logL_i = -\frac{k_i}{2}log(2\pi) - \frac{1}{2}log|\Sigma_i| - \frac{1}{2}(Y_i - \mu_i)^T{\Sigma_i}^{-1}(Y_i - \mu_i),$$

where k_i is the number of observed values for case i. For instance, $k_i = 2$ if there are three variables in the data where only two of them are observed on case i. μ_i indicates the mean vector and Σ_i a 2×2 covariance matrix of the observed variables for case i. In general, the cases in the data can be categorized according to their completeness on different variables. For the cases in each category the log-likelihood is calculated. The sample log-likelihood is the summation of all the individual log-likelihoods. An iterative optimization algorithm can be used to calculate the sample log-likelihood many times with different estimates of the population parameters. Eventually, a combination of parameter estimates which lead to the highest sample log-likelihood and consequently the best fit to the data is found. Obviously, the maximum likelihood method does not impute the missing values but uses the information on observed values to estimate the parameters of interest.

2.8 Variance estimation with replication methods: resampling

An important limitation of the single imputation methods described so far is that standard variance formulas, applied to the imputed data systematically underestimate the variance of estimates even if the model used to generate the imputations is correct. There are various methods that can, under circumstances, overcome this problem, for more details (see Little & Rubin, 2002). One of these methods, which is of most interest is the resampling technique. Two examples of this technique are bootstrap by (Efron, 1994) and jackknife by (Rao, 1996; Rao & Shao, 1992). In both techniques the sampling variance is computed by drawing samples from the original incomplete data. In bootstrap resampling, simple random samples are drawn repeatedly from the original incomplete sample, missing values are imputed using an appropriate single imputation method and a consistent estimate of the parameter of interest is calculated for each imputed sample. The bootstrap sample estimates are averaged to create the bootstrap estimate. For large samples and large numbers of bootstrap replications the estimated bootstrap variance of the bootstrap estimate is a consistent estimate of the variance of the sample estimate. Bootstrapping is however a computationally intensive technique because the imputation method is applied to each bootstrap

sample. Furthermore, it is important to use an imputation method, which leads to consistent bootstrap estimates. This ensures that the bootstrap confidence intervals have an appropriate confidence coverage and the statistical tests have the nominal significance level.

In the simple jackknife technique, the only difference is in the type of resampling. Samples are obtained by each time dropping an observation from the incomplete sample. Despite simple implementation, resampling techniques for incorporating the imputation uncertainty are computationally intensive.

2.9 Data augmentation and Gibbs sampling

For small sample sizes a Bayesian alternative approach to ML estimation for monotone patterns is to consider a prior distribution for the parameters of interest and then draw parameters from the posterior distribution. For general missing data patterns, drawing the parameters directly from the posterior distribution without iteration is not possible. Gibbs sampling (see Geman & Geman, 1984; Gelfand et al., 1990) is a Morkov Chain Monte Carlo algorithm which enables having a set of observations from a specified joint probability distribution of a set of variables when direct sampling is complicated. Other applications of Gibbs sampler are to approximate the marginal distribution of a variable, or a subset of unknown parameters or latent variables. The algorithm draws in an iterative process from the conditional distribution of each variable given all other variables. After enough number of iterations the stationary state of the markov chain is reached and Gibbs sampler converges to a draw from the joint distribution. For a set of q random variables $X_1, X_2, \dots, X_q$ and initial values, $x_1^0, x_2^0, \dots, x_q^0$, Gibbs sampling works as follows,

$$
\begin{aligned}
x_1^{i+1} &\sim P(x_1 | x_2^i, x_3^i, \dots, x_q^i) \\
x_2^{i+1} &\sim P(x_2 | x_1^{i+1}, x_3^i, \dots, x_q^i) \\
x_3^{i+1} &\sim P(x_3 | x_1^{i+1}, x_2^{i+1}, \dots, x_q^i) \\
&\vdots \\
x_q^{i+1} &\sim P(x_q | x_1^{i+1}, x_2^{i+1}, x_4^i, \dots, x_{q-1}^{i+1})
\end{aligned}
$$

if there are only two random variables with $X_1 = Y_{mis}$ and $X_2 = \theta$ and if the distributions of these two random variables are conditioned on the observed data Y_{obs}, Gibbs sampling is equivalent to data augmentation.

Data augmentation (Tanner & Wong, 1987) is a special case of Gibbs sampler which can be used to simulate the posterior distribution of random variables in an iterative process. The two steps of this process are the imputation step (I step) and the posterior step (P step). In each iteration `i`, with an initial draw, θ^0, the I step draws missing values from the posterior predictive distribution of missing values given the observed values and parameters drawn in the previous iteration. The P step draws the parameters from the posterior distribution of the parameters given the observed and imputed data from the previous iteration,

$$\begin{aligned} I &\quad Step: Y_{mis}^{i+1} \sim P(Y_{mis}|Y_{obs}, \theta^i) \\ P &\quad Step: \theta^{i+1} \sim P(\theta|Y_{obs}, Y_{mis}^{i+1}) \end{aligned}$$

if $i \to \infty$, the above data augmentation yields a draw from the joint distribution of Y_{mis} and θ given Y_{obs} namely, $P(Y_{mis}, \theta|Y_{obs})$. Drawing from the above conditional distributions is easier than drawing from the joint distribution or the conditional distributions, $P(Y_{mis}|Y_{obs})$ and $P(\theta|Y_{obs})$.

Data augmentation is the stochastic equivalent of the EM approach for ML estimation, where I step replaces the E step and the P step replaces the M step.

In order to ensure that the simulations lead to draws from an appropriate distribution, the number of Gibbs sampler or Data augmentation iterations must be adequately large. A general approach (Little & Rubin, 2002) to evaluate the convergence of these iterations to the target distribution is to simulate $D > 1$ sequences with different starting values spread all over the parameter space. The variations within and between these D sequences are monitored after each iteration until they are approximately equal. In other words, when the distribution of one sequence is close to the distribution of all D sequences together, this distribution represents the target distribution.

2.10 Multiple imputation

Replacing the missing values in an incomplete data set by means of the single imputation methods described so far has the disadvantage of not considering the uncertainty about the data generating process. Performing complete data analyses on imputed data sets is equivalent to treating them as real data, which is absolutely wrong. Even with an unbiased underlying imputation model the standard errors of the estimates will be underestimated because it is still based on a sample. Multiple imputation (MI) developed by Rubin (see Rubin, 1978, 1987a, 1996) creates different versions of imputed data

sets and the variation between these different data sets can be used as an estimate of the variation caused by imputation uncertainty and therefore the underestimation of the standard errors can be prevented using this method.

2.10.1 Multiple imputation framework

MI technique consists of three phases, imputation phase, analysis phase and pooling phase. The purpose of imputation phase is to fill in the missing values in the data set with the methods explained in Section 2.5, Section 2.6 and Section 2.9 in order to create multiple copies of complete data sets (e.g., m=5), each with different values for the missing data. In the analysis phase of MI, analyses of interest are performed on each imputed data set and for each imputed data set parameter estimates and standard errors are calculated. In the pooling phase, the model estimates and standard errors of all (m=5) imputed data set are combined to deliver a single MI estimate and standard error for each analysis model parameter. The pooling phase is performed using the Rubin's combining rules, explained in the next Subsection. In fact multiple version of imputed data sets accounts for the uncertainty due to missing values and reflects the information loss due to missing values correctly.

2.10.2 Rubin's combining rules

The estimates of the quantity of interest obtained from complete data analyses, performed on each imputed data set are combined to create a single inference using Rubin's combing rules (see Rubin, 1987a).

Suppose m versions of imputed data sets are created. For each imputed version j, $j = 1, \ldots, m$, $\widehat{Q}^j$ denotes the estimate of the quantity of interest Q and $\widehat{var}(\widehat{Q}^j)$ denotes the corresponding estimated variance.

According to Rubin's combing rules MI estimator is calculated by taking the average of $\widehat{Q}^j$ over all versions of imputed data sets as follows,

$$\widehat{Q}_{MI} = \frac{1}{m} \sum_{j=1}^{m} \widehat{Q}^j.$$

The total estimated variance of $\widehat{Q}_{MI}$ is defined as follows,

$$T = W + \left(1 + \frac{1}{m}\right) B,$$

W is the variance within the imputed data sets and is defined as follows,

$$W = \frac{1}{m}\sum_{j=1}^{m} W^j = \frac{1}{m}\sum_{j=1}^{m} \widehat{var}(\widehat{Q}^j)$$

where W^j is the estimated variance-covariance matrix of $\widehat{Q}^j$ and indicates the contribution of sampling variance within each imputed data set without considering the existence of missing values.

On the other hand B incorporates the variation caused by the missing values and indicates the variation of $\widehat{Q}^j$ among the different versions of imputed data with respect to the MI estimate $\widehat{Q}_{MI}$ and is defined as follows,

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\widehat{Q}^j - \widehat{Q}_{MI})'(\widehat{Q}^j - \widehat{Q}_{MI}).$$

the term B/m in the total variance is an adjustment for finite m.

For a scalar Q, the fraction of information about Q missing due to nonresponse (see Rubin, 1987a) is defined by,

$$\gamma = \frac{\nu+1}{\nu+3}\lambda + \frac{2}{\nu+3} \tag{2.1}$$

where ν is the degrees of freedom of the t-distribution considered for interval estimates and significance tests of Q,

$$(Q - \widehat{Q}_{MI})\,T^{-1/2} \sim t_\nu$$

and

$$\lambda = \frac{B + B/m}{T}$$

is the proportion of variance assignable to missing data. It can take value zero in a rather unlikely situation that missing data do not cause any extra variation to the sampling variance and can take value one if all the variation is caused by the missing data.

For large degrees of freedom, the fraction of missing information γ is equal to the variance ratio λ but the two values can be remarkably different for small values of ν. The degrees of freedom of the t-distribution, ν, assumed for the quantity of interest Q is defined as follows,

$$\nu^* = \frac{\nu\nu_{obs}}{\nu + \nu_{obs}}$$

ν^* introduced by (Barnard & Rubin, 1999) for small samples is an adapted version of ν,

$$\nu = \frac{m-1}{\lambda^2}$$

and

$$\nu_{obs} = \frac{\nu_{com}+1}{\nu_{com}+3}\nu_{com}(1-\lambda)$$

is the estimated observed degrees of freedom that account for the missing information, where $\nu_{com} = n-k$ is the degrees of freedom of the hypothetical complete data.

Another useful variance ratio is r which is the relative increase in variance due to nonresponse and is defined by,

$$r = (1+m^{-1})\frac{B}{W}.$$

The relation between r and γ is defined by,

$$\gamma = \frac{r + {}^{2}/_{(\nu+3)}}{1+r}.$$

The fraction of missing information shows the increase in variance of the estimates caused by missing data relative to the variance of the estimates in complete data. Therefore this quantity is considered as a measure of efficiency of SQDs and is used frequently in the analysis throughout this thesis.

If Q is a vector of dimension k, the ratios λ and r are averaged over all elements of Q as follows,

$$\lambda = (1+m^{-1}tr(BT^{-1})/k$$

$$r = (1+m^{-1}tr(BW^{-1})/k$$

The magnitude of the fraction of missing information can define the approximate number of imputations needed to obtain stable MI-estimates with the corresponding total estimated MI-variance which is close enough to the total variance that would be obtained if $m \to \infty$. (Raghunathan, 2015) refers to the equation

$$\nu = \frac{m-1}{\lambda^2}$$

and suggests that for the same value of ν a parameter with larger λ, and accordingly γ, needs larger number of imputations. In general he suggests to set m to the largest fraction of missing information times 100.

(van Buuren, 2012) suggests setting m to the average proportion of missing data. He also suggests setting $m = 5$ for the initial model building and increase it in the final round of imputations.

2.10.3 Benefits of multiple imputation

Among the methods explained so far for handling missing values in data only the ML approaches, resampling techniques and MI can lead to unbiased variance estimates. Although ML is a prominent approach for handling missing data, since it is an extremely model-driven approach it is not suitable to apply it on empirical data. The resampling technique is also computationally intensive. MI is the most practical technique for handling missing values in empirical data. Moreover, MI performs well when the analyses objectives are not clear in advance. Because of the advantages of MI over other methods, it is not only used for the previous publications on SQD's but it was also the choice to handle the planned missing values in this thesis.

Chapter 3

Missing Data Patterns

3.1 Introduction

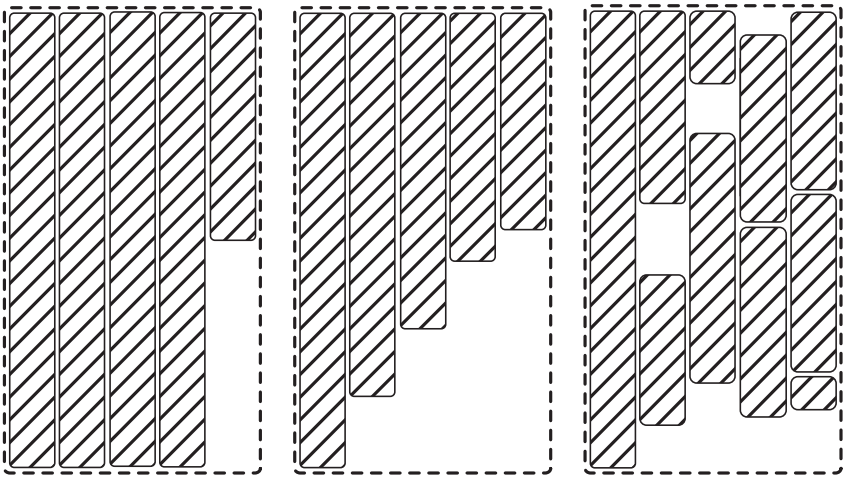
This chapter provides a brief review of missing data patterns. Missing data patterns must be distinguished from the missing data mechanism, which defines the probability of missingness in the data. A missing data pattern illustrates the observed and missing values in a data set and describes the location of missing values in a data set. These patterns can be classified in two categories, patterns containing planned missingness which is induced by the designer prior to data collection and patterns with missing values that are the result of, e.g. a sample individual's participation in a survey or a participant's decision to answer a particular question. My thesis is about the application of split questionnaire designs which belong to the category of planned missing data patterns. In the following several missing data patterns and appropriate methods to handle them are discussed. This will give the reader a better understanding of the behavior of SQD's in comparison to other missing data patterns.

In the next Section general archetypes of missing data are illustrated. In Section 3.3, first we discuss about statistical matching, a procedure of combining data from different sources leading to a missing by design pattern. After wards a comprehensive description of the origins and the first applications of multiple matrix sampling design (MMS) or the split questionnaire survey design (SQD) in their different forms is illustrated in detail.

3.2 General 'archetypes'

Figure 3.1 illustrates several missing data patterns which may occur in the data. The columns are variables and the rows are sample individuals. Shaded

areas represent the observed values and the rest represent the missing values on the corresponding variables. Illustrating the missing data pattern is important for figuring out the appropriate methods for handling missing data and the subsequent analysis of the data.



(a) Univariate (b) Monotone (c) General

Figure 3.1: Examples of nonresponse patterns

3.2.1 Univariate and multivariate patterns

The univariate missing data pattern, Figure 3.1a, is a pattern with missing values in only one of its variables. This pattern is one of the easiest patterns to deal with and is the basis for the more general multivariate patterns, where missingness is not limited to one variable and can occur anywhere in the data. The rest of the subfigures in Figure 3.1 are examples of multivariate patterns. Little & Rubin (2002, Chapter 2) discuss several methods for handling univariate missing data patterns. These methods are generalized for the more complicated case of multivariate patterns throughout the same book.

3.2.2 Item nonresponse and unit nonresponse

Item nonresponse refers to the case, where the sampled individuals take part in the survey but do not answer all the questions in the questionnaire. As a result some of the values of particular variables are missing for some individuals. The missing pattern for item nonresponse is haphazard as in Figure 3.1c and does not follow any particular pattern. Item nonresponse is primarily

addressed by imputation methods (see Little & Rubin, 2002, Chapter 4). In contrast, unit nonresponse occurs when a subset of individuals in the data set is missing on a subset of questions. For example, consider a longitudinal study, where a sample of individuals are surveyed every year. A subset of sample individuals may participate in the survey for a couple of waves and then for some reason like noncontact or refusal they can not be surveyed anymore. For these nonrespondents some variables on survey design are available and the rest of the variables are missing. There are several methods for dealing with unit nonresponse in surveys. One method is to assign weights to all the sample individuals including nonrespondents according to the information available from survey design variables. These weights can be used for further analysis to reduce nonresponse bias. Another method is the substitution method, where unit nonresponse in a survey is replaced by a possibly similar non-sampled unit. In this case caution should be taken for further analysis since the substituted units are in fact imputed. Another method for handling unit nonresponse are pattern-set mixture models. For further information (see Little & Rubin, 2002, Chapter 3).

3.2.3 Monotone and non-monotone patterns

In a monotone pattern (see Rubin, 1987a, p.171-172) as shown in Figure 3.1b, the variables can be ordered according to the number of their observed values. Variables with higher rates of observed values are set first. Furthermore, the cases which are missing for the first time on a variable are also missing for all the next variables. This can happen in longitudinal studies, where a subset of sample individuals may drop out of the study after several waves and never return. Information on this subset is not available from the time when they drop out of the study. Monotone patterns reduce the complexity of dealing with missing data problem with likelihood-based methods and multiple imputation substantially. For data with monotone pattern maximum likelihood estimates and Bayesian inferences can be conducted without iteration. In Schafer (1997, Section 6.5) maximum likelihood estimation and Bayesian inferences for multivariate normal data with monotone pattern are discussed in detail. In some cases the data are not monotone but by filling in a small proportion of missing values they become monotone. For this so called near monotone pattern *monotone data augmentation* is used to fill the missing values in order to have a monotone pattern (see Schafer, 1997).

3.2.4 General patterns

In contrast to a monotone pattern, the missing values in a general pattern as shown in Figure 3.1c, are distributed in the whole data set without any regular configuration. The general pattern can still be divided into several unique missing data patterns. Likelihood-based methods and multiple imputation are developed to deal with general pattern of missing data. For estimating maximum likelihood and for Bayesian inferences, iterative algorithms are needed.

3.2.5 Connected patterns

In connected patterns as described by van Buuren (2012), all the observed cells in the data can be reached from any other observed cell by horizontal or vertical moves through the cells. A connected pattern is needed for estimating the unknown parameters which involve several variables.

3.3 Missing by design

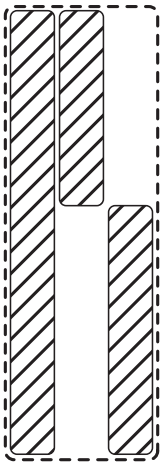
3.3.1 Statistical matching and identification problem

Statistical matching also called data fusion, see Figure 3.2, uses data from two different sources and matches them to use the information from the variables which are not available in both data sets, (see Rubin, 1986; Rässler, 2002). Typically, the units of the two data sets are matched together according to the common variables by means of a distance function. The distance function finds the so-called nearest neighbors and matches them together. The units of the data from one source are the donors who lend their observed information to the recipients (their nearest neighbors in the other data set). This creates an artificial data set for recipients which contains the variables of both data sources.

According to Rubin (1974) the conditional association of the variables which are not jointly observed given common variables are not estimable. Data fusion makes an implicit conditional independence assumption (CIA). CIA was first mentioned by Sims (1972) in his comments on Okner (1972), where he pointed out the potential risks of statistical matching because of the implicit strong CIA.

Only if the common variables which are observed in both data sets can determine the variables of both data sets, the variables which are not jointly observed are guaranteed to be conditionally independent given the common variables observed in both data sets.

Figure 3.2: Data fusion



The unconditional associations for not jointly observed variables given the conditional independence can be derived by a model for joint distribution of all the variables with a common covariance matrix. If the common variables are highly correlated with the variables of either data sets the CIA is more or less satisfied and the estimates of unobserved associations are valid. If the conditional independence assumption is violated and if the considered underlying joint model for the data is not the correct one, estimators for the association parameters can not be identified and there can be more than one feasible estimate for the unknown association parameters. This is called the identification problem. The identification problem may be tackled by means of some informative prior distributions.

Alternatively, statistical matching can be considered as a nonresponse phenomenon. The missing data mechanism is considered at least MAR and ignorable because, although the missingness is not created at random, it is induced by the design. Rässler (2002) has proposed a model based multiple imputation technique for the special case of statistical matching which considers inestimability of association parameters for unobserved associations to fill in the missing values in the matched data set. Explicit models are considered for the joint distribution of data without the assumption of conditional independence to create estimates for conditional associations by getting advantage of the prior information. In the next step the bounds of unconditional associations are calculated for the imputed data set. Based on these bounds a measure of explanatory power of common variables, and thus the validity of statistical matching, is given.

The `mice` algorithm by van Buuren & Groothuis-Oudshoorn (2011), for example, offers multiple imputation techniques for statistical matching which

according to Rässler (2002) works well under multivariate normal distribution and CIA. By violation of CIA, prior information by means of an external data source can be incorporated. Other algorithms which can be used for statistical matching are `IVEware` (see Raghunathan et al., 2002) and `BaBooN` (see Meinfelder & Schnapp, 2015).

Identification problem in split designs would cause some parameters to be inestimable which can be explained with the following example: Assume the parameter of interest to be estimated includes variables X, Y and Z. Furthermore, assume that variable X,Y and Z are not jointly observed in any split but the variables X and Z are jointly observed in one split and variables Y and Z are jointly observed in another split. Considering a multivariate normal distribution for the three variables the covariance of not jointly observed X and Y variables is calculated by, $\mathsf{cov}(\mathsf{X},\mathsf{Y}) = \mathsf{cov}(\mathsf{X},\mathsf{Y}|\mathsf{Z}) + \mathsf{cov}(\mathsf{X},\mathsf{Z})'\mathsf{var}(\mathsf{Z})^{-1}\mathsf{cov}(\mathsf{Y},\mathsf{Z})$, where $\mathsf{cov}(\mathsf{X},\mathsf{Y}|\mathsf{Z})$,the covariance matrix of X and Y given the common variable Z, is inestimable according to Rubin (1974). Under the conditional independence assumption of X and Y given Z the equation reduces to $\mathsf{cov}(\mathsf{X},\mathsf{Y}) = \mathsf{cov}(\mathsf{X},\mathsf{Z})'\mathsf{var}(\mathsf{Z})^{-1}\mathsf{cov}(\mathsf{Y},\mathsf{Z})$ with all terms known from the incomplete data collected by the split design. This rather strong assumption must hold for the (multiple) imputation of the missing parts and also for any kind of analysis which is performed on the completed data afterward. Without the CIA the inestimable parameter $\mathsf{cov}(\mathsf{X},\mathsf{Y}|\mathsf{Z})$ leads to a non-positive definite matrix for $\mathsf{cov}(\mathsf{X},\mathsf{Y},\mathsf{Z})$ which can be overcome using prior information as suggested by Rässler (2002).

However, in constructing the split designs studied in this thesis the identification problem is avoided from the beginning.[1] For instance, assuming that up to trivariate associations are of interest the split design is constructed in a way that there is always a subsample of individuals who receive all trivariate combinations of variables.

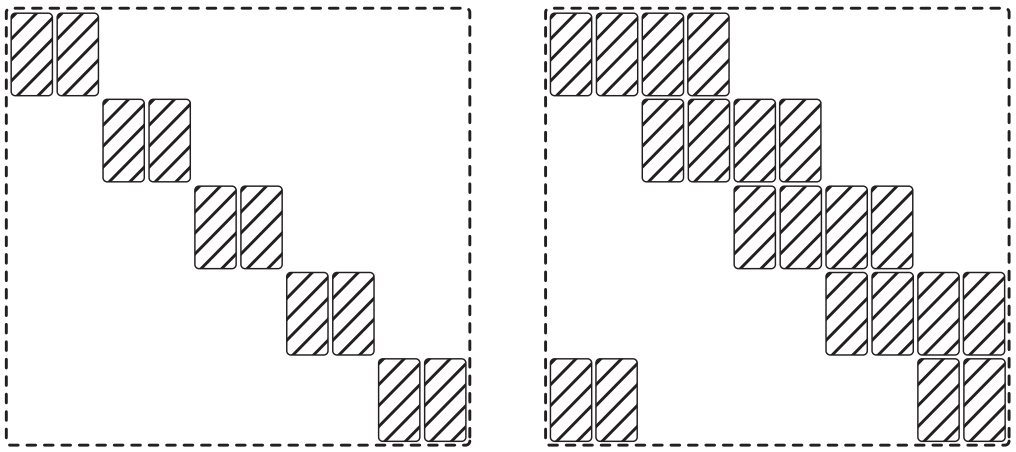
3.3.2 Historic roots of MMS

According to Shoemaker (1973), the statistical procedures were derived for MMS for the first time by psychometrician Lord in the 1950's. An application of MMS to psychometric problems is performed by Lord and Novick in 1968.

In the current subsection a summary of the text book (Shoemaker, 1973) is given. Shoemaker describes, studies, and unifies all the materials about MMS developed until that time in this textbook. He explains MMS known back then as item-examinee sampling, as replicating the process of selecting `t` subsets of items of size `K/t` by random sampling or stratified random

[1] Unless an analysis objective occurs which was not accounted for by the design.

sampling from the population of K items and administering each subset to only a subgroup of examinees selected randomly from the population of N examinees, where the subset of items which are assigned to each subsample can overlap with other subset of items as in Figure 3.3a or not as is illustrated in Figure 3.3b. Statistics obtained from MMS is used to estimate the parameters of the hypothetical N by K data matrix which would be generated if the entire N examinees were administered the complete set of K test items.



(a) Non-overlapping MMS (b) An example of overlapping MMS

MMS was proved either from a theoretical or an empirical point of view to be a good assessment tool for between group differences, where the differences between the individuals were of less importance. Lord and Novick in their application of MMS in 1968 figured out that the standard error of the group mean test score estimated from the data collected by MMS is less than that of an examinee sampling (ES). For the latter, all K items are administered to a random sample from the population of examinees.

To clarify this point, Shoemaker has performed a simulation study to compare ES with MMS. For simulating MMS, he divides K items into t non-overlapping subsets and administers each subset to a random subsample of the population. The mean score test estimate is calculated by pooling the t estimates obtained from all t subsets. This procedure is repeated 1000 times and the standard deviation of all 1000 pooled estimates is considered as an approximation for the standard error of the population mean.

The result of this simulation study demonstrates a smaller sampling variance for MMS comparing to ES, where the estimated sample mean is slightly more accurate for ES than MMS. Apart from this, he mentions other advantages of using MMS like shorter testing time per examinee and costs of

scoring each examinee. However, he does not suggest using MMS when the examinee's answer to a question is not independent from the context of the test.

Furthermore, he addresses some practical concerns in implementing MMS. For a sampling plan denoted by (t/k/n), t is the number of subtests, k is the number of items per subtest and n is the number of examinees who receive a subtest. For a sampling plan whose tk is equal to K, each subtest is a random sample of test items without replacement. In contrast, if tk is a multiple of K, each subtest is a random sample of test items with replacement, where in each subtest an item can be included only once. tk can only be set to an integer multiple of K to avoid any remarkable increase in standard error of the estimate. Shoemaker also suggests a minimum number of N = tnk observations which are needed for a plan to work well.

For choosing the right sampling plan several ex-post investigations have been performed by Shoemaker. These studies try to find a sampling plan with the smallest standard errors for the parameters of interest by varying the number of observations, sampling plan parameters (t, n, k), variance of item difficulty indices, the coefficient of reliability and the skewness of the test score distributions. The results of these studies suggest that a higher number of observations leads to smaller standard errors. Increasing the number of examinees per subgroup has not a considerable effect. Increasing the number of items within each subtest has a positive effect for test scores with a normal distribution, where, in contrast, for test score distributions with negative skewness increasing the number of subtests has the most positive effect. Furthermore, for high reliability and difficulty indices the number of needed observations is larger.

An application of the MMS in Shoemaker (1973) is within the framework of a spelling program for kindergarten students. 50 words are chosen for examining the spelling ability of the children. The children were familiar with these words through a reading program. These 50 words are divided into five subtests with simple random sampling, where each subtest contains ten words. The students are divided into five groups and each subtest is administered to a random group of students. Several statistics such as the population mean test score, the variance of test scores and the coefficient of the reliability are estimated for the population by combining or pooling the estimates obtained from each subtest i.

From subset i, The population parameters, the mean test score (μ_i), the variance of test scores (σ_i) and the coefficient of reliability (α_{21_i}) are estimated as follows,

$$\hat{\mu}_i = \frac{K\,\bar{T}_i}{k_i}$$
$$\hat{\sigma}_i^2 = \frac{n_i K(K-1){s_i}^2 - (K-k_i)\sum_{i=1}^{k_i} v_i}{k_i(k_i-1)(n_i-1)}$$
$$\hat{\alpha}_{21_i} = \frac{K}{K-1}\Big[1 - \frac{\hat{\mu}_i - \frac{\hat{\mu}_i^2}{K}}{\hat{\sigma}_i^2}\Big].$$

where $\bar{T}_i$ is the mean test score of subset `i`, ${s_i}^2$ is the variance of test scores of subset `i` and $\sum_{i=1}^{k_i} v_i$ is the sum of the k_i item variances in subset `i`.The pooled estimates are defined as follows,

$$o_i = n_i k_i$$
$$\hat{\mu}_{pooled} = \frac{\sum_{i=1}^{t} o_i \hat{\mu}_i}{\sum_{i=1}^{t} o_i}$$
$$\hat{\sigma}^2_{pooled} = \frac{\sum_{i=1}^{t} o_i {\hat{\sigma}_i}^2}{\sum_{i=1}^{t} o_i}$$
$$\hat{\alpha}_{21} = \frac{K}{K-1}\Big[1 - \frac{\hat{\mu}_{pooled} - \frac{\hat{\mu}^2_{pooled}}{K}}{\hat{\sigma}^2_{pooled}}\Big].$$

Furthermore, the frequency distribution of test scores that would have been achieved if all the students would have answered all the questions is estimated. He uses the negative hypergeometric distribution to estimate the distribution of the test scores, where the test score is the number of correct answers. The negative hypergeometric distribution is a function of mean and variance of test scores which are correspondingly substituted by $\hat{\mu}_{pooled}$ and $\hat{\sigma}^2_{pooled}$ and the total number of items K.

Then he explains a few applications of MMS, for example designing the pre-post paradigm experiments to evaluate the instructional programs. In such cases instead of asking the same questions from the same individuals before and after the instruction, random questions are asked from the individuals and the parameters of interest for pre and post tests are estimated by MMS. This is in line with the final goal of the researcher which is the evaluation of the group behavior rather than individual behaviors.

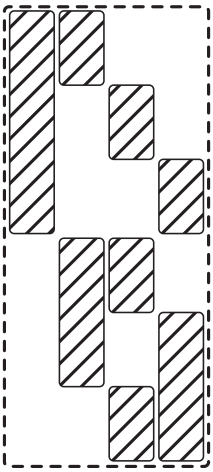
Estimating the covariance and correlation matrices is another application of MMS that he mentions, where every two items should be assigned together

to at least one subsample of individuals. This allows the estimation of every item-pair association for the population of examinees.

Shoemaker also suggests the application of MMS in questionnaires and surveys to possibly increase the response rate by reducing the questionnaire's length. For example an eight-page questionnaire for elementary school teachers of a city could be divided to eight one-page questionnaires and each could be assigned to a subsample of teachers. According to him 'a little data from a large number of teachers is better than a lot of data from few teachers'.

Sirotnik & Wellington (1977) generate estimators for moments and their error variances for any sampling plan from the population. Based on Hooke's approach in 1956, they explore the estimation procedure for different sampling designs, e.g. $\mathtt{n} \times \mathtt{k}$ single matrix sample, non-overlapping $\mathtt{n} \times \mathtt{k}$ MMS, overlapping $\mathtt{n} \times \mathtt{k}$ MMS and BIB (balanced incomplete block) designs. In a BIB design, see Figure 3.4, all pairs of items are assigned simultaneously to subsamples of the same size. The BIB design is described later in this subsection.

Figure 3.4: Balanced incomplete block design (BIB)



The limited knowledge about the sampling distributions of sample's generalized symmetric means (gsm), and the rapid increase in complexity of potential gsm's, as the number of factors in various designs is increased, are two obstacles mentioned for their approach.

Munger & Loyd (1988) investigate the use of MMS in mail surveys in the field of educational studies. In order to compare the response rate for a full questionnaire and the response rate for a shorter MMS questionnaire, they performed an empirical study. 100 public school principals receive a five page instrument with 61 items and 53% of them respond. 207 different schools

receive a two page instrument with 27 items and 63% of them respond. The responses in both cases are compared to figure out the potential presence of bias between the responses of the two groups. They find out that using MMS for a mail survey with lengthy questionnaire with a large number of items is a reasonable treatment. They also find out that among the ten items under test only two about the length of the questionnaire show a significant difference between the respondents of the two groups.

Application of MMS in NAEP

- Beaton & Zwick (1992) overview the application of MMS in NAEP. NAEP is the American national assessment of educational progress which is designed to measure the educational progress of the same students over the years. NAEP's goal is rather reporting the population characteristics than those of specific individuals. NAEP has a large pool of assessment items to administer to the students. Since the participation in NAEP assessment is voluntary, in order to keep the response rate high, a response time of about one hour is considered for each student. Since the amount of time needed for each student to answer all the items in the item pool is much more than one hour, MMS is used to reduce the number of items to assign to each student. NAEP in its early assessments used a non-overlapping MMS by assigning distinct subset of items in a booklet and administering each booklet to a random subsample of students. This method of MMS couldn't estimate the correlation of items in different booklets. From 1984 NAEP started using BIB design. The BIB design was constructed as follows, from a pool of cognitive items, item blocks were constructed so that each item block needed about one third of the available response time. Three item blocks were assigned to each booklet. Furthermore, each booklet contained two sessions of questions about the student's background and attitudes. The BIB design guaranties that every possible pair of item blocks is included in some booklets and therefore allows the estimation of correlation between any item pair in a subject area. Each block is included only once in a booklet. As many booklets as possible were then distributed to each participating school, so that only a few students in each school received the same booklet. On the other hand each item block was administered in many different schools to reduce the sampling variance and thus increase the sampling efficiency. Moreover each booklets was administered to approximately equal proportion of students. From 1988 each booklet includes item blocks of the same subject area.

- Mislevy et al. (1992) focus on the analytical and estimation part of the data collected by BIB designs in NAEP as is explained in detail in Beaton & Zwick (1992). NAEP's purpose based on estimating the nation-wide students' proficiency in different cognitive subject areas allows less need to obtain lots of data to first estimate individual students' proficiency in order to estimate the population characteristics. The population characteristics can be estimated without a need to collect a large amount of data to estimate the individual characteristics accurately.

MMS has found application in many different areas. See the application of MMS in US Bureau of the census by Navarro & Griffin (1993), in medicine by Wacholder et al. (1994), in educational research by Zeger & Thomas (1997) and Thomas & Gan (1997), in psychological research by Graham et al. (2006) and Littvay (2009). Furthermore, Gelman et al. (1998) suggest an MI method for questionnaires that do not ask all the questions from all the sample individuals. Their study was motivated by pre-election public opinion polls.

The process of splitting the long questionnaire involves the decision about the items that should be allocated together in a split, the number of split questionnaires which are created from the long questionnaire and the treatment that the collected data should get to be ready for further analyses.

Application of MMS by Raghunathan & Grizzle (1995)

A remarkable extension to MMS is the split questionnaire survey design (SQSD) developed by Raghunathan & Grizzle (1995) to decrease the interview time for the Cancer Risk Behavior Survey in order to reduce the respondent burden and increase the data quality. In their design, which was called split questionnaire survey design (SQSD), the items of the long questionnaire are divided into several components. Each component contains an approximately similar number of items. One of the components, called core component, contains items that should be asked from all the sample individuals. From the remaining components only a fraction of them are assigned to a random subsample of individuals. The assignment of items to components other than the core component is based on the partial correlation coefficients of the items. It is assumed that the partial correlation coefficient of items are available from a pilot study or previous surveys. Items with higher correlation are located to different components, so that items that can predict each other very well are not missing simultaneously. The random assignment of components to subsamples allows the assumption of MCAR for the nonresponse induced by the design.

Furthermore, the number of components which are administered to each subsample of individuals and the desired amount of reduction in the questionnaire length are defined according to the subsequent analyses of interest. For example in the first simulation study in this paper using an existing complete data from a cancer risk behavior survey, the interaction terms of third-order or higher are not of interest and the interview time should be reduced by 60%, therefore, the sample is divided to six subsamples of equal size and only 2 of 5 components are assigned to each subsample to avoid identification problem, see Subsection 3.3.1. Figure 3.5 illustrates a (4 choose 2) split questionnaire design, where only two of 4 subsample components are assigned to each subsample of individuals, this produces six different versions of split questionnaires which are administered to randomly chosen subsamples. The data obtained from SQD are analyzed once by available case method (the available case method uses the proportion of the sample for which the data on a particular variable or combination of variables in the analysis model are available and is only used for estimating population means and linear models), and once by multiple imputation. The estimates obtained from these two approaches are then compared with the estimates of the complete data obtained from administering the long questionnaire to the sample. The results indicate similar model estimate in all three cases, while the standard errors are larger for SQD relative to long questionnaire but smaller using MI method.

Figure 3.5: A (4 choose 2) SQD with six versions of split questionnaires

Questionnaire number	Core component	Split variables			
		Component 1	Component 2	Component 3	Component 4
1	■	■	■		
2	■	■		■	
3	■	■			■
4	■		■	■	
5	■		■		■
6	■			■	■

Chapter 4

A First Potential Application of SQD Strategies to NEPS Data Using an Ordered Design Variant

4.1 Introduction

In this chapter we develop and evaluate a method for creating a split design by combining the multiple matrix sampling design (MMS) and the split questionnaire design (SQD) introduced by Raghunathan & Grizzle (1995) in which a subset of items is administered to randomly selected respondents. This design is created in such a way that it includes items that are predictive of the excluded items, so that subsequent analysis based on multiple imputation can recover information about the excluded items more efficiently.

Similar to the design introduced by Raghunathan & Grizzle (1995) it is based on the correlation between the variables. Therefore, in order to be able to design a SQD, prior information about the correlation of the variables is needed. This information can be provided by data collected in surveys which are conducted on a regular basis with almost identical questions like panels or the considered tracking surveys; another possibility is to use data collected by pilot studies. Pilot studies are conducted prior to the main data collection on a small subsample of the population of interest to test the instruments and the course of the main data collection.

For illustration, the German National Educational Panel Study (NEPS) conducts pilot studies one year before the main data collections. Pilot data from one of the NEPS stages tracking a cohort of new entrants into higher education are used to illustrate the suggested SQD. The relationship between questions in the instrument is reflected in a structured correlation matrix for

the data. For example, there are five questions in the instrument about the academic self-concept and these five items are highly correlated. Moreover, there are five questions in relation to dropout intentions that are also highly correlated. Items that are highly correlated can be considered as a block. Hence, the correlation matrix structure of the pilot data features high within-block correlation and low between-block correlations for the data. Concerning the widespread use of such kind of instruments in surveys, a SQD for data with blockwise correlation matrix structure is designed.

Pilot data provided by NEPS consists of blocks, with each block containing several items. An interesting feature of the correlation matrix of the pilot data implied by the survey instrument is that the correlation between the items inside the blocks (within-block correlations) is higher than the correlations between the items in different blocks (between-block correlations). Moreover, we consider the blocks with within-correlations of at least 0.35 as highly correlated blocks. The choice of 0.35 is based on experience and personal judgment regarding general associations among survey variables and can be hence adapted if necessary.

On the basis of the correlation matrix structure, a SQD is designed. The idea is to divide the sample into subsamples and assign a fraction of items of each block to a random subsample. To ensure that the items of each block are administered simultaneously to a subsample of individuals, each random subsample receives one of the blocks completely. In addition, each respondent receives a random set of items from the rest of the blocks, where the probability of choosing items from highly correlated blocks is lower than the probability of choosing items from the rest of the blocks. Applying this design to a complete data set, results in a data set with missing values for items which are not administered to individuals. As expected, the design imposes a higher ratio of missing values on items in highly correlated blocks, which is reasonable because the information lost by eliminated items can be recovered from those items with which they are highly correlated by means of subsequent multiple imputation process.

For evaluation purposes, based on the correlation matrix structure of the pilot data, a data set is simulated. The implementation of the design on the simulated data set is described in Section 4.2, including a description of the data-generating process in Subsection 4.2.1, the application of the design on the simulated data set in Subsection 4.2.2 and the multiple imputation technique used to fill the missing values induced by the design, in Subsection 4.2.3. The procedure tries to conserve the correlation matrix structure of the pilot data provided by the higher education stage of the NEPS. Different regression analyses of interest are performed on the simulated data and the parameters are estimated for the data before and after implementing

the design. The point and interval estimates of the regression coefficients as well as their variance estimates and the coverage of their 95% confidence intervals between complete and split data set are compared and tabulated in Subsection 4.2.4. Further empirical illustration of our split design based on NEPS and ALLBUS data is provided in Section 4.3. Section 4.4 concludes.

The content of this chapter is mostly taken from Bahrami, Aßmann, Meinfelder, & Rässler (2014).

4.2 Evaluation of the split design using simulation

In order to assess the properties of estimators based on split design data, we generate a data source which reflects typical features of observed pilot data. This data set should reflect the characteristics of the pilot data with regard to its correlation matrix structure. We orientate the data-generating process towards the pilot data obtained from NEPS students cohort. Hence, a data set with 13 blocks and a total of 60 items from altogether 24 blocks of pilot data is considered. Five out of the 13 blocks are chosen as highly correlated blocks. The data-generating process aims to preserve the correlation matrix structure of these 13 item blocks and is discussed in more detail in the following subsection.

4.2.1 Data-generating process

The simulated data set consists of 70 variables including 40 metric variables, 10 binary variables, 10 ordinal variables and 10 multinomial variables with three categories each. To simulate all the variables with the exception of multinomial variables we use the correlation matrix that we have extracted from 13 blocks containing 60 variables of the pilot data. We start by simulating 60 variables from a multivariate normal distribution,

$$X = (X_1, X_2, \ldots, X_{60}), \quad X \sim \mathcal{N}(\mu, \Sigma),$$

where, for the first 40 elements of mean vector μ, arbitrary (non-zero) values are defined and the remaining 20 elements are considered zero. The covariance matrix Σ of the multivariate normal distribution is specified using the correlation matrix extracted from pilot data. The correlation matrix of the pilot data ρ_p was converted to the covariance matrix according to the following relation,

$$\Sigma = \rho_p \otimes (\sigma\sigma'), \tag{4.1}$$

where σ denotes the vector of standard deviations and is defined as follows. We specify arbitrary values for standard deviations of the first 40 variables and 1 for the remaining 20 variables. In the next step, we consider the first 40 variables with arbitrary mean values and standard deviations as our metric variables,

$$Y^{\text{met}} = (X_1, X_2, \dots, X_{40}),$$

the remaining 20 variables with mean zero and standard deviation of one are used to create binary and ordinal variables. Using binary probit models, ten binary variables are created. To create binary variables, 10 normally distributed latent variables are defined as follows,

$$Y_i^* = Z_i'\beta_i + \epsilon_i, \quad \epsilon_i \in (X_{41}, \dots, X_{50}), \quad i = 41, \dots, 50, \quad Z_i' \subseteq Y^{\text{met}},$$

and the subsequent binary variables are given as follows,

$$Y_i^{\text{bin}} = \begin{cases} 1 & \text{if} \quad Y_i^* \geq 0, \\ 0 & \text{if} \quad Y_i^* < 0. \end{cases}$$

The explanatory variables in the linear predictor of the latent variables are chosen from our metric variables Y^{met} and the error terms are set to $(X_{41}, \dots, X_{50})$. For example, to create the first latent variable Y_{41}^*, ϵ_{41} is set to X_{41}, for Y_{42}^*, ϵ_{42} is set to X_{42}, etc. The regression coefficients in β_i including an intercept are specified arbitrarily.

The same procedure is used to create ten ordinal variables using ordinal probit models as follows,

$$Y_i^* = Z_i'\beta_i + \epsilon_i, \quad \epsilon_i \in (X_{51}, \dots, X_{60}), \quad i = 51, \dots, 60, \quad Z_i' \subseteq Y^{\text{met}},$$

$$Y_i^{\text{ord}} = \begin{cases} 1 & \text{if} \quad -\infty < Y_i^* \leq \mu_{i,1}, \\ 2 & \text{if} \quad \mu_{i,1} < Y_i^* \leq \mu_{i,2}, \\ 3 & \text{if} \quad \mu_{i,2} < Y_i^* < +\infty. \end{cases}$$

In this case, the error terms of the normal latent variables are set to $(X_{51}, ..., X_{60})$. Considering three levels for the ordinal variables, two cut-points $\mu_{i,1}$ and $\mu_{i,2}$ for each variable are specified. The regression coefficients in β_i are specified again arbitrarily. The ordered probit model is identified either by setting $\mu_{i,1} = 0$ or $\beta_{i,0} = 0$, where we chose the latter. For ordered probit model, the

probability of being in category $j = 1, 2, 3$ is,

$$\Pr(Y_{ord,i} = j) = \begin{cases} \Phi(-Z_i'\beta) & \text{for} \quad j = 1 \\ \Phi(\mu_{j-1} - Z_i'\beta) - \Phi(-Z_i'\beta) & \text{for} \quad j = 2 \\ 1 - \Phi(\mu_{j-1} - Z_i'\beta) & \text{for} \quad j = 3, \end{cases}$$

where ϕ is again the cumulative standard normal distribution probability, and μ_{j-1}'s are the cut-points respectively. The linear predictor and the cut-points are chosen in a way that $\Pr(Y_{ord,i} = j)$ is not close to zero or one.

Furthermore, 10 unordered categorical variables are generated using multinomial logit models. Assuming that the multinomial variables we want to generate have three categories labeled as 0,1 and 2. The first value, here 0, is designated as the reference category. The probability of membership in other categories is compared to the probability of membership in the reference category. Hence,

$$\ln \frac{\Pr(Y_i = k)}{\Pr(Y_i = 0)} = \beta_{0k} + \beta_{1k}X_{i1} + \beta_{2k}X_{i2} = z_{ki}$$

with $i = 1, \ldots, n$ and $k = 0, 1, \ldots, K$, where n is the number of observations and K is the number of categories.

$$\Pr(Y_i = k) = \begin{cases} \frac{\exp(z_{ki})}{1+\sum_{h=1}^{K}\exp(z_{hi})}, & k = 1, \ldots, K \\ \frac{1}{1+\sum_{h=1}^{K}\exp(z_{hi})}, & k = 0 \end{cases}$$

with $z_{ki} = \beta_{0k} + \beta_{1k}X_{i1} + \beta_{2k}X_{i2}$. For a multinomial variable with $K = 3$ categories, calculation of $K - 1 = 2$ equations are needed to describe the relationship between the multinomial variable and the independent variables. So we need to define two equations for z_{1i} and z_{2i} in which X_{i1}, X_{i2} are chosen from metric variables and the two sets of coefficients $\beta_{01}, \beta_{11}, \beta_{21}$ and $\beta_{02}, \beta_{12}, \beta_{22}$ are defined arbitrarily, similar to the binary and ordinal cases, carefully to avoid the problem of perfect prediction. Small values of $\exp(z_{ki})$ would cause probabilities close to zero.

In the next step, for each observation an independent random numbers u is drawn from a standard uniform distribution, consequently Y_i takes values 0, 1, and 2 if u belongs to the ranges shown below,

$$Y_i^{\text{cat}} = \begin{cases} 0 & \text{if} \quad 0 < u \leq \Pr(Y_i = 0), \\ 1 & \text{if} \quad \Pr(Y_i = 0) < u \leq \Pr(Y_i = 0) + \Pr(Y_i = 1), \\ 2 & \text{if} \quad \Pr(Y_i = 0) + \Pr(Y_i = 1) < u \leq 1. \end{cases}$$

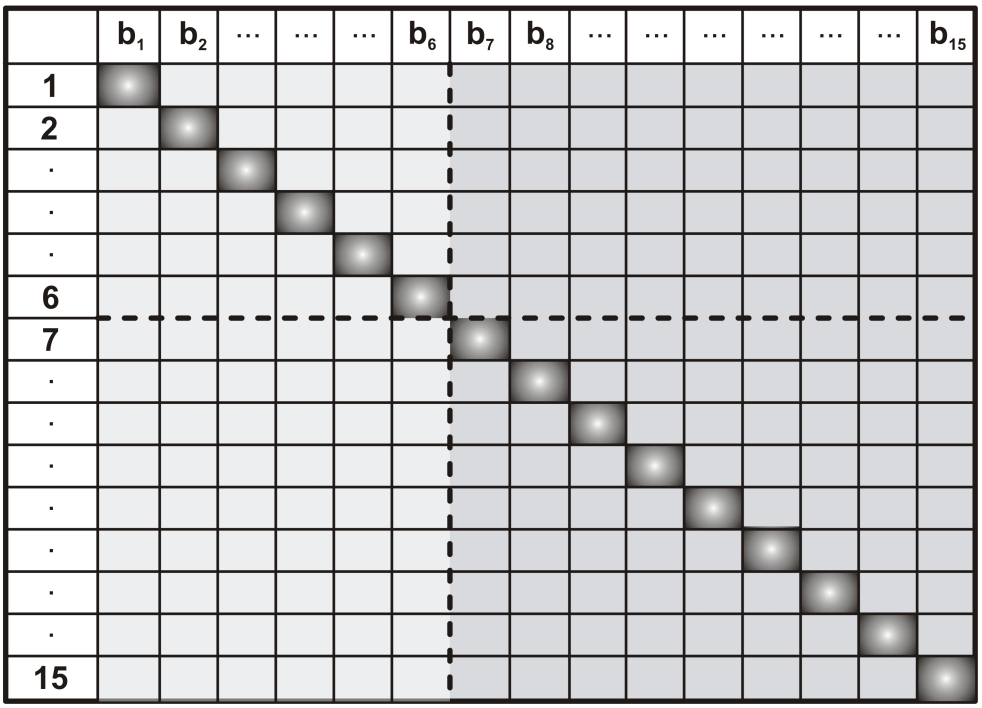
Note that different from the metric, binary, and ordinal variables the nominal variables depend on other variables only via specified conditional mean function. This approach provides a data set in which the correlation matrix structure of the pilot data is preserved. Using different measures of association like Tetrachoric Coefficient between binary variables, Polychoric Coefficient between ordinal variables, Biserial Coefficient between binary and metric variables and Spearman Rank-Order Coefficient between metric and ordinal variables confirms the preservation of association between the variables of the generated data set, although the considered measures of correlation are not directly designed for variables with nominal scales.

4.2.2 Application of the design

In this subsection we explain how we apply the design to the simulated data set. The 70 variables in the simulated data set can be categorized into 15 blocks, each block containing four to five variables. In general, the correlation of variables inside the blocks are higher than the correlation of variables between different blocks. Furthermore, six blocks have relatively higher within-correlations than the other nine blocks. In this study we want to simulate a case in which only half of the items are assigned to each individual, thus each individual should receive only 35 items. For this reason we divide the sample into as many subsamples as the number of blocks, in our case 15 subsamples. Each subsample receives one of the blocks completely. In addition to a complete block (four to five items), each individual randomly receives 30 to 31 items from the rest of the blocks, so that the probability of receiving items from blocks with higher within-correlation is lower than the corresponding probability for blocks with smaller within-correlations. In our design, we set the probability of missigness in highly correlated blocks to 0.55 and the probability of missingness in the rest of the blocks to 0.45 respectively.

Figure 4.1 shows the application of the design on the data schematically. The columns $b_1, \ldots, b_{15}$ are the blocks, where $b_1, \ldots, b_6$ are those with high within correlations, i.e. all bivariate correlations within a block exceed .35. The rows are the subsamples and the shaded diagonal squares represent the complete blocks received by each subsample. Applying the design on the complete data set induces missing values to the data. The next step is to deal with the missing values.

Figure 4.1: Implementation of the design on the simulated data set



4.2.3 Multiple imputation to recover missing values induced by split design

In order to analyze the split data, the missing values induced by split design are treated using multiple imputation methods as discussed in detail in Chapter 2. Multiple imputation (MI) is a general-purpose procedure to analyze incomplete data sets, where the uncertainty due to missingness is represented by different data sets that vary in the imputed part. The `R` package `mice` by van Buuren & Groothuis-Oudshoorn (2011) implements MI via chained equations. The imputation method used for metric and binary variables is predictive mean matching (PMM) as described in 2.6. To impute the ordinal and multinomial variables, the ordered logistic regression and polytomous logistic regression are used. Given the simulation setup, the PMM method for imputing the binary variables yields better MI estimates for the binary probit model compared to the logistic regression method recommended by `mice` for imputing binary variables.[1] The missing values are imputed m=10 times.

[1]This was figured out by means of a small simulation study.

4.2.4 Simulation results

The next step is to evaluate the design by means of the generated data set. For this purpose several Monte Carlo simulations are performed. In each simulation study several analysis models are considered and the models are estimated once for the complete data and once for the data deleted by design. The general process of the simulation studies is as follows, first a sample of size 50,000 is generated using the method described in Subsection 4.2.1. This sample is considered as the population. The regression coefficients of the analysis models are then estimated for this sample and considered as the true parameters.

In the next step 1,000 samples (simple random sampling without replacement), each of size 5,000 are drawn from the hypothetical population. The models are then estimated for these samples before and after design implementation and the following MI. Before deletion model estimates and their estimated variances specified by BD in the table are the average model estimates and their estimated variances over all 1,000 samples before design-implementation.

Multiple imputation estimates and their corresponding variances which are specified by MI in the table are calculated as follows: the design is applied to each sample, missing values caused by design are imputed using MI. In our case, MI produces m=10 imputed data sets. The models are estimated for each imputed data set. The model estimates and their estimated variances of all m=10 imputed data sets are combined using Rubin's combining rules as described in Subsection 2.10.2. MI variances in the table indicate the average total variances of $\widehat{\beta}_{MI}$ over all 1,000 samples drawn from the population.

Coverage or the coverage probability of a confidence interval indicates the proportion of samples for which the 95% confidence interval of the corresponding model estimate contains the true parameter. BD coverage indicates the 95% coverage before design-implementation and MI coverage indicates the 95% coverage after design implementation and the following MI. Furthermore, the fraction of missing information γ due to multiple imputation, is calculated for each estimator.

In the first simulation study several empirical distribution parameters such as arithmetic mean and probability quantiles for the metric variable X_{21}, arithmetic mean for the binary variable Y_3 and proportions for ordinal variable Y_{14} and the multinomial variable Y_{22} are estimated (these variables are selected randomly). The results presented in Table 4.1 reveal that all MI-coverages meet the expected 95% level, except for the second parameter of the ordinal variable and the first parameter of the multinomial variable.

In the second simulation study three analysis models are selected, namely a

binary probit model,[2] where the proportion of missing values for the variables selected for this model are 60% after implementing the design,

$$Y_4 = \beta_0 + \beta_i X_i, \qquad X_i \in (X_{27}, X_{30})$$

an ordered probit model,[3]where the proportion of missing values of dependent variable, Y_{13} is 40% and those of the independent variables are 60% after implementing the design,

$$Y_{13} = \beta_0 + \beta_i X_i, \qquad X_i \in (X_{29}, X_{32})$$

and a multinomial logit model,[4]where the proportion of missing values of the independent variable, X_{15} is 40% and those of Y_{24} and X_{27} are 60% after implementing the design,

$$Y_{24} = \beta_0 + \beta_i X_i, \qquad X_i \in (X_{15}, X_{27})$$

In the MMS design each random sample individual receives a random set of questions (here, 50% of the questions). Table 4.2 represents the comparison between the block design and the MMS design for three mentioned analysis models. Apart from the design used for inducing planned missingness, the results reveal that most coverages meet the expected 95% level, except those for the intercepts of the multinomial logit model. We assume that these findings result from the data generating process used to create the simulated data, where the nominal variables depend on any other variable only via the specified conditional mean function. This provides less association with other variables to be exploited within the MI procedure. Further, the chosen correlation measure is not directly designed for variables with nominal scale. Given this, we consider the high coverage for the other parameters of the multinomial logit model as evidence in favor of the robustness of the suggested SQD. For the third coefficient of the binary probit model and the coefficients of the ordered probit model, the coverages are slightly under the expected 95% level. This may be addressed by increasing the number of imputations.[5]

[2] `glm` function with a probit link in `R` is used for estimating the binary probit model.

[3] `polr` function in `R` package `MASS` (Venables & Ripley, 2002) is used for estimating the ordered logistic model.

[4] `multinom` function in `R` package `nnet` (Venables & Ripley, 2002) is used for estimating the multinomial model.

[5] Performing further simulations were not possible due to deadline pressure.

	metric						binary	ordinal			multinomial		
	mean	q(2.5%)	q(10%)	q(50%)	q(90%)	q(97.5%)	*mean*	p_1	p_2	p_3	p_1	p_2	p_3
Estimate													
BD	.049	-1.929	-1.230	.051	1.324	2.023	.478	.386	.176	.438	.162	.417	.420
MI	.049	-1.975	-1.267	.051	1.369	2.078	.479	.396	.167	.437	.181	.406	.413
Variance													
BD	.0010	.0087	.0030	.0016	.0034	.0091	.00025	.00024	.00014	.00024	.00014	.00024	.00024
MI(T)	.0018	.0148	.0058	.0027	.0067	.0162	.00040	.00033	.00021	.00033	.00021	.00031	.00027
Coverage													
BD	.96	.97	.96	.96	.96	.97	.96	.96	.95	.95	.96	.95	.97
MI	.96	.96	.94	.96	.90	.96	.97	.97	.88	.97	.79	.95	.97
γ	.40	.37	.38	.36	.40	.36	.37	.27	.34	.26	.28	.21	.10

Table 4.1: Comparison of split and complete data estimates of empirical distribution parameters and their variances as well as the coverage of their 95% confidence intervals for a data set on 1000 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table.

	binary probit			ordered probit		multinomial logit					
	β_0	β_1	β_2	β_1	β_2	β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}
Parameter	.112	.198	-.210	.307	-.189	.980	.203	-.154	.998	.197	-.160
Estimate											
BD	.112	.198	-.209	.308	-.189	.982	.204	-.153	1.001	.197	-.158
MI(MMSD)	.091	.191	-.194	.294	-.177	.921	.186	-.135	.945	.178	-.139
MI(BlockD)	.086	.190	-.190	.296	-.180	.918	.190	-.131	.944	.181	-.134
Variance											
BD	.00078	.00021	.00015	.00010	.00010	.00186	.00058	.00087	.00185	.00058	.00087
MI(MMSD)	.00207	.00067	.00049	.00029	.00029	.00253	.00132	.00199	.00230	.00121	.00178
MI(BlockD)	.00214	.00067	.00052	.00028	.00028	.00255	.00118	.00204	.00231	.00109	.00182
Coverage											
BD	.95	.97	.96	.96	.94	.96	.96	.96	.95	.96	.96
MI(MMSD)	.94	.94	.90	.77	.85	.81	.94	.94	.78	.94	.93
MI(BlockD)	.92	.94	.87	.89	.91	.79	.94	.94	.80	.94	.92
γ(MMSD)	.60	.66	.66	.64	.64	.28	.54	.55	.21	.51	.50
γ(BlockD)	.61	.67	.68	.62	.62	.28	.49	.56	.22	.46	.51

Table 4.2: Comparison of the complete and split (caused by block and MMS) data estimates of regression coefficients and their variances as well as the coverage of their 95% confidence intervals for three different models for a data set on 1,000 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table.

Furthermore for some coefficients the MMS design work better than the block design. This can be addressed by comparing the proportion of missing values caused by the block design and the MMS design to the variables used in these models. For example, for the ordered probit model, where the dependent variable has a lower proportion of missing values (40%) in the block design comparing to (50%) in MMS design, the coverages are higher for the block design than the MMS design. Similarly, for the second coefficient of the multinomial logit model the coverages are higher in the block design.

4.3 Empirical applications

The SQD is illustrated once by means of surveyed data taken from the ALLBUS 2009 survey and once by surveyed data taken from the student cohort of NEPS in 2010.

4.3.1 Empirical application using ALLBUS data

The ALLBUS data set resembles the correlation structure found in the NEPS pilot data. For this reason, 63 variables are selected from ALLBUS data and combined to create a smaller data set. The data set contains variables about personal qualities, political participation and attitudes towards the political system, world view and value orientations as well as attitudes to different ethnic groups in Germany, citizenship, national identity, social inequality and religion.

The variables are divided into 12 blocks, each containing five to six variables. The within-block correlations are in general higher than between-block correlations, four blocks have higher within correlations than the other blocks.

In this study we use two empirical models to evaluate the design, a binary probit model and a multinomial logit model. In the multinomial logit model, we specify the opinion on ethnically mixed neighbourhoods with three levels, yes, no and no preferences, as the dependent variable. As explanatory variables, we use opinion on intermarriage and how pleasant it is to have foreign neighbours. Both explanatory variables are ordinal and contain six levels. In the binary probit model, we specify opinion about dual citizenship permission as the dependent variable, and the variables, importance of peace and order and self-classification in left-right political spectrum, as explanatory variables. Both explanatory variables are again ordinal with five and ten levels respectively.

The ALLBUS data set includes approx. 3,400 observations. For simulation purposes, this sample is considered as population and model estimates

	binary probit			multinomial logit					
	β_0	β_1	β_2	β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}
Parameter									
	-.073	-.116	.099	1.670	-.310	-.530	.940	-.120	-.270
Estimate									
BD	-.078	-.116	.101	1.666	-.316	-.532	.932	-.120	-.267
MI	-.117	-.101	.099	.962	-.189	-.367	1.000	-.100	-.247
Variance									
BD	.026	.0006	.0015	.27	.0066	.0085	.1166	.0023	.0031
MI	.063	.0015	.0032	.49	.0133	.0147	.2869	.0071	.0076
Coverage									
BD	.97	.97	.96	.98	.98	.97	.99	.96	.98
MI	.97	.97	.97	.87	.86	.77	.95	.95	.95
γ	.62	.72	.50	.66	.68	.62	.60	.59	.62

Table 4.3: Comparison of split and complete data estimates of regression coefficients and their variances as well as the coverage of their 95% confidence intervals for two different models for *Allbus* Data on 1,000 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table.

for this sample are considered as true parameters. 1,000 samples each of size 1,000 are drawn from the population, the models are estimated for each sample before implementing the design (BD) and after implementing the design[6] and the following multiple imputation. Their associated variances, the 95% coverage and the fraction of missing information (γ) due to MI are provided in Table 4.3 The results suggest that the model estimates are quite unbiased and the coverages are quite high which meet our expectations regarding the application of SQD for long questionnaires. Again, some caveats apply concerning the coefficients of the multinomial logit model, where again for the given choice of the correlation measure we interpret the results in favor of robustness of the suggested approach. Future research could address the effects of correlation measures directly designed for variables with nominal scale.

4.3.2 Empirical application using NEPS data

For the second empirical application of the design, data from Wave 2 of Starting Cohort 5 (SC5) of the National Educational Panel Study (NEPS) is used,

[6]The proportion of missing values forced by the design to all three variables used in the binary probit model are 43%. In the multinomial logit model the dependent variable recieves 43%, the first independent variable 56% and the second dependent variable 42% missing values.

see Blossfeld et al. (2011). The data refers to the Scientific Use File (SUF; DOI:10.5157/NEPS:SC5:3.0.0). SC5 follows the pathway of the first year students entering German universities through the job market. Due to the reduction of the length of the questionnaire in the main study, from 104 variables in the pilot study with a block wise correlation matrix structure only 68 could be retrieved in the main study. However this has not influenced the block wise structure of the correlation matrix of the remaining variables. Similar to the simulation studies performed for simulated data, in this study the effect of including a core component to the SQD is compared to the case of absence of a core component. A core component with the variables which are assigned to all the observations is specified. The core component contains 29 variables including socio-demographic variables such as gender, age, migration background, nationality, religion, marital status, household size and other variables such as preparation for the academic studies, change of field of study, learning environment, employment outlook, personal importance of final degree, importance of status preservation and parents education and employment status.

68 variables in split components and 29 variables of the core component are observed on about 12,300 individuals, where about 1,500 observations contributed with skip patterns and 500 observations with aborted interviews are removed from the data set. The 0.5% missingness in the remaining data set with 97 variables on 10,300 observations due to item nonresponse is imputed with single imputation in `mice`.

For implementing the SQD on the data, 68 variables of the complete data are divided into 20 blocks, each block containing three to four variable. Nine blocks with minimum correlation of 0.40 are considered as highly correlated blocks. The sample individuals are divided to 20 subsamples and the data are deleted according to the SQD explained in subsection 4.2.2.

Considering that the variables in the complete data are ordinal with 4, 5 or 7 levels, an ordered logistic regression model is fitted to the data to evaluate the design.

The variable drop-out intention (*abint1*) with four levels is considered as dependent variable in this model. The independent variables in the model are academical self-concept (*aask1*) with seven levels, fulfillment of achievement claims (*acom8*) with four levels, being recognized by lecturers (*lerw6*) with four levels, helplessness to get better grades (*shilf1*) with five levels, little fun with studies (*sintl3*) with five levels, clear definition of educational objectives (*ssco_st21*), cooperativeness of lecturers (*ssco_su15*), non-supportive behavior of the students (*ssco_su22*) and examination burden (*ssco_c14*) all with five levels.

In the block design, the proportion of missing values induced in the men-

	Ordered logit								
	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
Parameter									
	-.270	.898	-.397	.199	-.089	-.087	-.060	-.110	.059
Estimate									
BD	-.273	.901	-.398	.200	-.089	-.086	-.061	-.113	.059
MI(with core)	-.251	.895	-.350	.200	-.090	-.091	-.050	-.094	.055
MI(wout core)	-.257	.914	-.352	.205	-.095	-.085	-.039	-.096	.056
Variance									
BD	.0020	.0022	.0044	.0024	.0059	.0029	.0041	.0031	.0022
MI(with core)	.0103	.0081	.0226	.0114	.0230	.0110	.0209	.0100	.0074
MI(wout core)	.0106	.0082	.0232	.0117	.0240	.0113	.0219	.0113	.0074
Coverage									
BD	.97	.97	.96	.96	.97	.95	.97	.96	.94
MI(with core)	.96	.94	.96	.97	.96	.96	.97	.96	.96
MI(wout core)	.95	.92	.96	.96	.97	.96	.96	.96	.96
$\boldsymbol{\gamma}$(with core)	.78	.71	.79	.77	.72	0.71	.79	.69	.69
$\boldsymbol{\gamma}$(wout core)	.78	.70	.78	.76	.73	0.71	.79	.70	.68

Table 4.4: Comparison of split and complete data estimates of regression coefficients and their variances as well as the coverage of their 95% confidence intervals for an ordered logistic model for NEPS main survey Data on 2,500 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table.

tioned variables are as follows, *abint1* : 56%, *aask1* : 57%, *acom8* : 58%, *lerw6* : 41%, *shilf1* : 58% , *sintl3* : 42%, *ssco_st21* : 42%, *ssco_su15* : 60%, *ssco_su22* : 41% and *ssco_c14* : 40%. The simulation results are provided in Table 4.5.

For simulation purposes, the sample with 10,300 observations is considered as population and model estimates for this sample are considered as true parameters. 500 samples each of size 2,500 are drawn from the population, the model is estimated for each sample before implementing the design (BD) and after implementing the design and the following multiple imputation (MI), once with the core component and once without the core component. The number of multiple imputations is set m=10. The associated variances, the 95% coverage and the fraction of missing information (γ) due to MI are provided in Table 4.4. The results do not show a significant difference between the the design with and without a core component.

In another simulation study, the block design without the core component is compared with the MMS design. The analysis model used in this simulation study is an ordered logistic regression with the variable drop-out intention (*abint1*) with four levels is considered as dependent variable in this model.

	Ordered logit				
	β_1	β_2	β_3	β_4	β_5
Parameter					
	-.275	.918	-.410	.211	-.164
Estimate					
BD	-.278	.920	-.411	.211	-.161
MI(BlockD)	-.265	.920	-.385	0.214	-0.158
MI(MMSD)	-.258	.932	-.384	0.212	-0.160
Variance					
BD	.0020	.0022	.0044	.0023	.0050
MI(BlockD)	.0098	.0104	.0163	.0111	.0179
MI(MMSD)	.0068	.0068	.0149	.0075	.0173
Coverage					
BD	.96	.95	.97	.95	.98
MI(BlockD)	.95	.95	.94	.95	.97
MI(MMSD)	.94	.93	.94	.96	.97
γ(BlockD)	.79	.79	.73	.79	.71
γ(MMSD)	.70	.68	.71	.69	.70

Table 4.5: Comparison of split and complete data estimates of regression coefficients and their variances as well as the coverage of their 95% confidence intervals for an ordered logistic model for NEPS main survey Data on 2,500 observations. Furthermore, the fraction of missing information (γ) for the individual estimates are represented in this table.

The independent variables in the model are academical self-concept (*aask1*) with seven levels, fulfillment of achievement claims (*acom8*) with four levels, being recognized by lecturers (*lerw6*) with four levels, helplessness to get better grades (*shilf1*) with five levels, little fun with studies (*sintl3*) with five levels.

In MMS design each sample individual receives only 50% of the whole 68 items randomly. In the block design, the proportion of missing values induced in the mentioned variables are as follows, *abint1* : 56%, *aask1* : 57%, *acom8* : 58%, *lerw6* : 41%, *shilf1* : 58% and *sintl3* : 42%. The simulation results are provided in Table 4.5.

The results do not show a significant difference between the two designs. This can again be explained by the higher proportion of missing values (56%) induced by the block design to the dependent variable compared to MMS design.

4.4 Discussion

In this study we have explained how we designed a particular kind of SQD for items with a particular correlation matrix structure available from a pilot study. Due to the fact that the small number of observations in the empirical data would lead to inaccurate MI estimates, generating a data set with a sufficient sample size that could reproduce the correlation matrix structure of the pilot data could help us to evaluate our design.

The results shown earlier in this chapter illustrate how using this design could affect the regression coefficient estimates of the analysis models. The estimates are unbiased and the coverages are high, although further improvements may be gained via consideration of richer dependence structure for the variables with nominal scale and other measure of dependence between the variables. Increasing the number of multiple imputations may also lead to better results. By implementing the design on empirical data, we demonstrated that this design could work very well for data with this particular correlation matrix structure.

Chapter 5

Relaxing the Ordered Design in SQD

5.1 Introduction

In this chapter we have developed a method to find an optimal split questionnaire design among a large amount of possible designs. The method uses the genetic algorithm (GA) for deriving the optimal split design. (See Mitchell, 1996) for an introduction to genetic algorithms. The potential advantage of this approach over the traditional approaches by Raghunathan & Grizzle (1995) Adigüzel & Wedel (2008) [1] is mainly its flexibility. For instance the items which should always be asked simultaneously can be considered as a gene and the content of a gene would never come apart in the algorithm. GA's belong to the category of evolutionary computation which involve natural selection followed by some sort of variation by means of crossover, mutation and inversion in the hope of finding solutions to complicated problems which can best adapt to their environments. GA's were introduced for the first time by John Holland in the 1960s and were further developed later in the 1970s by John Holland and his coworkers and students (see Holland, 1975). GA describes a technique which evolves from one population of chromosomes to a new one by selecting pairs of chromosomes and producing their offspring by operators inspired by the nature namely, crossover, mutation and inversion.

In our approach GA starts with a solution space which is a random population of SQDs. Subsequently GA evaluates the individual solutions by means of a fitness function and sorts them according to their fitness. The solution space evolves to a fitter one by means of GA operators, and after an ade-

[1] The latter approach relies on the Kullback-Leibler distance which compares (multivariate) densities using standard distributional assumptions for data.

quate number of iterations it results in an optimal solution with respect to the fitness function. In this study the fitness function is considered to be the fraction of missing information (FMI) (see Equation 2.1) of one or more model estimates of one's choice, where the degrees of freedom introduced by (Barnard & Rubin, 1999) for small samples is used for calculating FMI. The average FMI due to planned missingness induced by the design is the least for the model estimates in the optimal SQD.

The structure of this chapter is as follows: in Section 5.2 the steps toward an optimal design by means of GA are described in detail. These steps include the selection of an initial population by defining specific characteristics for the SQDs, a detailed description of GA, its parameters and its application in our approach. In Subsection 5.2.4 multiple imputation of missing values induced by the design is discussed. In Section 5.3, first a small simulation study is performed to test the approach and in a second simulation study a larger data set has been used. It is always assumed that a training data set is available prior to the construction of an optimal SQD which reveals the characteristics of items used in the main survey. In each simulation study a data set with a predefined correlation matrix is simulated. Appropriate GA parameters are set and several statistical models are defined for evaluation purposes. The simulation results are illustrated in Section 5.4 which includes the representation of the optimal SQD found by GA and its comparison with a random MMS design. An empirical application of this approach on NEPS data is illustrated in Section 5.5. The chapter ends with a discussion in Section 5.6.

5.2 Steps towards an optimal SQD

The first step towards an optimal SQD is the construction of the initial solution space or the initial population of SQDs. According to ones' preferences, there are several parameters which should be defined beforehand for the construction of an SQD, i.e., the number of sub-questionnaires or reduced questionnaires which are planned to be created, the number of questions which should be assigned to each sub-questionnaire, and the fraction of missingness which should be induced to each SQD as a whole. Apart from these parameteres the items in the questionnaire which can not be separated for some logical reasons must be taken into account.

The first parameter, the number of sub-questionnaires, depends on the data collection method. For a computer assisted data collection no constraint is required, as long as there are adequate number of individuals who receive a sub-questionnaire, whereas a pencil and paper questionnaire can only have

a limited number of sub-questionnaires due to layout costs. In the former case the fraction of sample individuals in the training data who receive each sub-questionnaire according to the eventual optimal design can be used as a measure for assigning them to the target sample. This leads to varying subsample sizes for each sub-questionnaire. If fix number of sub-questionnaires or fix subsample sizes are desired a constraint is set. Computation time, as in our case, is another reason for setting a constraint. The second parameter, the number of questions of each sub-questionnaire, can vary across the sub-questionnaires or can be set as a fix number. A minimum and a maximum value can be set for the number of questions which are assigned to each sub-questionnaire. The third parameter, the fraction of overall missingness can be set as desired.

To avoid the identification problem, the algorithm applies the constraint that all the variables which are used in constructing the analysis model must be observed for a subsample of individuals.

5.2.1 Design-generating algorithm

To find the optimal SQD among a vast number of possible solutions, we use a GA. GAs are highly suitable for dealing with complex, large-scale optimization problems in an efficient way. GA starts from an initial population of so-called "chromosomes". A chromosome specifies a candidate solution in GA. Each chromosome consists of a set of genes which delivers information about the solution which it represents. Genes are either single bits (e.g. zeros or ones) or small blocks of bits that are used to encode elements of chromosomes. In most cases the genes are denoted as 0's and 1's and therefore the chromosomes are represented as binary strings. Permutation encoding, value encoding, and tree encoding are other types of GA encoding (see Whitley, 1994).

GA mimics natural selection by applying a process of selecting the chromosomes which are allowed to reproduce. Two chromosomes are usually selected for reproduction. On average, the "fitter" chromosomes are more likely to reproduce than the less fit chromosomes. "Crossover" describes the process of exchanging parts of the selected parent chromosomes to produce an offspring. Crossover can have different forms, such as single point, two point, uniform, and arithmetic crossover. Crossover probability is the probability that crossover takes place. Mutation is the process of replacing a gene in the offspring chromosome. The mutation probability is set inversely proportional to the chromosome length l. Mutation ensures the genetic diversity in GA. Inversion is rarely used in GA. Elitism is another operator in GA which allows transferring some of the best solutions directly to the next

generation.

GA moves from one population of chromosome to another one. Each population is a search space containing a set of chromosomes or possible solutions to the problem. A fitness function is needed to rate the solutions and give each solution a score, giving a chance to the fitter chromosomes with higher scores to move to the next population or reproduce for the next population. This is seen in the nature as the fittest organisms are given more chance to spread their genetic material to the next generations. However the simple rules of selection followed by variation such as crossover and mutation lead to complex variety in the nature.

Furthermore, in searching for a solution by GA, there is no need to find all possible solutions and search among them. Only a fraction of possible solutions are needed to be evaluated.

5.2.2 Steps of a genetic algorithm

The following pseudo code by Xuan et al. (2011) describes the steps of genetic algorithm. The algorithm tries to find an optimal solution in form of bit strings to a given problem by following steps,

- Generate an initial population of solutions in form of l-bit chromosomes (chromosomes of length l). The initial population can either be generated randomly or a more suitable initial population can be suggested by the user. The size of the population, N, called population size, is set according to the length of the chromosome. For lengthy chromosomes the population size grows to ensure the inclusion of the optimal solution in the population.

- Repeat the following steps until the generation of a new population of the same size,

    1. Select a fraction of the fittest chromosomes (β) to be copied to the new population as "elitism".
    2. Select two chromosomes from the population with a selection probability which is directly proportional to the fitness of the chromosomes. Sampling with replacement increases the probability of choosing fitter chromosome more than once for the reproduction. A "child" solution typically shares many of the characteristics of its parents.
    3. In the case of single point crossover, choose a crossover point randomly and exchange the parts of parent chromosomes in this point.

The part of the first parent solution from the beginning to the crossover point is combined with the part of the second parent solution from the crossover point to the end. If the crossover point turns out to be the first or the last bit of the chromosome, no crossover is performed and instead one of the parent chromosomes is copied correspondingly to next population.

4. Each bit in an l-bit chromosome is mutated with a mutation probability of $\gamma = 1/(1 + l)$.

- Replace the current population with the new population.
- Iterate the process by going to step 2. These processes eventually yield the next population generation of solutions that is different from the previous generation. In general the overall fitness increases by this procedure for the population, fitter individuals have a higher probability of reproduction. Just like real-world evolution the process is probabilistic, which ensures the genetic diversity in the gene pool of the subsequent generation of children.

The iterations are performed until an optimal solution is found. The algorithm stops if no considerable improvement is recognized for the fitness or no further resources or computation time are available.

Algorithm 1 Pseudo code for Genetic Algorithm

Input:

instance Π,
size N of population,
rate β of elitism,
rate γ of mutation,
numbers δ of iterations

Output:
Solution X

`//Initialization`

1: generate N feasible solutions of length I randomly;
2: save them in the population Pop;

`//Loop until the terminal condition`

3: **for** i=1 to δ **do**

`//Elitism based selection`

4: number of elitism $ne = N.\beta$;
5: select the best ne solutions in Pop and save them in Pop_1;

`//Crossover`

6: number of crossover $nc = (N - ne)/2$;

7: **for** j=1 to nc **do**

8: randomly select two solutions X_A and X_B from Pop;
9: generate X_C and X_D by one-point crossover to X_A and X_B;
10: save X_C and X_D to Pop_2;

11: **end for**

Algorithm 1 Pseudo code for Genetic Algorithm (continued)

```
    //Mutation
12:     for j=1 to nc do
13:         select solution X_j from Pop_2;
14:         mutate each bit of X_j under the rate γ and generate a new solution X_j';
15:         if X_j' is unfeasible then
16:             update X_j' with a feasible solution b repairing X_j';
17:         end if
18:         update X_j with X_j' in Pop_2;
19:     end for
    //Updating
20:     update Pop = pop_1 + Pop_2;
21: end for
    //Returning the best solution
22: return the best solution X in Pop;
```

5.2.3 Application of GA to find an optimal split questionnaire design

An optimal split questionnaire design which we try to find by means of GA, assigns a fraction of questions in the questionnaire to each sample individual, with a minimum loss of information in comparison to the full questionnaire. Data from a pilot study or data collected in previous waves of a panel can be used to find the optimal SQD. The configuration of the split design depends on the structure of the long questionnaire. Therefore, prior to the search for an optimal split design, decisions must be taken about e.g., the set of items that are desired to be assigned simultaneously (items of similar context) or the number of sub-questionnaires which are created to assign to each subsample of individuals.

In the following, the steps of GA for finding an optimal SQD are illustrated,

1. In the first step a genetic representation of the candidate solutions is needed. In our case, each chromosome represents a split questionnaire design. Suppose we have an $n \times p$ data matrix Y, containing the data values on p variables for all n observations in the data set. This can be a data set available from a pilot study for which we are supposed to design a split questionnaire. We define an indicator matrix R as an $n \times p$, $0-1$ matrix for representing a SQD. The elements of Y and R are denoted by y_{ij} and r_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. If y_{ij} is missing due to the design, $r_{ij} = 0$ and otherwise $r_{ij} = 1$. So each chromosome in the population represents an SQD in vectorized form. If the number of sub-questionnaires, s, is set in advance, $i = 1, \ldots, s$ in r_{ij}. If the items are allocated in b blocks, $j = 1, \ldots, b$ in r_{ij}.

2. In the next step a population of chromosomes (SQDs) is created either randomly or according to a criterion set by the programmer. The SQDs must fulfill the constraints, which are set on the SQD in advance. This constraints are number of sub-questionnaires, subsample sizes, number of questions of each sub-questionnaire and the fraction of overall missingness. The population size varies with the length of the chromosomes. For lengthy chromosomes, the population size should be large enough to guarantee the inclusion of adequate variety in the candidate solutions.

3. The main task in using GA is to find an appropriate fitness function to evaluate the split designs. The choice of the fitness function is highly dependent on the subsequent analysis which are planned to be performed on the collected data. In the simulation studies and empirical

data applications, illustrated in following sections, several statistical models of interest are built. The fitness function, correspondingly evaluates the FMI of the model parameters for each SQD. A SQD with the minimum FMI is considered to be the optimal design. The FMI, Equation 2.1 as is explained in detail in Subsection 2.10.2, shows the increase in variance of the estimates caused by missing data, relative to the variance of the estimates in complete data and is a measure of efficiency of SQDs. The evaluated chromosomes are ordered according to their fitness.

4. Apart from a proportion of fittest chromosomes which is copied to the next generation, GA operators namely, selection, crossover and mutation, as described in detail in subsection 5.2.2, are applied to the population to generate offspring for the next generation.

5. GA is terminated when no considerable improvement is seen in the FMI of the model parameters.

The genetic algorithm, used in this study, is based on the `rbga.bin` function in `R` package `genalg`. Several functions and constraints are adapted to this function in order to generalize it for our purpose. Parallel computation is performed by the `R` package `parallel`.

5.2.4 Multiple imputations to recover missing values induced by split design

In order to analyze the split data, the missing values induced by split design are treated using Multiple Imputation (MI) methods as discussed in detail in Section 2.10. Multiple Imputation is a general-purpose procedure to analyze incomplete data sets, where the uncertainty due to missingness is represented by different data sets that vary in the imputed part. The `R` package `mice` by van Buuren & Groothuis-Oudshoorn (2011) is used in this chapter to multiply impute the missing values.

Genetic algorithm searches among plenty of possible SQDs to find the optimal SQD by means of a fitness function. The fitness function that is used in this study to evaluate the SQDs is the FMI which is induced by SQD, see Subsection 2.10.2 for details. The fitness function calculates the FMI for the parameters of statistical models of interest eligible for the further analysis of the data. The average FMI of all the analysis model parameters of interest is considered as the evaluation value. The optimal SQD is the one which has the smallest evaluation value.

To be able to compare the SQDs to find the optimal design, each SQD should have a unique evaluation value which is reproducible after each time applying multiple imputation. This is possible for a training dataset which is observed on a relatively large sample which ensures that each version of split questionnaire is assigned to a subsample of adequate size. This in combination with an adequate number of multiple imputations, m, will deliver unique evaluation values for each SQD.

For training data of smaller size, repeating multiple imputation creates more stable values of FMI which in turn leads to high computation time of the Monte Carlo studies.

5.3 Simulation study

To evaluate our proposal of using genetic algorithm to find an optimal SQD among a large number of possible SQDs two simulation studies are performed. The approaches used in these simulation studies and the corresponding results are described in this section.

5.3.1 First simulation study

We start with a small data set of 210 observations on four variables. The sample is drawn from a multivariate normal distribution.[2]

$$X = (X_1, \dots, X_4), \quad X \sim \mathcal{N}(\mu, \Sigma),$$

where all four elements of mean vector μ are set to two and the standard deviation of all four variables are set to one. The covariance matrix Σ is set to the following correlation matrix according to Equation 4.1,

$$\rho = \begin{pmatrix} 1 & .25 & .25 & .25 \\ .25 & 1 & .80 & .80 \\ .25 & .80 & 1 & .60 \\ .25 & .80 & .60 & 1 \end{pmatrix}.$$

The sample is divided into subsamples which means that the number of sub-questionnaires which are created is set to a fix number. For the data set with four variables five subsamples are sufficient. A larger number of

[2]The `mvrnorm` function in `R` package `MASS` (Venables & Ripley, 2002) is used for simulating the data.

subsamples would only produce similar sub-questionnaire while increasing the GA computation time. With four items $2^4 = 16$ combination of items are possible. In other words 16 different sub-questionnaires can be created [3], From the possible item combinations some are not desirable, e.g. in this study the one with no items. The overall missingness is set to 50% MCAR. As described in the preceding sections GA starts with a population of chromosomes, where the chromosomes are SQDs. To decrease the computation time, for creating the GA population the possible sub-questionnaires are labeled with indexes from one to 15. Each chromosome (SQD) is a random draw of size $I = 5^4$ from these labels. In other words each gene represents a possible sub-questionnaire. GA operators, i.e. selection, crossover and mutation are applied to this population. The labels are replaced with the corresponding sub-questionnaire and evaluated by means of the fitness function.

GA parameters used for this study are as follows,

- $N = 100$
- $I = 5$
- $\gamma = \frac{1}{1+I} = .17$
- $\beta \approx N \times .05 = 5$
- $\delta = 40$

The fitness function evaluates the SQDs by means of the average FMI of the following parameters: the arithmetic mean of all four variables and the parameters of the linear model, $\mathbf{E}(X_3) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4$.[5]Thus, $\gamma_{fit} = [\hat{\gamma}_{\mu_1}, \dots, \hat{\gamma}_{\mu_4}, \hat{\gamma}_{\beta_0}, \dots, \hat{\gamma}_{\beta_3}]$ consists of the corresponding 8 FMI's. Convergence is assumed if the average of γ_{fit} ($\bar{\gamma}_{fit}$) could be decreased after a maximum of 40 iterations. The optimal SQD found by GA is illustrated in Table 5.1.[6] Table A.1 in Appendix A shows the best evaluation value in each iteration and the average fitness value over all the population members in each iteration.

[3]Each sub-questionnaire contains a subset of items which are supposed to be assigned to each subsample of individuals.

[4]Because there are five subsamples.

[5]In a slight abuse of notation we denote for regression models throughout this chapter the left-hand side with an unconditional expectational value rather than a conditional one.

[6]In MI-Step of the fitness function we set m=90 and repeat MI 40 times to obtain stable fitness values. The method used for MI is mice's built-in method `'norm'` [7] which uses a Bayesian linear regression model for imputation. Each GA-iteration takes 51 minutes.

Subsample	Item 1	Item 2	Item 3	Item4
1-42		■		
43-84			■	
85-126				■
127-168	■	■	■	
169-210	■	■	■	■

Table 5.1: The optimal design suggested by GA for the data set with 4 Items.

5.3.2 Second simulation study

For the second simulation study, a larger data set is simulated. The data set consists of eight variables and 400 observations. The variables are drawn from a multivariate normal distribution,

$$X = (X_1, X_2, \dots, X_8), \quad X \sim \mathcal{N}(\mu, \Sigma),$$

where all eight elements of mean vector μ are again set to two and the standard deviation of all eight variables are set to one. The covariance matrix Σ is set to the following correlation matrix according to Equation 4.1, [8]

$$\rho = \begin{pmatrix} 1.0 & 0.80 & 0.70 & 0.60 & 0.50 & 0.40 & 0.30 & 0.20 \\ 0.8 & 1.00 & 0.75 & 0.65 & 0.55 & 0.45 & 0.35 & 0.25 \\ 0.7 & 0.75 & 1.00 & 0.70 & 0.60 & 0.50 & 0.40 & 0.30 \\ 0.6 & 0.65 & 0.70 & 1.00 & 0.65 & 0.55 & 0.45 & 0.35 \\ 0.5 & 0.55 & 0.60 & 0.65 & 1.00 & 0.60 & 0.50 & 0.40 \\ 0.4 & 0.45 & 0.50 & 0.55 & 0.60 & 1.00 & 0.55 & 0.45 \\ 0.3 & 0.35 & 0.40 & 0.45 & 0.50 & 0.55 & 1.00 & 0.50 \\ 0.2 & 0.25 & 0.30 & 0.35 & 0.40 & 0.45 & 0.50 & 1.00 \end{pmatrix}$$

In order to set a fix number of sub-questionnaires and also to decrease the computation time, the sample is divided into ten subsamples. For the data set of size 400 with eight variables ten subsamples are created. This ensures the inclusion of more sub-questionnaires for each SQD, since for eight items, $2^8 = 256$ combinations are possible.[9] However some combinations are not

[8] The original correlation matrix is slightly altered by `nearPD` function in `R` package `Matrix` by Bates & Maechler (2016), to make it positive-definite.

[9] Each sub-questionnaire contains a subset of items which are supposed to be assigned to each subsample of individuals.

desirable, e.g. in this study, one with no items and those eight that assign only one item to each subsample. The overall missingness is set to 50% MCAR. Similar to the previous simulation study, to decrease the computation time, for creating the GA population the possible item combinations are labeled with indexes from one to 247. Each chromosome is a random draw of size ten [10]from these labels.

Furthermore, GA parameters, used in this simulation study are set as follows,

- $N = 252$
- $I = 10$
- $\gamma \approx \frac{1}{1+10} = .09$
- $\beta \approx N \times .05 = 12$
- $\delta = 40$

The fitness function evaluates the SQDs by means of the average FMI of the following parameters: the arithmetic mean of all eight variables and the linear model, $\mathbf{E}(X_8) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$. The fitness value ($\bar{\gamma}_{fit}$) is thus the average over $\gamma_{fit} = [\hat{\gamma}_{\mu_1}, \dots, \hat{\gamma}_{\mu_8}, \hat{\gamma}_{\beta_0}, \dots, \hat{\gamma}_{\beta_7}]$. Convergence is assumed if after 40 iterations no significant improvement of fintness value is observed. The optimal SQD found by GA is illustrated in Table 5.2.[11] Furthermore, Table A.2 in Appendix A shows the best fitness value in each iteration and the average fitness value over all the population members.

5.4 Simulation results

To evaluate the optimal SQD found by GA in the simulation studies performed so far, the optimal SQD is compared with a multiple matrix sampling (MMS) design. For this purpose a Monte Carlo simulation is performed. The results are illustrated in the next section. The evaluation steps and the corresponding results are described for both simulation studies in the following subsections.

[10]Because there are ten subsamples.

[11]In the MI-Step of the fitness function we set m=100 and repeat MI 50 times to obtain stable fitness values. The method used for MI is mice's built-in method `'norm'` [12] which uses a Bayesian linear regression model for imputation. Each GA-iteration takes approx. 9 hours.

Subsample	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8
1-40					■		■	■
41-80			■		■			
81-120		■		■				
121-160		■	■	■	■	■	■	■
161-200	■						■	■
201-240	■					■		■
241-280	■			■				■
281-320	■	■						
321-360	■	■	■		■	■	■	■
361-400	■	■	■	■	■	■	■	■

Table 5.2: The optimal design suggested by GA for the data set with 8 Items.

5.4.1 First simulation study

The optimal SQD found by GA in the first simulation study, Table 5.1, is compared with an MMS design, see Section 3.3 for detailed information about MMS. To provide fair evaluation conditions for creating an MMS design, the distribution of items assigned to random sample individuals in the MMS design is set similar to the GA-optimal design. Figure 5.1 displays the frequency of items received by random individuals in the MMS design. In order to be consistent with the optimal design, the overall reduction of items amounts to 50%.

A Monte Carlo study is performed as follows: In each simulation run a (210×4) data set is drawn from a multivariate normal distribution with the mean vector and covariance matrix, mentioned in Subsection 5.3.1. At the next step, both designs are applied to this data set. The analysis models which were used in the fitness function of GA as well as an additional arbitrary linear model which was not accounted for by the fitness function, $\mathbf{E}(X_4) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2$ are estimated, once before applying the design (BD), and once after applying the design and the subsequent multiple imputation (MI), where m=80. The method used for MI is the `mice` built-in method `'norm'` which uses a Bayesian linear regression model for imputation. The simulation is repeated $2,000$ times.

Table 5.3 shows the simulation results averaged over all Monte Carlo studies for the two linear models: Linear model 1 is used in the fitness function to find the optimal design and linear model 2 is used to examine the behavior of the optimal SQD design toward an arbitrary analysis model which was

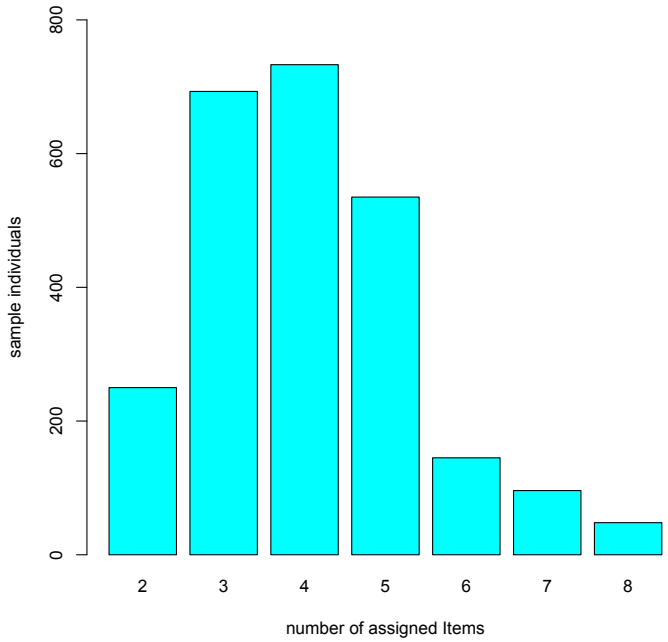


Figure 5.1: Distribution of items on individuals in the MMS design

not considered by the fitness function and compare the behavior of the optimal SQD and a random MMS design toward an arbitrary analysis model. The true model parameters as well as the average model estimates, relative bias and the coverage are demonstrated for $2,000$ samples before deletion (BD), after imposing the optimal SQD found by GA and subsequent MI, and after imposing a random MMS design and subsequent MI. The results indicate absence of bias for most of the estimates. More bias can be seen in the estimates of the two linear models under MMS design in comparison to optimal design which can be referred to identification problem. Furthermore, coverages meet the expected 95% level.

Figure 5.2[13] and Figure 5.3 show the confidence interval (CI) widths of the estimates based on linear model 1 and linear model 2 for the complete data before deletion and after applying GA-optimal design and a random MMS design respectively. The CI width is defined by the corresponding lower and upper bound estimates averaged over all $2,000$ iterations.

Figure 5.2 and Figure 5.3 display much smaller confidence intervals for the

[13]All the figures in this section which demonstrate the confidence interval of model estimates are created by `plotCI` function in `R` package `plotrix` (Lemon, 2006)

	linear model 1				linear model 2		
	β_0	β_1	β_2	β_3	γ_0	γ_1	γ_2
Parameter Estimate	.358	.060	.881	-.119	0.320	.053	.787
BD	.360	.060	.880	-.121	.321	.053	.787
MI(GA)	.375	.058	.872	-.118	.354	.049	.775
MI(MMS)	.348	.050	.897	-.121	.328	.034	.804
Bias							
BD	.002	.001	-.001	-.002	.001	-.001	.000
MI(GA)	.017	-.002	-.008	.002	.034	-.004	-.012
MI(MMS)	-.010	-.010	.017	-.001	.008	-.020	.017
Coverage							
BD	.949	.952	.952	.950	.952	.950	.951
MI(GA)	.955	.958	.949	.941	.953	.951	.960
MI(MMS)	.996	.993	.999	.996	.981	.982	.976

Table 5.3: Comparison of complete data and data reduced by GA-optimal SQD and a random MMS design based on the estimates of two linear models, their relative bias and coverage for 2,000 samples of size 210.

estimates of both linear models for the GA-optimal design in comparison to a random MMS design. This result meets our expectation that the optimal design performs better and illustrates more efficient estimates than MMS design for an analysis model which is used in the fitness function to optimize the SQD. Despite better performance of GA-optimal design for an arbitrary linear model, the analyst cannot bank on a general superiority of the GA solution over plain MMS for general analysis objectives.

5.4.2 Second simulation study

The optimal SQD found by GA in the second simulation study, Table 5.2, is compared with a random MMS design. The steps of the Monte Carlo simulation is the same as in the first simulation study. In each simulation run a (400×8) data set is drawn from a multivariate normal distribution with the mean vector and covariance matrix, mentioned in Subsection 5.3.2. At the next step, both designs are applied to this data set. The analysis models which were used in the fitness function of GA with an additional arbitrary linear model, $\mathsf{E}(X_5) = \beta_0 + \beta_1 X_7 + \beta_2 X_2 + \beta_3 X_4$, are estimated, once before applying the design (BD), and once after applying the design and the subsequent multiple imputation (MI), where m=80. The method used for MI

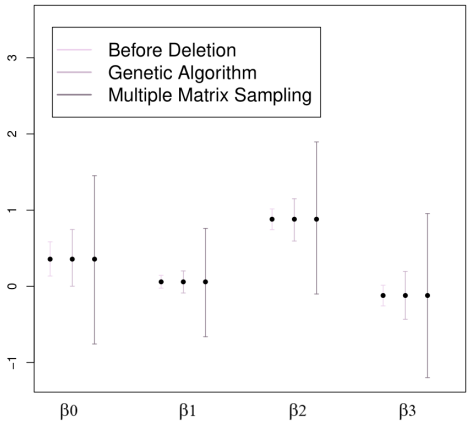


Figure 5.2: CIs of the estimates of the optimized linear model

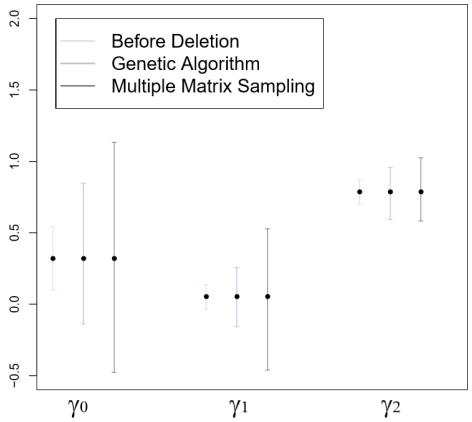


Figure 5.3: CIs of the estimates of an arbitrary linear model

is mice's built-in method `'norm'`. The simulation is repeated $1,000$ times.[14]

Table 5.4 shows the simulation results for the two linear models. The true model parameters as well as the average model estimates, the bias, and the coverage for $1,000$ samples before deletion (BD), after imposing the two designs and the subsequent MI are represented in the table. The results indicate absence of bias for most of the estimates. A few estimates in the first linear model indicate small bias under MMS design and a few estimates indicate small bias in the the second linear model under GA-optimal design. 95% coverage for all the estimates meet the expected 95% level for both designs.

Figure 5.4 and Figure 5.5 show the CI width of the estimates based on the two linear models for the complete data before deletion and after applying the two designs respectively. The CI width is defined by the corresponding lower and upper bound estimates averaged over all $1,000$ iterations.

Figure 5.4 displays much smaller confidence intervals for estimates of the first linear model for the GA-optimal design than for a random MMS design. This result meets our expectation that the optimal design performs better and illustrates more efficient estimates than MMS design for an analysis model which is used in the fitness function.

Figure 5.5 displays smaller confidence intervals for the estimates of the arbitrary model for the GA-optimal design than for a random MMS design.

[14]The number of iterations is less compared to the first simulation study to reduce the computation time.

	linear model 1								linear model 2			
	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_0	β_1	β_2	β_3
Parameter Estimate	.740	-.078	-.006	.025	.058	.110	.193	.328	.292	.246	.196	.412
BD	.742	-.076	-.010	.027	.057	.112	.192	.326	.292	.245	.196	.412
MI(GA)	.739	-.074	-.011	.025	.066	.107	.191	.326	.312	.249	.199	.395
MI(MMS)	.708	-.093	.002	.028	.046	.111	.193	.359	.303	.240	.195	.412
Bias												
BD	.002	.002	-.003	.002	-.001	.003	-.001	-.002	.000	-.001	.000	.000
MI(GA)	-.001	.004	-.004	.001	.008	-.002	-.002	-.002	.020	.003	.002	-.017
MI(MMS)	-.032	-.016	.008	.003	-.012	.002	.000	.031	.011	-.006	-.001	.000
Coverage												
BD	.951	.948	.946	.957	.944	.952	.946	.954	.954	.960	.955	.949
MI(GA)	.952	.950	.951	.960	.963	.957	.951	.956	.961	.960	.960	.964
MI(MMS)	.982	.982	.973	.990	.984	.985	.990	.988	.960	.958	.956	.955

Table 5.4: Comparison of complete data and data reduced by GA-optimal SQD and a random MMS design based on the estimates of two linear models, their bias and coverage for $1,000$ samples of size 400.

The extent to which the CI width of the arbitrary model estimates differ under two designs are however less prominent in the second model.

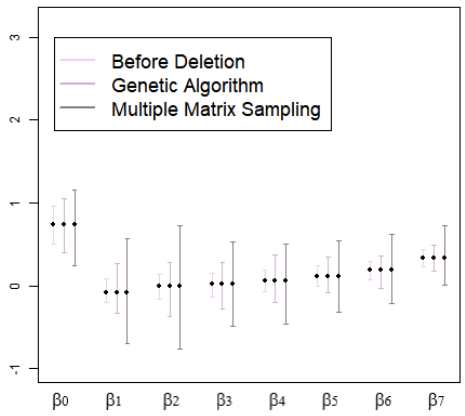


Figure 5.4: CIs of the estimates of the optimized linear model

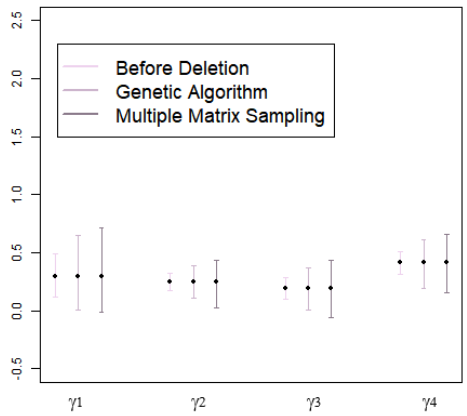


Figure 5.5: CIs of the estimates of an arbitrary linear model

5.5 Empirical application

In this section an application of GA on empirical data is illustrated. For this purpose data from Wave 4 of Starting Cohort 3 (SC3) of the National Educational Panel Study (NEPS) is used (see Blossfeld et al., 2011). The data refers to the Scientific Use File (SUF; DOI:10.5157/NEPS:SC3:6.0.1). SC3 follows the pathway of the students in Grade 5 through lower secondary education. The original sample is made of a main sample of students in Grade 5 in regular schools and special-needs schools and a supplement sample of students with a migration background from Turkey and former Soviet Union. Due to the Federal-State-specific timing in transition in lower secondary education in regular schools a refreshment sample of students in Grade 7 was drawn two years later.

In this study a sample of $2,500$ students on 48 variables are selected randomly from the original sample. The original sample excluding students of special-needs schools consists of 6239 students. The rate of missing values are ranged from 0.0% to 0.08% in the selected 48 variables. In order to create an artificial complete population missing values are pre-imputed using a single imputation with the built-in method `'pmm'` in `R` package `mice`.

The variables of the sample consist of:

- 8 ordinal variables with four levels about student's motivation in subjects 'German' and 'Mathematics'.
- 12 ordinal variables with four levels about the motivational reason of studying for school such as "good career opportunities" or "better achievements".
- 9 ordinal variables with ten levels about satisfaction with life, possessions, health, family, friends, school, having say in family, school, and society. Three ordinal variables with four levels about parental support.
- 2 categorical variables with seven levels about idealistic and realistic educational aspiration which indicate the school certificate aspired by the students correspondingly.
- 14 ordinal variables with six levels about students occupational orientation,

see Table A.4 in Appendix A for more details about the variable descriptions.

Furthermore, Table A.5, Table A.6, and Table A.7 in Appendix A represent the correlation matrix of the original data. Every couple of variables belong to a common subject and their correlations are represented by submatrices bordered by blue rectangles. The correlations within the rectangles are higher relative to the correlations outside the rectangles. To speed up GA in finding the optimal SQD, the variables are categorized in eight blocks, each block containing six variables. The variables with high correlation are assigned to different blocks, see Table A.8 in Appendix A for details. Instead of assigning individual items to each subsample of individuals, a whole block is assigned to each subsample of individuals. The number of subsamples is set to twenty. This ensures more variation in the choice of sub-questionnaires in each SQD. Since for eight blocks, $2^8 = 256$ block combinations are possible, excluding undesirable combinations (one with no blocks and eight with only one block) 247 sub-questionnaires can be produced. Similar to the simulation studies discussed in Section 5.3.1 and Section 5.3.2, to decrease the computation time, for creating the GA population the desirable combinations (sub-questionnaires) are labeled with indexes from one to 247. Each chromosome is a random draw of size twenty from these labels.

Furthermore, the GA parameters used for this study are set as follows,

- $N = 160$
- $I = 20$

- $\gamma \approx \frac{1}{1+20} = .05$
- $\beta \approx N \times .05 = 8$
- $\delta = 41$

The fitness function evaluates the SQDs by means of the average FMI of following parameter estimates, the arithmetic mean of all 48 variables, estimates of an ordered logistic regression model, [15,16]

$$\mathbf{E}(V_9^*) = \beta_0^{OL} + \beta_i^{OL} V_i, \quad V_i \in (V_{13}, V_{16}, V_{17}, V_{20}, V_{26}, V_{31}, V_{34})$$

$$V_9 = \begin{cases} 1 & \text{if} \quad -\infty < V_9^* \leq \mu_{i,1}, \\ 2 & \text{if} \quad \mu_{i,1} < V_9^* \leq \mu_{i,2}, \\ 3 & \text{if} \quad \mu_{i,2} < V_9^* < +\infty. \end{cases}$$

and two linear models,

$$\mathbf{E}(V_{22}) = \beta_0^{L1} + \beta_i^{L1} V_i$$
$$V_i \in (V_1, V_4, V_6, V_{31}, V_{34}, V_{38}, V_{41}, V_{48})$$

$$\mathbf{E}(V_{26}) = \beta_0^{L2} + \beta_i^{L2} V_i$$
$$V_i \in (V_5, V_9, V_{13}, V_{17}, V_{20}, V_{32}, V_{33}, V_{43})$$

The model estimates are displayed in Table A.10 in Appendix A.

After 41 iterations no significant improvement can be recognized in the SQD suggested by GA and therefore GA is terminated. Table A.3 shows the best evaluation value in each iteration as well as the average evaluation value over all the population members in each iteration. The optimal SQD found by GA is illustrated in Table 5.5.[17] [18]

[15] `polr` function in `R` package `MASS` (Venables & Ripley, 2002) is used for estimating the ordered logistic model.

[16] The independent variable in the ordered logistic model `v9` has three levels `2,3,4`. Originally this variable had four levels, where the first two were combined to avoid empty cells.

[17] In MI-Step of the fitness function we set m=60 and repeat MI 2 times to obtain stable evaluation values. The method used for MI is mice's built-in method `'pmm'` which uses a predictive mean matching method for imputation.

[18] The number of imputations and the number of times MI is repeated is set according to small simulation studies which were performed beforehand.

Subsample	Bl1	Bl2	Bl3	Bl4	Bl5	Bl6	Bl7	Bl8
1-125			■	■			■	
126-250			■	■			■	■
251-375		■	■		■	■		
376-500	■					■	■	
501-625	■			■			■	
626-750	■			■		■		
751-875	■			■		■	■	
876-1000	■		■					■
1001-1125	■		■		■	■		
1126-1250	■		■		■	■		■
1251-1350	■		■	■		■		
1351-1500	■	■					■	
1501-1625	■	■	■		■			■
1626-1750	■	■	■	■			■	■
1751-1875	■	■	■	■			■	■
1876-2000	■	■	■	■		■		
2001-2125	■	■	■	■		■	■	
2126-2250	■	■	■	■	■			
2251-2375	■	■	■	■	■			■
2376-2500	■	■	■	■	■	■	■	■

Table 5.5: The optimal design suggested by GA for the NEPS data.

The last row of the optimal design in Table 5.5 indicates a subset of individuals who receive all the items. As mentioned before this can lead to response burden and consequently to measurement error and potential MNAR dropouts. Therefore, caution must be taken in reality to avoid undesirable results caused by assigning the full questionnaire to a subsample. This can be carried out by simply implementing a constraint to the algorithm.

To evaluate the optimal design found by GA (see Table 5.5) a Monte carlo study is performed to compare the GA-optimal design with the complete data as well as an MMS design. To provide fair evaluation conditions, for creating an MMS design the distribution of items assigned to random sample individuals in the MMS design is set similar to the GA-optimal design. Figure 5.6 displays the frequency of items received by random individuals in the MMS design. In order to be consistent with the optimal design, the

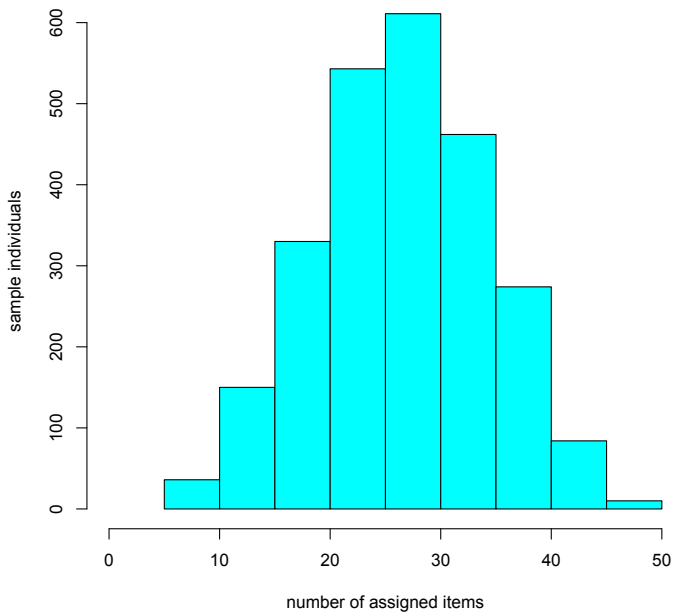


Figure 5.6: Distribution of items on individuals in the MMS design

overall reduction of items amounts to 44%. For simulation purposes, the original data of size 6239 is considered as the population. In each simulation run a random sample of size 2500 is drawn without replacement from the population. The proportion of the sample to the population is rather large which can lead to small population bias. Since this problem did not affect the purpose of our work we proceeded with drawing large samples.At the next step, both designs are applied to this data set. The analysis models which were used in the fitness function of GA as well as two additional arbitrary models are estimated, once before applying the design (BD), and once after applying the design and the subsequent multiple imputation (MI), where m=56.[19] The method used for MI is mice's built-in method `'pmm'`, which uses a predictive mean matching model for imputation. The simulation is repeated 800 times.[20] The additional arbitrary models used for simulation

[19]Since the number of available cores on linux cluster were 28 the number of imputation is set to a multiple of it.

[20]The number of Monte Carlo iterations could not be larger due to computation time restrictions.

study are as follows, an ordered logistic regression model,[21]

$$\mathbf{E}(V_{11}^{*}) = \gamma_0^{OL} + \gamma_i^{OL} V_i, \quad V_i \in (V_7, V_{14}, V_{19}, V_{21}, V_{27}, V_{30}, V_{36})$$

$$V_{11} = \begin{cases} 1 & \text{if} \quad -\infty < V_{11}^{*} \leq \mu_{i,1}, \\ 2 & \text{if} \quad \mu_{i,1} < V_{11}^{*} \leq \mu_{i,2}, \\ 3 & \text{if} \quad \mu_{i,2} < V_{11}^{*} < +\infty. \end{cases}$$

and a linear model,

$$\mathbf{E}(V_{28}) = \gamma_0^{L} + \gamma_i^{L} V_i$$
$$V_i \in (V_3, V_6, V_{12}, V_{18}, V_{30}, V_{35}, V_{42}, V_{46})$$

Relative bias and 95% coverage are illustrated in Table 5.6 for the linear models and in Table 5.7 for the ordered logistic models.

The first and the second linear models in Table 5.6 are those that are used in the fitness function to find the optimal design. The arbitrary linear model was not implemented in the fitness function and is only used to examine the optimal SQD with an arbitrary analysis model and compare the optimal SQD and MMS design for such arbitrary analysis models. The table illustrates the relative bias and the 95% coverage for 800 samples after imposing the optimal SQD found by GA and subsequent MI and after applying a random MMS design and subsequent MI. The results indicate that despite very good coverage high relative bias arises for some coefficients. This can partially be due to the observations with high leverage (high cook's distance) which cause the variation of parameter estimates iteration after iteration. this in combination with very small parameter values can lead to a high relative bias.

[21]The independent variable in the ordered logistic model `v11` has three levels `1,2,3,4`.

Table 5.6: Relative bias and coverage for linear model estimates.

	First linear model in fitness function				Second linear model in fitness function				An arbitrary linear model			
	GA-design		MMS design		GA-design		MMS design		GA-design		MMS design	
	bias	coverage	bias	coverage	bias	coverage	bias	coverage	bias	coverage	bias	coverage
β_0^{L1}	-.013	.92	.005	.97	β_0^{L2} -.136	.92	-.069	.93	γ_0^L -.062	.98	-.038	.97
β_1^{L1}	-.054	.94	-.130	.92	β_1^{L2} -.021	.99	-.027	.95	γ_1^L .072	.97	.068	.96
β_2^{L1}	-.221	.95	-.225	.96	β_2^{L2} -.032	.99	-.091	.92	γ_2^L -.262	.92	-.154	.96
β_3^{L1}	-.059	.98	.057	.94	β_3^{L2} .011	.91	-.050	.97	γ_3^L -.018	.92	-.024	.95
β_4^{L1}	.023	.99	-.017	.95	β_4^{L2} -.119	.98	-.037	.97	γ_4^L .024	.94	.021	.94
β_5^{L1}	.080	.93	.021	.98	β_5^{L2} .319	.94	-.361	.92	γ_5^L .173	.93	.164	.95
β_6^{L1}	-.271	.94	-.240	.97	β_6^{L2} .203	.96	.370	.97	γ_6^L -.066	.95	-.090	.86
β_7^{L1}	-.086	.87	-.104	.98	β_7^{L2} .237	.98	.389	.94	γ_7^L .899	.98	.414	.92
β_8^{L1}	.015	.96	-.040	.98	β_8^{L2} .002	.99	-.203	.88	γ_8^L -.522	.98	-.117	.94

Table 5.7: Relative bias and coverage for ordered logit model estimates.

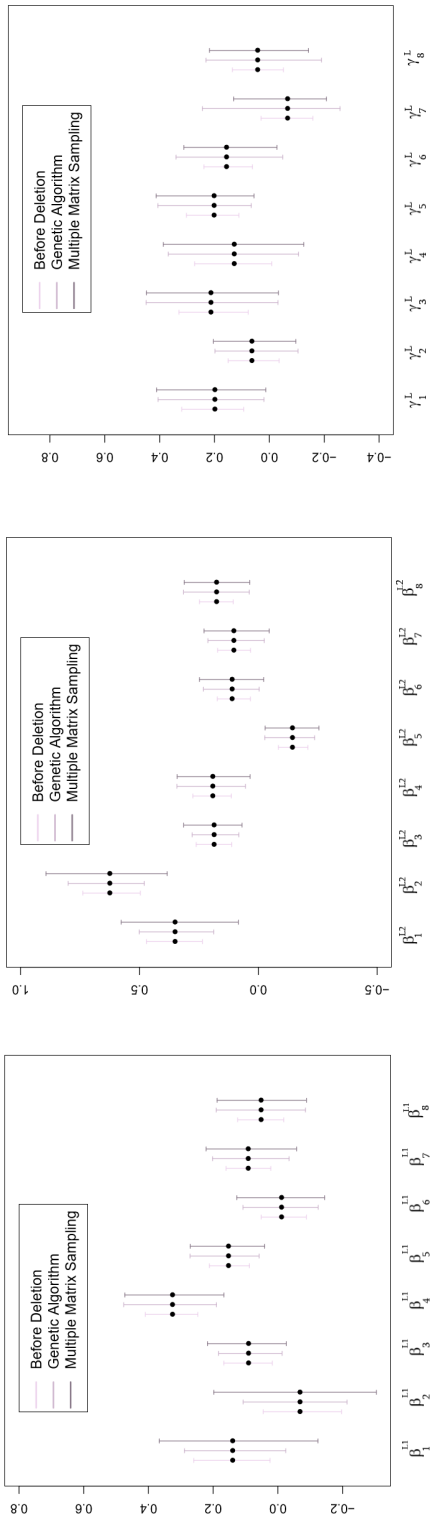
	Model used in fitness func.					An arbitrary model			
	GA-design		MMS design			GA-design		MMS design	
	bias	coverage	bias	coverage		bias	coverage	bias	coverage
β_1^{OL}	-.048	.99	-.064	.97	γ_1^{OL}	.037	.96	-.073	.94
β_2^{OL}	-.052	.99	.074	.99	γ_2^{OL}	-.043	.97	-.024	.98
β_3^{OL}	-.000	.90	.039	.93	γ_3^{OL}	-.037	.98	-.029	.93
β_4^{OL}	-.039	.92	-.132	.94	γ_4^{OL}	-.024	.99	-.155	.94
β_5^{OL}	-.058	.98	-.113	.95	γ_5^{OL}	-.009	.99	-.022	.98
β_6^{OL}	-.154	.94	.027	.94	γ_6^{OL}	-.105	.95	-.037	.94
β_7^{OL}	-.003	.98	-.003	.97	γ_7^{OL}	-.126	.98	-.071	.94

Figures 5.7 and 5.8 show the confidence interval widths of the estimates based on the models implemented in the fitness function and arbitrary models for the complete data before deletion and after applying GA-optimal design and a random MMS design respectively. The CI width is defined by the corresponding lower and upper bound estimates averaged over all 800 iterations. Tables A.11 and A.12 in Appendix A display the CI widths of all model parameters.

Figures 5.7a and 5.7b display smaller confidence intervals for seven out of eight estimates of the first and the second linear model for the GA-optimal design than for a random MMS design. This result meets our expectation that the optimal design performs better and illustrates more efficient estimates than MMS design for an analysis model which is used in the fitness function to optimize the SQD.

Figure 5.7c displays smaller confidence interval for some of the estimates of the arbitrary model for the GA-optimal design than for a random MMS design.

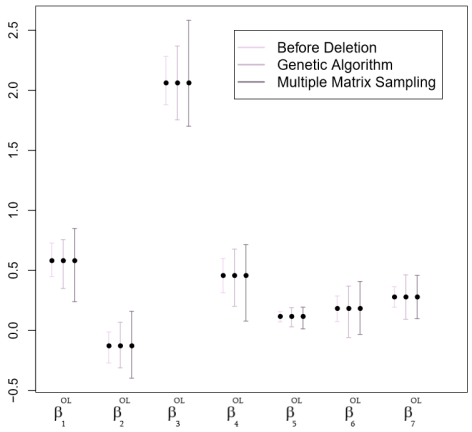
Figure 5.8a displays smaller confidence intervals for five out of seven estimates for GA-optimal design and equal confidence intervals for the other two estimates for both designs. Figure 5.8b displays smaller confidence intervals for six out of seven estimates and equal confidence interval for the GA-optimal design comparing to the MMS design.



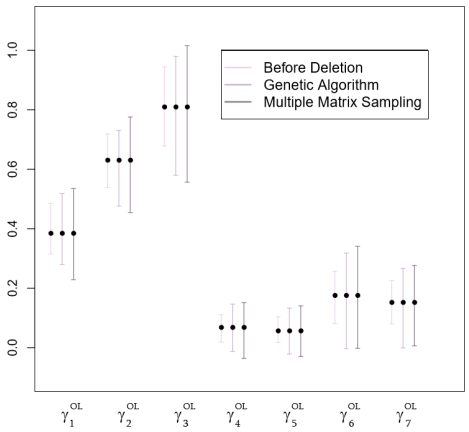
(a) Fitness model 1 (b) Fitness model 2 (c) Arbitrary model

Figure 5.7: CI widths of the linear model estimates under two designs

Figure 5.8: CI widths of the OL model estimates under two designs



(a) Fitness model



(b) Arbitrary model

5.6 Discussion

We have described the applicability of the GA for splitting a questionnaire based on simulated data, with the FMI as the fitness function for identifying the optimal design. The simulation study suggests that for an established analysis model, the GA can find an optimal design which performs better than the MMS design. This is particularly useful if there is a set of analyses which will be conducted in every wave, such as a standardized analysis report. Of course, these analysis objectives could be used for the derivation of the fitness function which was represented in our study by a linear regression as well as an ordered probit model.

Our current experiences with the GA suggest that results and run-time are sensitive to the choice of its parameters, and computing time in particular is an issue that has yet to be improved. One aspect that has not been covered yet due to run-time constraints, is if the GA solution identifies generally superior patterns that can be transferred to other samples or if the more efficient estimates can only be found for the training sample. However, the final solutions are far from 'ordered' patterns and suggest that at least the benefits of MMS discussed in the introduction section – avoiding the identification problem for high-dimensional analyses – also apply to the GA solution. While the various parameters make it difficult to find an optimal combination of settings, thus somewhat hampering the generalizations of our findings, the flexibility of the GA is on the other hand very helpful in customizing it to particular questionnaires. Unlike most other methods for finding optimal designs, constraints such as logically connected items can be easily implemented into the GA method.

Chapter 6

Summary-Outlook

The main objective of this thesis is to develop and evaluate new data collection methods to reduce the respondent burden and improve the data quality by administering only a fraction of the long questionnaire to a subsample of individuals while the least amount of information is lost due to non-asked questions. Two main studies performed in these thesis are as follows:

In the first study (see Chapter 4) a split questionnaire design has been applied to NEPS data. The data consist of item blocks which contain several questions about the same subject. A pilot data collected prior to the main data collection reveals the approximate correlation of variables in the data. The within block correlations are relatively higher than the between block correlations. Accordingly, a split questionnaire is designed to assign the most suitable combination of items to each subsample of individuals to avoid as much information loss as possible. To evaluate our design a data set with the same characteristics corresponding the correlation matrix of the pilot data is simulated. The data set contains metric, ordinal and nominal variables. The data set reduced by the design is imputed using multiple imputation. Several statistical models such as linear, ordered probit and multinomial logit are fitted to data reduced by our design and multiply imputed and the model estimates and the corresponding coverages are compared to model estimates of complete data before deletion by design and data reduced by a random multiple matrix sampling design. The results illustrate how using this design could affect the regression coefficient estimates of the analysis models. The estimates are unbiased and the coverages are high, although further improvements may be gained via consideration of richer dependence structure for the variables with nominal scale and other measure of dependence between the variables. This design is implemented on empirical data with similar correlation matrix structure and the results confirm that this design could work very well for data with this particular correlation matrix structure.

We chose an ordered design in the first study which only works for a particular data structure, but can be applied to pen and paper questionnaires, because the number of different questionnaires is small. The GA is an efficient approach that enables choosing the best combination of possible sub-questionnaires and gives us a better opportunity to draw more information from the design but can only be used in a sensible way if the questionnaire is computer-based, as in CAPI or CATI setting. As explained in the second study in Chapter 5 GA starts with an initial population of different combinations of sub-questionnaires and evolves generation by generation to a fitter population by means of GA operators such as crossover and mutation. The applicability of the GA for splitting a questionnaire is explained based on simulated data and empirical data, where the FMI is used as the fitness function for identifying the optimal design. The study suggests that for an established analysis model, the GA can find an optimal design which performs better than the MMS design in most cases. This is particularly useful if there is a set of analyses which will be conducted in every wave. The fitness function can then be defined based on these analysis objectives.

GA method is introduced as an alternative to the ordered SQSD by Raghunathan & Grizzle (1995) and a more flexible approach in comparison to the SQSD by Adigüzel & Wedel (2008). Since GA is based on step by step evolution in the solution space by means of FMI as the fitness function there is no limitation on the level of measurement of the variables of the training data set and GA can easily be applied to mixed data. Currently, the most notable limitation of our approach is due to high computation time. On this account, some constraints are imposed. For instance an initial population of solutions is set for GA to force it to choose solely from this solution space. Without the limitations the initial solution space could be made of any random MMS designs. Another example is the constraint on the number of different versions of sub-questionnaires which should be administered to each subsample of individuals. This constraint is set to decrease the length of solutions in the solution space to enable smaller population size and consequently decrease the computation time. The number of questions which are assigned to each sub-questionnaire is another constraint which is set due to time limitation. After all, we expect that it will take some time until our approach can be widely used in this field. Furthermore, prior information about the correlation of the variables are still required. Apart from that, further research is needed in respect to parameter setting since this approach is very sensitive to the parameter selection such as the initial population size, crossover and mutation chance probabilities.

References

Adams, L. L. M., & Gale, D. (1982). Solving the quandary between questionnaire length and response rate in educational research. *Research in Higher Education*, *17*(3), 231-240. Retrieved from http://www.jstor.org/stable/40195497

Adigüzel, F., & Wedel, M. (2008). Questionnaire survey design for massive surveys. *Journal of Marketing Research*, *XLV*, 608-–617.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*(278), 200–203.

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International statistical review = Revue internationale de statistique*, *78*(1), 40–64.

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

Barnard, J., & Rubin, D. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*, 948-955.

Bates, D., & Maechler, M. (2016). Matrix: Sparse and dense matrix classes and methods [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=Matrix (R package version 1.2-7.1)

Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, *37*(1), 129–145.

Beaton, A., & Zwick, R. (1992). Overview of the national assessment of educational progress. *Journal of Educational Statistics.*, *17*, 95-109.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special Issue]: Zeitschrift für Erziehungswissenschaft* (Vol. 14). Wiesbaden: VS Verlag für Sozialwissenschaften.

Blumberg, H., Fuller, C., & Hare, A. (1974). Response rates in postal surveys. *Public Opinion Quarterly*, *38*, 113-123.

Brand, J. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. The Author. Retrieved from https://books.google.de/books?id=-YOTywAACAAJ

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Dillman, D., Sinclair, M., & Clark, J. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*, *57*, 289-304.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, *89*, 463-478.

Enders, C. K. (2010). *Applied missing data analysis.* NewYork: Guilford Press.

Gelfand, A., Hills, S., Racine-Poon, A., & Smith, A. (1990). Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, *85*, 972–985.

Gelman, A., King, G., & Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, *93*(443).

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*, 721-741.

Gonzalez, J. M., & Eltinge, J. L. (2007). Multiple matrix sampling: A review. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 3069-75.

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

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on responsequality. *Public Opinion Quarterly*, *45*, 549—559.

Holland, J. H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor: University of Michigan Press.

Kennickell, A. B. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Asso- ciation*, 1–10.

Kim, J., & Curry, J. (1997). The treatment of missing data in multivariate analysis. *Sociol. Meth. Res.*, *6*, 215–240.

Leifeld, P. (2013). texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, *55*(8), 1–24. Retrieved from http://www.jstatsoft.org/v55/i08/

Lemon, J. (2006). Plotrix: a package in the red light district of r. *R-News*, *6*(4), 8-12.

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

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data.* (2nd ed.). New York: Wiley.

Littvay, L. (2009). Questionnaire design considerations with planned missing data. *Review of psychology*, *16*(2), 103-114.

Meinfelder, F., & Schnapp, T. (2015). Baboon: Bayesian bootstrap predictive mean matching - multiple and single imputation for discrete data [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=BaBooN

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133-161. Retrieved from http://www.jstor.org/stable/1434599

Mitchell, M. (1996). *An introduction to genetic algorithms.* Cambridge: MA: MIT Press.

Munger, G. F., & Loyd, B. H. (1988). The use of multiple matrix sampling for survey research. *The Journal of Experimental Education*, *56*(4), pp. 187-191. Retrieved from http://www.jstor.org/stable/20151742

Navarro, A., & Griffin, R. (1993). Matrix sampling designs for the year 2000 census. *American Statistical Association*, 480-485.

Okner, B. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, *1*, 325-362.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raghunathan, T. (2015). *Missing data analysis in practice*. Chapman & Hall /CRC.

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

Raghunathan, T., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, *27*(1), 85–96.

Raghunathan, T., Solenberger, P. W., & Hoewyk, J. V. (2002). IVEware: Imputation and variance estimation software, user guide [Computer software manual].

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, *91*(434), 499-506. doi: 10.1080/01621459.1996.10476910

Rao, J. N. K., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, *79*(4), 811-822. Retrieved from http://biomet.oxfordjournals.org/content/79/4/811.abstract doi: 10.1093/biomet/79.4.811

Rässler, S. (2002). *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* New York: Springer. (ISBN 9780387955162)

Roszkowski, M. J., & Bean, A. G. (1990). Believe it or not! longer questionnaires have lower response rates. *Journal of Business and Psychology*, *4*, 495-509.

Rubin, D. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, *69*, 467-474.

Rubin, D. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Asso- ciation*, 20–34.

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

Rubin, D. (1987a). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, *91*(434), 473–489.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton and London and New York and Washington D.C.: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Sharp, L., & Frankel, J. (1983). Respondent burden: A test of some common assumptions. *Public Opinion Quarterly*, *47*, 36–53.

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

Sims, C. (1972). Comment on okner. *Annals of Economic and Social Measurement*, *1*, 343 - 345.

Sirotnik, K., & Wellington, R. (1977). Incidence sampling: An integrated theory for matrix sampling. *Journal of Educational Measurement*, *14*, 343-399.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, *82*, 528–550.

Thomas, N., & Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics*, *22*(4), 425–445.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. Retrieved from http://www.jstatsoft.org/v45/i03/

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall /CRC.

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

van Buuren, S., & Oudshoorn, C. (2000). *Multivariate imputation by chained equations: Mice v1. 0 users's manual*. TNO Prevention and Health, Public Health.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4 (ISBN 0-387-95457-0)

Wacholder, S., Carroll, R., Pee, D., & Gail, M. (1994). The partial questionnaire design for casecontrol studies (with discussion). *Statistics in Medicine*, *13*, 623-649.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, *4*, 65-85.

Xuan, J., Jiang, H., & Ren, Z. (2011). *Pseudo code of genetic algorithm and multi-start strategy based simulated annealing algorithm for large scale next release problem," technical report.*

Zeger, L. M., & Thomas, N. (1997). Efficient matrix sampling for correlated latent traits: Examples from the national assessment of educational progress. *Journal of the American Statistical Association*, *92*, 416-425.

Appendix A

Tables

iter. No.	best eval. value	mean eval. value
1	0.569	0.696
2	0.566	0.695
3	0.558	0.693
4	0.548	0.685
5	0.545	0.687
6	0.546	0.685
7	0.543	0.681
8	0.544	0.671
9	0.542	0.667
10	0.543	0.651
11	0.538	0.653
12	0.524	0.647
13	0.523	0.645
14	0.524	0.649
15	0.521	0.635
16	0.521	0.641
17	0.521	0.634
18	0.522	0.645
19	0.521	0.637
20	0.522	0.645
21	0.520	0.657
22	0.521	0.653
23	0.520	0.644
24	0.519	0.639
25	0.520	0.625
26	0.520	0.631
27	0.519	0.630
28	0.517	0.634
29	0.519	0.625
30	0.520	0.615
31	0.519	0.622
32	0.519	0.615
33	0.518	0.619
34	0.519	0.609
35	0.520	0.614
36	0.519	0.610
37	0.518	0.627
38	0.518	0.626
39	0.518	0.629
40	0.518	0.628

Table A.1: Best and average fitness values for each GA iteration.

iter. No.	best eval. value	mean eval. value
1	0.596	0.672
2	0.551	0.666
3	0.551	0.660
4	0.549	0.661
5	0.549	0.663
6	0.549	0.660
7	0.549	0.664
8	0.549	0.662
9	0.549	0.656
10	0.549	0.662
11	0.549	0.659
12	0.549	0.662
13	0.549	0.656
14	0.549	0.653
15	0.549	0.647
16	0.549	0.646
17	0.549	0.645
18	0.544	0.647
19	0.539	0.644
20	0.539	0.638
21	0.539	0.639
22	0.538	0.630
23	0.537	0.629
24	0.529	0.628
25	0.529	0.618
26	0.529	0.617
27	0.523	0.620
28	0.523	0.618
29	0.523	0.616
30	0.523	0.613
31	0.523	0.612
32	0.523	0.618
33	0.508	0.611
34	0.505	0.611
35	0.505	0.618
36	0.505	0.617
37	0.505	0.608
38	0.500	0.614
39	0.500	0.612
40	0.500	0.600

Table A.2: Best and average fitness values for each GA iteration.

iter. No.	best eval. value	mean eval. value
1	0.457	0.501
2	0.452	0.500
3	0.448	0.501
4	0.443	0.500
5	0.443	0.501
6	0.433	0.500
7	0.433	0.501
8	0.433	0.500
9	0.403	0.494
10	0.408	0.476
11	0.396	0.462
12	0.399	0.451
13	0.398	0.445
14	0.394	0.444
15	0.387	0.441
16	0.394	0.441
17	0.386	0.438
18	0.377	0.435
19	0.365	0.437
20	0.375	0.425
21	0.371	0.416
22	0.368	0.421
23	0.369	0.418
24	0.365	0.419
25	0.359	0.411
26	0.359	0.411
27	0.357	0.408
28	0.361	0.408
29	0.356	0.408
30	0.358	0.411
31	0.359	0.404
32	0.359	0.407
33	0.366	0.411
34	0.354	0.404
35	0.362	0.438
36	0.371	0.421
37	0.367	0.415
38	0.363	0.427
39	0.370	0.413
40	0.366	0.415
41	0.363	0.415

Table A.3: Best and average fitness values for each GA iteration.

Table A.4: Variables selected for constructing the empirical data.

Variable	**label**	**description**
V_1	t66400a	Motivation German: Enjoyment of content
V_2	t66400b	Motivation German: Content reflects personal inclinations
V_3	t66400c	Motivation German: Important content
V_4	t66400d	Motivation German: Interested in content
V_5	t66401a	Motivation Math: Enjoyment of content
V_6	t66401b	Motivation Math: Content reflects personal inclinations
V_7	t66401c	Motivation Math: Important content
V_8	t66401d	Motivation Math: Interested in content
V_9	t66402a	Motivation: Obtain school-leaving certificate
V_{10}	t66402b	Motivation: Do well
V_{11}	t66402c	Motivation: Success is very important to me
V_{12}	t66402d	Motivation: Do well in examinations
V_{13}	t66403a	Motivation: Be one of the best
V_{14}	t66403b	Motivation: More intelligent than others
V_{15}	t66403c	Motivation: Excellent achievement
V_{16}	t66403d	Motivation: Do better than others in examinations
V_{17}	t66404a	Motivation: Good career opportunities
V_{18}	t66404b	Motivation: Financial security
V_{19}	t66404c	Motivation: Well-paid career
V_{20}	t66404d	Motivation: Increase chances of getting a job
V_{21}	t514001	Satisfaction with life
V_{22}	t514002	Satisfaction with possessions
V_{23}	t514003	Satisfaction with health
V_{24}	t514004	Satisfaction with family
V_{25}	t514005	Satisfaction with acquaintances and friends
V_{26}	t514006	Satisfaction with school
V_{27}	t517200	Satisfaction: having say in family
V_{28}	t517201	Satisfaction: having say in class/school
V_{29}	t517202	Satisfaction: having say in society
V_{30}	t28430b	Student: Parental support: Recitations/ presentations
V_{31}	t28430c	Student: Parental support: Talk about topics
V_{32}	t28430d	Student: Parental support: Talk about problems
V_{33}	t31035c	Idealistic educational aspiration - highest school-leaving qualification
V_{34}	t31135c	Realistic educational aspiration - highest school-leaving qualification
V_{35}	t66210a	Occup. orientation: Learning
V_{36}	t66210b	Occup. orientation: Good working atmosphere
V_{37}	t66210c	Occup. orientation: Opportunities for advancement

Table A.4: Variables selected for constructing the empirical data.

Variable	**label**	**description**
V_{38}	t66210d	Occup. orientation: Good working hours
V_{39}	t66210e	Occup. orientation: Variety
V_{40}	t66210f	Occup. orientation: Interesting work
V_{41}	t66210g	Occup. orientation: Job security
V_{42}	t66210h	Occup. orientation: Good pay
V_{43}	t66210i	Occup. orientation: Match with skills
V_{44}	t66210k	Occup. orientation: Autonomy
V_{45}	t66210l	Occup. orientation: useful work
V_{46}	t66210m	Occup. orientation: authority to decide
V_{47}	t66210n	Occup. orientation: helping others
V_{48}	t66210p	Occup. orientation: doing useful things

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20
v1	1.00	0.68	0.56	0.67	0.13	0.13	0.17	0.12	0.20	0.25	0.25	0.20	0.21	0.13	0.24	0.14	0.16	0.13	0.10	0.14
v2	0.68	1.00	0.60	0.68	0.17	0.23	0.21	0.18	0.17	0.23	0.25	0.20	0.22	0.16	0.26	0.17	0.15	0.12	0.10	0.16
v3	0.56	0.60	1.00	0.62	0.19	0.22	0.35	0.21	0.22	0.28	0.29	0.24	0.24	0.17	0.27	0.17	0.19	0.16	0.12	0.19
v4	0.67	0.68	0.62	1.00	0.18	0.20	0.22	0.22	0.17	0.24	0.25	0.20	0.22	0.15	0.25	0.16	0.13	0.09	0.08	0.15
v5	0.13	0.17	0.19	0.18	1.00	0.79	0.64	0.82	0.14	0.19	0.22	0.17	0.24	0.20	0.24	0.20	0.13	0.11	0.10	0.17
v6	0.13	0.23	0.22	0.20	0.79	1.00	0.67	0.80	0.13	0.19	0.24	0.17	0.25	0.23	0.26	0.22	0.12	0.13	0.11	0.19
v7	0.17	0.21	0.35	0.22	0.64	0.67	1.00	0.68	0.19	0.23	0.27	0.22	0.24	0.18	0.26	0.18	0.19	0.16	0.15	0.21
v8	0.12	0.18	0.21	0.22	0.82	0.80	0.68	1.00	0.14	0.20	0.24	0.18	0.25	0.22	0.26	0.22	0.12	0.11	0.10	0.18
v9	0.20	0.17	0.22	0.17	0.14	0.13	0.19	0.14	1.00	0.49	0.42	0.42	0.32	0.11	0.30	0.15	0.58	0.44	0.43	0.38
v10	0.25	0.23	0.28	0.24	0.19	0.19	0.23	0.20	0.49	1.00	0.56	0.55	0.46	0.27	0.51	0.31	0.43	0.36	0.34	0.35
v11	0.25	0.25	0.29	0.25	0.22	0.24	0.27	0.24	0.42	0.56	1.00	0.49	0.46	0.32	0.57	0.36	0.35	0.34	0.31	0.35
v12	0.20	0.20	0.24	0.20	0.17	0.17	0.22	0.18	0.42	0.55	0.49	1.00	0.37	0.22	0.44	0.33	0.40	0.34	0.38	0.37
v13	0.21	0.22	0.24	0.22	0.24	0.25	0.24	0.25	0.32	0.46	0.46	0.37	1.00	0.56	0.58	0.57	0.25	0.25	0.22	0.29
v14	0.13	0.16	0.17	0.15	0.20	0.23	0.18	0.22	0.11	0.27	0.32	0.22	0.56	1.00	0.51	0.70	0.10	0.18	0.14	0.24
v15	0.24	0.26	0.27	0.25	0.24	0.26	0.26	0.26	0.30	0.51	0.57	0.44	0.58	0.51	1.00	0.54	0.25	0.27	0.25	0.33
v16	0.14	0.17	0.17	0.16	0.20	0.22	0.18	0.22	0.15	0.31	0.36	0.33	0.57	0.70	0.54	1.00	0.13	0.18	0.15	0.26
v17	0.16	0.15	0.19	0.13	0.13	0.12	0.19	0.12	0.58	0.43	0.35	0.40	0.25	0.10	0.25	0.13	1.00	0.57	0.62	0.45
v18	0.13	0.12	0.16	0.09	0.11	0.13	0.16	0.11	0.44	0.36	0.34	0.34	0.25	0.18	0.27	0.18	0.57	1.00	0.66	0.50
v19	0.10	0.10	0.12	0.08	0.10	0.11	0.15	0.10	0.43	0.34	0.31	0.38	0.22	0.14	0.25	0.15	0.62	0.66	1.00	0.48
v20	0.14	0.16	0.19	0.15	0.17	0.19	0.21	0.18	0.38	0.35	0.35	0.37	0.29	0.24	0.33	0.26	0.45	0.50	0.48	1.00

Table A.5: Correlation matrix for for variables v1 to v20 of data used for empirical application of GA.

	v17	v18	v19	v20	v21	v22	v23	v24	v25	v26	v27	v28	v29	v30	v31	v32
v17	1.00	0.57	0.62	0.45	0.17	0.14	0.12	0.16	0.10	0.19	0.17	0.13	0.13	0.13	0.13	0.13
v18	0.57	1.00	0.66	0.50	0.12	0.10	0.08	0.09	0.07	0.12	0.11	0.10	0.09	0.09	0.11	0.11
v19	0.62	0.66	1.00	0.48	0.12	0.10	0.09	0.11	0.09	0.11	0.12	0.09	0.08	0.09	0.10	0.10
v20	0.45	0.50	0.48	1.00	0.13	0.10	0.07	0.11	0.05	0.11	0.13	0.09	0.07	0.10	0.12	0.12
v21	0.17	0.12	0.12	0.13	1.00	0.53	0.48	0.61	0.50	0.50	0.55	0.45	0.43	0.19	0.18	0.18
v22	0.14	0.10	0.10	0.10	0.53	1.00	0.41	0.50	0.42	0.34	0.49	0.39	0.38	0.17	0.19	0.17
v23	0.12	0.08	0.09	0.07	0.48	0.41	1.00	0.46	0.41	0.36	0.43	0.38	0.37	0.11	0.10	0.10
v24	0.16	0.09	0.11	0.11	0.61	0.50	0.46	1.00	0.51	0.38	0.66	0.39	0.39	0.27	0.25	0.26
v25	0.10	0.07	0.09	0.05	0.50	0.42	0.41	0.51	1.00	0.38	0.44	0.47	0.42	0.13	0.14	0.12
v26	0.19	0.12	0.11	0.11	0.50	0.34	0.36	0.38	0.38	1.00	0.41	0.44	0.37	0.12	0.14	0.09
v27	0.17	0.11	0.12	0.13	0.55	0.49	0.43	0.66	0.44	0.41	1.00	0.49	0.46	0.23	0.27	0.26
v28	0.13	0.10	0.09	0.09	0.45	0.39	0.38	0.39	0.47	0.44	0.49	1.00	0.61	0.12	0.16	0.11
v29	0.13	0.09	0.08	0.07	0.43	0.38	0.37	0.39	0.42	0.37	0.46	0.61	1.00	0.14	0.15	0.13
v30	0.13	0.09	0.09	0.10	0.19	0.17	0.11	0.27	0.13	0.12	0.23	0.12	0.14	1.00	0.50	0.40
v31	0.13	0.11	0.10	0.12	0.18	0.19	0.10	0.25	0.14	0.14	0.27	0.16	0.15	0.50	1.00	0.51
v32	0.13	0.11	0.10	0.12	0.18	0.17	0.10	0.26	0.12	0.09	0.26	0.11	0.13	0.40	0.51	1.00

Table A.6: Correlation matrix for variables v17 to v32 of data used for empirical application of GA.

	v30	v31	v32	v33	v34	v35	v36	v37	v38	v39	v40	v41	v42	v43	v44	v45	v46	v47	v48
v30	1.00	0.50	0.40	0.04	0.08	0.13	0.12	0.09	0.07	0.12	0.09	0.11	0.04	0.09	0.11	0.10	0.09	0.12	0.09
v31	0.50	1.00	0.51	0.05	0.10	0.17	0.11	0.08	0.03	0.12	0.11	0.08	0.02	0.11	0.11	0.14	0.11	0.12	0.10
v32	0.40	0.51	1.00	0.01	0.03	0.14	0.12	0.07	0.03	0.10	0.10	0.11	0.03	0.08	0.10	0.15	0.09	0.14	0.11
v33	0.04	0.05	0.01	1.00	0.69	0.11	0.09	0.14	0.09	0.10	0.13	0.07	0.10	0.17	0.11	0.05	0.17	0.02	0.11
v34	0.08	0.10	0.03	0.69	1.00	0.10	0.06	0.11	0.06	0.09	0.12	0.05	0.05	0.15	0.09	0.03	0.14	0.01	0.09
v35	0.13	0.17	0.14	0.11	0.10	1.00	0.50	0.43	0.28	0.45	0.47	0.37	0.32	0.45	0.34	0.54	0.45	0.41	0.42
v36	0.12	0.11	0.12	0.09	0.06	0.50	1.00	0.52	0.43	0.45	0.50	0.43	0.43	0.43	0.34	0.45	0.52	0.43	0.45
v37	0.09	0.08	0.07	0.14	0.11	0.43	0.52	1.00	0.50	0.45	0.52	0.48	0.58	0.45	0.40	0.30	0.57	0.34	0.46
v38	0.07	0.03	0.03	0.09	0.06	0.28	0.43	0.50	1.00	0.42	0.47	0.45	0.53	0.37	0.34	0.20	0.41	0.27	0.37
v39	0.12	0.12	0.10	0.10	0.09	0.45	0.45	0.45	0.42	1.00	0.64	0.40	0.34	0.43	0.39	0.38	0.44	0.47	0.46
v40	0.09	0.11	0.10	0.13	0.12	0.47	0.50	0.52	0.47	0.64	1.00	0.47	0.44	0.48	0.39	0.36	0.50	0.47	0.50
v41	0.11	0.08	0.11	0.07	0.05	0.37	0.43	0.48	0.45	0.40	0.47	1.00	0.48	0.40	0.36	0.33	0.41	0.43	0.44
v42	0.04	0.02	0.03	0.10	0.05	0.32	0.43	0.58	0.53	0.34	0.44	0.48	1.00	0.39	0.35	0.19	0.41	0.24	0.40
v43	0.09	0.11	0.08	0.17	0.15	0.45	0.43	0.45	0.37	0.43	0.48	0.40	0.39	1.00	0.47	0.35	0.46	0.33	0.43
v44	0.11	0.11	0.10	0.11	0.09	0.34	0.34	0.40	0.34	0.39	0.39	0.36	0.35	0.47	1.00	0.27	0.46	0.29	0.44
v45	0.10	0.14	0.15	0.05	0.03	0.54	0.45	0.30	0.20	0.38	0.36	0.33	0.19	0.35	0.27	1.00	0.36	0.53	0.40
v46	0.09	0.11	0.09	0.17	0.14	0.45	0.52	0.57	0.41	0.44	0.50	0.41	0.41	0.46	0.46	0.36	1.00	0.36	0.43
v47	0.12	0.12	0.14	0.02	0.01	0.41	0.43	0.34	0.27	0.47	0.47	0.43	0.24	0.33	0.29	0.53	0.36	1.00	0.49
v48	0.09	0.10	0.11	0.11	0.09	0.42	0.45	0.46	0.37	0.46	0.50	0.44	0.40	0.43	0.44	0.40	0.43	0.49	1.00

Table A.7: A Submatrix of Correlation matrix for variables v30 to v48 of data used for empirical application of GA.

Table A.8: Variables assigned to each block in empirical data application of GA.

Blocks	*Variables*					
Block 1	v1	v9	v17	v25	v33	v41
Block 2	v5	v13	v21	v29	v37	v45
Block 3	v2	v10	v18	v26	v34	v42
Block 4	v6	v14	v22	v30	v38	v46
Block 5	v3	v11	v19	v27	v35	v43
Block 6	v7	v15	v23	v31	v39	v47
Block 7	v4	v12	v20	v28	v36	v44
Block 8	v8	v16	v24	v32	v40	v48

Table A.9: Models used in fitness function of GA to find the optimal SQD.

	ordered logit	linear 1	linear 2
v13	**.581**		**.186**
	(.045)		(.034)
v16	**−.128**		
	(.042)		
v17	**2.061**		**.192**
	(.065)		(.062)
v20	**.458**		**−.144**
	(.046)		(.043)
v26	**.117**		
	(.014)		
v31	**.183**	**.326**	
	(.035)	(.026)	
v34	**.279**	**.153**	
	(.028)	(.020)	
v1		**.140**	
		(.038)	
v4		−.068	
		(.039)	
v6		**.091**	
		(.024)	
v38		−.011	
		(.023)	
v41		**.092**	
		(.022)	
v48		**.052**	
		(.023)	
v5			**.350**
			(.028)
v9			**.625**
			(.058)
v32			**.110**
			(.032)
v33			**.103**
			(.028)
v43			**.176**
			(.027)
AIC	7599.182		

Table A.9: Models used in fitness function of GA to find the optimal SQD.

	ordered logit	linear 1	linear 2
BIC	7659.829		
Log Likelihood	-3790.591		
Deviance	7581.182		
Num. obs.	6239	6239	6239
R^2		.056	.110
Adj. R^2		.055	.109
RMSE		1.744	2.129

Coefficients with $p < 0.05$ in **bold**.

Note: This regression model is calculated with the `polr` function in `MASS` package Venables & Ripley (2002) in R (R Core Team, 2015). The independent variable in the first and the second linear models are `v22` and `v26` respectively. [1]

Table A.10: polytomous regression Model used in fitness function of GA to find the optimal SQD.

	v33:4	v33:5	v33:6
v2	−.**267**	−.**363**	−.**317**
	(.087)	(.103)	(.083)
v25	.033	.014	.045
	(.033)	(.039)	(.031)
v30	.110	.080	.**163**
	(.070)	(.082)	(.066)
v37	.**158**	.**377**	.**422**
	(.051)	(.067)	(.049)
AIC	11872.249	11872.249	11872.249
BIC	11973.327	11973.327	11973.327
Log Likelihood	-5921.124	-5921.124	-5921.124
Deviance	11842.249	11842.249	11842.249
Num. obs.	6239	6239	6239

Coefficients with $p < 0.05$ in **bold**.

Note: This regression model is calculated with the `multinom` function in `R` package `nnet` (Venables & Ripley, 2002)

[1] This representation of model estimates is created by texreg function in texreg-package Leifeld (2013) in R

Table A.11: CI widths for linear model estimates.

First linear model in fitness function			Second linear model in fitness function			An arbitrary linear model		
	GA-design	MMS design		GA-design	MMS design		GA-design	MMS design
β_0^{L1}	(4.686;6.315)	(4.702;6.506)	β_0^{L2}	(-0.044;2.714)	(0.113;2.765)	γ_0^L	(3.150;5.647)	(3.426;5.596)
β_1^{L1}	(-.0245;0.289)	(-0.123;0.367)	β_1^{L2}	(0.198;0.487)	(0.189;.493)	γ_1^L	(0.020;0.406)	(0.013;0.411)
β_2^{L1}	(-.214;0.107)	(-0.304;0.198)	β_2^{L2}	(0.252;0.957)	(0.172;.965)	γ_2^L	(-0.104;0.198)	(-0.096;0.204)
β_3^{L1}	(-.013;0.183)	(-0.026;0.218)	β_3^{L2}	(0.018;0.358)	(-0.017;.371)	γ_3^L	(-0.031;0.449)	(-0.033;0.448)
β_4^{L1}	(.190;0.476)	(0.166;0.473)	β_4^{L2}	(-0.180;0.517)	(-0.233;.602)	γ_4^L	(-0.107;0.369)	(-0.125;0.387)
β_5^{L1}	(.058;0.271)	(0.041;0.271)	β_5^{L2}	(-0.389;0.193)	(-0.366;.183)	γ_5^L	(0.066;0.406)	(0.056;0.413)
β_6^{L1}	(-.124;0.108)	(-0.144;0.127)	β_6^{L2}	(-0.068;0.333)	(-0.044;.346)	γ_6^L	(-0.049;0.340)	(-0.028;0.312)
β_7^{L1}	(-.035;0.202)	(-0.058;0.222)	β_7^{L2}	(-0.059;0.314)	(-0.029;.315)	γ_7^L	(-0.258;0.244)	(-0.208;0.130)
β_8^{L1}	(-.085;0.190)	(-0.088;0.188)	β_8^{L2}	(-0.003;0.355)	(-0.027;.307)	γ_8^L	(-0.190;0.231)	(-0.143;0.218)

Table A.12: CI widths for ordered logit model estimates.

Model used in fitness func.			An arbitrary model		
	GA-design	MMS design		GA-design	MMS design
β_1^{OL}	(0.350;0.756)	(0.240;0.849)	γ_1^{OL}	(0.279;0.518)	(0.228;0.535)
β_2^{OL}	(-0.310;0.068)	(-0.397;0.160)	γ_2^{OL}	(0.476;0.731)	(0.455;0.776)
β_3^{OL}	(1.754;2.368)	(1.701;2.582)	γ_3^{OL}	(0.580;0.980)	(0.556;1.016)
β_4^{OL}	(0.202;0.677)	(0.078;0.715)	γ_4^{OL}	(-0.014;0.146)	(-0.036;0.151)
β_5^{OL}	(0.030;0.190)	(0.013;0.194)	γ_5^{OL}	(-0.021;0.133)	(-0.030;0.140)
β_6^{OL}	(-0.059;0.369)	(-0.033;0.409)	γ_6^{OL}	(-0.004;0.318)	(-0.003;0.341)
β_7^{OL}	(0.093;0.463)	(0.098;0.458)	γ_7^{OL}	(-0.001;0.267)	(0.006;0.276)